# Nearly Optimal Sparse Group Testing

Venkata Gandikota, Elena Grigorescu, Sidharth Jaggi, Samson Zhou

***Abstract*—*Group testing is the process of pooling arbitrary subsets from a set of $n$ items so as to identify, with a minimal number of disjunctive tests, a "small" subset of $d$ defective items. In "classical" non-adaptive group testing, it is known that when $d = o(n^{1-\delta})$ for any $\delta > 0$, $\theta(d \log(n))$ tests are both information-theoretically necessary, and sufficient to guarantee recovery with high probability. Group testing schemes in the literature meeting this bound require most items to be tested $\Omega(\log(n))$ times, and most tests to incorporate $\Omega(n/d)$ items.*

*Motivated by physical considerations, we study group testing models in which the testing procedure is constrained to be "sparse". Specifically, we consider (separately) scenarios in which (a) items are finitely divisible and hence may participate in at most $\gamma$ tests; and (b) tests are size-constrained to pool no more than $\rho$ items per test. For both scenarios we provide information-theoretic lower bounds on the number of tests required to guarantee high probability recovery. In particular, one of our main results shows that $\gamma$-finite divisibility of items forces any group testing algorithm with probability of recovery error at most $\epsilon$ to perform at least $\Omega(\gamma d(n/d)^{(1-2\epsilon)/((1+2\epsilon)\gamma)})$ tests. Analogously, for $\rho$-sized constrained tests, we show an information-theoretic lower bound of $\Omega(n \log(n/d)/(\rho \log(n/\rho d)))$. In both scenarios we provide both randomized constructions (under both $\epsilon$-error and zero-error reconstruction guarantees) and explicit constructions of computationally efficient group-testing algorithms (under $\epsilon$-error reconstruction guarantees) that require a number of tests that are optimal up to constant factors in some regimes of $n, d, \gamma$ and $\rho$. We also investigate the effect of unreliability/noise in test outcomes.*

## I. Introduction

Group testing deals with identifying a relatively small number of defective items among a large population via non-linear "grouped" tests. The model was introduced by Dorfman [1] in 1943, motivated by the task of identifying syphilitic individuals among military inductees during World War II. Individual blood tests for syphilis were expensive, so multiple blood samples could be pooled and tested simultaneously. It was desirable to minimize the number of tests, while correctly identifying the disease status of every individual.

This paper studies group testing with two potential types of constraints. First, we consider a model where each item can be tested a limited number of times (*e.g.* due to a limited amount of blood that can be taken an individual). Second, we consider a model where each test can have a limited number of items (*e.g.* equipment limitations impose a maximum on the number of objects that can be simultaneously tested).

## II. Model

There is a set $\mathcal{S}$ that contains $n$ items, including an unknown subset $\mathcal{D}$ of size $d$ which are said to be "defective". Here $d$ is considered to be "small" with respect to $n$ – perhaps as small as a constant, but at any rate no larger than $O(n^{1-\epsilon})$ for some $\epsilon > 0$.[1] We wish to identify these items through a series of *group tests*, which take as input a subset (group) of the $n$ items, and outputs whether or not there exists at least one defective item in the subset/group.[2]

Group testing may be *adaptive* (the set of items to be tested in a group may be a function of prior test outcomes) or *non-adaptive* (all group tests have to be chosen independently of prior test outcomes). The advantage of non-adaptive group-tests is that they allow for parallel testing, and can use off-the shelf hardware, and hence we focus on non-adaptive group testing in this paper.

The goal of non-adaptive group testing is to correctly identify the exact set of defective items with a minimal number of non-adaptive group tests $T$. The correctness may be required either with high probability over the identity of the set $\mathcal{D}$ of $d$ defective items assumed to be uniformly distributed over all $\binom{n}{d}$ such sets (*$\epsilon$-error group testing*) or with probability 1 (*zero-error group testing*).

We use binary *test matrices* to represent non-adaptive group tests. For a given $T \times n$ test matrix $M$, there is a 1 in the $(i, j)$th location if item $j$ is tested in test $i$, and 0 otherwise.

The weight-$d$ binary *input vector* $X \in \{0, 1\}^n$ represents the set $\mathcal{S}$, and contains 1's in the locations corresponding to the items of $\mathcal{D}$. The locations with 1s are said to be *defective* while others are said to be *non-defective*. The outcomes of the tests correspond to the *result vector* $Y \in \{0, 1\}^T$, with a 1 in the $i$th location if and only if the $i$th test has at least one defective item, *i.e.*, the OR of the components of $X$ restricted to the support of the $i$th row of $M$ is 1.

The decoder then estimates the locations of the defective items and outputs an *estimate vector* $\widehat{X} \in \{0, 1\}^n$, with 1's in the locations where the group testing algorithm estimates the defective items to be. The probability of error of any

Department of Computer Science, Purdue University, West Lafayette, IN. Email: vgandiko@purdue.edu, elena-g@purdue.edu, samsonzhou@gmail.com.

Department of Information Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. Email: jaggi@ie.cuhk.edu.hk.

[1] The regime where $d = \theta(n)$ is much less well-studied.

[2] We assume that the true value of $d$ is known *a priori*.

group testing algorithm is defined as the probability over the input vector $X$ that the estimate vector $\widehat{X}$ differs from $X$ in any location. Thus, $\epsilon$-error group testing requires a test matrix $M$ and a corresponding decoding procedure such that $\underset{X}{\mathbf{Pr}}\big[\widehat{X} \neq X\big] < \epsilon$ over all possible sets of defectives $\mathcal{D}$, where each set of defectives may occur uniformly[3] with probability $1/\binom{n}{d}$. In contrast, zero-error group testing requires a test-matrix $M$ and a decoding procedure such that for all $d$-sparse inputs $X$, the decoding procedure outputs $\widehat{X} = X$. Some authors [2], [3] also consider "noisy" tests, in which with probability $\sigma$, test outcomes are misreported. In such models, an $\epsilon$-error reconstruction guarantee is desired, where the probability is over randomness in $X$ and in the noise process that converts $Y$ to a noisy vector $\hat{Y}$ (via, for instance a binary-symmetric channel with crossover probability $\sigma$). We briefly consider such models in Section VII.

We define the $\gamma$-*divisible* model, where each item can be tested at most $\gamma$ times, and so each column of $M$ contains at most $\gamma$ 1's. Similarly, we define the $\rho$-*sized* model, where each test can include at most $\rho$ items, and so each row of $M$ contains at most $\rho$ 1's.

All logarithms in this paper are base 2. The function $H(X)$ denotes the entropy of the (vector valued) random variable $X$, $H(p)$ denotes the binary entropy function, $H(X|Y)$ denotes the conditional entropy of $X$ given $Y$, and $I(X;Y)$ the mutual information between $X$ and $Y$.

### A. Related work

While there is significant literature on multiple alternative models of group testing [3]–[11], the focus of this work is primarily on non-adaptive group testing, under $\epsilon$-error and zero-error reconstruction guarantee metrics. We thus restrict the discussion of prior work to the literature on lower bounds and algorithms (both deterministic and randomized) for $\epsilon$-error and zero-error non-adaptive group testing.

Du and Hwang [12] show through disjunct matrices that $O(d^2 \log n)$ tests suffice for zero-error group testing, while Porat and Rothschild [13] provide an explicit NAGT algorithm with $O(d^2 \log n)$ tests, almost matching the best known lower bound of $\Omega\left(d^2 \log n / \log d\right)$ [14].

The lower bound of $\Omega((1-\epsilon)d\log(n/d)/(1-H(\sigma))$ for $\epsilon$-error group testing, [15] is met (up to constant factors) by [11], [16], [17].

In all the works mentioned above there are no *a priori* constraints on the group tests themselves. In classical group testing algorithms that meet the information-theoretic lower bound of $\theta(d\log(n))$ tests for $\epsilon$-error reconstruction, each item is tested $\Omega(\log(n/d))$ times.

## III. Results

### A. $\gamma$-divisible items

*Small-error:* Suppose there are $n$ items, including $d$ defective items. If each item can be tested at most $\gamma$ times, then to identify the $d$ defective items with probability at least $1-\epsilon$ in the non-adaptive group testing model:

**Theorem III.1** (Section V-A)**.** *For $\gamma = o(\log n)$, at least $\Omega\left(\gamma d \left(\frac{n}{d}\right)^{\frac{1-2\epsilon}{(1+2\epsilon)\gamma}}\right)$ tests are needed.*

**Theorem III.2** (Section V-B)**.** *There exists a randomized algorithm using $T = O\left((\gamma d)\left(\frac{n-d}{\epsilon}\right)^{1/\gamma}\right)$ tests.*

**Theorem III.3** (Section V-C)**.** *There exists a deterministic algorithm using $T = \frac{d^2\gamma}{\epsilon}\left(\frac{n\epsilon}{d^2}\right)^{1/\gamma}$ tests.*

*Zero-error:* We also study the zero-error model where the $d$ out of $n$ defective items have to identified without error. We show the following results for $\gamma$-divisible group testing:

**Theorem III.4** (Section VI-A)**.** *There exists a randomized algorithm for $\gamma$-divisible group testing with zero errors using $O\left(\gamma d\left(n\left(\frac{n}{d}\right)^d\right)^{\frac{1}{\gamma}}\right)$ tests.*

*Noisy tests:* Finally, we consider the case where individual tests can fail with probability $\sigma$. We show:

**Theorem III.5** (Section VII)**.** *It is not possible to recover the set of defective items with probability at least $1-\epsilon$ for arbitrary $0 < \epsilon < 1$ and $\gamma = o(\log n)$, when each item can be tested at most $\gamma$ times.*

### B. $\rho$-sized tests

*Small-error:* Suppose there are $n$ items, including $d$ defective items. If each test can contain at most $\rho$ items, then to identify the $d$ defective items with probability at least $1-\epsilon$ in the non-adaptive group testing model:

**Theorem III.6.** *At least $\Omega\left(\frac{n}{\rho}\frac{\log\left(\frac{n}{d}\right)}{\log\left(\frac{n}{\rho d}\right)}\right)$ tests are needed.*

**Theorem III.7.** *There exists a randomized algorithm using $T = O\left(\frac{n}{\rho}\log\left(\frac{n}{\epsilon}\right)\right)$ tests.*

**Theorem III.8.** *For $\rho = n^{1-1/k}$, for some integer $k \geq 2$, there exists a deterministic algorithm using $T = \frac{n}{\rho}\frac{d^2\log n}{\epsilon\log\left(\frac{n}{p}\right)}$ tests.*

[3] One may also consider slightly different distributions over $\mathcal{D}$, as other authors do. For instance, it may be of interest to consider a uniform distribution over all $\sum_{i=0}^{d}\binom{n}{i}$ subsets of size *at most* $d$ (rather than *exactly* $d$, as we do in our model). Alternatively, one may consider a model in which items are defective i.i.d. Bernoulli($d/n$), leading to an expected group size of $d$. It turns out that these model perturbations do not substantially change our results, and hence we focus on just the model wherein each set of $d$ items may equal $\mathcal{D}$ with probability $1/\binom{n}{d}$.

<u>*Zero-error:*</u> For $\rho$-sized group testing without error, we show that

**Theorem III.9.** *There exists a randomized algorithm for $\rho$-sized group testing with zero-errors using $O\left(\frac{n}{\rho}\log\left(n(\frac{n}{d})^d\right)\right)$ tests.*

<u>*Noisy tests:*</u>

**Theorem III.10.** *There exists a randomized algorithm for $\rho$-sized group testing with $kT$ tests, where $T = O\left(\frac{n}{\rho}\log\left(\frac{n}{\epsilon}\right)\right)$ and $k = O\left(\log\left(\frac{n}{\rho}\log\left(\frac{n}{\epsilon}\right)\right)\right)$, which will recover the set of defective items with probability at least $1-\epsilon$.*

For detailed proofs to these results, we refer to the full version of the paper [18]. Table I contains a summary of prior work and our results.

## IV. Proof Overview

For our information-theoretic lower bounds on $\epsilon$-error non-adaptive group testing, we start with the "folklore" observation that for any group testing procedure to succeed, the entropy of the test outcome vector $Y$ must almost equal the entropy of the input vector $X$ (which has entropy $\log\left(\binom{n}{d}\right)$, which is approximately $d\log(n/d)$ for $d = O(n^{1-\epsilon})$. Indeed, in classical group testing, this is a design principle for the test matrices $M$, leading to designs such that the probability of test outcomes being either positive or negative should ideally be close to $1/2$ (and hence the entropy of each individual test should be close to 1). [4] This design principle implies that each test should include about $g^* = (n/d)\ln(2)$ items, since then the probability of a negative test outcome can then be shown to be $\approx 1/2$. This density of items per test (corresponding to the density of items per row of $M$) coupled with the desire to use only an information-theoretically optimal number of tests of about $\theta(d\log(n/d))$ (hence restricting $M$ to have $\theta(d\log(n/d))$ rows), induces the fact that each column of $M$ should have on average about $\theta(\log(n/d))$ items.

But for "sparse" matrices, for instance when tests are size constrained to $\rho = o(n/d)$, it may be impossible to meet this design principle. This implies a fundamental upper bound on the entropy that can be "squeezed" out of each test $Y_j$. Coupled with the need to squeeze a total of $d\log(n/d)$ bits of entropy out of the test vector $Y$, and "standard" information-theoretic techniques (such as Fano's inequality) relating entropic quantities to probability of error give us non-trivial lower bounds on the number of tests required in the $\rho$-sized constrained model.

Similar techniques also work in Section V-A to provide lower bounds in the case when the testing procedure involves $\gamma$-divisible items – this puts a fundamental upper bound on $\gamma$ on the number of 1's in any column of the testing matrix $M$. This implies a constraint on the *average* density of each row in $M$. However, more care is required in this model, since there may be a few test rows of $M$ with "high" weight. Our bounding technique therefore proceeds by choosing a threshold above which we consider a test to be "heavy". We then do a two-stage approximation to obtain an upper bound on the entropy of the test outcome vector $Y$, and the rest of the proof is similar to the one on $\rho$-sized tests.

As an explicit example of the type of results obtainable via these lower bounding techniques, we can show that to detect a single defective ($d = 1$) out of $n$ items, with a constraint that each item may be tested at most twice ($\gamma = 2$), it must be the case that group testing procedure has at least about $\sqrt{n}$ tests. (Compared with $\log(n)$ tests, which would suffice in the unconstrained case.) To gain further intuition on why such a lower bound might be tight, consider the following testing algorithm. The $n$ items are arranged into a $\sqrt{n} \times \sqrt{n}$ grid, as in Figure 1. The test matrix then comprises of $2\sqrt{n}$ rows, corresponding to the $\sqrt{n}$ sets of $\sqrt{n}$ items in each column of this grid, and the $\sqrt{n}$ sets of $\sqrt{n}$ items in each row of this grid. The unique defective item then must correspond to the item sitting at the intersection of the single column and the single row that return positive test outcomes.

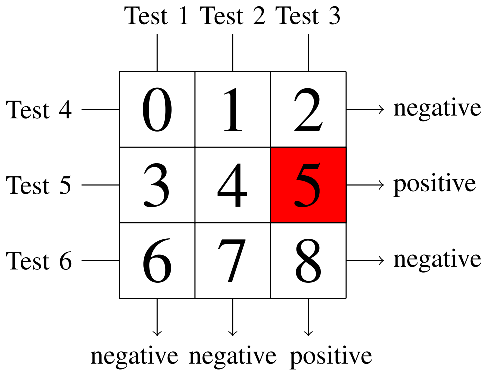

Fig. 1: If $n = 9, \gamma = 2, d = 1$, the above test uniquely determines that item 5 is defective.

Generalizing this design to general item and test constrained settings takes more work. We provide explicit constructions that use the toy example ($d = 1, \gamma = 2$) and generalize to

[4] Note that this is not a sufficient condition to guarantee low-error reconstructability of $X$ from $Y$, merely a necessary one. For instance, consider a test matrix $M$ such that the first test $Y_1$ has entropy 1 bit, and each of the remaining $d\log(n/d)$ rows are identical to this first row. So while the sum of the entropies of individual tests is large, the overall entropy of the test outcome vector is *just* 1 bit. This is due to the *extreme* correlation across tests. So really, one needs to design a matrix $M$ which not only has high entropy per tests, but also high entropy for most collections of tests. Nonetheless, as a lower-bounding technique, bounding the entropy of individual tests often provides a reasonable first-order approximation, as indeed seems to be the case in this work.

| Model | | $\epsilon$-error | | 0-error | |
|---|---|---|---|---|---|
| | | Lower Bound | Upper Bound | Lower Bound | Upper Bound |
| General | Randomized | $\Omega\left((1-\epsilon)d\log\frac{n}{d}\right)$ [15] | $O((1-\epsilon)d\log(n/d)$ [11], [16], [17] | $\Omega\left(d^2\frac{\log n}{\log d}\right)$ [14] | $O(d^2\log n)$ [12] |
| | Explicit | Same as above | $O\left(d\frac{\log n}{\log d}\log(\frac{n}{\epsilon})\right)$ [19] | Same as above | $O(d^2\log n)$ [9], [13], [20], [21] |
| $\gamma$-divisible items | Randomized | $\Omega\left(\gamma d\left(\frac{n}{d}\right)^{\frac{1-2\epsilon}{(1+2\epsilon)\gamma}}\right)$ [Thm III.1] | $O\left((\gamma d)\left(\frac{n-d}{\epsilon}\right)^{1/\gamma}\right)$ [Thm III.2] | $\Omega\left(\gamma d\left(\frac{n}{d}\right)^{\frac{1}{\gamma}}\right)$ [Thm III.1] | $O\left(\gamma d\left(n\left(\frac{n}{d}\right)^{d}\right)^{\frac{1}{\gamma}}\right)$ [Thm III.4] |
| | Explicit | Same as above | $O\left(\frac{d^2\gamma}{\epsilon}\left(\frac{n\epsilon}{d^2}\right)^{1/\gamma}\right)$ [Thm III.3] | Same as above | $O\left(n^{\frac{1}{\gamma^{1/d}}}\right)$ [22] |
| $\rho$-sized tests | Randomized | $\Omega\left(\frac{n}{\rho}\frac{\log\left(\frac{n}{d}\right)}{\log\left(\frac{n}{\rho d}\right)}\right)$[Thm III.6] | $O\left(\frac{n}{\rho}\log\left(\frac{n}{\epsilon}\right)\right)$ [Thm III.7] | $\Omega\left(\frac{n}{\rho}\frac{\log\left(\frac{n}{d}\right)}{\log\left(\frac{n}{\rho d}\right)}\right)$[Thm III.6] | $O\left(\frac{n}{\rho}\log\left(n(\frac{n}{d})^{d}\right)\right)$ [Thm III.9] |
| | Explicit | Same as above | $O\left(\frac{n}{\rho}\frac{d^2\log n}{\epsilon\log\left(\frac{n}{p}\right)}\right)$ [Thm III.8] | Same as above | $O\left(\frac{n}{\rho}\left(\frac{\log n}{\log(n/\rho)}\right)^{d}\right)$ [22] |

TABLE I: A summary of non-adaptive group testing results.

arbitrary $d$, and $\gamma$ or $\rho$, via a "divide-and-conquer" approach. Details are provided in Section V-C.

We also provide randomized designs that draw intuition from the analysis of "classical" (unconstrained) group testing schemes. We analyze the probability that randomly chosen matrices chosen from suitable ensemble of matrices with either $\rho$-sparse rows or $\gamma$-sparse columns (for the two models considered) have a "reasonable" probability of success, by analyzing the probability that a non-defective item is "masked" by the set of $d$ defective items. These results are outlined in Section III.

We extend these randomized constructions in two ways, paralleling the development of the literature in unconstrained group testing. In Section VI we consider the randomized design of zero-error group testing schemes. The techniques here are relatively straightforward – essentially, we take the corresponding matrices ensembles of matrices in Section V-B that guarantee $\epsilon$-error reconstructability for $\gamma$-divisible item models and $\rho$-sized test models respectively, and union bound over all $\binom{n}{d}$ possible sets of defectives. Finally, in Section VII we examine the effect of $\sigma$-noise (say BSC($\sigma$) noise for concreteness) in test outcomes $Y_i$ on the reconstructability of $X$ – interestingly, while non-trivial achievability schemes exist in the $\rho$-test size constrained setting with $\sigma$-noisy test outcomes (for instance by repeating each test an appropriate number of times and taking the majority), in the $\gamma$-divisible item scenario any non-trivial amount of noise renders *any* group testing algorithm unable to reconstruct $X$ with a vanishing probability of error. This latter impossibility result stems from the fact that if the columns of $M$ are sufficiently sparse ($o(\log(n))$, then with non-trivial probability $(1-\sigma^{o(log(n))})^n$, all information about the status of at least one item will be completely masked by the noise in the tests in which the item participates.

## V. $\gamma$-Divisible Items

### A. Theorem III.1: Information-Theoretic Lower Bounds

In classical group testing, each item is tested $\Omega\left(\log\frac{n}{d}\right)$ times. Thus, we consider the regime where $\gamma = o\left(\log\frac{n}{d}\right)$. We now provide information-theoretic lower bounds on the number of tests required to guarantee high probability reconstruction of the set of defectives items in a model with column constraints (*i.e.*, each item can be tested at most $\gamma$ times).

**Proof of Theorem III.1:** Recall that $X$ is the input vector, $Y$ is the result vector, and $\widehat{X}$ is the estimate vector so that $X\to Y\to\widehat{X}$ forms a Markov chain. From standard information-theoretic definitions, we have

$$H(X) = H(X|\widehat{X}) + I(X;\widehat{X}), \tag{1}$$

where $H(X)$ is the binary entropy of the length-$n$ binary vector $X$, and $I(X;\widehat{X})$ is the mutual information between $X$ and $\widehat{X}$. Since $X$ is uniformly distributed over $\mathcal{X}$, the set of all length-$n$, $d$-sparse binary vectors, we have

$$H(X) = \log|\mathcal{X}| = \log\binom{n}{d} \tag{2}$$

We now upper bound each of the terms in RHS of Equation 1 separately. By Fano's Inequality, $H(X|\widehat{X}) \le H(\epsilon) + \epsilon\log(|\mathcal{X}|-1)$. Note that for $\epsilon < \frac{1}{2}$,

$$H(\epsilon) < -2\epsilon\log\epsilon. \tag{3}$$

Also, by the data processing inequality and standard information theoretic inequalities, $I(X;\widehat{X}) \le I(X;Y) = H(Y) - H(Y|X) \le H(Y)$. The bound on $H(Y)$ follows from Lemma V.1 which will be proved later.

**Lemma V.1.**

$$H(Y) \leq (1+2\epsilon)\gamma d \log\left(\frac{T}{\gamma d}\right)$$

Plugging in the value of $H(X)$ from Equation 2 and the inequalities from Equation 3 and Lemma V.1 in Equation 1,

$$\begin{aligned} H(X) &= H(X|\widehat{X}) + I(X;\widehat{X}) \\ &\leq H(\epsilon) + \epsilon\log(|\mathcal{X}|-1) + H(Y) \\ \log\left(\binom{n}{d}\right) &\leq -2\epsilon\log\epsilon + \epsilon\log\left(\binom{n}{d}\right) \\ &\quad + (1+2\epsilon)\gamma d\log\left(\frac{T}{\gamma d}\right). \end{aligned}$$

By reordering the terms we get a lower bound on the number of tests as

$$\begin{aligned} T &\geq \gamma d e^{\frac{(1-\epsilon)\log\left(\binom{n}{d}\right)+2\epsilon\log\epsilon}{(1+2\epsilon)\gamma d}} \\ &\geq \gamma d e^{\frac{(1-2\epsilon)\log\left(\binom{n}{d}\right)}{(1+2\epsilon)\gamma d}} \qquad \text{(for sufficiently large } n\text{)} \\ &\geq \gamma d \binom{n}{d}^{\frac{1-2\epsilon}{(1+2\epsilon)\gamma d}} \\ &\geq \gamma d\left(\frac{n}{d}\right)^{\frac{(1-2\epsilon)(1+\epsilon)}{(1+2\epsilon)\gamma}} \qquad \text{(by Sterling's approximation)} \end{aligned}$$

to ensure a probability of reconstruction error of at most $\epsilon$. Hence, $T = \Omega\left(\gamma d\left(\frac{n}{d}\right)^{\frac{1-2\epsilon}{(1+2\epsilon)\gamma}}\right)$ tests are needed. We remark that the same inequalities hold for adaptive group testing. □

**Proof of Lemma V.1:** Let $Y = (Y_1, Y_2, \ldots, Y_T)$, where

$$Y_i = \begin{cases} 1 & \text{if test } i \text{ is negative} \\ 0 & \text{if test } i \text{ is positive.} \end{cases}$$

By the chain rule, $H(Y) \leq \sum_{i=1}^{T} H(Y_i)$. We partition the tests $T$ into sets $S_1$ and $S_2$, where $i \in S_1$ if test $i$ includes less than $\frac{n}{\epsilon d\log\left(\frac{T}{\gamma d}\right)}$ items, and $i \in S_2$ otherwise (that is, test $i$ includes at least $\frac{n}{\epsilon d\log\left(\frac{T}{\gamma d}\right)}$ items). [5]

Since there are at most $\gamma n$ items that can be tested in total and the entropy of each test outcome binary variable $Y_i$ is at most 1, then

$$\sum_{i\in S_2} H(Y_i) \leq |S_2| \leq \frac{\gamma n}{\left(\frac{n}{\epsilon d\log\left(\frac{T}{\gamma d}\right)}\right)} = \epsilon\gamma d\log\left(\frac{T}{\gamma d}\right).$$

For $i \in S_1$, test $i$ includes $g_i$ items, where $g_i < \frac{n}{\epsilon d\log\left(\frac{T}{\gamma d}\right)}$. Then the probability $p_i^-$ that $Y_i$ is negative is

$$\begin{aligned} \binom{n-g_i}{d}\Big/\binom{n}{d} &= \frac{(n-d)!(n-g_i)!}{(n-d-g_i)!n!} \\ &\geq \left(1 - O\left(\frac{1}{n-d-g_i}\right)\right) \\ &\quad \times \frac{(n-d)^{n-d+\frac{1}{2}}(n-g_i)^{n-g_i+\frac{1}{2}}}{n^{n+\frac{1}{2}}(n-d-g_i)^{n-d-g_i+\frac{1}{2}}} \\ &\quad \text{(by Sterling's approximation)} \\ &= (1-\delta)\left(1-\frac{d}{n}\right)^{n-d+\frac{1}{2}} \\ &\quad \times \left[\frac{\left(1-\frac{g_i}{n}\right)^{n-g_i+\frac{1}{2}}}{\left(1-\frac{d}{n}-\frac{g_i}{n}\right)^{n-d-g_i+\frac{1}{2}}}\right] \\ &\quad \text{(for any } \delta > 0 \text{ and sufficiently large } n\text{)} \end{aligned}$$

Since $\left(1-\frac{d}{n}-\frac{g_i}{n}\right) < \left(1-\frac{d}{n}\right)\left(1-\frac{g_i}{n}\right)$, then

$$\begin{aligned} &\binom{n-g_i}{d}\Big/\binom{n}{d} \geq (1-\delta)\left(1-\frac{d}{n}\right)^{n-d+\frac{1}{2}} \\ &\times \left[\frac{\left(1-\frac{g_i}{n}\right)^{n-g_i+\frac{1}{2}}}{\left(1-\frac{d}{n}\right)^{n-d-g_i+\frac{1}{2}}\left(1-\frac{g_i}{n}\right)^{n-d-g_i+\frac{1}{2}}}\right] \\ &= (1-\delta)\left(1-\frac{d}{n}\right)^{g_i}\left(1-\frac{g_i}{n}\right)^{d} \\ &\geq (1-\delta)\left(1-\frac{d}{n}\right)^{g_i}\left(1-\frac{dg_i}{n}\right) \\ &\text{(by Bernoulli's Inequality)} \\ &\geq (1-\delta)\left(1-\frac{d}{n}\right)^{g_i}\left(1-\frac{1}{\epsilon\log\left(\frac{T}{\gamma d}\right)}\right) \\ &\text{(since } g_i \text{ corresponds to a "light" test)} \\ &\geq (1-2\delta)\left(1-\frac{d}{n}\right)^{g_i} \\ &\text{(for sufficiently large } n\text{)} \end{aligned}$$

where the last inequality comes from the observation that in the regime where $\gamma = o\left(\log\left(\frac{n}{d}\right)\right)$ and $T = \Omega\left(d\log\left(\frac{n}{d}\right)\right)$, then $\lim_{n\to\infty}\frac{T}{\gamma d} = \infty$. (See the beginning of the section for a discussion of why we consider this regime.) By the Arithmetic-Geometric Mean Inequality,

$$\begin{aligned} &(1-2\delta)\frac{1}{|S_1|}\sum_{i\in S_1}\left(1-\frac{d}{n}\right)^{g_i} \geq \\ &\qquad (1-2\delta)\left(\prod_{i\in S_1}\left(1-\frac{d}{n}\right)^{g_i}\right)^{\frac{1}{|S_1|}}. \end{aligned}$$

**Claim V.2.** $\frac{\sum_{i\in S_1} g_i}{|S_1|} \leq \frac{\gamma n}{T}$

[5] Roughly speaking, tests in set $S_1$ are "light" (test "few" items per test) and hence have a "high" probability of being negative, and thus "low" entropy (significantly less than 1 bit per test). Conversely, tests in set $S_2$ are "heavy" (test "many items per test) and may potentially have "high" entropy (as much as 1 bit per test) - however, there cannot be too many heavy tests, due to the constraint that each item is tested at most $\gamma$ times.

*Proof:* Since $i \in S_1$ for $g_i < \frac{n}{\epsilon d \log\left(\frac{T}{\gamma d}\right)}$ and $i \in S_2$ for $g_i \geq \frac{n}{\epsilon d \log\left(\frac{T}{\gamma d}\right)}$, then $\frac{\sum_{i\in S_1} g_i}{|S_1|} < \frac{n}{\epsilon d \log\left(\frac{T}{\gamma d}\right)} \leq \frac{\sum_{i\in S_2} g_i}{|S_2|}$. But then

$$\begin{aligned}
|S_2| \sum_{i \in S_1} g_i &\leq |S_1| \sum_{i \in S_2} g_i \\
|S_1| \sum_{i \in S_1} g_i + |S_2| \sum_{i \in S_1} g_i &\leq |S_1| \sum_{i \in S_1} g_i + |S_1| \sum_{i \in S_2} g_i \\
(|S_1| + |S_2|) \sum_{i \in S_1} g_i &\leq |S_1| \sum_{i \in (S_1 \cup S_2)} g_i
\end{aligned}$$

$$|T| \sum_{i \in S_1} g_i \leq |S_1| \gamma n, \quad \frac{\sum_{i \in S_1} g_i}{|S_1|} \leq \frac{\gamma n}{T}$$

■

We now use the bound in Claim V.2 and properties of the binary entropy function to bound the entropy of tests in $S_1$. Since $\frac{\sum_{i\in S_1} g_i}{|S_1|} \leq \frac{\gamma n}{T}$, then

$$\begin{aligned}
&(1-2\delta) \left( \prod_{i \in S_1} \left(1 - \frac{d}{n}\right)^{g_i} \right)^{\frac{1}{|S_1|}} \\
&\quad = (1-2\delta)\left(1 - \frac{d}{n}\right)^{\frac{\sum_{i \in S_1} g_i}{|S_1|}} \geq (1-2\delta)\left(1 - \frac{\gamma d}{T}\right),
\end{aligned}$$

by Bernoulli's Inequality. Note that in the $\gamma = o\left(\log \frac{n}{d}\right)$ regime, $(1-\delta)\left(1 - \frac{\gamma d}{T}\right) \geq \frac{1}{2}$ since $T = \Omega\left(d \log \frac{n}{d}\right)$. Specifically,

$$1 \geq \prod_{i \in S_1} \left( \frac{\binom{n-g_i}{d}}{\binom{n}{d}} \right)^{\frac{1}{|S_1|}} \geq (1-2\delta)\left(1 - \frac{\gamma d}{T}\right) \geq \frac{1}{2}. \tag{4}$$

Since the binary entropy function $H(x)$ is monotonically decreasing for $x \in (1/2, 1)$, using Equation 4 above we have that:

$$\begin{aligned}
H\left(1 - \frac{\gamma d}{T}\right) &\geq H\left( \left( \prod_{i \in S_1} \frac{\binom{n-g_i}{d}}{\binom{n}{d}} \right)^{\frac{1}{|S_1|}} \right) \\
&\geq H\left( \frac{1}{|S_1|} \sum_{i \in S_1} \frac{\binom{n-g_i}{d}}{\binom{n}{d}} \right) = H\left( \frac{1}{|S_1|} \sum_{i \in S_1} p_i^- \right).
\end{aligned}$$

Furthermore, in the $\gamma = o\left(\log \frac{n}{d}\right)$ regime, $(1-2\delta)\left(1 - \frac{\gamma d}{T}\right)$ approaches $1 - \epsilon$ (since $T = \Omega(d \log(n/d))$ even in the unconstrained group testing case). This implies that $H(p_i^-) \leq -(1+2\delta)(1-p_i^-)\log(1-p_i^-)$ for all sufficiently large $n$. Hence,

$$H\left( \sum_{i \in S_1} p_i^- \right) \leq \left( \frac{\gamma d}{T} + 3\delta \right) \log \frac{T}{\gamma d}.$$

Therefore,

$$\begin{aligned}
H(Y) &= \sum_{i \in S_1} H(Y_i) + \sum_{i \in S_2} H(Y_i) \\
&\leq T\left( \frac{\gamma d}{T} + 3\delta \right) \log\left( \frac{T}{\gamma d} \right) + \epsilon \gamma d \log\left( \frac{T}{\gamma d} \right) \\
&\leq (1 + 2\epsilon) \gamma d \log\left( \frac{T}{\gamma d} \right)
\end{aligned}$$

(for appropriate choice of $\delta$).

□

### *B. Theorem III.2: Randomized Construction of Test Matrices*

We describe a randomized construction of a $T \times n$ test matrix $M$, where $T = O\left( (\gamma d) \left( \frac{n-d}{\epsilon} \right)^{1/\gamma} \right)$. We pick each column of $M$ uniformly at random from the vectors of $\{0,1\}^T$ with support size $\gamma$. Now, we describe how to recover the estimate vector $\widehat{X}$ from the test results.

*1) The Column Matching Algorithm (CoMa):* To obtain the estimate vector $\widehat{X}$ from result vector $Y$, the Column Matching algorithm (CoMa) from [15] uses the tests which have positive outcomes to identify all defective items, while declaring all other items to be non-defective. Namely, the algorithm marks item $i$ defective if every test in which $i$ is included is positive. Note that CoMa cannot incorrectly mark positive items. CoMa can only incorrectly designate a non-defective item as defective if the item is not tested, or is only tested in positive tests (*i.e.,* every test it occurs in has at least one defective item). If $M$ is chosen to have enough rows and $d = o(n)$, then with significant probability, each non-defective item should appear in at least one negative test, and hence will be appropriately marked non-defective.

*2) Analysis:* Since each of the $d$ defective items can be tested at most $\gamma$ times, the maximum number of tests which are positive is at most $d\gamma$. Now, an item will be marked by CoMa as defective if all the tests which pick this particular item are positive. Therefore for a fixed item $i$, the probability that it is incorrectly marked defective is the probability that $i$ is always tested with one of the $d$-defective items which is given as $\binom{d\gamma}{\gamma} / \binom{T}{\gamma}$. Taking a union bound over the $(n-d)$ non-defective items, we require $(n-d)\binom{d\gamma}{\gamma} / \binom{T}{\gamma} < \epsilon$. Since $\binom{d\gamma}{\gamma} < (ed)^\gamma$ and $\binom{T}{\gamma} > \left(\frac{T}{\gamma}\right)^\gamma$, then this certainly occurs if $(ed)^\gamma (n-d) < \epsilon \left(\frac{T}{\gamma}\right)^\gamma$. Thus, we see that

$$T > (e\gamma d) \left( \frac{n-d}{\epsilon} \right)^{1/\gamma}$$

suffices to ensure correct recovery of the set of defective items with a probability of error of at most $\epsilon$.

### C. Theorem III.3: Explicit Construction of Test Matrices

Recall that we seek to identify the $d$ defective items among all $n$ items, where each item may be tested at most $\gamma$ times. We first attempt to generalize the grid construction in Section IV, and point out a shortcoming in a naïve implementation.

*1) First Tool: $\gamma$-Dimensional Hypergrid:* For ease of presentation, define $b = n^{1/\gamma}$ and assume $b$ to be an integer. We represent each item $i \in \{0, \dots, n-1\}$ by its base-$b$ representation $(x_\gamma \dots x_2 x_1)_b$, so that each $x_j \in \{0, 1, \dots, b-1\}$ and

$$i = \sum_{j=1}^{\gamma} x_j b^{j-1}.$$

For test $j$, where $j = \alpha b + k$, where $\alpha \in \{0, 1, \dots, \gamma - 1\}$ and $k \in \{0, 1, \dots, b-1\}$, we include exactly the items whose $(\alpha+1)$th coordinate is $k$, *i.e.*, $x_{\alpha+1} = k$. Hence, there are $\gamma b = \gamma n^{1/\gamma}$ tests in total. See Figure 2, for an example.

Intuitively, test $j = \alpha b + k$ returns whether there exists a defective item $i$ whose base-$b$ representation has $x_{\alpha+1} = k$. Note that a defective item $i \in [n]$ will cause exactly $\gamma$ tests to be defective, corresponding to when each of its coordinates is tested. Thus, if there exists a unique defective item, it can be successfully recovered from its unique base-$b$ representation.

However, with multiple defective items, we may not be able to uniquely determine each item. For example, for $n = 9$, $d = 2$ and $\gamma = 2$, if items 2 and 4 are defective, then positive tests will tell us that there exist defective items with $x_1 = 1$ (corresponding to item 4), $x_1 = 2$ (corresponding to item 2), $x_2 = 0$ (corresponding to item 2) and $x_2 = 1$ (corresponding to item 4). However, another pair of defective items which return the same positive test results are items 1 and 5. Thus, we cannot uniquely recover all defective items, unless there is only one defective item. See Figure 2 for more details.

*2) Block Algorithm: Divide and Conquer:* We now provide an explicit construction a $T \times n$ test matrix $M$, where $T = \frac{d^2\gamma}{\epsilon}\left(\frac{n\epsilon}{d^2}\right)^{1/\gamma}$, using the previous ideas. The key observation is that the first previous algorithm succeeds if there is a low number of defective items. Thus, we split $[n]$ into $cd^2$ blocks, where $c = \frac{1}{\epsilon}$ and run the previous algorithm on each block of size $n' = n/cd^2$. (See Figure 3, for an example.) Then the probability that no two defective items fall into the same block is

$$\begin{aligned} &1\left(1 - \frac{1}{cd^2}\right)\left(1 - \frac{2}{cd^2}\right)\cdots\left(1 - \frac{d-1}{cd^2}\right) \\ &\geq \left(1 - \frac{d}{cd^2}\right)^d = \left(1 - \frac{1}{cd}\right)^d \\ &\geq 1 - \frac{1}{c} = 1 - \epsilon \qquad \text{(by Bernoulli's Inequality)} \end{aligned}$$

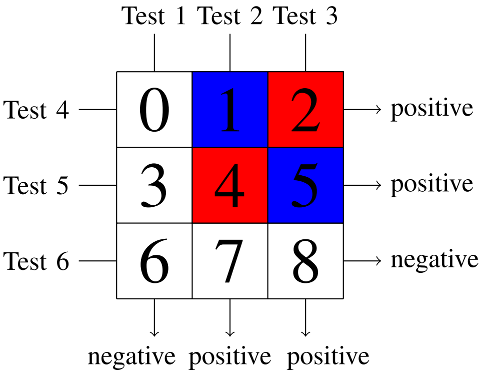

Fig. 2: If $n = 9, \gamma = 2, d = 2$, the above test cannot distinguish whether the red items or the blue items are defective. However, if there were only one defective item, the series of tests would uniquely identify the defective item.

Thus, with probability at least $1 - \epsilon$, the maximum number of defective items in a single block is 1, so we can also successfully identify the $d$ defective items with the probability at least $1 - \epsilon$ using the previous algorithm for $n'$ items. Since there are $\frac{d^2}{\epsilon}$ blocks, each requiring $\gamma\left(\frac{n\epsilon}{d^2}\right)^{1/\gamma}$ tests, for a total of $T = \frac{d^2\gamma}{\epsilon}\left(\frac{n\epsilon}{d^2}\right)^{1/\gamma}$ tests.

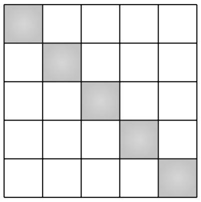

Fig. 3: The test matrix for the block algorithm, where each gray block represents the test matrix for the first part.

## VI. Zero-Error Tests

A matrix $M$ is called *$d$-disjunct* if the union of any $d$ columns does not contain any other column. It is well known [20] that a $T \times n$ binary $d$-disjunct matrix corresponds to an efficient non-adaptive group testing algorithm to identify the $d$ defects among $n$ items with $T$ tests. In this section we focus on construction of $\gamma$-divisible matrices.

D'yachkov and Rykov [23] showed that there are at most $T$ columns in a $T \times n$ $d$-disjunct matrix of weight at most $d$. This gives a trivial lower bound of $n$ tests for $\gamma$-divisible non-adaptive group tests if $\gamma \leq d$. $d$-disjunct matrices have been well-studied by [24], [25] under the name of superimposed codes. Macula [22] gave a deterministic construction of $T \times n$ binary $d$-disjunct matrix which is simultaneously $\gamma$-divisible and $\rho$-sized with $t = O(\gamma n^{1/\gamma^{\frac{1}{d}}})$ rows. We therefore give an efficient randomized procedure to construct a $T \times n$ binary matrix $M$ which is $d$-disjunct with high probability for sufficiently large $T$ and satisfies the row (or column) constraints.

### A. Theorem III.4 : $\gamma$-divisible $d$-disjunct matrices

In this section, we give a randomized construction of $d$-disjunct matrices with each column having weight at most $\gamma$.

Define a random $T \times n$ binary $d$-disjunct matrix $M$ as follows: For each column $C_j$, $j \in [n]$ of $M$, sample $\gamma$ rows, $R \subseteq [T], |R| = \gamma$ with replacement and set $M_{i,j} = 1$ for $i \in R$. We now show that for sufficiently large number of rows, $M$ will be $d$-disjunct with high probability.

**Lemma VI.1.** *For $0 \leq \epsilon \leq 1$ and $T \geq \gamma d \left( \frac{n}{\epsilon} \left( \frac{en}{d} \right)^d \right)^{\frac{1}{\gamma}}$, then $M$ is a $d$- disjunct matrix with probability at least $1 - \epsilon$.*

*Proof:* Let $S \subseteq [T]$ be the set of positive test outcomes. Since each item is tested at most $\gamma$ times, there are at most $\gamma d$ positive test outcomes. Note that $M$ cannot distinguish a column $C_j$ from the defectives if support$(C_j) \subseteq S$ For any fixed column $C_j$, this probability is at most

$$\mathbf{Pr}\left[\text{support}(C_j) \subseteq S\right] = \frac{|S|}{T} \leq \left( \frac{\gamma d}{T} \right)^{\gamma}$$

Taking a union bound over all possible columns corresponding to $(n - d)$ non-defective items and over all possible choices of $d$ defective items, we get

$$\begin{aligned} \mathbf{Pr}\left[M \text{ is not } d - \text{disjunct}\right] &\leq \binom{n}{d} (n - d) \left( \frac{\gamma d}{T} \right)^{\gamma} \\ &\leq n \left( \frac{en}{d} \right)^d \left( \frac{\gamma d}{T} \right)^{\gamma}. \end{aligned}$$

Therefore, if $T \geq \gamma d \left( \frac{n}{\epsilon} \left( \frac{en}{d} \right)^d \right)^{\frac{1}{\gamma}}$, then $M$ is $d$-disjunct with probability at least $1 - \epsilon$. ∎

## VII. Impact of Noisy Tests

### A. Theorem III.5: $\gamma$-Divisible Items

We consider the noisy setting, where each test can be incorrect with probability $\sigma$, for $\gamma$-divisible tests. Since the Coupon Collector Algorithm collects certificates for non-defective items, and each item is tested at most $\gamma$ times, there is probability $\sigma^{\gamma}$ that a non-defective item will not have a certificate (i.e., all of the tests for which it is included are erroneous). Note that for $\gamma = o(\log n)$, this probability is $\omega(1/n)$. Thus it is not possible to recover the set of defective items with probability at least $1 - \epsilon$ for arbitrary $0 < \epsilon < 1$.

## Acknowledgements

This research was initiated at the ICERM Workshop on Algorithmic Coding Theory, 2016. We would like to thank ICERM and the organizers for their hospitality. We would also like to thank Abhinav Ganesan and Gu Wenyuan for fruitful discussions.

## References

[1] Robert Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics*, 14:436–440, 1943.

[2] Leonardo Baldassini, Oliver Johnson, and Matthew Aldridge. The capacity of adaptive group testing. In *Proceedings of the IEEE International Symposium on Information Theory*, pages 2676–2680, 2013.

[3] Mahdi Cheraghchi, Ali Hormati, Amin Karbasi, and Martin Vetterli. Group testing with probabilistic tests: Theory, design and application. *IEEE Trans. Information Theory*, 57(10):7057–7067, 2011.

[4] Annalisa De Bonis. Efficient group testing algorithms with a constrained number of positive responses. In *Combinatorial Optimization and Applications, COCOA*, pages 506–521, 2014.

[5] Sheng Cai, Mohammad Jahangoshahi, Mayank Bakshi, and Sidharth Jaggi. GROTESQUE: noisy group testing (quick and efficient). In *51st Annual Allerton Conference on Communication, Control, and Computing, Allerton*, pages 1234–1241, 2013.

[6] Chun Lam Chan, Sheng Cai, Mayank Bakshi, Sidharth Jaggi, and Venkatesh Saligrama. Stochastic threshold group testing. In *IEEE Information Theory Workshop, ITW*, pages 1–5, 2013.

[7] Mahdi Cheraghchi, Amin Karbasi, Soheil Mohajer, and Venkatesh Saligrama. Graph-constrained group testing. *IEEE Trans. Information Theory*, 58(1):248–262, 2012.

[8] Amin Emad, Jun Shen, and Olgica Milenkovic. Symmetric group testing and superimposed codes. In *Information Theory Workshop (ITW), 2011 IEEE*, pages 20–24. IEEE, 2011.

[9] Piotr Indyk, Hung Q. Ngo, and Atri Rudra. Efficiently decodable non-adaptive group testing. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 1126–1142, 2010.

[10] Tongxin Li, Chun Lam Chan, Wenhao Huang, Tarik Kaced, and Sidharth Jaggi. Group testing with prior statistics. In *IEEE International Symposium on Information Theory*, pages 2346–2350, 2014.

[11] Jonathan Scarlett and Volkan Cevher. Phase transitions in group testing. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, pages 40–53, 2016.

[12] Ding-Zhu Du and Frank K. Hwang. *Combinatorial Group Testing and Its Applications*. Singapore: World Scientific, 2nd edition, 2000.

[13] Ely Porat and Amir Rothschild. Explicit non-adaptive combinatorial group testing schemes. In *Automata, Languages and Programming, 35th International Colloquium, ICALP*, pages 748–759, 2008.

[14] Paul Erdös, Peter Frankl, and Zoltán Füredi. Families of finite sets in which no set is covered by the union of $r$ others. *Israel J. Math*, 51:79–89, 1985.

[15] Chun Lam Chan, Sidharth Jaggi, Venkatesh Saligrama, and Samar Agnihotri. Non-adaptive group testing: Explicit bounds and novel algorithms. *IEEE Trans. Information Theory*, 60(5):3019–3035, 2014.

[16] András Sebö. On two random search problems. *Journal of Statistical Planning and Inference*, 11:23–31, 1985.

[17] Matthew Aldridge, Oliver Johnson, and Jonathan Scarlett. Improved group testing rates with constant column weight designs. *CoRR*, 2016.

[18] Venkata Gandikota, Elena Grigorescu, Sidarth Jaggi, and Samson Zhou. Nearly optimal sparse group testing. In preparation.

[19] Arya Mazumdar. Nonadaptive group testing with random set of defectives via constant-weight codes. *CoRR*, 2015.

[20] W. Kautz and R. Singleton. Nonrandom binary superimposed codes. *IEEE Transactions on Information Theory*, 10(4):363–377, Oct 1964.

[21] Mahdi Cheraghchi. Noise-resilient group testing: Limitations and constructions. In *International Symposium on Fundamentals of Computation Theory*, pages 62–73. Springer, 2009.

[22] Anthony J Macula. A simple construction of d-disjunct matrices with certain constant weights. *Discrete Mathematics*, 162(1):311–312, 1996.

[23] Arkadii Georgievich D'yachkov and Vladimir Vasil'evich Rykov. Bounds on the length of disjunctive codes. *Problemy Peredachi Informatsii*, 18(3):7–13, 1982.

[24] A. G. D'yachkov, I. V. Vorob'ev, N. A. Polyansky, and V. Yu. Shchukin. Bounds on the rate of disjunctive codes. *Problems of Information Transmission*, 50(1):27–56, 2014.

[25] Arkadii Georgievich D'yachkov and Vladimir Vasil'evich Rykov. A survey of superimposed code theory. *Problems of Control and Information Theory*, 12(4):1–13, 1983.