Non-Adaptive Randomized Algorithm for Group Testing

Nader H. Bshouty   BSHOUTY@CS.TECHNION.AC.IL
Dept. of Computer Science
Technion, Haifa, 32000

Nuha Diab   NUHA15.3.98@GMAIL.COM
Sisters of Nazareth High School
12th Grade
P.O.B. 9422, Haifa, 35661

Shada R. Kawar   SHADAKAWAR.135@GMAIL.COM
Nazareth Baptist High School
11th Grade
P.O.B. 20, Nazareth, 16000

Robert J. Shahla   SHAHLA.ROBERT@GMAIL.COM
Sisters of Nazareth High School
11th Grade
P.O.B. 9422, Haifa, 35661

Abstract

We study the problem of group testing with a non-adaptive randomized algorithm in the random incidence design (RID) model where each entry in the test is chosen randomly independently from $\{0, 1\}$ with a fixed probability p.

The property that is sufficient and necessary for a unique decoding is the separability of the tests, but unfortunately no linear time algorithm is known for such tests. In order to achieve linear-time decodable tests, the algorithms in the literature use the disjunction property that gives almost optimal number of tests.

We define a new property for the tests which we call semi-disjunction property. We show that there is a linear time decoding for such test and for $d \to \infty$ the number of tests converges to the number of tests with the separability property. Our analysis shows that, in the RID model, the number of tests in our algorithm is better than the one with the disjunction property even for small d.

Keywords: Group Testing, Non-adaptive Algorithm, Randomized Algorithm, RID Algorithm.

1. Introduction

As quoted in Wikipedia., "Robert Dorfman's paper in 1943, Dorfman (1943), introduced the field of Group Testing. The motivation arose during the Second World War when the United States Public Health Service and the Selective service embarked upon a large scale

project. The objective was to weed out all syphilitic men called up for induction. However, syphilis testing back then was expensive and testing every soldier individually would have been very cost heavy and inefficient. A basic breakdown of a test is: Draw sample from a given individual, perform required tests and determine the presence or absence of syphilis. Suppose we have n soldiers. Then this method of testing leads to n tests. Our goal is to achieve effective testing in a scenario where it does not make sense to test $100,000$ people to get (say) 10 positives. The feasibility of a more effective testing scheme hinges on the following property. We can combine blood samples and test a combined sample together to check if at least one soldier has syphilis".

Group testing was originally introduced as a potential approach to the above economical mass blood testing, Dorfman (1943). However it has been proven to be applicable in a variety of problems, including quality control in product testing, Sobel and Groll (1959), searching files in storage systems, Kautz and Singleton (1964), sequential screening of experimental variables, Li (1962), efficient contention resolution algorithms for multiple-access communication, Kautz and Singleton (1964); Wolf (1985), data compression, Hong and Ladner (2002), and computation in the data stream model, Cormode and Muthukrishnan (2005). See a brief history and other applications in Cicalese (2013); Du and Hwang (2000, 2006); Hwang (1972); Macula and Popyack (2004); Ngo and Du (1999) and references therein.

We now formalize the problem. Let S be the set of the n *items* (soldiers, clones) and let $I \subseteq S$ be the set of *defective items* (the sick soldiers, positive clone). Suppose we know that the number of defective items, $|I|$, is bounded by some integer d. A *test* or a *query* (pool or probing) is a set $J \subset S$. The answer to the test is $T(I, J) = 1$ if $I \cap J \neq \emptyset$ and 0 otherwise. Here J is the set of soldiers for which their blood samples is combined. The problem is to find the defective items (sick soldiers) with a minimum number of tests. We will assume that $S = [n] := \{1, 2, \ldots, n\}$ and will identify the test $J \subseteq S$ with an *assignment* $a^J \in \{0, 1\}^n$ where $a_i^J = 1$ if and only if $i \in J$, i.e., $a_i^J = 1$ signifies that item i is in test J. Then the answer to the test a^J is $T(I, a^J) = 1$ (*positive*) if there is $i \in I$ such that $a_i^J = 1$ and 0 (*negative*) otherwise.

In the *adaptive algorithm*, the tests can depend on the answers to the previous ones. In the *non-adaptive algorithm* they are independent of the previous one and; therefore, one can do all the tests in one parallel step. In all the above applications, non-adaptive algorithms are most desirable since minimizing time is of utmost importance. The set of tests in any non-adaptive deterministic algorithm (resp. randomized) can be identified with a (resp. random) $m \times n$ *test matrix* M (pool design) such that its rows are all the assignments a^J of all the tests J in the algorithm. The above problem is also equivalent to the problem of non-adaptive deterministic (resp. randomized) learning the class of boolean disjunction of up to d variables from membership queries Bshouty (2013, 2017).

It is known that any non-adaptive *deterministic* algorithm must do at least $\Omega(d^2 \log n / \log d)$ tests D'yachkov and Rykov (1982); Füredi (1996); Porat and Rothschild (2011); Ruszinkó (1994). This is $O(d/\log d)$ times more than the number of tests of the known non-adaptive randomized algorithms that do only $O(d \log n)$ tests. In this paper we study the case where the number of items n is very large and the cost of the test is "very expensive" and in that case, deterministic algorithms are impractical. Also, algorithms that run in super-linear time are not practical for group testing. So we need our algorithms to do a

minimum number of tests and run in quasi-linear time, preferably, linear time in the *size of all the tests*, that is, $O(dn \log n)$.

Non-adaptive randomized algorithms have been proposed in Balding et al. (1995); Bruno et al. (1995); Du and Hwang (2006); Erdös and Rényi (1963); Hwang (2000); Hwang and Liu (2001). The first folklore non-adaptive randomized algorithm known from the literature simply chooses a random $m \times n$ test matrix M, for sufficiently large m, where each entry of M is chosen randomly independently to be 0 with probability p and 1 with probability $1 - p$. This type of algorithm is called a *random incidence design* (RID) and is the simplest type of randomized algorithm. In this paper we will study this type of algorithms.

If $I \subseteq S = [n]$ is the set of defective items then the vector of answers to the tests (rows of M) is $T(I, M) := \vee_{i \in I} M^{(i)}$ where $\vee$ is bitwise "or" and $M^{(i)}$ is the ith column of M. The above parameters, m and p, are chosen so that the probability of success of recovering the defective items (decoding) is maximal. If the number of defective items is bounded by $|I| \leq d$ then in order to find the defective items with probability at least $1 - \delta$, we need that with probability at least $1 - \delta$, M satisfies the following property: For every $J \subseteq [n]$, $|J| \leq d$ and $J \neq I$ we have $T(J, M) \neq T(I, M)$. This basically says that the only set J of up to d items that is consistent with the answers of the tests $\wedge_{i \in I} M^{(i)}$ is I. We call such a property (I, d)-*separable property* and M is called an (I, d)-*separable test matrix*. Unfortunately, although this property is what we need to solve the problem, there is no known linear time (or even polynomial time) algorithm that finds the defective items from the answers of an (I, d)-separable test matrix. The only known algorithm is the trivial one that exhaustively goes over all possible up to d sets J and verifies whether $T(J, M)$ is equal to the answers of the tests. This is one of the reasons why the *separability property* has not been much studied in the literature.

Sebö, Sebö (1985), studies the easy case when the number of defective items is *exactly d*. He shows that the best probability for the random test matrix (i.e., that gives a minimum number of tests) is $p = 1 - 1/2^{1/d}$. In this paper, we study the general case when the number of defective items is at most d. We give a new analysis for the general case and show that the probability that gives minimum number of tests is $p = 1 - 1/d$. In particular, we show that

$$\lim_{n \to \infty} \frac{m_{\text{RID}}}{\ln n} \leq \gamma_{d-1} + O\left(\frac{1}{d}\right) := ed - \frac{e+1}{2} + O\left(\frac{1}{d}\right) \tag{1}$$

where m_{RID} is the minimum number of tests in the RID algorithms, $e = 2.71828\cdots$ is the Euler's number and γ_d is defined in (2). But again, no polynomial time algorithm is known for such matrix even for the special case when the number of the defective items is exactly d.

To be able to find the defective items in *linear time* in the size of all the tests, that is, $O(dn \log n)$, and asymptotically optimal $O(d \log n)$ tests, algorithms in the literature use a relaxed property for M. The folklore non-adaptive randomized algorithm randomly chooses M such that with probability at least $1 - \delta$ the following relaxed property holds: for each non defective item $i \in S \backslash I$ there is a test that contains it but does not contain the defective items. Such a test is surely negative (gives answer 0) and is a witness for the fact that item i is not defective. We call such a property an (I, d, i)-*disjunct* property where I is the set of defective elements. If the test matrix is (I, d, i)-disjunct for every

non-defective item i then we call the test matrix (I, d)-*disjunct*. When the test matrix is (I, d)-disjunct, the algorithm that finds the defective items simply starts with $X = S$ and then for every negative answer, $T(I, a) = 0$, removes from X all the items i where $a_i = 1$. From the property of (I, d)-disjunct all the non-defective items are removed, and X will eventually contain the defective items. This can be done in linear time in the size of M. This is why the property of *disjunctness* is well studied in the literature, Balding et al. (1995); Bruno et al. (1995); Du and Hwang (2006); Hwang (2000); Hwang and Liu (2001, 2003, 2004). It is well known (and very easy to see) that if M is (I, d)-disjunct then M is (I, d)-separable, and therefore, this algorithm is required to do more tests to get this property (with probability at least $1 - \delta$). For completeness, we show in Section 2 that the best probability for this property is $p = 1 - 1/(d + 1)$ and the number of tests m satisfies $\lim_{n\to\infty} m/\ln n = \gamma_d = ed + (e - 1)/2 + O(1/d)$ (see the definition of γ_d in (2)) which is greater than the minimum number of test in (1) by the constant e.

In this paper we give a new algorithm that runs in linear time. Our test matrix M has the following property with probability at least $1 - \delta$: It is (I, d)-separable test matrix and (I, d, i)-disjunction for at least $n - n^{1/d}$ non-defective items. Therefore one can eliminate all the non-defective items except at most $n^{1/d}$ items. Then, since M is (I, d)-separable, the exhaustive search of the defective items in $n^{1/d}$ items takes linear time. We call a test matrix with such a property, a *semi-disjunct test matrix*. We show that the number of tests m in this algorithm satisfies $\lim_{n\to\infty} m/\ln n = \gamma_{d-1} + O(1/d)$ which is asymptotically the same as the one for the (I, d)-separable test matrix. Our analysis shows that the number of tests in our algorithm is better than the one with the disjunction property even for small d.

There are also other types of randomized algorithms in the literature, and they all study the property of disjunctness. For example, the *Random r-size design* (RrSD) are algorithms where each row in M is a random vector $a \in \{0, 1\}^n$ of weight (number of ones in a) r. Due to the messiness of the analysis of those algorithms, only limited results have been obtained. See for example Balding et al. (1995); Bruno et al. (1995); Hwang (2000); Hwang and Liu (2001, 2003, 2004). In this paper we show that, for large n, the optimal number of tests in the RrSD algorithms converges to the optimal number of tests in the RID model.

One advantage of the RID and RrSD algorithms over the other randomized algorithm is that, in parallel machines, the tests can be generated by different processors (or laboratories) without any communication between them. In this model all the machines uses the same distribution, draw a sample and do the test. Those algorithms are *strongly non-adaptive* in the sense that the rows of the matrix M can be also non-adaptively generated in one parallel step. Therefore, our algorithm can also find the defective items in parallel in time $O(d \log d \log n)$ and one round of parallel queries.

There are other relaxations of the separability in the literature that give other algorithms in different models that do $O(d \log n)$ tests. Aldridge et al., Aldridge et al. (2017), and De Bonis and Vaccaro Bonis and Vaccaro (2015) define ϵ-almost d-separable matrices. This gives a deterministic (non-polynomial time) algorithm that does $O(d \log n)$ tests that for $1 - \epsilon$ fraction of the sets of defective items I succeeds to detect I. In Barg and Mazumdar (2015), Barg and Mazumdar defines the (d, ϵ)-disjunct matrices that identify a uniform random set I of defective items with false positive probability of an item at most ϵ.

Our paper is organized as follows. In Section 2 we give the analysis for the (I, d)-disjunct test matrix and give the folklore algorithm. In Section 3 we give the analysis for the (I, d)-

separable test matrix. In Section 4 we give our new algorithm. Section 5 contains some open problems.

2. (I, d)-Disjunct Matrix

In this section, we find the probability p for which a random test matrix M has a minimum number of rows (tests) and is (I, d)-disjunct with probability at least $1 - \delta$. The results in this section can be found in Balding et al. (1995); Du and Hwang (2006); Hwang and Liu (2004). We give it here for completeness.

We may assume w.l.o.g that the set of defective items is $I = \{1, 2, \ldots, d'\}$ for some $d' \le d$. Recall that M is (I, d)-disjunct if for every non-defective item $i \in [d' + 1, n]$ there is a row j in M such that $M_{j,1} = \cdots = M_{j,d'} = 0$ and $M_i = 1$. For a random $m \times n$ test matrix M with entries that are chosen independently where each entry is 0 with probability p and 1 with probability $1 - p$ we have

$$
\begin{aligned}
\Pr[M \quad &\text{is not } (I, d)\text{-disjunct}] \\
&= \Pr[(\exists i \in [d' + 1, n])(\forall j) \text{ not } (M_{j,1} = \cdots = M_{j,d'} = 0, M_{j,i} = 1)] \\
&\le n(1 - p^{d'}(1 - p))^m \le n(1 - p^d(1 - p))^m.
\end{aligned}
$$

To have a success probability at least $1 - \delta$, we need that $n(1 - p^d(1 - p))^m \le \delta$ and therefore (here and throughout the paper $\ln(x)^{-1} = \ln(1/x)$ and not $1/\ln(x)$)

$$
m \ge \frac{\ln n + \ln(1/\delta)}{\ln \left(1 - p^d(1 - p)\right)^{-1}}.
$$

To minimize m we need to minimize $1 - p^d(1 - p)$. From the first derivative, we get that the optimal m is obtained when $p = d/(d + 1)$. See also Balding et al. (1995), Theorem 3.6 in Hwang and Liu (2004) and Theorem 5.3.6 in Du and Hwang (2006). Therefore

Theorem 1 *Balding et al. (1995) Let $p = 1 - 1/(d + 1)$ and*

$$
m = \gamma_d \cdot (\ln n + \ln(1/\delta))
$$

where

$$
\gamma_d = \frac{1}{\ln \left(1 - \frac{1}{d}\left(1 - \frac{1}{d+1}\right)^{d+1}\right)^{-1}}. \tag{2}
$$

With probability at least $1 - \delta$ the $m \times n$ test matrix M is (I, d)-disjunct.

In particular,

Theorem 2 *There is a linear time, $O(dn \log n)$, randomized algorithm that does*

$$
m = \gamma_d \cdot (\ln n + \ln(1/\delta))
$$

tests and with probability at least $1 - \delta$ finds the defective items.

The algorithm is in Figure 1.

To determine the asymptotic behavior of m we use (11) and (8) in Lemma 6 in Appendix A and get

$$\lim_{n \to \infty} \frac{m}{\ln n} = \gamma_d = ed + \frac{e-1}{2} - \frac{e^2+2}{24 \cdot e} \cdot \frac{1}{d} + O\left(\frac{1}{d^2}\right). \tag{3}$$

Folklore Random Group Test(n, d, δ)

1) $m \leftarrow \gamma_d(\ln n + \ln(1/\delta))$
2) Let M be $m \times n$ test matrix with entries 1 with
   probability $1/(d+1)$ and 0 with probability $1 - 1/(d+1)$.
3) $X \leftarrow [n]$.
4) For $i = 1$ to m do (in parallel)
   4.1) Test row i in $M \to \text{ANSWER}_i$
   4.2) If the $\text{ANSWER}_i = 0$ then
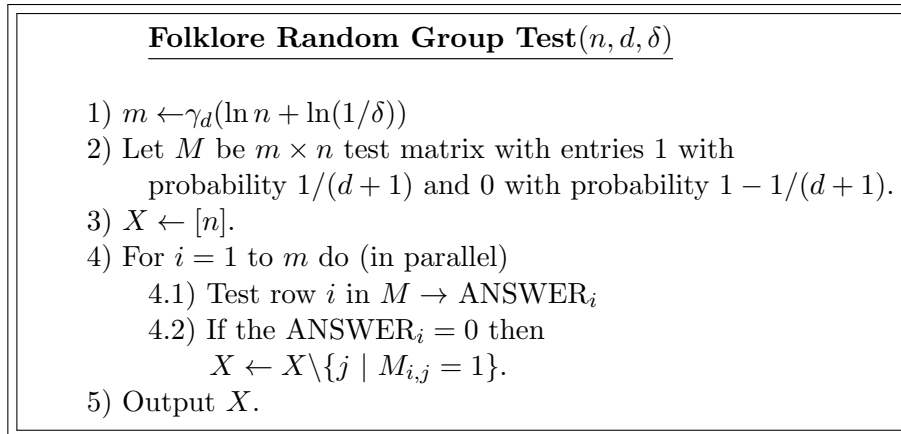 $X \leftarrow X \setminus \{j \mid M_{i,j} = 1\}$.
5) Output X.

Figure 1: An algorithm using disjunct test matrix.

3. (I, d)-Separable Matrix

In this section, we find the probability p for which a test matrix M with a minimum number of rows (tests) is (I, d)-separable with probability at least $1 - \delta$. We show that, using the union bound method, the optimal number of tests is obtained when $p = 1 - 1/d$ and the number of tests is (see (2))

$$m = \gamma_{d-1}(\ln n + \ln(1/\delta) + O(d)).$$

Notice that separability is the minimal condition we need to make sure that answers to the tests uniquely determine the defective items. Unfortunately, there is no polynomial time algorithm to find the defective items from separable matrices. In the next section, we show that the same number of tests (up to additive $o(d)$ in γ_{d-1}) can be achieved with what we will call the semi-disjunct matrices and with such matrix one can find the defective items in linear time. So the goal of this section is to give the minimum number of tests that is needed, and the next section will show that it is asymptotically achievable with linear time decoding.

Let $I \subseteq [n]$. Recall that for an assignment $b \in \{0, 1\}^n$, $T(I, b) = 1$ if $b_j = 1$ for some $j \in I$ and 0 otherwise. This is the result of the test of b when I is the set of defective items. For an $m \times n$ test matrix M, let $T(I, M)$ the column vector $(T(I, M_i))_{i=1,\ldots,m}$ where M_i is the ith row of M. Recall that M is (I, d)-separable matrix if for any set $J \neq I$ of up to d items $T(I, M) \neq T(J, M)$.

Suppose $I = \{1, 2, \ldots, d_1\}$ are the defective items where $d_1 \leq d$. For $d_2 \leq d$ and $0 \leq k \leq \min(d_1, d_2)$ let $\mathcal{B}(n, d_2, k)$ be the set of all sets $J \subseteq [n]$ of size $|J| = d_2$ where $|I \cap J| = k$. The number of such sets is

$$|\mathcal{B}(n, d_2, k)| = \binom{d_1}{k}\binom{n - d_1}{d_2 - k}.$$

Now let $a \in \{0, 1\}^n$ be a random assignment where $a_i = 0$ with probability p and 1 with probability $1 - p$. The probability that $T(I, a) \neq T(J, a)$ is the probability that $a_i = 0$ for all $i \in I \cap J$ and either (1) $a_i = 0$ for all $i \in I \backslash J$ and $a_j = 1$ for some $j \in J \backslash I$ or (2) $a_i = 0$ for all $i \in J \backslash I$ and $a_j = 1$ for some $j \in I \backslash J$. Therefore

$$
\begin{aligned}
\Pr[T(I, a) = T(J, a)] &= 1 - \left((1 - p^{d_1 - k})p^{d_2} + (1 - p^{d_2 - k})p^{d_1} \right) \\
&= 1 - p^{d_2} - p^{d_1} + 2p^{d_1 + d_2 - k}. \quad\quad (4)
\end{aligned}
$$

Therefore for a random $m \times n$ test matrix M, the probability that M is not (I, d)-separable is the probability that $T(I, M) = T(J, M)$ for some $J \neq I$. This is

$$
\begin{aligned}
\Pr[\ (\exists d_2, k) \ \ (\exists J \in \mathcal{B}(n, d_2, k)) \ \ &T(I, M) = T(J, M)] \\
\leq \quad & \sum_{d_2, k} \binom{d_1}{k}\binom{n - d_1}{d_2 - k}\left(1 - p^{d_2} - p^{d_1} + 2p^{d_1 + d_2 - k} \right)^m. \\
\leq \quad & d^2 2^d \max_{d_1, d_2, k} n^{d_2 - k}\left(1 - p^{d_2} - p^{d_1} + 2p^{d_1 + d_2 - k} \right)^m \\
\leq \quad & d^2 2^d \max_{d_1, d_2, k} P(n, d_1, d_2, k, p)
\end{aligned}
$$

where

$$P(n, d_1, d_2, k, p) := n^{d_2 - k}\left(1 - p^{d_2} - p^{d_1} + 2p^{d_1 + d_2 - k} \right)^m.$$

Denote

$$\Pi = \max_{d_1, d_2, k} P(n, d_1, d_2, k, p).$$

The above implies

Lemma 3 *If $\Pi \leq \delta/(d^2 2^d)$ then with probability at least $1 - \delta$ the test matrix M is (I, d)-separable.*

To find Π, we prove in Appendix A that

Lemma 4 *We have*

$$
\begin{aligned}
\Pi &= \max_{d_1, d_2, k} P(n, d_1, d_2, k, p) \\
&= \max_{0 \leq w \leq d - 1} \max\{(P(n, d, d, w, p), P(n, w, d, w, p)\} \\
&= \max_{0 \leq w \leq d - 1} \max\{n^{d - w}(1 - 2p^d + 2p^{2d - w})^m, n^{d - w}(1 + p^d - p^w)^m\}
\end{aligned}
$$

Now by Lemma 3 and 4 it immediately follows

Lemma 5 *If*

$$m = \max_{0 \le w \le d-1} \frac{(d-w)\ln n + \ln(1/\delta) + \ln(d^2 2^d)}{\min(\ln(1 - 2p^d + 2p^{2d-w})^{-1}, \ln(1 + p^d - p^w)^{-1})}$$

then the probability that random $m \times n$ test matrix M is (I, d)-separable is at least $1 - \delta$.

In particular, for the minimum number of tests m we have

$$\lim_{n \to \infty} \frac{m}{\ln n} = \min_p \max_{0 \le w \le d-1} \max\left(T_1(d, w, p), T_2(d, w, p)\right)$$

where

$$T_1(d, w, p) = \frac{d-w}{\ln(1 + p^d - p^w)^{-1}}, \quad T_2(d, w, p) = \frac{d-w}{\ln(1 - 2p^d + 2p^{2d-w})^{-1}}.$$

Therefore, our goal now is to find the probability p that minimizes

$$T = \max_{0 \le w \le d-1} \max\left(T_1(d, w, p), T_2(d, w, p)\right).$$

The probability that gives the minimum T is denoted by p^*. It is easy to see that

$$T_1^*(d, w) := \min_{0 \le p \le 1} T_1(d, w, p) = \frac{d-w}{\ln(1 + p_{1,w}^d - p_{1,w}^w)^{-1}}$$

with the global minimum point

$$p_{1,w} = \left(\frac{w}{d}\right)^{\frac{1}{d-w}}$$

and

$$T_2^*(d, w) := \min_{0 \le p \le 1} T_2(d, w, p) = \frac{d-w}{\ln(1 - 2p_{2,w}^d + 2p_{2,w}^{2d-w})^{-1}}$$

with the global minimum point

$$p_{2,w} = \left(\frac{d}{2d-w}\right)^{\frac{1}{d-w}}.$$

To find p^* we will first use the following simple fact

Lemma 6 *Let $f_0, f_1, \ldots, f_t$ be functions $[0, 1] \to \Re \cup \{\infty\}$. If x_0 is a global minimum point for f_0 and $f_0(x_0) > f_i(x_0)$ for all $i = 1, \ldots, t$ then the global minimum point of $f = \max_{0 \le i \le t} f_i$ is x_0 and $\min_{0 \le x \le 1} f = f_0(x_0)$.*

Proof From the definition of f we have $f(x) \ge f_0(x) \ge f_0(x_0)$ for all x. On the other hand since $f_0(x_0) > f_i(x_0)$ for all $i = 1, \ldots, t$ we have $f(x_0) = \max_{0 \le i \le t} f_i(x_0) = f(x_0)$. Therefore x_0 is global minimum of f. ∎

In particular, we now show that

Lemma 7 *We have $p^* = p_{1,d-1} = 1 - 1/d$ and*

$$
\begin{aligned}
\lim_{n \to \infty} \frac{m}{\ln n} &= \min_{p} T = T_1^*(d, d-1) = T_1(d, d-1, p^*) \\
&= \frac{1}{\ln\left(1 - \frac{1}{d}\left(1 - \frac{1}{d}\right)^{d-1}\right)^{-1}} = \frac{1}{\ln\left(1 - \frac{1}{d-1}\left(1 - \frac{1}{d}\right)^{d}\right)^{-1}} = \gamma_{d-1}.
\end{aligned}
$$

Proof We use Lemma 6. The point $p^* = p_{1,d-1} = 1 - 1/d$ is a global minimum point of $T_1(d, d-1, p)$. We now show that $T_1(d, w, p^*) \le T_1(d, d-1, p^*)$ and $T_2(d, w, p^*) \le T_1(d, d-1, p^*)$ for all $w = 0, 1, \ldots, d-1$.

The proof of the fact $T_1(d, w, p^*) \le T_1(d, d-1, p^*)$ is in Lemma 8 in Appendix A and the proof of the fact $T_2(d, w, p^*) \le T_1(d, d-1, p^*)$ is in Lemma 9 in Appendix A. ■

This implies

Theorem 8 *Let $p = 1 - 1/d$ and*

$$
m = \gamma_{d-1}(\ln n + \ln(1/\delta) + O(d))
$$

where

$$
\gamma_{d-1} = \frac{1}{\ln\left(1 - \frac{1}{d-1}\left(1 - \frac{1}{d}\right)^{d}\right)^{-1}}.
$$

With probability at least $1 - \delta$ the $m \times n$ test matrix M is (I, d)-separable.

By (3) we get

$$
\gamma_{d-1} = ed - \frac{e+1}{2} + O\left(\frac{1}{d}\right). \tag{5}
$$

The following table gives the coefficient of $\ln n + \ln(1/\delta)$

d	(I, d)-Disjunct	(I, d)-Seperable
2	6.2366	3.4761
3	8.9722	6.2366
4	11.6999	8.9722
5	14.4241	11.6999
6	17.1465	14.4241
7	19.8678	17.1465

The difference of the coefficients in the sizes is

$$
\gamma_d - \gamma_{d-1} = e + O\left(\frac{1}{d}\right) \longrightarrow e.
$$

4. An Algorithm with Linear Decoding Time

In this section we give our algorithm, show that the number of tests in this algorithm is asymptotically equal to the one with the (I, d)-separability property and runs in linear time.

We say that a matrix M is (I, d)-*semi-disjunct matrix* if it is (I, d)-separable and (I, d, i)-disjunct for at least $n - n^{1/d}$ items i.

We prove

Theorem 9 *Let $p = 1 - 1/d$ and*

$$m = \frac{\left(1 - \frac{1}{d}\right) \ln n + \ln \frac{1}{\delta} + O(d)}{\ln \left(1 - \frac{1}{d} \left(1 - \frac{1}{d}\right)^d\right)^{-1}}.$$

With probability at least $1 - \delta$ the $m \times n$ test matrix M is (I, d)-semi-disjunct matrix.

In particular we prove

Theorem 10 *There is a linear time non-adaptive randomized algorithm that does*

$$m = \frac{\left(1 - \frac{1}{d}\right) \ln n + \ln \frac{1}{\delta} + O(d)}{\ln \left(1 - \frac{1}{d} \left(1 - \frac{1}{d}\right)^d\right)^{-1}}$$

tests and with probability at least $1 - \delta$ finds the defective items.

In particular,

$$\frac{m}{\ln n} = \frac{1 - \frac{1}{d}}{\ln \left(1 - \frac{1}{d} \left(1 - \frac{1}{d}\right)^d\right)^{-1}} = \gamma_{d-1} + O\left(\frac{1}{d}\right) \to \gamma_{d-1}$$

is asymptotically equal to the constant that is achieved by the (I, d)-separability property.

We first prove Theorem 9
Proof of Theorem 9
Proof In Lemma 10 in Appendix A, we show that

$$\frac{\left(1 - \frac{1}{d}\right)}{\ln \left(1 - \frac{1}{d} \left(1 - \frac{1}{d}\right)^d\right)^{-1}} \geq \gamma_{d-1}.$$

Therefore $m \geq \gamma_{d-1}(\ln n + \ln(1/\delta) + O(d))$ and by Theorem 8, with probability at least $1 - (\delta/2)$, M is (I, d)-separable. Note that $\delta/2$ adds a constant to the second term that is swallowed by $O(d)$.

Let w.l.o.g $I = \{1, \ldots, d'\}$ where $d' \leq d$. Define for every $k \notin I$, a random variable X_k that takes the values $\{0, 1\}$ and is equal to 0 if there is a row j in the test matrix such that $M_{j,i} = 0$ for all $i \in I$ and $M_{j,k} = 1$. That is, M is (I, d, k)-disjunct. If so then we know that $k \notin I$. Let $D = \{k | X_k = 1\}$. The set D is the set of items that are not (I, d, k)-disjunct. We now find the expected size of D. We first have

$$\mathbf{E}[X_k] \leq (1 - p^d(1 - p))^m \leq \frac{\delta}{2n^{1 - \frac{1}{d}}}.$$

10

Then

$$\mathbf{E}[|D|] = \mathbf{E}\left[\sum_{i \notin I} X_k\right] = \sum_{i \notin I} \mathbf{E}[X_k] \le (\delta/2)n^{1/d}.$$

By Markov bound

$$\Pr[|D| \ge n^{1/d}] \le \frac{\delta}{2}$$

and therefore with probability at least $1 - (\delta/2)$ we have that $|D| \le n^{1/d}$.

Therefore, with probability at least $1 - \delta$ the test matrix M is (I, d)-separable and (I, d, i)-disjunct for at least $|[n]\backslash D| \ge n - n^{1/d}$ items. ∎

Theorem 10 is proved in the next section.

In the table we compare the size of the (I, d)-separable, (I, d)-disjunct and (I, d)-semi-disjunct.

d	(I, d)-Disjunct	(I, d)-Separable	(I, d)-Semi-Disjunct
2	6.2366	3.4761	3.7444
3	8.9722	6.2366	6.4109
4	11.6999	8.9722	9.1013
5	14.4241	11.6999	11.8025
6	17.1465	14.4241	14.5093
7	19.8678	17.1465	17.2193

For the (n, d)-semi-disjunct matrix, the constant in the table is

$$\begin{aligned}
\frac{m}{\ln n} &= \frac{\left(1 - \frac{1}{d}\right)}{\ln\left(1 - \frac{1}{d}\left(1 - \frac{1}{d}\right)^d\right)^{-1}} \\
&= ed - \frac{e+1}{2} - \frac{e^2 - 12e + 2}{24e}\frac{1}{d} + O\left(\frac{1}{d^2}\right) = \gamma_{d-1} + O\left(\frac{1}{d}\right)
\end{aligned}$$

and therefore the constant for the (I, d)-Separable (the third column in the table) converges to the constant of the (I, d)-Semi-Disjunct (the forth column in the table).

4.1. The Algorithm

In our algorithm, we choose $p = 1 - 1/d$ and

$$m = \frac{\left(1 - \frac{1}{d}\right)\ln n + \ln \frac{1}{\delta} + O(d)}{\ln\left(1 - \frac{1}{d}\left(1 - \frac{1}{d}\right)^d\right)^{-1}}.$$

We now prove

Theorem 5 *There is a linear time, $O(dn \log n)$, randomized algorithm that does (Here $o(1) = O(1/d)$)*

$$m = \frac{\left(1 - \frac{1}{d}\right)\ln n + \ln \frac{1}{\delta} + O(d)}{\ln\left(1 - \frac{1}{d}\left(1 - \frac{1}{d}\right)^d\right)^{-1}} = (\gamma_{d-1} + o(1)) \cdot \ln n + \gamma_d \ln(1/\delta) + O(\gamma_d d)$$

tests and with probability at least $1 - \delta$ finds the defective items.

Proof Consider the random $m \times n$ test matrix M with entries 0 with probability $p = 1 - 1/d$ and 1 with probability $1/d$. Let I be set of defective items. If $|I| < d$ then since $m \geq \gamma_{d-1}(\ln n + \ln(1/\delta) + O(d))$, by Theorem 8, with probability at least $1 - \delta$ the test matrix M is $(I, d-1)$-disjunct test matrix. Therefore, if the number of defective items is less than d then with probability at least $1 - \delta$ the algorithm in Section 2 finds the defective items. Therefore we may assume that the number of defective items is exactly d.

Now, if the number of defective items is d then, by Theorem 9, with probability at least $1 - \delta$, the test matrix M is (I, d)-semi-disjunct. Therefore by the algorithm in Section 2 all the non-defective items are eliminated except at most $n^{1/d}$ items. Since M is also (n, I)-separable matrix the only set J that satisfies $T(J, M) = T(I, M)$ is I. So we can exhaustively search for the d defective items in $n^{1/d}$ items. This takes time

$$\binom{n^{1/d}}{d} dm \leq \left(\frac{en^{1/d}}{d}\right)^d d^2 \log n = e\left(\frac{e}{d}\right)^{d-1} dn \log n = O(dn \log n)$$

and therefore linear time. ∎

The algorithm is in Figure 2. In the algorithm, $M^{(i)}$ is the ith column of M. ANSWER$_i$ is 0 if the test is negative and 1 if it is positive. The correctness of the algorithm follows from the above analysis.

Random Group Test(n, d, δ)

1) $m \leftarrow \dfrac{\left(1 - \frac{1}{d}\right)\ln n + \ln\frac{1}{\delta} + d\ln 2 + 2\ln d}{\ln\left(1 - \frac{1}{d}\left(1 - \frac{1}{d}\right)^d\right)^{-1}}.$

2) Let M be $m \times n$ test matrix with entries 1 with probability $1/d$ and 0 with probability $1 - 1/d$.

3) $X \leftarrow [n]$.

4) For $i = 1$ to m do (in parallel)
    4.1) Test row i in $M \rightarrow$ ANSWER$_i$
    4.2) If the ANSWER$_i = 0$ then
        $X \leftarrow X \backslash \{j \mid M_{i,j} = 1\}$.

5) If $|X| \leq d$ then return X and halt
    * Otherwise $|X| = d$ *\

6) For every $Y \subseteq X$ s.t. $|Y| = d$
    6.1) If for all i, $\vee_{j \in Y} M_i^{(j)} =$ANSWER$_i$ then
        Output Y and halt.

Figure 2: An algorithm using semi-disjunct test matrix.

5. Open Problems

In this section we give some open problems

1. Our analysis for (I, d)-separable matrices gives the bound

$$\lim_{n\to\infty} \frac{m_{\text{RID}}}{\ln n} \leq \gamma_{d-1} := ed - \frac{e+1}{2} + O\left(\frac{1}{d}\right) \tag{6}$$

   We still do not know whether this bound is tight in the RID model. Also, no lower bound is known except the information theoretic lower bound $d \log n$.

2. We show in Section 2 that the best probability for the disjunctness property is $p = 1 - 1/(d+1)$. This is based on maximizing the expected number of elements that are eliminated for each test. This is not necessarily the best probability for obtaining a minimum number of tests. Find the probability that gives the minimum number of tests. Find a lower bound for the number of tests.

3. In Section 5 we show that the RrSD model converges to the RID model. Consider the most general strong non-adaptive model (RSN model) where each row in M is chosen randomly independently according to some distribution D over $\{0,1\}^n$. Show that the algorithm that do the minimum number of tests is the RsSD algorithm with $r = n/d$.

References

Matthew Aldridge, Leonardo Baldassini, and Karen Gunderson. Almost separable matrices. *J. Comb. Optim.*, 33(1):215–236, 2017. doi: 10.1007/s10878-015-9951-1. URL https://doi.org/10.1007/s10878-015-9951-1.

D. J. Balding, W. J. Bruno, E. Knill, and D. C. Torney. A comparative survey of non-adaptive pooling designs. genetic mapping and dna sequencing. *IMA volumes in mathematics and its applications. Springer-Verlag.*, pages 133–155, 1995.

Alexander Barg and Arya Mazumdar. Almost disjunct matrices from codes and designs. *CoRR*, abs/1510.02873, 2015. URL http://arxiv.org/abs/1510.02873.

Annalisa De Bonis and Ugo Vaccaro. ϵ-almost selectors and their applications. In *Fundamentals of Computation Theory - 20th International Symposium, FCT 2015, Gdańsk, Poland, August 17-19, 2015, Proceedings*, pages 255–268, 2015. doi: 10.1007/978-3-319-22177-9_20. URL https://doi.org/10.1007/978-3-319-22177-9_20.

W. J. Bruno, E. Knill, D. J. Balding, D. C. Bruce, N. A. Doggett, W. W. Sawhill, R. L. Stalling, C. C. Whitaker, and D. C. Torney. Efficient pooling designs for library screening. *Genomics*, pages 21–30, 1995.

Nader H. Bshouty. Exact learning from membership queries: Some techniques, results and new directions. In *Algorithmic Learning Theory - 24th International Conference, ALT 2013, Singapore, October 6-9, 2013. Proceedings*, pages 33–52, 2013. doi: 10.1007/978-3-642-40935-6_4. URL https://doi.org/10.1007/978-3-642-40935-6_4.

Nader H. Bshouty. Exact learning from an honest teacher that answers membership queries. *CoRR*, abs/1706.03935, 2017. URL http://arxiv.org/abs/1706.03935.

Ferdinando Cicalese. *Fault-Tolerant Search Algorithms - Reliable Computation with Unreliable Information.* Monographs in Theoretical Computer Science. An EATCS Series. Springer, 2013. ISBN 978-3-642-17326-4. doi: 10.1007/978-3-642-17327-1. URL https://doi.org/10.1007/978-3-642-17327-1.

Graham Cormode and S. Muthukrishnan. What's hot and what's not: tracking most frequent items dynamically. *ACM Trans. Database Syst.*, 30(1):249–278, 2005. doi: 10.1145/1061318.1061325. URL http://doi.acm.org/10.1145/1061318.1061325.

R. Dorfman. The detection of defective members of large populations. *Ann. Math. Statist.*, pages 436–440, 1943.

D. Du and F. K Hwang. Combinatorial group testing and its applications. *World Scientific Publishing Company.*, 2000.

D. Du and F. K Hwang. Pooling design and nonadaptive group testing: important tools for dna sequencing. *World Scientific Publishing Company.*, 2006.

A. G. D'yachkov and V. V. Rykov. Bounds on the length of disjunctive codes. *Problemy Peredachi Informatsii*, 18:7–13, 1982.

P. Erdös and A. Rényi. On two problems of information theory. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 8:241–254, 1963.

Zoltán Füredi. On r-cover-free families. *J. Comb. Theory, Ser. A*, 73(1):172–173, 1996. doi: 10.1006/jcta.1996.0012. URL https://doi.org/10.1006/jcta.1996.0012.

Edwin S. Hong and Richard E. Ladner. Group testing for image compression. *IEEE Trans. Image Processing*, 11(8):901–911, 2002. doi: 10.1109/TIP.2002.801124. URL https://doi.org/10.1109/TIP.2002.801124.

F. K. Hwang. A method for detecting all defective members in a population by group testing. *Journal of the American Statistical Association*, 67:605–608, 1972.

F. K. Hwang. Random k-set pool designs with distinct columns. *Probability in Engneering and Informational Sciences*, 14:49–56, 2000.

F. K. Hwang and Y. C. Liu. The expected number of unresolved positive clones in various random pool designs. *Probability in Engineering and Informational Sciences*, 15:57–68, 2001.

F. K. Hwang and Y. C. Liu. A general approach to compute the probabilities of unresolved clones in random pooling designs. *Probability in the Engineering and Informational Sciences*, 18:161–183, 2004.

Frank K. Hwang and Y. C. Liu. Random pooling designs under various structures. *J. Comb. Optim.*, 7(4):339–352, 2003. doi: 10.1023/B:JOCO.0000017382.83399.0b. URL https://doi.org/10.1023/B:JOCO.0000017382.83399.0b.

William H. Kautz and Richard C. Singleton. Nonrandom binary superimposed codes. *IEEE Trans. Information Theory*, 10(4):363–377, 1964. doi: 10.1109/TIT.1964.1053689. URL https://doi.org/10.1109/TIT.1964.1053689.

C. H. Li. A sequential method for screening experimental variables. *J. Amer. Statist. Assoc.*, 57:455–477, 1962.

Anthony J. Macula and Leonard J. Popyack. A group testing method for finding patterns in data. *Discrete Applied Mathematics*, 144(1-2):149–157, 2004. doi: 10.1016/j.dam.2003.07.009. URL https://doi.org/10.1016/j.dam.2003.07.009.

Hung Q. Ngo and Ding-Zhu Du. A survey on combinatorial group testing algorithms with applications to DNA library screening. In *Discrete Mathematical Problems with Medical Applications, Proceedings of a DIMACS Workshop, December 8-10, 1999*, pages 171–182, 1999.

Ely Porat and Amir Rothschild. Explicit nonadaptive combinatorial group testing schemes. *IEEE Trans. Information Theory*, 57(12):7982–7989, 2011. doi: 10.1109/TIT.2011.2163296. URL https://doi.org/10.1109/TIT.2011.2163296.

Miklós Ruszinkó. On the upper bound of the size of the r-cover-free families. *J. Comb. Theory, Ser. A*, 66(2):302–310, 1994.

A. Sebö. On two random search problems. *J. of statistical planning and inference*, 11:23–31, 1985.

M. Sobel and P. A. Groll. Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Tech. J.*, 38:1179–1252, 1959.

Wikipedia. Group testing. URL https://en.wikipedia.org/wiki/Grouptesting.

Jack K. Wolf. Born again group testing: Multiaccess communications. *IEEE Trans. Information Theory*, 31(2):185–191, 1985. doi: 10.1109/TIT.1985.1057026. URL https://doi.org/10.1109/TIT.1985.1057026.

Appendix A. Proofs

Proof of Lemma 4

We need to prove

$$
\begin{aligned}
\Pi &= \max_{d_1, d_2, k} P(n, d_1, d_2, k, p) \\
&= \max_{0 \leq w \leq d-1} \max\{(P(n, d, d, w, p), P(n, w, d, w, p)\}.
\end{aligned}
$$

Proof If $1 - 2p^{d_2-k} \geq 0$ then $d_2 - k \geq 1$ and

$$
\begin{aligned}
P(n, d_1, d_2, k, p) &= n^{d_2-k} \left(1 - p^{d_2} - p^{d_1} + 2p^{d_1+d_2-k} \right)^m \\
&= n^{d_2-k} \left(1 - p^{d_2} - p^{d_1}(1 - 2p^{d_2-k}) \right)^m \\
&\leq n^{d_2-k} \left(1 - p^{d_2} - p^{d}(1 - 2p^{d_2-k}) \right)^m \\
&\leq n^{d_2-k} \left(1 - p^{d} - p^{d}(1 - 2p^{d_2-k}) \right)^m \\
&= n^{d_2-k} \left(1 - 2p^{d} + 2p^{d+d_2-k} \right)^m \\
&= P(n, d, d, d - (d_2 - k), p) \\
&\leq \max_{0 \leq w \leq d-1} P(n, d, d, w, p).
\end{aligned}
$$

If $1 - 2p^{d_2-k} < 0$ and $d_2 > k$ then

$$
\begin{aligned}
P(n, d_1, d_2, k, p) &= n^{d_2-k} \left(1 - p^{d_2} - p^{d_1} + 2p^{d_1+d_2-k} \right)^m \\
&= n^{d_2-k} \left(1 - p^{d_2} - p^{d_1}(1 - 2p^{d_2-k}) \right)^m \\
&\leq n^{d_2-k} \left(1 - p^{d_2} - p^{k}(1 - 2p^{d_2-k}) \right)^m \\
&= n^{d_2-k} \left(1 + p^{d_2} - p^{k} \right)^m \\
&= n^{d_2-k} \left(1 + p^{k}(p^{d_2-k} - 1) \right)^m \\
&\leq n^{d_2-k} \left(1 + p^{k+d-d_2}(p^{d_2-k} - 1) \right)^m \\
&= n^{d_2-k} \left(1 + p^{d} - p^{d-(d_2-k)} \right)^m \\
&= P(n, d - (d_2 - k), d, d - (d_2 - k), p) \\
&\leq \max_{0 \leq w \leq d-1} P(n, w, d, w, p).
\end{aligned}
$$

If $d_2 = k$ then $d_1 > d_2$ (otherwise $J = I$) and

$$
\begin{aligned}
P(n, d_1, d_2, k, p) &= \left(1 - p^{d_2} + p^{d_1} \right)^m \\
&\leq (1 - p^{d_1-1} + p^{d_1})^m \\
&= (1 - p^{d_1-1}(1 - p))^m \\
&\leq (1 - p^{d-1}(1 - p))^m \\
&= P(n, d, d - 1, d - 1, p) \\
&\leq P(n, d - 1, d, d - 1, p) \\
&\leq \max_{0 \leq w \leq d-1} P(n, w, d, w, p).
\end{aligned}
$$

∎

The following are used to estimate some of the expressions in the paper

Lemma 6

$$\left(1 - \frac{1}{d}\right)^d = \frac{1}{e} - \frac{1}{2e}\frac{1}{d} - \frac{5}{24e}\frac{1}{d^2} + O\left(\frac{1}{d^3}\right) \tag{7}$$

$$\left(1 - \frac{1}{d+1}\right)^{d+1} = \frac{1}{e} - \frac{1}{2e}\frac{1}{d} + \frac{7}{24e}\frac{1}{d^2} + O\left(\frac{1}{d^3}\right) \quad and$$

$$\left(1 - \frac{1}{d+1}\right)^{-(d+1)} = e + \frac{e}{2}\frac{1}{d} + \frac{e}{24}\frac{1}{d^2} + O\left(\frac{1}{d^3}\right). \tag{8}$$

$$\frac{1}{d+1} = \frac{1}{d} - \frac{1}{d^2} + \frac{1}{d^3} - \cdots \quad and \quad \frac{1}{d-1} = \frac{1}{d} + \frac{1}{d^2} + \frac{1}{d^3} + \cdots \tag{9}$$

$$\ln(1-x)^{-1} = x + \frac{1}{2}x^2 + \frac{1}{3}x^3 + O(x^4). \tag{10}$$

$$\frac{1}{\ln(1-x)^{-1}} = \frac{1}{x} - \frac{1}{2} - \frac{1}{12}x - \frac{1}{24}x^2 - \frac{19}{720}x^3 + O(x^4). \tag{11}$$

$$\frac{1}{A + Bx + Cx^2 + O(x^3)} = \frac{1}{A} - \frac{B}{A^2}x + \frac{B^2 - CA}{A^3}x^2 + O(x^3). \tag{12}$$

$$e^{A+Bx+Cx^2+O(x^3)} = e^A\left(1 + Bx + \left(\frac{B^2}{2} + C\right)x^2 + O(x^3)\right). \tag{13}$$

The following lemma will be frequently used in the sequel

Lemma 7 *Let $0 < z_1 < z_2 < 1$ then*

$$\frac{\ln(1-z_1)^{-1}}{\ln(1-z_2)^{-1}} \leq \frac{z_1}{z_2}$$

and

$$\frac{\ln(1-z_2)^{-1}}{\ln(1-z_1)^{-1}} \leq \frac{z_2}{z_1}.$$

Proof By (10) in Lemma 6 we have

$$\frac{\ln(1-z_1)^{-1}}{\ln(1-z_2)^{-1}} = \frac{z_1}{z_2}\frac{\sum_{i=1}^{\infty}\frac{1}{i}z_1^{i-1}}{\sum_{i=1}^{\infty}\frac{1}{i}z_2^{i-1}} \leq \frac{z_1}{z_2}.$$

∎

Lemma 8 *We have $T_1(d, w, p^*) \leq T_1(d, d-1, p^*)$.*

Proof By Lemma 7, we have

$$(d-w)\frac{T_1(d,d-1,p^*)}{T_1(d,w,p^*)} = \frac{\ln\left(1+\left(\frac{d-1}{d}\right)^d - \left(\frac{d-1}{d}\right)^w\right)^{-1}}{\ln\left(1+\left(\frac{d-1}{d}\right)^d - \left(\frac{d-1}{d}\right)^{d-1}\right)^{-1}}$$

$$\geq \frac{\left(\frac{d-1}{d}\right)^w - \left(\frac{d-1}{d}\right)^d}{\left(\frac{d-1}{d}\right)^{d-1} - \left(\frac{d-1}{d}\right)^d} = (d-1)\left(\left(1+\frac{1}{d-1}\right)^{d-w} - 1\right)$$

$$\geq (d-1)\left(\left(1+\frac{d-w}{d-1}\right) - 1\right) = d-w.$$

■

Lemma 9 *We have $T_2(d,w,p^*) \leq T_1(d,d-1,p^*)$.*

Proof Since $f(x) = (1-x/z)^{1/x}$ is monotonically decreasing function for $x \in [0,z]$ we have, for $w \in [1, d-1]$,

$$\left(1 - \frac{d-w}{2(d-1)}\right)^{\frac{1}{d-w}} \geq 1 - \frac{1}{2(d-1)} \geq 1 - \frac{1}{d}.$$

This is equivalent to

$$2\left(1 - \left(1 - \frac{1}{d}\right)^{d-w}\right) \geq \frac{d-w}{d-1}. \tag{14}$$

Now by (14) and Lemma 7 we have

$$(d-w)\frac{T_1(d,d-1,p^*)}{T_2(d,w,p^*)} = \frac{\ln\left(1 - 2\left(\frac{d-1}{d}\right)^d + 2\left(\frac{d-1}{d}\right)^{2d-w}\right)^{-1}}{\ln\left(1 - \frac{1}{d-1}\left(1-\frac{1}{d}\right)^d\right)^{-1}}$$

$$= \frac{\ln\left(1 - 2\left(1 - \left(1-\frac{1}{d}\right)^{d-w}\right)\left(1-\frac{1}{d}\right)^d\right)^{-1}}{\ln\left(1 - \frac{1}{d-1}\left(1-\frac{1}{d}\right)^d\right)^{-1}}$$

$$\geq \frac{\ln\left(1 - \frac{d-w}{d-1}\left(1-\frac{1}{d}\right)^d\right)^{-1}}{\ln\left(1 - \frac{1}{d-1}\left(1-\frac{1}{d}\right)^d\right)^{-1}} \geq d-w$$

■

Lemma 10 *We have*

$$\frac{\left(1 - \frac{1}{d}\right)}{\ln\left(1 - \frac{1}{d}\left(1 - \frac{1}{d}\right)^d\right)^{-1}} \geq \gamma_{d-1}.$$

Proof By Lemma 7 we have

$$
\frac{1/\gamma_{d-1}}{\ln\left(1 - \frac{1}{d}\left(1 - \frac{1}{d}\right)^d\right)^{-1}} = \frac{\ln\left(1 - \frac{1}{d-1}\left(1 - \frac{1}{d}\right)^d\right)^{-1}}{\ln\left(1 - \frac{1}{d}\left(1 - \frac{1}{d}\right)^d\right)^{-1}}
$$

$$
\geq \frac{\frac{1}{d-1}}{\frac{1}{d}} = \frac{1}{1 - \frac{1}{d}}.
$$

∎

The following lemma can be proved by induction

Lemma 11 *For $0 \leq x_i \leq 1$ and $w_i > 0$, $i = 1, \ldots, n$ we have*

$$\prod_{i=1}^{n}(x_i - w_i) \geq \left(\prod_{i=1}^{n} x_i\right) - \sum_{i=1}^{n} w_i.$$

Appendix B. Equivalent Models

In this section we show that, for large n, the number of tests in this RrSD algorithm converges to the number of tests in the RID algorithm. This shows that our RID algorithm in this paper gives the best number of tests in the strong non-adaptive models. I.e., in the models where the rows of the matrix M is chosen randomly independently.

We now prove

Theorem 12 *In the RrSD model, for large enough n, the algorithm that gives the minimum number of tests is the one with $r = n/d$. The number of tests converges to the number of tests of the best RID algorithm with $p = 1/d$.*

Proof Following the analysis in Section 3, let $a \in \{0, 1\}^n$ be a random uniform assignment of weight r. The probability that $T(I, a) \neq T(J, a)$ is the probability that $a_i = 0$ for all $i \in I \cap J$ and either (1) $a_i = 0$ for all $i \in I \setminus J$ and $a_j = 1$ for some $j \in J \setminus I$ or (2) $a_i = 0$ for all $i \in J \setminus I$ and $a_j = 1$ for some $j \in I \setminus J$. Therefore

$$\Pr[T(I, a) = T(J, a)] = 1 - \frac{\binom{n-d_1-k}{r} + \binom{n-d_2-k}{r} - 2\binom{n-d_1-d_2-k}{r}}{\binom{n}{r}}.$$

Now since

$$q_x := \frac{\binom{n-x}{r}}{\binom{n}{r}} = \prod_{i=1}^{x}\left(1 - \frac{r}{n-i+1}\right) \leq \left(1 - \frac{r}{n}\right)^x$$

and for large enough n, by Lemma 11,

$$
\begin{aligned}
q_x = \prod_{i=1}^{x}\left(1 - \frac{r}{n-i+1}\right) &\geq \prod_{i=1}^{x}\left(\left(1-\frac{r}{n}\right) - \frac{2(i-1)r}{n^2}\right) \\
&\geq \left(1-\frac{r}{n}\right)^x - \sum_{i=1}^{x}\frac{2(i-1)r}{n^2} \\
&\geq \left(1-\frac{r}{n}\right)^x - \frac{x^2 r}{n^2}
\end{aligned}
$$

we have

$$
\begin{aligned}
\Pr[T(I,a) = T(J,a)] &= 1 - q_{d_1+k} - q_{d_2+k} + 2q_{d_1+d_2-k} \\
&= 1 - p^{d_2} - p^{d_1} + 2p^{d_1+d_2-k} + O\left(\frac{d^2(1-p)}{n}\right)
\end{aligned}
$$

where $p = 1 - r/n$.

Notice that this converges to the same probability in (4) for the RID algorithm. We have shown that the value of p that gives the best number of tests is $p = 1 - 1/d$ and therefore $r = n/d$. ∎