Undergraduate Research Opportunity Program
(UROP) Project Report

Sparse Group Testing: Bounds and Algorithms

By

Tan Thong Cai, Nelvin

Department of Computer Science

School of Computing

National University of Singapore

2019/2020 Semester 2

Undergraduate Research Opportunity Program
(UROP) Project Report

Sparse Group Testing: Bounds and Algorithms

By

Tan Thong Cai, Nelvin

Department of Computer Science

School of Computing

National University of Singapore

2019/2020 Semester 2

Project No: U240020
Advisor: Prof. Jonathan Scarlett
Deliverables:
Report: 1 Volume

Abstract

In group testing, the goal is to identify a subset of defective items within a larger set of items based on tests whose outcomes indicate whether any defective item is present. This problem is relevant in areas such as medical testing, data science, communications, and many more. Motivated by physical considerations, we consider a sparsity-based constrained setting (Gandikota et al., 2019) where the testing procedure is subject to one of the following two constraints: items are finitely divisible and thus may participate in at most γ tests; or tests are size-constrained to pool no more than ρ items per test.

While-information theoretic limits and algorithms are known for the non-adaptive setting, relatively little is known in the adaptive setting. We address this gap by providing an information theoretic converse that holds even in the adaptive setting, as well as a near-optimal noiseless adaptive algorithm for γ-divisible items. In broad scaling regimes, our upper and lower bounds asymptotically match up to a factor of e. We also present an adaptive algorithm for ρ-sized tests.

Under the non-adaptive setting, we generalize both constraints into a general constraint and provide information theoretic converse for both constraints. For the γ-divisible items constraints, we use the Definite Defectives (DD) decoding algorithm and study bounds on the required number of tests for vanishing error probability under the near-constant random test per item designs. We show that the number of tests required is less than the Combinatorial orthogonal matching pursuit (COMP) decoding algorithm, and is even order optimal for some scaling regimes.

Subject Descriptors:
G.3 Probability and Statistics
E.4 Coding and Information Theory

Keywords:
Information Theory, Probability Theory, Statistics, and Combinatorics

Implementation Software and Hardware:
R

Acknowledgement

I am deeply indebted to my supervisor Prof. Jonathan Scarlett for advising me on this project. He is very supportive and patient with me during the entire course of this research. Under his guidance, I have learnt a great amount from him on numerous aspects of academic research. Also, it was only through him that I discovered my love for information theory and statistics. It has truly been a great pleasure to research and study under him, and I could not have asked for a better role model, teacher, and supervisor.

I would also like to extend my gratitude to the amazing teachers who have helped developed my foundations in Computer Science and Statistics. In particular, I would like to thank Prof. Kim Cuc (Daisy) Pham, Prof. Wanjie Wang, Prof. Vik Gopal, Prof. Kwok Pui Choi, and Prof. Wynne Hsu for their superb teaching.

Thanks should also go to my bright and motivated friends who have supported me along the way, especially Andrew Tan, Gary Goh, Nicholas Teh, Kang Tze Ng, Heng Guang Lim, Sebastian Lie, and Kenneth Goh for stimulating discussions and a source of encouragement.

Lastly, I am extremely grateful to my family for their omnipresent warmth and support, especially my mum, who is the most important person in my life.

List of Figures

Table of Contents

Chapter 1

Introduction

Group testing originated in the United States in 1943, where a large number of conscripted soldiers were required to be screened for syphilis. Due to the existence of an accurate test (the Wassermann test), a naive option would be to test each soldier's blood sample individually. However, syphilis being a rare disease will result in most of the tests being negative. Since most tests have a high probability of being negative, under an information theoretic viewpoint, the tests are not very informative.

Our goal is to reduce the number of tests required to identify the soldiers with syphilis, using only the list of tests and their corresponding outcomes. Robert Dorfman [11] discovered that the number of tests required can be dramatically reduced by pooling samples. In other words, we can take blood samples from a "pool" of many soldiers, mix the samples, and perform the syphilis test on the pooled sample. We assume that the test is reliable which means that if the outcome is positive, it indicates that there is at least one soldier from the "pool" that has syphilis. If the outcome is negative, no soldier from the "pool" has syphilis. This led to the birth of group testing, which can be thought of as a discrete sparse inference problem.

Our idealized model where tests are reliable is known as the *standard noiseless group testing*. Generally, we have n items (number of soldiers) of which d are defective (number that has syphilis). The central problem of group testing is as follows: Given the number of items n and the number of defectives d, how many tests T are required to accurately discover the defective items, and how can this be achieved? The answer to this central problem depends on the assumptions on the mathematical model used. Below are some important distinctions [3] in the assumptions:

Adaptive vs. non-adaptive: Under adaptive testing, the test pools are designed sequentially, and each one can depend on the previous test outcomes. Under non-adaptive testing, the test pools are designed in advance before the testing process. This makes parallel implementation of the tests more viable.

Zero error probability vs. small error probability: Under the zero error probability criterion, we want to be certain that we will accurately recover the defective set. Under the small error probability criterion, we want to accurately recover the defective set with high probability.

Exact recovery vs. partial recovery: Under the exact recovery criterion, we require that every defective item is correctly classified as defective, and every non-defective item is correctly classified as non-defective. Under the partial recovery criterion, we are more lenient and allow a small number of incorrectly classified items.

Noiseless vs. noisy testing: Under noiseless testing, we are guaranteed that the test procedure works perfectly: We get a negative test outcome if all items in the testing pool are non-defective, and a positive outcome if at least one item in the pool is defective. This is equivalent to the reliable assumption made earlier in the syphilis example. Under noisy testing, errors can occur, either according to some specified random model or in an adversarial manner.

Binary vs. non-binary outcomes: Under binary outcomes, tests are either positive or negative. Under the non-binary outcomes, there might be a wider range of outcomes. For example, based on the number of defective and non-defective items, we can have different degrees of positivity (weak to strong) .

Throughout this report, our focus will be on small error probability, exact recovery, noiseless testing, and binary outcomes. We will study both the adaptive and non-adaptive settings. There are also further distinctions regarding the assumed distribution of the defective items among all items, and the decoder's knowledge (or lack of knowledge). These are explained below.

Combinatorial vs. i.i.d. prior: Under the combinatorial prior, there is a fixed number of defectives, and the defective set is uniformly random among all sets of this size. Under the i.i.d.

prior, each item is defective independently with the same fixed probability.

Known vs. unknown number of defectives: This distinguishes algorithms that need to be given the true number of defectives (or require estimation of the number of defectives first), and those that do not.

1.1 Model Setup

We will now introduce the model mathematically. Let n be the number of items, which we label as $\{1, 2, \ldots, n\}$. Let $\mathcal{D} \subset \{1, 2, \ldots, n\}$ be the set of defective items, and $d = |\mathcal{D}|$ be the number of defective items. We write $u_i = 1$ to denote that item $i \in \mathcal{D}$ is defective, and $u_i = 0$ to denote that $i \notin \mathcal{D}$ is non-defective. In other words, u_i is the indicator function $u_i = \mathbb{1}\{i \in \mathcal{D}\}$. We then write $\mathbf{u} = (u_i) \in \{0, 1\}^n$ for the defectivity vector.

In this report, we consider only the case where n is large, and d comparatively is small. Hence, we are interested in the *sparse* regime $d \in o(n)$. Also let $T = T(n)$ be the *number of tests* performed and label the tests $\{1, 2, \ldots, T\}$. To keep track of the design of the test pools in the non-adaptive setting, we write $x_{ti} = 1$ to denote that item $i \in \{1, 2, \ldots, n\}$ is in the pool for test $t \in \{1, 2, \ldots, T\}$, and $x_{ti} = 0$ to denote that item i is not in the pool for test t. This can be represented by the matrix $\mathsf{X} \in \{0, 1\}^{T \times n}$, known as the *testing matrix* or *test design.* A visual representation of an example is shown in Figure 1.1 (top) where $x_{ti} = 1$ is represented by a shaded box, and $x_{ti} = 0$ is represented by a white box.

It is useful to think of group testing as a channel coding framework, where the particular defective set $\mathcal{D}$ acts like the source message, finding the defective set can be thought of as decoding, and the matrix X acts like the codebook. This is shown in Figure 1.1 (bottom). Due to the previous success of randomized codes in channel communication [21] (*e.g.* turbo codes and LDPC codes), it is natural to consider randomized matrix designs. We use a capital X_{ti} to denote the random entries of a random testing matrix. In this report, we are interested in the following design:

Near-constant tests-per-item design: In this design, each item is included in some fixed number L of tests. The L tests for each item are chosen uniformly at random with replacement.

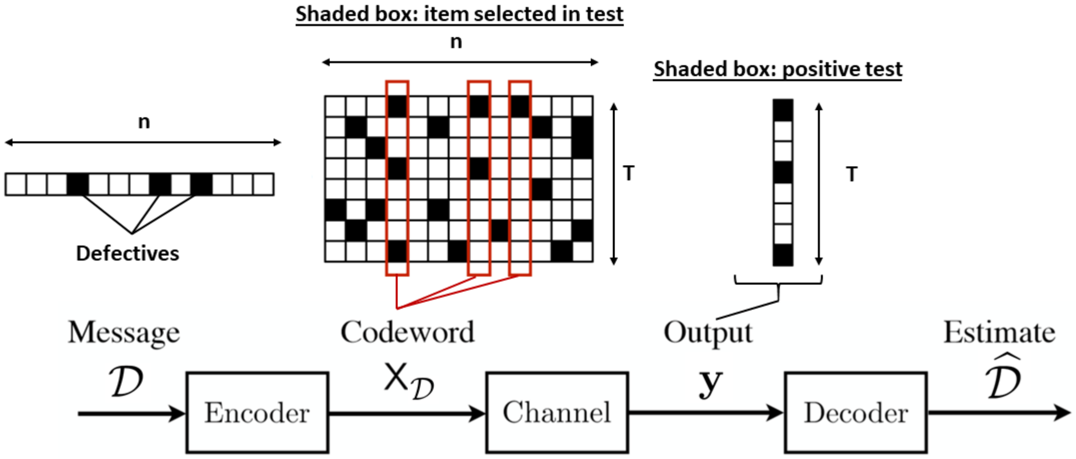

Figure 1.1: Group testing interpreted as a channel coding problem. The notation $\mathrm{X}_{\mathcal{D}}$ denotes the $T \times d$ sub-matrix of X obtained by keeping only the d columns indexed by $\mathcal{D}$, and the output $\mathbf{y}$ is the 'OR' of these d columns.

More specifically, L entries of each column are selected uniformly at random with replacement and set to one. The remaining entries are set to zero. Since we are selecting with replacement, some item(s) may be in fewer than L tests, hence the terminology "near-constant". This is a mathematical convenience that makes the analysis more tractable.

Now let $y_t \in \{0,1\}$ be the *outcome* of the test $t \in \{1,2,\dots,T\}$, where $y_t = 1$ denotes a positive outcome and $y_t = 0$ a negative outcome. Hence, we have $\mathbf{y} = (y_t) \in \{0,1\}^T$ for the vector of test outcomes. Using the OR (or disjunction) operator $\bigvee$, we have

$$y_t = \bigvee_{i \in \mathcal{D}} x_{ti}, \tag{1.1}$$

and using the definition of u_i above, we have

$$y_t = \bigvee_{i} x_{ti} u_i, \tag{1.2}$$

A *decoding* (or detection) *algorithm* is a (possibly randomized) function $\widehat{D} : \{0,1\}^{T\times n} \times \{0,1\}^T \to \mathcal{P}(\{1,2,\dots,n\})$, where the power-set $\mathcal{P}(\{1,2,\dots,n\})$ is the collection of the subsets of items. Given the tests and their outcomes, the decoding algorithm outputs an estimate vector $\widehat{\mathbf{u}} \in \{0,1\}^n$, representing an estimate of the defectivity vector of the population.

1.2 Applications

In this section, we provide some applications of group testing. In recent times, group testing has been abstracted into a combinatorial and algorithmic problem. This has led to many uses in other domains. Some examples are as follows:

Biology: In DNA testing (as described in [24]), modern sequencing methods search for particular subsequences of the genome in relatively short fragments of DNA. This results in the samples from individuals to be easily mixed. Hence, we can apply group testing to reduce the number of tests to isolate individuals (treated as defectives) with rare genetic conditions.

Furthermore, group testing can used to count defective items. Here, instead of identifying the defective set, we just want to estimate the number of defective items. This proves to be useful in situations where there is no need to distinguish the defective items (*e.g.*, insects in [26]) and when there is an intention to protect the privacy of the defective items (*e.g.*, patients with HIV/AIDS in [14]).

Communications: In multiple access channel [28], which is a channel where many users can communicate with a single receiver, at any one time, a small subset of users (treated as defectives) will have messages to transmit. Group testing can be applied to identify those users.

In cognitive radio networks, where "secondary users" can opportunistically transmit on frequency bands which are unoccupied by primary users, unoccupied bands are treated as defectives. We scan combinations of several bands at the same time (equivalent to pooling) and detect if any signal is being transmitted across any of them.

Information technology: In cybersecurity, there is an important problem known as the file comparison problem. The goal is to efficiently determine which computer files have changed based on a collection of various combinations of files. Here we treat the modified files as defectives and the combine hash acting as a testing pool. Non-adaptive group testing is then used to solve the problem, as described in [19].

Data science: Group testing is used in a variety of sparse inference problems, including streaming algorithms and learning sparse linear functions [15]. Group testing is also used in classical problems in computer science, including estimation of high degree vertices in hidden bipartite graphs [27] and pattern matching [17].

1.3 Constraints

Our focus in this report will be on sparse group testing introduced in [13] where the testing procedure is subjected to one of two constraints:

1. Items are *finitely divisible* and thus may participate in at most γ tests.

2. Tests are *size-constrained* and thus contain no more than ρ items per test.

Referring to our syphilis example in the introduction section, the γ-divisible items constraint can arise in the situation where there are limitations on the volume of blood provided by each soldier for the tests limiting the number of tests that each soldier can participate in. The ρ-sized tests constraint can arise in the situation where there are limitations on the number of blood samples that the machine can accept.

1.4 Previous Work on Sparse Group Testing

In the standard group-testing setting, in the absence of testing constraints, $T > (1-\epsilon)(d\log(\frac{n}{d}))$ tests are necessary to identify all defectives with error probability at most ϵ [2, 6]. Hence, the same is certainly true in the constrained setting. The same goes for the *strong converse*, which improves the preceding bound to $T > (1-o(1))(d\log(\frac{n}{d}))$ for any fixed $\epsilon \in (0,1)$ [5, 18]. A matching upper bound is known for all $d = o(n)$ in the unconstrained adaptive setting [16], whereas matching this lower bound non-adaptively is only possible non-adaptively in certain sparser regimes [7, 23].

It is well-known that if each test comprises of $\Theta(\frac{n}{d})$ items, then $\Theta(d\log n)$ tests suffice for group-testing algorithms with vanishing error probability [2, 6]. Hence, the parameter regime of primary interest in the size-constrained setting is $\rho \in o(\frac{n}{d})$. By a similar argument, the parameter regime of primary interest in the finitely divisible setting is $\gamma \in o(\log(\frac{n}{d}))$. Combined with the condition $T \in \Omega(d \log(\frac{n}{d}))$, the latter scaling regime implies that

$$\frac{T}{\gamma d} \to \infty \tag{1.3}$$

as $n \to \infty$, which will be useful in our proofs.

For the non-adaptive setting, Gandikota *et al.* [13] proved the following results for the γ-divisible setting.

Theorem 1.4.1. [13] *For any sufficiently large n, sufficiently small $\epsilon > 0$, $\gamma \in o(\log n)$, and $d \in \Theta(n^\theta)$ for some positive constant $\theta \in [0, 1)$, there exists a randomized design testing each item at most γ times that uses at most $\lceil e\gamma d(\frac{n}{\epsilon})^{1/\gamma} \rceil$ tests and ensures a reconstruction error of at most ϵ.*

Theorem 1.4.2. [13] *For any sufficiently large n, sufficiently small $\epsilon > 0$, $\gamma \in o(\log n)$, and $d \in \Theta(n^\theta)$ for some positive constant $\theta \in [0, 1)$, any non-adaptive group testing algorithm that tests each item at most γ times and has a probability of error of at most ϵ requires at least $\gamma d(\frac{n}{d})^{(1-5\epsilon)/\gamma}$ tests.*

For ρ-sized tests, the following achievability and converse results were also proved in [13].

Theorem 1.4.3. [13] *For any sufficiently large n, sufficiently small $\zeta > 0$, $\rho \in \Theta\big((\frac{n}{d})^\beta\big)$ (for some constant $\beta \in [0, 1)$), and $d \in \Theta(n^\theta)$ for some positive constant $\theta \in [0, 1)$, there exists a randomized non-adaptive group testing design that includes at most ρ items per test, using at most $\lceil \frac{1+\zeta}{(1-\alpha)(1-\beta)} \rceil \lceil \frac{n}{\rho} \rceil$ tests and ensuring a reconstruction error of at most $\epsilon = n^{-\zeta}$.*

Theorem 1.4.4. [13] *For any sufficiently large n, sufficiently small $\epsilon > 0$, $\rho \in \Theta\big((\frac{n}{d})^\beta\big)$ (for some constant $\beta \in [0, 1)$), and $d \in \Theta(n^\theta)$ for some positive constant $\theta \in [0, 1)$, any non-adaptive group testing algorithm that includes ρ items per test and has a probability of error of at most ϵ requires at least $\big(\frac{1-6\epsilon}{1-\beta}\big)\frac{n}{\rho}$ tests.*

We observe that under the ρ-sized test constraint, both the lower and upper bounds have the same leading order term $\frac{n}{\rho}$. Hence, there is not much of a gap between the lower and upper bounds. However, under the γ-divisible items constraint, the lower bound contains the term $(\frac{n}{d})^{(1-5\epsilon)/\gamma}$ while the upper bound contains the term $(\frac{n}{\epsilon})^{1/\gamma}$. Hence, there is significant gap between the lower and upper bounds; we will see that the gap can be made much smaller in the adaptive setting.

1.5 Overview of the Report

The structure of the report, as well as the main contributions, are outlined as follows:

- In Chapter 2, we consider the sparse adaptive setting and present an information theoretic lower bound for γ-divisible items, which strengthens the previous information theoretic lower bound in [13] for γ-divisible items by improving its dependence on error probability, as well as, extending its validity to the adaptive setting. Furthermore, we present adaptive algorithms for both γ-divisible items and ρ-sized tests and show that both algorithms recover the defective set with zero error probability. Moreover, we show that the algorithm for γ-divisible items is nearly optimal.

- In Chapter 3, we consider the sparse non-adaptive setting. By generalizing the γ-divisible items constraint and the ρ-sized tests constraint into a general constraint, we obtain information theoretic lower bounds for both settings which are consistent with previous information theoretic lower bounds in [13]. Furthermore, we extend an existing decoding algorithm to γ-divisible items and show that it performs better than a previously analyzed algorithm in [13].

- In Chapter 4, we review the main contributions of the report, and present various directions for future research.

1.6 Notation

Throughout the rest of the report, we will stay consistent with the notations introduced in Section 1.1.

We use bold symbols for vectors (*e.g.* $\mathbf{x}$), and we denote the corresponding i-th entry using a subscript (*e.g.* x_i). The natural logarithm is denoted by $\log(\cdot)$. We use the symbol $\sim$ to denote asymptotic equivalence.

We use standard notations in information theory where $H(X)$ denotes entropy, $H_2(X)$ denotes binary entropy, $H(Y|X)$ denotes conditional entropy and $I(X|Y)$ denotes mutual information.

We use standard notations in statistics and probability, where the symbol $\sim$ means "distributed as" (but sometimes also represents asymptotic equivalence; this will be clear from the context), $\mathbb{P}[\cdot]$ denotes the probability of an event, the hat symbol (*e.g.*, $\widehat{x}$) represents an estimator or an estimate, $\mathbb{E}_P[\cdot]$ denotes the expectation with respect to distribution P. When the probability distribution is understood from the context, we simply write $\mathbb{E}[\cdot]$. Also, we denote the indicator function of an event by $\mathbb{1}\{\cdot\}$. The ceiling function is denoted by $\lceil\cdot\rceil$.

To improve readability, we use Bachmann-Landau asymptotic notation (*i.e.*, O, o, Ω, ω, Θ) to specify parameter regimes. Any other notations required for specific proofs will be introduced when necessary.

Chapter 2

Sparse Adaptive Group Testing

2.1 Introduction

In this chapter, we seek information theoretic bounds and algorithms for the sparse adaptive setting.

Throughout the chapter, we consider the setup described in Section 1.1 and Section 1.3. More specifically, we consider the sparse adaptive group testing problem with small error probability, exact recovery, noiseless tests, binary outcomes, and combinatorial prior. Concretely, we target the error probability being bounded by some $\epsilon > 0$:

$$P_e = \mathbb{P}[\widehat{\mathbf{u}} \neq \mathbf{u}] \leq \epsilon, \tag{2.1}$$

where the probability is taken over the randomness of the set of defective items. Our main contributions are as follows:

- In Section 2.2, we provide information theoretic lower bound for γ-divisible items under the sparse adaptive setting.
- In Section 2.3, we provide an algorithm for γ-divisible items and study the number of tests for reliable recovery with zero error probability.
- In Section 2.4, we provide an algorithm for ρ-divisible tests and study the number of tests for reliable recovery with zero error probability.

Our analysis will make use of several techniques and results from probability theory and information theory.

Parts of this chapter were presented in [25].

2.2 Information Theoretic Lower Bound for γ-divisible Items

In this section, we present our information theoretic lower bounds for sparse group testing under the γ-divisible items. We first prove a counting bound which gives us an upper bound on the success probability $\mathbb{P}(\mathrm{suc}) = 1 - P_e$, following similar proof techniques as [5], with suitable refinements to account for the γ-divisibility constraint. Afterwards, we will use the bound on $\mathbb{P}(\mathrm{suc})$ to prove the converse result (lower bound on T).

Theorem 2.2.1. *Consider the case of n items with d defectives where each item can be tested at most γ times. Any algorithm (possibly adaptive) to recover the defective set $\mathcal{D}$ with T tests has success probability $\mathbb{P}(\mathrm{suc})$ satisfying*

$$\mathbb{P}(\mathrm{suc}) \leq \frac{\sum_{i=0}^{\gamma d} \binom{T}{i}}{\binom{n}{d}}. \tag{2.2}$$

Proof. See Section 2.5.1.

We now use the result in (2.2) to prove the following converse.

Theorem 2.2.2. *Fix $\epsilon \in (0,1)$, and suppose that $d \in o(n)$, $\gamma \in o(\log n)$, and $\gamma d \to \infty$ as $n \to \infty$. Then any non-adaptive or adaptive group testing algorithm that tests each item at most γ times and has a probability of error of at most ϵ requires at least $e^{-(1+o(1))}\gamma d\big(\frac{n}{d}\big)^{1/\gamma}$ tests.*

Proof. See Section 2.5.2.

Since ϵ only affects the $e^{o(1)}$ term, asymptotically, the number of tests required remains unchanged for any nonzero target success probability. This is in analogy with the strong converse results of [5, 18].

Theorem 2.2.2 strengthens the previous information theoretic lower bound in [13] for γ-divisible items (stating that $T \geq \gamma d(\frac{n}{d})^{(1-5\epsilon)/\gamma}$) by improving the dependence on ϵ, as well as

Algorithm 1 Adaptive algorithm for γ-divisible items

```
Require: Number of items $n$, number of defective items $d$, and divisibility of each item $\gamma$
Initialize $M \leftarrow (\frac{n}{d})^{\frac{\gamma-1}{\gamma}}$ and defective set $\mathcal{D} \leftarrow \emptyset$
Arbitrarily group the $n$ items into $\frac{n}{M}$ groups of size $M$
Test each group and discard any that return a negative outcome
Label the remaining groups incrementally as $G_j^{(0)}$, where $j = 1, 2, \dots$
for $i = 1$ to $\gamma - 1$ do
    for each group $G_j^{(i-1)}$ from the previous stage do
        Arbitrarily group all items in $G_j^{(i-1)}$ into $M^{1/(\gamma-1)}$ sub-groups of size $M^{1-i/(\gamma-1)}$
        Test each sub-group and discard any that return a negative outcome
        Label the remaining sub-groups incrementally as $G_j^{(i)}$
    end for
end for
Add the items in all the remaining groups $G_j^{(\gamma-1)}$ to $\mathcal{D}$
return $\mathcal{D}$
```

extending its validity to the adaptive setting (whereas [13] used an approach based on Fano's inequality that is specific to the non-adaptive setting).

2.3 Algorithm for γ-divisible Items

We first consider the recovery of the defective set given knowledge of the size d of the defective set. Afterwards, we consider the estimation of d.

2.3.1 Recovering the Defective Set

Our algorithm for the case that d is known is described in Algorithm 1, where we assume for simplicity that $(\frac{n}{d})^{1/\gamma}$ is an integer.[1] Using Algorithm 1, we have the following theorem, which is proved throughout the remainder of the subsection.

[1] Note that we assume $d \in o(n)$ and $\gamma \in o(\log(\frac{n}{d}))$, meaning that $(\frac{n}{d})^{1/\gamma} \to \infty$. Hence, the effect of rounding is asymptotically negligible, and is accounted for by the $1 + o(1)$ term in the theorem statement.

Theorem 2.3.1. *For $\gamma \in o(\log n)$, and $d \in o(n)$, there exists an adaptive group testing algorithm that tests each item at most γ times that uses at most $\gamma d(\frac{n}{d})^{1/\gamma}$ tests to recover the defective set exactly with zero error probability given knowledge of d.*

Proof. See Section 2.5.3.

Comparisons: Referring to Theorem 1.4.1, the upper bound for the non-adaptive algorithm of [13] using a randomized test matrix design is $T \leq \lceil e\gamma d(\frac{n}{\epsilon})^{1/\gamma} \rceil$. The non-adaptive algorithm has a $(\frac{n}{\epsilon})^{1/\gamma}$ term in the upper bound, while our adaptive algorithm has a $(\frac{n}{d})^{1/\gamma}$ term. Since ϵ is small but d is large, we see that our adaptive algorithm gives a significantly improved bound on the number of tests. Furthermore, the upper bound of our algorithm matches the information-theoretic lower bound in Theorem 2.2.2 up to a constant factor of $e^{1+o(1)}$. This proves that our algorithm is nearly optimal.

2.3.2 Estimating the Number of Defectives

Since each item can appear in at most γ tests, existing adaptive algorithms for estimating d that place items in $\Omega(\log \log d)$ tests [9, 12] are not suitable when $\gamma \ll \log \log d$, and may be wasteful of the budget γ even when $\gamma \gg \log \log d$.

To overcome this limitation, we introduce and evaluate two approaches to obtain a suitable input for d in Algorithm 1 given knowledge of an upper bound $d_{\max} \geq d$. The first approach uses $d_{\max}$ directly in Algorithm 1, while the second approach refines $d_{\max}$ by deriving an estimate $\widehat{d}$ that is passed to Algorithm 1. Note that we need $\widehat{d}$ to be an overestimate for the proof of Theorem 2.3.1 to still apply (with $\widehat{d}$ in place of d).

Using an $d_{\max}$ directly

Assuming that $(\frac{n}{d_{\max}})^{1/\gamma}$ is an integer, we first consider using $d_{\max}$ directly in Algorithm 1 (in place of d) to recover the defective set $\mathcal{D}$.

Analysis: Referring to Algorithm 1, this changes our initialization of M which becomes

Algorithm 2 Estimation of d

```
Require: Population of items, number of items $n$, upper bound $d_{\max} \geq d$, and a probability
    parameter $\beta_n$
1: Initialize number of bins $B \leftarrow d_{\max}/\beta_n$
2: Partition the items into $B$ bins of size $n/B$ each, uniformly at random
3: Test each bin and discard any with a negative test outcome
4: $\widehat{d} \leftarrow$ #positive bins$/(1 - \sqrt{\beta_n})$
5: return $\widehat{d}$
```

$(\frac{n}{d_{\max}})^{(\gamma-1)/\gamma}$. Substituting the updated value of M into (2.34), we obtain the following:

$$T \leq \frac{n}{(\frac{n}{d_{\max}})^{(\gamma-1)/\gamma}} + (\gamma - 1)d\Big[\Big(\frac{n}{d_{\max}}\Big)^{\frac{\gamma-1}{\gamma}}\Big]^{\frac{1}{\gamma-1}}, \tag{2.3}$$

which simplifies to

$$T \leq (d_{\max} - d + \gamma d)\Big(\frac{n}{d_{\max}}\Big)^{\frac{1}{\gamma}}. \tag{2.4}$$

Binning Method

We will show that the bound on T can be improved by forming a refined estimate of d using knowledge of $d_{\max}$, at the expense of having a non-zero (but asymptotically vanishing) probability of error.

Let β_n be a given parameter, which we will assume tends to zero as $n \to \infty$. We first run Algorithm 2 to obtain a new input $\widehat{d}$ to Algorithm 1. We then run Algorithm 1 with modified inputs (described in the following) to recover the defective set $\mathcal{D}$. Assuming that $(\frac{n}{d})^{1/\gamma}$ is an integer, we set the population of items in Algorithm 1 to be the remaining items left in the positive bins, the number of items as $d \times$ (bin size) $= d(\frac{\beta_n n}{d_{\max}})$, the (upper bound on the) number of defective items as $\widehat{d}$, and the divisibility of each item as $\gamma - 1$ (since each item is tested once in Algorithm 2).

Analysis: We first show that the probability of a particular defective item colliding with any other defective item (*i.e.*, falling in the same bin) tends to zero as $n \to \infty$. Referring to step 2 in Algorithm 2, conditioning on a particular item being in a particular bin, we see that the probability of another particular item being in the same bin is at most $1/B$. By the union bound, the probability of a particular defective item colliding with any of the other $d-1$ defective items is at most d/B, which behaves as

$$\frac{d}{B} = \frac{d}{d_{\max}/\beta_n} \leq \frac{d}{d/\beta_n} = \beta_n \to 0, . \tag{2.5}$$

Secondly, we show that with high probability as $n \to \infty$, $\widehat{d}$ overestimates d. From (2.5), we have

$$\mathbb{E}[\#\text{collisions}] \leq d\beta_n, \tag{2.6}$$

where #collisions refer to the number of items that are in the same bin as any of the other $d-1$ items. By Markov's inequality, we have

$$\mathbb{P}[\#\text{collisions} \geq d\sqrt{\beta_n}] \leq \sqrt{\beta_n}, \tag{2.7}$$

which implies the following:

$$\mathbb{P}[d - \#\text{collisions} \geq d - d\sqrt{\beta_n}] \geq 1 - \sqrt{\beta_n} \tag{2.8}$$

$$\implies \mathbb{P}\Big[\frac{d - \#\text{collisions}}{1-\sqrt{\beta_n}} \geq d\Big] \geq 1 - \sqrt{\beta_n}. \tag{2.9}$$

Since (#positive bins $\geq d -$ #collisions) always hold, we have $\mathbb{P}[\widehat{d} \geq d] \geq 1 - \sqrt{\beta_n}$, which tends to 1 because $\beta_n \to 0$.

Finally, we derive the new upper bound for T. After estimating d, we have used $B = d_{\max}/\beta_n$ number of tests and have a remaining budget of $\gamma - 1$ per item. We discard the bins (groups) that returned a negative outcome; instead of continuing with n items, we continue with less than or equal to ($d \times$ bin size) items. To simplify notation, our updated inputs (labeled with subscript "new") are

$$n_{\text{new}} = \frac{\beta_n dn}{d_{\max}}, \; d_{\text{new}} = \widehat{d}, \; \gamma_{\text{new}} = \gamma - 1. \tag{2.10}$$

We can then run Algorithm 1 to recover the defective set. Substituting our updated inputs into (2.34) and using $M = \big(\frac{\beta_n dn}{d_{\max}\widehat{d}}\big)^{\frac{\gamma-2}{\gamma-1}}$, we have the following bound for T:

$$T \leq \frac{d_{\max}}{\beta_n} + \frac{\beta_n dn}{d_{\max}\big(\frac{\beta_n dn}{d_{\max}\widehat{d}}\big)^{\frac{\gamma-2}{\gamma-1}}} + (\gamma-2)d\Big(\frac{\beta_n dn}{d_{\max}\widehat{d}}\Big)^{\frac{1}{\gamma-1}}, \tag{2.11}$$

which simplifies to

$$T \leq \frac{d_{\max}}{\beta_n} + (\widehat{d} - 2d + \gamma d)\Big(\frac{\beta_n d n}{d_{\max}\widehat{d}}\Big)^{\frac{1}{\gamma-1}} \tag{2.12}$$

$$\overset{(a)}{\leq} \frac{d_{\max}}{\beta_n} + \Big(\frac{d}{1-\sqrt{\beta_n}} - 2d + \gamma d\Big)\Big(\frac{\beta_n n}{d_{\max}}\Big)^{\frac{1}{\gamma-1}}, \tag{2.13}$$

where we used $d \leq \widehat{d} \leq \frac{d}{1-\sqrt{\beta_n}}$ in (a).

Comparisons: By using T within the derived upper bounds, the first approach recovers the defective set with zero error probability while the second approach recovers the defective set with a small error probability determined by the β_n parameter. Referring to (2.4) and (2.13), we consider two examples to compare the bounds on T. The first example is when $d_{\max} = d$, and the second example is when $\gamma d \ll d_{\max} \ll n$.

For $d_{\max} = d$, as we would naturally expect, (2.4) is the better bound; its leading term is $\gamma d\big(\frac{n}{d}\big)^{1/\gamma}$. In particular, we note the following two cases: (i) If $\beta_n \ll \frac{1}{\gamma(\frac{n}{d})^{1/\gamma}}$, then the $\frac{d_{\max}}{\beta_n}$ term in (2.13) is strictly higher than $\gamma d\big(\frac{n}{d}\big)^{1/\gamma}$; (ii) If $\beta_n \gg \frac{1}{\gamma(\frac{n}{d})^{1/\gamma}}$, then some simple algebra gives $\frac{\beta_n n}{d} \gg \frac{1}{\gamma}\big(\frac{n}{d}\big)^{(\gamma-1)/\gamma}$, which implies that the $\gamma d\big(\frac{\beta_n n}{d}\big)^{1/(\gamma-1)}$ term from (2.13) is strictly higher than $\gamma d\big(\frac{n}{d}\big)^{1/\gamma}$ (note that $\big(\frac{1}{\gamma}\big)^{1/(\gamma-1)} = \Theta(1)$).

For $\gamma d \ll d_{\max} \ll n$, the choice of β_n can impact which bound is smaller. First note that the dominating term in (2.4) is $d_{\max}\big(\frac{n}{d_{\max}}\big)^{1/\gamma}$. Since the dominating term $\max\big\{\frac{d_{\max}}{\beta_n}, \gamma d\big(\frac{\beta_n n}{d_{\max}}\big)^{1/(\gamma-1)}\big\}$ in (2.13) is not obvious, we consider both possibilities: (i) $d_{\max}\big(\frac{n}{d_{\max}}\big)^{1/\gamma} \gg \frac{d_{\max}}{\beta_n}$ whenever $\beta_n \gg \big(\frac{d_{\max}}{n}\big)^{1/\gamma}$; and (ii) $d_{\max}\big(\frac{n}{d_{\max}}\big)^{1/\gamma} \gg \gamma d\big(\frac{\beta_n n}{d_{\max}}\big)^{\frac{1}{\gamma-1}}$ whenever $\beta_n \ll \big(\frac{d_{\max}}{\gamma d}\big)^{\gamma-1}\big(\frac{d_{\max}}{n}\big)^{1/\gamma}$. Combining these cases, we see that if β_n is in the range $\big(\frac{d_{\max}}{n}\big)^{1/\gamma} \ll \beta_n \ll \big(\frac{d_{\max}}{\gamma d}\big)^{\gamma-1}\big(\frac{d_{\max}}{n}\big)^{1/\gamma}$, the dominating term in (2.4) is greater than the dominating term in (2.13).

Since we have assumed β_n to be decaying, we briefly discuss conditions under which the requirement $\big(\frac{d_{\max}}{n}\big)^{1/\gamma} \ll \beta_n$ is consistent with this assumption. While this lower bound on β_n may not always vanish as $n \to \infty$, it does so in broad scaling regimes, including the following: $\gamma \in \Theta((\log n)^c)$ for some $c \in [0,1)$, and $d_{\max} = d = \Theta(n^\theta)$ for some $\theta \in (0,1)$. To see this, note that

$$\lim_{n\to\infty} \log\Big(\frac{d_{\max}}{n}\Big)^{\frac{1}{\gamma}} = \lim_{n\to\infty} (\alpha - 1)(\log n)^{1-c} = -\infty, \tag{2.14}$$

and that taking $\exp(\cdot)$ on both sides gives the desired result.

Algorithm 3 Adaptive algorithm for ρ-sized tests

```
Require: Population of items, number of items $n$, number of defective items $d$, and test size restriction $\rho$
Initialize defective set $\mathcal{D} \leftarrow \emptyset$
Randomly group $n$ items into $n/\rho$ groups of size $\rho$
for each group $G_i$ where $i \in \{1, 2, \ldots, n/\rho\}$ do
    while testing $G_i$ returns a positive outcome do
        run Algorithm 4 on $G_i$ and add its one defective item output $d^*$ into $\mathcal{D}$
        $G_i \leftarrow G_i \setminus \{d^*\}$
    end while
end for
return $\mathcal{D}$
```

Algorithm 4 Binary splitting

```
Require: a group of items $G_i$
1: If $G_i$ consists of a single item, return that item.
2: Pick half of the items in $G_i$ and call this set $G'_i$. Perform a single test on $G'_i$.
3: If the test is positive, set $G_i \leftarrow G'_i$. Else, set $G_i \leftarrow G_i \setminus G'_i$. Return to step 1.
4: return $\mathcal{D}$
```

Hence, for β_n in the appropriate range, when $d_{\max}$ is close to d, using the upper bound directly in Algorithm 1 leads to a smaller T. On the other hand, when $\gamma d \ll d_{\max} \ll n$, using the binning method before Algorithm 1 leads to a smaller T.

2.4 Algorithm for ρ-sized Tests

We state our algorithm as shown in Algorithm 3. Our adaptive algorithm under the ρ-sized test constraint is a modification of Hwang's generalized binary splitting algorithm [16] where we divide the n items into $\frac{n}{\rho}$ groups of size ρ, instead of d groups of size $\frac{n}{d}$ as in the original algorithm.

Analysis: Let d_i be the number of defective items in each of the initial $\frac{n}{\rho}$ groups. Note that since $\rho \in o(\frac{n}{d})$ implies $d \in o(\frac{n}{\rho})$, most groups will not have a defective item. In the binary splitting stage of the algorithm, we can round the halves in either direction if they are not an integer. Hence, for each of the initial $\frac{n}{\rho}$ groups, we take at most $\lceil \log_2 \rho \rceil$ adaptive tests to find a defective item, or one test to confirm that there are no defective item. Therefore, for each of the initial $\frac{n}{\rho}$ groups, we need $\max\{1, d_i \log_2 \rho + O(d_i)\}$ tests to find d_i defective items. Summing across all $\frac{n}{\rho}$ groups, we need a total of $T = \sum_{i=1}^{n/\rho} \max\{1, d_i \log_2 \rho + O(d_i)\}$ tests. This has the following upper bound:

$$T \leq \frac{n}{\rho} + d \log_2 \rho + O(d) \tag{2.15}$$

$$\overset{(a)}{=} \frac{n}{\rho}(1 + o(1)) + d \log_2 \rho, \tag{2.16}$$

where (a) uses $d \in o\big(\frac{n}{\rho}\big)$. With the further condition $\rho \in O\big(\frac{n}{d\log(n/d)}\big)$, we have $\frac{n}{\rho} \in \Omega\big(d \log\left(\frac{n}{d}\right)\big)$ and $d \log \rho \in o\big(d \log\left(\frac{n}{d}\right)\big)$. Thus, we can further simplify to get

$$T \leq \frac{n}{\rho}(1 + o(1)). \tag{2.17}$$

This upper bound is tight in the sense that attaining vanishing error probability trivially requires a fraction $1 - o(1)$ of the items to be tested at least once, which implies $T \geq \frac{n}{\rho}(1 - o(1))$ by the ρ-sized test constraint.

2.5 Proofs

2.5.1 Proof of Theorem 2.2.1 (Counting Bound)

Given a population of n objects, we write $\Sigma_{n,d}$ for the collection of subsets of size d from the population. Furthermore, we write $\mathcal{D}$ for the true defective set.

We follow the steps of [5] as follows: The testing procedure defines a mapping $\theta : \Sigma_{n,d} \to \{0,1\}^T$. Given a putative defective set $S \in \Sigma_{n,d}$, $\theta(S)$ is the vector of test outcomes, with positive tests represented as 1s and negative tests represented as 0s. For each vector $\mathbf{y} \in \{0,1\}^T$, we write $\mathcal{A}_{\mathbf{y}} \subseteq \Sigma_{n,d}$ for the inverse image of $\mathbf{y}$ under θ,

$$\mathcal{A}_{\mathbf{y}} = \theta^{-1}(\mathbf{y}) = \{S \in \Sigma_{n,d} : \theta(S) = \mathbf{y}\}. \tag{2.18}$$

The role of an algorithm that decodes the outcome of the tests is to mimic the effect of the inverse image map θ^{-1}. Given a test output $\mathbf{y}$, the optimal decoding algorithm would use a lookup table to find the inverse image $\mathcal{A}_{\mathbf{y}}$. If this inverse image $\mathcal{A}_{\mathbf{y}} = \{S\}$ has size $|\mathcal{A}_{\mathbf{y}}| = 1$, we can be certain that the defective set was S. In general, if $|\mathcal{A}_{\mathbf{y}}| \geq 1$, we cannot do better than pick uniformly among $\mathcal{A}_{\mathbf{y}}$, with success probability $\frac{1}{|\mathcal{A}_{\mathbf{y}}|}$ (We can ignore empty $\mathcal{A}_{\mathbf{y}}$, since we are only concerned with vectors $\mathbf{y}$ that occur as a test output).

Hence, overall, the probability of recovering a defective set S is $\frac{1}{|\mathcal{A}_{\theta(S)}|}$, depending only on $\theta(S)$. We can write the following expression for the success probability, conditioning over all the equiprobable values of the defective set:

$$
\begin{aligned}
\mathbb{P}(\mathrm{suc}) &\overset{(a)}{=} \sum_{S\in\Sigma_{n,d}} \mathbb{P}(\mathrm{suc}|\mathcal{D}=S)\frac{1}{\binom{n}{d}} && (2.19)\\
&= \frac{1}{\binom{n}{d}} \sum_{S\in\Sigma_{n,d}} \sum_{\mathbf{y}\in\{0,1\}^T} \mathbb{1}(\theta(S)=\mathbf{y})\mathbb{P}(\mathrm{suc}|\mathcal{D}=S) && (2.20)\\
&= \frac{1}{\binom{n}{d}} \sum_{S\in\Sigma_{n,d}} \sum_{\mathbf{y}\in\{0,1\}^T:|\mathcal{A}_{\mathbf{y}}|\geq 1} \mathbb{1}(\theta(S)=\mathbf{y})\frac{1}{|\mathcal{A}_{\mathbf{y}}|} && (2.21)\\
&= \frac{1}{\binom{n}{d}} \sum_{\mathbf{y}\in\{0,1\}^T:|\mathcal{A}_{\mathbf{y}}|\geq 1} \frac{1}{|\mathcal{A}_{\mathbf{y}}|}\Big(\sum_{S\in\Sigma_{n,d}} \mathbb{1}(\theta(S)=\mathbf{y})\Big) && (2.22)\\
&= \frac{1}{\binom{n}{d}} \sum_{\mathbf{y}\in\{0,1\}^T:|\mathcal{A}_{\mathbf{y}}|\geq 1} \frac{1}{|\mathcal{A}_{\mathbf{y}}|}|\mathcal{A}_{\mathbf{y}}| && (2.23)\\
&= \frac{|\{\mathbf{y}\in\{0,1\}^T : |\mathcal{A}_{\mathbf{y}}|\geq 1\}|}{\binom{n}{d}} && (2.24)\\
&\overset{(b)}{\leq} \frac{|\{\mathbf{y} \text{ with} \leq \gamma d \text{ ones}\}|}{\binom{n}{d}} = \frac{\sum_{i=0}^{\gamma d}\binom{T}{i}}{\binom{n}{d}}, && (2.25)
\end{aligned}
$$

where (a) uses the law of total probability and the uniform prior on $\mathcal{D}$, and (b) uses the fact that at most γd test outcomes can be positive, even in the adaptive setting. This is because adding another defective always introduces at most γ additional positive tests.

2.5.2 Proof of Theorem 2.2.2 (Converse for γ-divisible Items)

From the counting bound in (2), we upper bound the sum of binomial coefficients [4, Section 4.7.] to obtain

$$\mathbb{P}(\mathrm{suc}) \leq \frac{e^{TH_2(\frac{\gamma d}{T})}}{\binom{n}{d}} \equiv \delta, \tag{2.26}$$

where $H_2(\cdot)$ is the binary entropy function in nats. From (2.26), we have $e^{TH_2(\frac{\gamma d}{T})}/\binom{n}{d} = \delta$, which implies that

$$\log\left(\delta\binom{n}{d}\right) = TH_2\Big(\frac{\gamma d}{T}\Big) \tag{2.27}$$

$$= \gamma d \log \frac{T}{\gamma d} + (T - \gamma d)\log\frac{1}{1-\frac{\gamma d}{T}} \tag{2.28}$$

$$\overset{(a)}{=} \gamma d \log \frac{T}{\gamma d} + \gamma d(1 + o(1)), \tag{2.29}$$

where (a) uses a Taylor expansion and the fact that $\frac{\gamma d}{T} \in o(1)$ from (1.3). Hence, we have $(1-\frac{\gamma d}{T})^{-1} = \exp(\frac{\gamma d}{T})(1+o(1))$ which is used to obtain the simplification. Rearranging (2.29), we obtain

$$\gamma d \log \frac{T}{\gamma d} = \log\left(\delta\binom{n}{d}\right) - \gamma d(1+o(1)) \tag{2.30}$$

$$\implies \log\frac{T}{\gamma d} = \frac{1}{\gamma d}\log\left(\delta\binom{n}{d}\right) - (1+o(1)), \tag{2.31}$$

which gives

$$T = e^{-(1+o(1))}\gamma d\left(\delta\binom{n}{d}\right)^{\frac{1}{\gamma d}} \tag{2.32}$$

$$\overset{(a)}{\geq} e^{-(1+o(1))}\gamma d \delta^{\frac{1}{\gamma d}}\left(\frac{n}{d}\right)^{\frac{1}{\gamma}}, \tag{2.33}$$

where (a) follows from the fact that $\binom{n}{d} \geq \left(\frac{n}{d}\right)^d$.

The proof is completed by noting that for a fixed target success probability $\delta = 1 - \epsilon$, $\delta^{1/(\gamma d)} \to 1$ as $\gamma d \to \infty$.

2.5.3 Proof of Theorem 2.3.1 (Adaptive Algorithm Performance)

Similar to Hwang's generalized binary splitting algorithm [16], the idea behind the parameter M in Algorithm 1 is that when d becomes large, having large groups during the initial splitting stage is wasteful, as it results in each test having a very high probability of being positive (not very informative). Hence, we want to find the appropriate group sizes that result in more informative tests to minimize the number of tests. Each stage (outermost for-loop in Algorithm 1) here refers to the process where all groups of the same sizes are split into smaller groups (as seen in Figure 2.1). We let M be the group size at the initial splitting stage of the algorithm. The

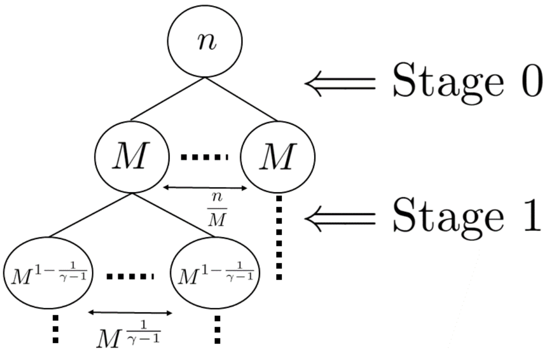

Figure 2.1: Visualization of splitting in the adaptive algorithm.

algorithm first tests n/M groups of size M each,[2] then steadily decrease the sizes of each group down the stages: $M \to M^{1-1/(\gamma-1)} \to M^{1-2/(\gamma-1)} \to \cdots \to 1$ (see Figure 2.1 for visualization). Hence, we have n/M groups in the initial splitting and $M^{\frac{1}{\gamma-1}}$ groups in all subsequent splits.

With the above observations, we can derive an upper bound on the total number of tests needed. We have n/M tests in the first stage. Since we have d defectives and split into $M^{\frac{1}{\gamma-1}}$ sub-groups in subsequent stages, the number of smaller groups that each stage can produce is at most $dM^{\frac{1}{\gamma-1}}$. This implies that the number of tests conducted at each stage is at most $dM^{\frac{1}{\gamma-1}}$. This gives us the following bound on T:

$$T \leq \frac{n}{M} + (\gamma - 1)dM^{\frac{1}{\gamma-1}}. \tag{2.34}$$

We optimize with respect to M by differentiating the upper bound and setting it to zero. This gives us $M = (\frac{n}{d})^{\frac{\gamma-1}{\gamma}}$. Substituting $M = (\frac{n}{d})^{\frac{\gamma-1}{\gamma}}$ into the general upper bound in (2.34), we have the following upper bound:

$$T \leq \frac{n}{(\frac{n}{d})^{\frac{\gamma-1}{\gamma}}} + (\gamma - 1)d\left[\left(\frac{n}{d}\right)^{\frac{\gamma-1}{\gamma}}\right]^{\frac{1}{\gamma-1}} = \gamma d\left(\frac{n}{d}\right)^{\frac{1}{\gamma}}. \tag{2.35}$$

[2] Note that $\frac{n}{M}$ is an integer for our chosen M below, which gives $\frac{n}{M} = d(\frac{n}{d})^{1/\gamma}$, and $(\frac{n}{d})^{1/\gamma}$ was assumed to be an integer earlier.

Chapter 3

Sparse Non-adaptive Group Testing

3.1 Introduction

In this chapter, we seek information theoretic bounds and algorithms for the sparse non-adaptive setting.

Throughout the chapter, we consider the setup described in Section 1.1 and Section 1.3. More specifically, we consider the sparse non-adaptive group testing problem with small error probability, exact recovery, noiseless tests, binary outcomes, and combinatorial prior. Concretely, we target the error probability being bounded by some $\epsilon > 0$:

$$P_e = \mathbb{P}[\hat{\mathbf{u}} \neq \mathbf{u}] \leq \epsilon, \tag{3.1}$$

where the probability is taken over the randomness of the set of defective items. Our main contributions are as follows:

- In Section 3.2, we provide preliminary definitions and results that will be used in later parts of this chapter.
- In Section 3.3, we generalize the constraints of γ-divisible items and ρ-sized tests into a general constraint and use it to provide information theoretic lower bounds for both constraints.
- In Section 3.4, we apply an existing algorithm to the case of γ-divisible items and study the number of tests for reliable recovery.

Our analysis will make use of several techniques and results from probability theory and information theory.

3.2 Preliminary Definitions and Results

We first introduce some definitions below which will be useful to us in this chapter.

Definition 3.2.1. Consider an item i and a set of items $\mathcal{L}$ not including i. We say that item i is *masked* by $\mathcal{L}$ if every test that includes i, also includes at least one member of $\mathcal{L}$.

Definition 3.2.2. The number of *collisions* between item i and a set of items $\mathcal{L}$ refers to the number of test(s) that include item i, and also includes at least one member of $\mathcal{L}$.

Next, we introduce some lemmas which will be used for some parts of our proofs later.

Lemma 3.2.1. *For $\gamma \in \Theta\big((\log n)^c\big)$ for some $c \in [0,1)$, and $d \in \Theta(n^\theta)$ for some $\theta \in (0,1]$, we have $\big(1 \pm \frac{1}{d^{1/\gamma}}\big)^\gamma = 1 \pm o(1)$.*

Proof. Note that since $\big(1 \pm \frac{1}{d^{1/\gamma}}\big)^\gamma \to 1$ if $d^{1/\gamma} \gg \gamma$, it suffices to show that $d^{1/\gamma} \gg \gamma$. We have

$$\theta(\log n)^{1-c} \gg c \log\log n \tag{3.2}$$

$$\implies \frac{\theta}{(\log n)^c}(\log n) \gg \log(\log n)^c \tag{3.3}$$

$$\implies \log(n^{\theta/(\log n)^c}) \gg \log(\log n)^c \tag{3.4}$$

$$\implies \quad n^{\theta/(\log n)^c} \gg (\log n)^c. \tag{3.5}$$

Since $d \in \Theta(n^\theta)$ and $\gamma \in \Theta\big((\log n)^c\big)$, by substitution in the above equation, we get $d^{1/\gamma} \gg \gamma$ which completes the proof. □

Let $W^{(\mathcal{D})}$ be the total number of positive tests containing at least one item from $\mathcal{D}$. To understand the distribution of this quantity, it is helpful to think of the process by which elements of the columns are sampled as a *coupon collector* problem, where each coupon corresponds to one of the T tests. For a single defective item, $W^{(\{i\})}$ is the number of distinct coupons selected when γ coupons are chosen uniformly at random from a population of T coupons. In general, for the defective set $\mathcal{D}$ of size d, the independence of distinct columns means that $W^{(\mathcal{D})}$ is the number of distinct coupons collected when choosing γd coupons uniformly at random from a population of T coupons. We now give a concentration measure result for $W^{(\mathcal{D})}$ around its mean.

Lemma 3.2.2. *When making $\gamma d \in o(T)$ draws with replacement from a total of T coupons, the total number of distinct coupons $W^{(\mathcal{D})}$ satisfies*

$$\mathbb{P}[|W^{(\mathcal{D})} - \gamma d(1-\delta_n)| \geq (\gamma d)^{2/3}] \leq 2\exp(-2(\gamma d)^{1/3}), \tag{3.6}$$

where $\delta_n \in O\big(\frac{\gamma d}{T}\big)$.

Proof. For any coupon, the probability of not being selected is $1 - \big(1-\frac{1}{T}\big)^{\gamma d}$. This gives us

$$\mathbb{E}[W^{(\mathcal{D})}] = \left(1 - \left(1 - \frac{1}{T}\right)^{\gamma d}\right)T \tag{3.7}$$

$$\overset{(a)}{=} \left(1 - \left(1 - \frac{\gamma d}{T} + O\Big(\Big(\frac{\gamma d}{T}\Big)^2\Big)\right)\right)T \tag{3.8}$$

$$= \left(\frac{\gamma d}{T} - O\Big(\Big(\frac{\gamma d}{T}\Big)^2\Big)\right)T \tag{3.9}$$

$$\overset{(b)}{=} \gamma d(1-\delta_n), \tag{3.10}$$

where (a) is due to second order binomial approximation using Taylor series, and we introduce $\delta_n \in O\big(\frac{\gamma d}{T}\big)$ in (b). Let $Y_1, Y_2, \dots, Y_{\gamma d}$ be the labels of the selected coupons and $W(\gamma d) = f(Y_1, Y_2, \dots, Y_{\gamma d})$ be the number of distinct coupons. We have the bounded property difference property

$$|f(Y_1, \dots, Y_j, \dots, Y_{\gamma d}) - f(Y_1, \dots, \widehat{Y}_j, \dots, Y_{\gamma d})| \leq 1 \tag{3.11}$$

for any $j, Y_1, Y_2, \dots, Y_{\gamma d}$, and $\widehat{Y}_j$ since the largest difference we can make is swapping a distinct coupon Y_j for a non-distinct coupon $\widehat{Y}_j$, or vice versa. McDiarmid's inequality [20] gives

$$\mathbb{P}(|f(Y_1, Y_2, \dots, Y_{\gamma d}) - \mathbb{E}[f(Y_1, Y_2, \dots, Y_{\gamma d})]| \geq \delta) \leq 2\exp\Big(-\frac{2\delta^2}{\gamma d}\Big). \tag{3.12}$$

Setting $\delta = (\gamma d)^{2/3}$, we get the desired result. □

Let $W^{(\mathcal{D}\setminus i)}$ and $W^{(\mathcal{D}\setminus i,j)}$ be the total number of positive tests containing at least one item in $\mathcal{D} \setminus \{i\}$, and the total number of positive tests containing at least one item in $\mathcal{D} \setminus \{i,j\}$ respectively. We then have the following two corollaries.

Corollary 3.2.2.1. *When making $\gamma(d-1) \in o(T)$ draws with replacement from a total of T coupons, the total number of distinct coupons $W^{(\mathcal{D}\setminus i)}$ satisfies*

$$\mathbb{P}[|W^{(\mathcal{D}\setminus i)} - \gamma(d-1)(1-\delta_n^{(1)})| \geq (\gamma(d-1))^{2/3}] \leq 2\exp(-2(\gamma(d-1))^{1/3}), \tag{3.13}$$

where $\delta_n^{(1)} \in O\big(\frac{\gamma d}{T}\big)$.

Corollary 3.2.2.2. *When making $\gamma(d-2) \in o(T)$ draws with replacement from a total of T coupons, the total number of distinct coupons $W^{(\mathcal{D}\setminus i,j)}$ satisfies*

$$\mathbb{P}[|W^{(\mathcal{D}\setminus i,j)} - \gamma(d-2)(1-\delta_n^{(2)})| \geq (\gamma(d-2))^{2/3}] \leq 2\exp(-2(\gamma(d-2))^{1/3}), \tag{3.14}$$

where $\delta_n^{(2)} \in O\big(\frac{\gamma d}{T}\big)$.

3.3 Information Theoretic Lower Bound

In this section, we present our information theoretic lower bounds for sparse group testing under the γ-divisible items constraint and the ρ-sized tests constraints.

3.3.1 γ-divisible Items Case

Theorem 3.3.1. *Under the combinatorial prior, for any sufficiently large n, sufficiently small $\epsilon > 0$, and $d \in \Theta(n^\theta)$ for some positive constant $\theta \in (0,1)$, any non-adaptive group testing algorithm that tests each item at most $\gamma \in o(\log n)$ times and has a probability of error of at most ϵ requires*

$$T \geq \gamma d \Big(\frac{n}{d}\Big)^{\frac{1-\epsilon}{\gamma}(1+o(1))}(1+o(1)). \tag{3.15}$$

Proof. See Section 3.5.1.

We will now introduce both the Combinatorial Orthogonal Matching Pursuit (COMP) algorithm and the Smallest Satisfying Set (SSS) algorithm which will be used in the next theorem.

Definition 3.3.1. The COMP algorithm for noiseless non-adaptive group testing is given as follows: mark each item that appears in a negative test as non-defective, and refer to every other item as a possibly defective. We write $\mathcal{PD}$ for the set of such items. Mark every item in $\mathcal{PD}$ as defective.

We observe that the COMP algorithm fails when at least one non-defective item is masked by $\mathcal{D}$ (equivalently, it succeeds when zero non-defective items are masked). For the SSS algorithm, we state a key definition first before describing the algorithm.

Definition 3.3.2. We say that a putative defective set $\mathcal{J}$ is *satisfying* if:

1. No negative test contains a member of $\mathcal{J}$.

2. Every positive test contains at least one member of $\mathcal{J}$.

Definition 3.3.3. The SSS algorithm for noiseless non-adaptive group testing is given as follows: find the smallest satisfying set (breaking ties arbitrarily), and take that as the estimate $\widehat{\mathcal{D}}_{\mathrm{SSS}}$.

Note that the true defective set $\mathcal{D}$ is certainly a satisfying set, and hence SSS is guaranteed to return a set of no larger size, giving us $|\widehat{\mathcal{D}}_{\mathrm{SSS}}| \leq |\mathcal{D}|$. However, it may not be the case that $\widehat{\mathcal{D}}_{\mathrm{SSS}} \subseteq \mathcal{D}$. We can identify a particular failure event for SSS: If a defective item $i \in \mathcal{D}$ is masked by the other defective items $\mathcal{D} \setminus \{i\}$, then $\mathcal{D} \setminus \{i\}$ will be a smaller satisfying set, so SSS is certain to fail. We study the probability of this failure event in detail in the proof of our next result, Theorem 3.3.2.

The analysis of both algorithms will be important in proving the following result, which is obtained by looking at a specific test design X.

Theorem 3.3.2. *Consider a near-constant column weight design, with $d \in \Theta(n^\theta)$ for some positive constant $\theta \in (0,1)$, any non-adaptive group testing algorithm that tests each item at most $\gamma \in \Theta\big((\log n)^c\big)$ times for some $c \in [0,1)$ with tests*

$$T \leq \gamma d^{\frac{1}{\gamma}}(d-1)(1+o(1)), \tag{3.16}$$

has an error probability bounded away from zero.

Proof. See Section Section 3.5.2.

This theorem states that by considering a near-constant column weight design, the converse in Theorem 3.3.2 is stronger (lower bound on T increases) than the converse in Theorem 3.3.1 when d is "large", or specifically when the sparsity parameter θ is large.

3.3.2 ρ-sized Tests Case

Theorem 3.3.3. *Under the combinatorial prior, for any sufficiently large n, sufficiently small $\epsilon > 0$, and $d \in \Theta(n^{\theta})$ for some positive constant $\theta \in (0,1)$, any non-adaptive group testing algorithm that includes $\rho \in \Theta((\frac{n}{d})^{\beta})$ (for some constant $\beta \in [0,1)$) items per test and has a probability of error of at most ϵ requires at least $\frac{1-\epsilon}{1-\beta}(\frac{n}{\rho})(1+o(1))$ tests.*

Proof. See Section 3.5.3.

Our bounds obtained are consistent with the bounds in [13]. Note that the Fano's inequality approach in the proofs does not carry through for the adaptive setting, since our bounding of entropy $H(y_t)$ of individual testing outcomes critically relies on the test matrix X being independent of $\mathbf{u}$.

3.4 Algorithm for γ-divisible Items

We focus on the near-constant tests per item design for the γ-divisible items constraint, where $\gamma \in o\big(\log\big(\frac{n}{d}\big)\big)$ tests for each item are chosen uniformly at random with replacement.

We will use the Definite Defectives (DD) decoding algorithm which is defined as follows.

Definition 3.4.1. The Definite Defectives (DD) algorithm for noiseless non-adaptive group testing has two keys steps.

1. Since $y_t = 1$ if and only if the test pool contains a defective item, we can be sure that each item that appears in a negative test is not defective. We form a list of such items from all the negative tests, which we refer to as the guaranteed non-defective ($\mathcal{ND}$) set and the rest of the items $\mathcal{PD} := \{1,\dots,n\} \setminus \mathcal{ND}$ are considered in the possibly defective ($\mathcal{PD}$) set.

2. Since every positive test must contain at least one defective item, if a test with $Y = 1$ contains exactly one item from $\mathcal{PD}$, then we can be certain that the item in question is defective. The DD algorithm estimates $\mathcal{D}$ using $\widehat{\mathcal{D}}$ to be the set of $\mathcal{PD}$ items which appear in a positive test with no other $\mathcal{PD}$ item.

Note that the first step is just our COMP algorithm. The first step makes no mistake in adding to $\mathcal{ND}$ (items are correctly marked as non-defective), and the second step also makes no mistake in adding to $\widehat{\mathcal{D}}$ (items are correctly marked as defective). Hence, any errors due to DD come from marking a true defective as non-defective in the second step, meaning that our estimate $\widehat{\mathcal{D}}$ satisfies $\widehat{\mathcal{D}} \subseteq \mathcal{D}$. The choice to mark all remaining items as non-defective is motivated by the sparsity of the problem (recall that $d \in o(n)$), since *a priori* an item is much less likely to be defective than non-defective.

Using DD, we have the following theorem.

Theorem 3.4.1. *For $\gamma \in \Theta\big((\log n)^c\big)$ for some $c \in [0,1)$, $d \in \Theta(n^\theta)$ for some $\theta \in (0,1)$, $\alpha_2 \in (0,1)$, and any function β_n decaying as n increases, there exists a randomized design testing each item at most γ times that uses at most*

$$T = \gamma d \max\left\{2^{\frac{1}{\alpha_2}H_2(\max\{\alpha_2,\frac{1}{2}\})}\Big(\frac{d}{\beta_n}\Big)^{\frac{1}{\alpha_2\gamma}}, 2^{1/\gamma}\Big(\frac{n-d}{d}\Big)^{\frac{1}{\gamma}}\Big(\frac{d}{\beta_n}\Big)^{\frac{1}{(1-\alpha_2)\gamma^2}}\right\}. \tag{3.17}$$

tests and ensures a reconstruction error of at most

$$\exp\left(-\frac{3d}{16}\Big(\frac{\beta_n}{d}\Big)^{\frac{1}{(1-\alpha_2)\gamma)}}\right) + 2\exp(-2(\gamma d)^{1/3}) + 2\beta_n(1+o(1)). \tag{3.18}$$

Proof. See Section 3.5.4.

We now provide an interpretation for the bound on T. We consider two scaling regimes below, where we assume that β_n is a slowly decaying term (*e.g.*, log factors only). Hence, we will omit β_n from our asymptotic bounds on T. The regimes are

1. Large γ: $\gamma \in \Theta((\log n)^c)$ for some $c \in (0,1)$, and $d \in \Theta(n^\theta)$ for some $\theta \in (0,1)$

2. Constant γ: $\gamma \in O(1)$, and $d \in \Theta(n^\theta)$ for some $\theta \in (0,1)$.

Let us first introduce a variable η, defined as

$$\eta = \lim_{n\to\infty} \frac{\log(\frac{n}{d})}{\gamma\log(\frac{T}{\gamma d})}. \tag{3.19}$$

which will be used in our plots later.

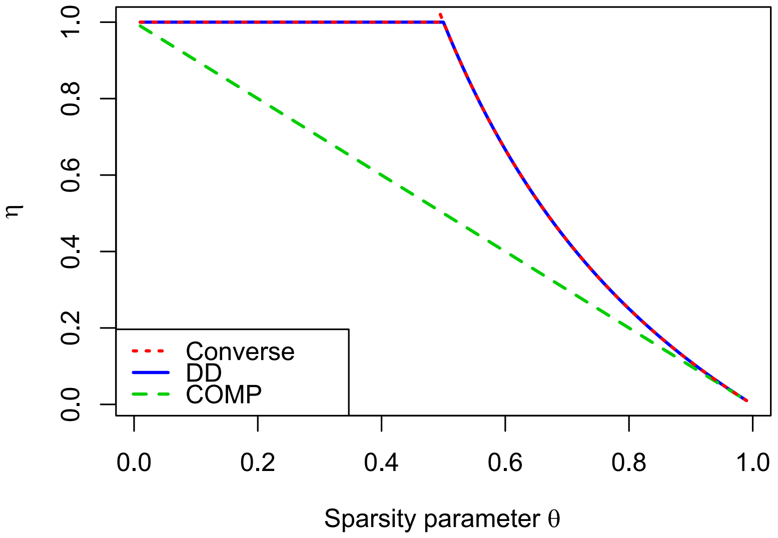

Figure 3.1: Plots of the variable η, such that $T \sim \gamma d\big[\big(\frac{n}{d}\big)^{1/\gamma}\big]^{(1+o(1))/\eta}$, against the sparsity parameter θ for the converse, DD algorithm, and COMP algorithm, when $n \to \infty$ and $\gamma = (\log n)^c$ for some $c \in (0,1)$.

For regime 1 (large γ), When α_2 is a fixed constant close to 1, we have $2^{\frac{1}{\alpha_2}H_2(\max\{\alpha_2,\frac{1}{2}\})} \approx 1$, and $\big(\frac{n-d}{d}\big)^{1/\gamma} d^{\frac{1}{(1-\alpha_2)\gamma^2}} = \big(\frac{n}{d}\big)^{(1-o(1))/\gamma}$. To see the latter, note that

$$\left(\frac{n-d}{d}\right)^{\frac{1}{\gamma}} d^{\frac{1}{(1-\alpha_2)\gamma^2}} = \left(\frac{n-d}{d^{1-O(1/\gamma)}}\right)^{\frac{1}{\gamma}} \tag{3.20}$$

$$= \left(\frac{n^{\frac{1}{(1-O(1/\gamma)}}(1-O(d/n))^{\frac{1}{1-O(1/\gamma)}}}{d}\right)^{\frac{1-O(1/\gamma)}{\gamma}} \tag{3.21}$$

$$= \left(\frac{n}{d}\right)^{\frac{1-o(1)}{\gamma}} \tag{3.22}$$

By substituting the scaling regimes into the bound in Theorem 3.4.1 and omitting β_n, we have

$$T = \tilde{\Omega}\Big(\gamma d \max\{n^{\theta}, n^{1-\theta}\}^{\frac{1}{\gamma}}\Big). \tag{3.23}$$

We plot η against $\theta \in (0,1)$ in Figure 3.1 to show how the asymptotic bound of the DD algorithm compares to the converse and the COMP algorithm's bound in Theorem 1.4.1. Note that for the COMP algorithm, we have omitted ϵ in our asymptotic bound, giving us $T = \tilde{\Omega}(\gamma d n^{1/\gamma})$. From Figure 3.1, we see that the DD algorithm performs better than the COMP algorithm, and

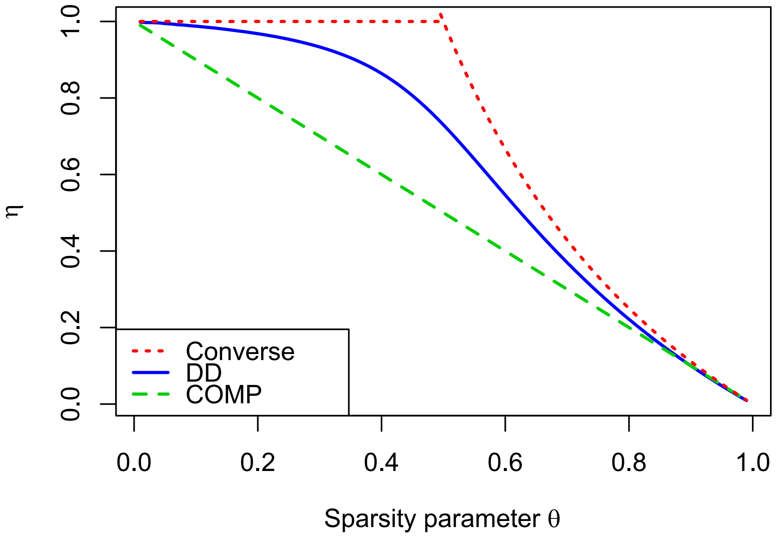

Figure 3.2: Plots of the variable η, such that $T \sim \gamma d\big[\big(\frac{n}{d}\big)^{1/\gamma}\big]^{(1+o(1))/\eta}$, against the sparsity parameter θ for the converse, DD algorithm, and COMP algorithm, when $n \to \infty$ and $\gamma = 10$.

achieves the optimal limit η when $\theta \in (0, 0.5]$.

For regime 2 (constant γ), we substitute the scaling regimes and omitting β_n to get

$$T = \tilde{\Omega}\Big(\gamma d \max\Big\{n^{\frac{\theta}{\alpha_2 \gamma}}, n^{\frac{1-\theta}{\gamma} + \frac{\theta}{(1-\alpha_2)\gamma^2}}\Big\}\Big). \tag{3.24}$$

We numerically optimize with respect to α_2 to obtain our bound on T. Figure 3.2 shows how the asymptotic bound of the DD algorithm compares to the converse and the COMP algorithm's bound in Theorem 1.4.1 (ϵ is again omitted from the asymptotic bound), when $\gamma = 10$. This is done by plotting η against $\theta \in (0, 1)$ like before. From Figure 3.2, we see that the DD algorithm performs better than the COMP algorithm.

3.5 Proofs

3.5.1 Proof of Theorem 3.3.1 (Converse for γ-divisible Items)

Our proof is similar to the proof presented in [13]. Let g_t be the number of items in the t-th test and p_t^- be the probability that the t-th test outcome y_t is negative. Our key insight is that both constraints can be generalized to a general constraint. We observe that the γ-divisible items constraint restricts the number of ones in the testing matrix X to be less than or equal to γd. Similarly, the ρ-sized tests constraint restricts the number of ones in the testing matrix X to be less than or equal to ρT. Hence, we can work with the general constraint, which is a restriction on the number of ones in X. Our proof treats both constraints as a general constraint which is a constraint on the number of ones in the testing matrix X. More formally, we want to find the converse for #ones $\leq c$ subjected to $\sum_{t=1}^{T} g_t \leq c$. Under the combinatorial prior, we have the following expression for p_t^-:

$$p_t^- = \frac{\binom{n-g_t}{d}}{\binom{n}{d}} = \frac{(n-d)!}{(n-d-g_t)!} \cdot \frac{(n-g_t)!}{n!} \tag{3.25}$$

$$= \frac{(n-d)(n-d-1)\dots(n-d-g_t+1)}{n(n-1)\dots(n-g_t+1)} \tag{3.26}$$

$$= \prod_{k=0}^{g_t-1} \left(1 - \frac{d}{n-k}\right) \leq \left(1 - \frac{d}{n}\right)^{g_t}. \tag{3.27}$$

Note that when $p_t^- > \frac{1}{2}$, (3.27) implies that $\left(1 - \frac{d}{n}\right)^{g_t} > \frac{1}{2}$, which simplifies to give:

$$g_t < \frac{-\log 2}{\log\left(1 - \frac{d}{n}\right)} \overset{(a)}{=} \frac{n \log 2}{d}(1 + o(1)), \tag{3.28}$$

where (a) follows from the fact that $\frac{d}{n} \in o(1)$ implying $\left(1 - \frac{d}{n}\right) = \exp\left(-\frac{d}{n}\right)(1 + o(1))$. Using the bound on g_t, we partition tests T into sets S_l (light set), and S_h (heavy set) where $t \in S_l$ if $g_t < \frac{n \log 2}{d}(1 + o(1))$, and $t \in S_h$ otherwise.

<u>For the light set S_l</u>, we have

$$p_t^- \overset{(a)}{=} \prod_{k=0}^{g_t-1} \left(1 - \frac{d}{n-k}\right) \geq \left(1 - \frac{d}{n-g_t}\right)^{g_t} \tag{3.29}$$

$$\overset{(b)}{\geq} 1 - \frac{d g_t}{n - g_t} \overset{(c)}{=} 1 - \frac{d g_t(1 + o(1))}{n}, \tag{3.30}$$

where (a) is from (3.27), (b) is by Bernoulli's identity [22], and (c) follows from the fact that we have $g_t \in o(n)$ since $g_t < \frac{n \log 2}{d}(1 + o(1))$. We want to bound $\sum_{t=1}^{|S_l|} H(y_t)$. Since $1 - p_t^- < \frac{1}{2}$

for $t \in S_l$, we bound the entropy of each test as follows:

$$\begin{aligned} H(y_t) = H_2(p_t^-) = H_2(1-p_t^-) &\leq H_2\Big(\frac{dg_t(1+o(1))}{n}\Big) && (3.31)\\ &= \frac{dg_t(1+o(1))}{n}\log\Big(\frac{n}{dg_t(1+o(1))}\Big) && (3.32)\\ &\quad - \Big(1-\frac{dg_t(1+o(1))}{n}\Big)\log\Big(1-\frac{dg_t(1+o(1))}{n}\Big) && (3.33)\\ &\overset{(a)}{\leq} \frac{dg_t(1+o(1))}{n}\log\Big(\frac{n}{dg_t(1+o(1))}\Big) + \frac{dg_t(1+o(1))}{n} && (3.34)\\ &= \frac{dg_t(1+o(1))}{n}\Big[\log\Big(\frac{n}{dg_t(1+o(1))}\Big)+1\Big], && (3.35)\end{aligned}$$

where (a) uses the fact that $-\log(1-x) \leq \frac{x}{1-x}$ for $x < 1, x \neq 0$. This gives us $H(y_t) \leq f(g_t)$ where $f(g_t) = \frac{dg_t(1+o(1))}{n}[\log(\frac{n}{dg_t(1+o(1))})+1]$. Using the above results, we have

$$\begin{aligned} \sum_{t=1}^{|S_l|} H(y_t) \leq \frac{|S_l|}{|S_l|}\sum_{t=1}^{|S_l|} f(g_t) &= |S_l|\sum_{t=1}^{|S_l|}\frac{1}{|S_l|}f(g_t) && (3.36)\\ &\overset{(a)}{\leq} |S_l| f\Big(\sum_{t=1}^{|S_l|}\frac{1}{|S_l|}g_t\Big) \leq |S_l| f\Big(\frac{c}{|S_l|}\Big), && (3.37)\end{aligned}$$

where (a) is by Jensen's inequality since $f''(g_t) = -\frac{d(1+o(1))}{ng_t} < 0$ (*i.e.*, concave), $\frac{1}{|S_l|} \geq 0$, and $\sum_{t=1}^{|S_l|}\frac{1}{|S_l|} = 1$.

For the heavy set S_h, we naively bound each entropy by 1:

$$\begin{aligned} \sum_{t=1}^{|S_h|} H(y_t) \leq |S_h| &\overset{(a)}{\leq} \frac{c}{\frac{n\log 2}{d}(1+o(1))} && (3.38)\\ &= \frac{cd}{n\log 2}(1+o(1)). && (3.39)\end{aligned}$$

where (a) is because $g_t \geq \frac{n\log 2}{d}(1+o(1))$. Combining (3.37) and (3.39), we have

$$\begin{aligned} \sum_{t=1}^{T} H(y_t) &\leq |S_l| f\Big(\frac{c}{|S_l|}\Big) + \frac{cd}{n\log 2}(1+o(1)) && (3.40)\\ &= \frac{cd(1+o(1))}{n}\Big[\log\Big(\frac{n|S_l|}{cd(1+o(1))}\Big)+1\Big] + \frac{cd}{n\log 2}(1+o(1)) && (3.41)\\ &\leq \frac{cd(1+o(1))}{n}\Big[\log\Big(\frac{nT}{cd(1+o(1))}\Big)+1+\frac{1}{\log 2}\Big]. && (3.42)\end{aligned}$$

Note that $\mathbf{u} \to \mathbf{y} \to \widehat{\mathbf{u}}$ forms a Markov chain. By Fano's inequality [8], we have

$$H(\mathbf{u}|\widehat{\mathbf{u}}) \leq H_2(P_e) + P_e\log(|\mathcal{U}|-1), \qquad (3.43)$$

where $\mathbf{u}$ is uniformly distributed over $\mathcal{U}$, the length of all length-n, d-sparse binary vectors. Weakened and rearranged, we have

$$P_e \geq 1 - \frac{I(\mathbf{u};\widehat{\mathbf{u}}) + \log 2}{\log |\mathcal{U}|} \tag{3.44}$$

$$\geq 1 - \frac{\sum_{t=1}^{T} H(y_t) + \log 2}{\log |\mathcal{U}|}, \tag{3.45}$$

because

$$I(\mathbf{u};\widehat{\mathbf{u}}) \overset{(a)}{\leq} I(\mathbf{u};\mathbf{y}) \overset{(b)}{=} H(\mathbf{y}) - H(\mathbf{y}|\mathbf{u}) \overset{(c)}{\leq} H(\mathbf{y}) \overset{(d)}{\leq} \sum_t H(y_t), \tag{3.46}$$

where (a) by the data processing inequality, (b) by the definition of mutual information, (c) by the non-negativity of entropy (conditional) and (d) by the subadditivity of entropy. Since $c = n\gamma$ for γ-divisible items, we substitute $c = n\gamma$ into (3.42) and further substitute the result into (3.45). Thus, we have

$$\epsilon \geq 1 - \frac{\gamma d(1+o(1))[\log(\frac{T}{\gamma d(1+o(1))}) + 1 + \frac{1}{\log 2}] + \log 2}{\log \binom{n}{d}}, \tag{3.47}$$

which simplifies to

$$T \geq \gamma d \exp\left[\frac{(1-\epsilon)\log(\frac{n}{d}) - \frac{\log 2}{d}}{\gamma(1+o(1))} - 1 - \frac{1}{\log 2}\right](1+o(1)) \tag{3.48}$$

$$= \gamma d \exp\left[\frac{1-\epsilon}{\gamma}(1+o(1))\log\left(\frac{n}{d}\right)\right](1+o(1)) \tag{3.49}$$

$$= \gamma d \left(\frac{n}{d}\right)^{\frac{1-\epsilon}{\gamma}(1+o(1))}(1+o(1)). \tag{3.50}$$

3.5.2 Proof of Theorem 3.3.2 (Converse for γ-divisible Items)

For the SSS algorithm, we look at a failure event where any defective item $i \in \mathcal{D}$ is masked by the other defective items $\mathcal{D}\setminus\{i\}$. We denote the error probability of such an event by $\mathbb{P}^{\text{SSS}}(\text{err})$. For the COMP algorithm, we look at a failure event where at least one non-defective item is masked by $\mathcal{D}$. We denote the error probability of such an event by $\mathbb{P}^{\text{COMP}}(\text{err})$. Here, we write A_i for the event that item $i \in \mathcal{D}$ is masked by $\mathcal{D}\setminus\{i\}$. Using union bound followed by de Caen's lower bound on union from [10], we get

$$\mathbb{P}^{\text{SSS}}(\text{err}) \geq \mathbb{P}\left(\bigcup_{i\in\mathcal{D}} A_i\right) \geq \sum_{i\in\mathcal{D}} \frac{\mathbb{P}(A_i)^2}{\mathbb{P}(A_i) + \sum_{j\in\mathcal{D}\setminus\{i\}}\mathbb{P}(A_i \cap A_j)}. \tag{3.51}$$

It was proved in [1] that if $\mathbb{P}^{\text{SSS}}(\text{err}) + \mathbb{P}^{\text{COMP}}(\text{err}) > 1 + \epsilon$ for some $\epsilon > 0$ that remains bounded away from zero as $n \to \infty$, then the error probability is also bounded away from zero for an arbitrary algorithm. Hence, it suffices to show that $\mathbb{P}^{\text{SSS}}(\text{err})$ is bounded away from zero, and $\mathbb{P}^{\text{COMP}}(\text{err}) \to 1$. We will bound the right hand side of (3.51) by bounding the numerator and denominator separately, before combining to bound $\mathbb{P}^{\text{SSS}}(\text{err})$. Afterwards, we will bound $\mathbb{P}^{\text{COMP}}(\text{err})$.

We claim that it suffices to show that the error probability is bounded away from zero some T satisfying

$$T = \gamma d^{\frac{1}{\gamma}}(d-1)(1+o(1)). \tag{3.52}$$

Although in the theorem we have $T \le \gamma d^{1/\gamma}(d-1)(1+o(1))$, we can choose $T \in \Theta(\gamma d \cdot d^{1/\gamma})$ because if the error probability is already bounded away from zero at the bound, we cannot hope to do any better with fewer tests.

Bounding Masking Error Probability - Numerator Term

Fixing the index i of some defective item, we note that conditioned on $W^{(\mathcal{D}\setminus i)} = w$, the event A_i occurs if each test that item i occurs in is contained in the w "already hit" tests. Hence, for any constant $c_1 > 0$, we have

$$\begin{aligned}
\mathbb{P}(A_i) &= \sum_w \mathbb{P}(A_i | W^{(\mathcal{D}\setminus i)} = w)\mathbb{P}(W^{(\mathcal{D}\setminus i)} = w) && (3.53)\\
&= \sum_w \left(\frac{w}{T}\right)^\gamma \mathbb{P}(W^{(\mathcal{D}\setminus i)} = w) && (3.54)\\
&\ge \sum_{w \ge c_1\gamma(d-1)} \left(\frac{w}{T}\right)^\gamma \mathbb{P}(W^{(\mathcal{D}\setminus i)} = w) && (3.55)\\
&\ge \sum_{w \ge c_1\gamma(d-1)} \left(\frac{c_1\gamma(d-1)}{T}\right)^\gamma \mathbb{P}(W^{(\mathcal{D}\setminus i)} = w) && (3.56)\\
&= \left(\frac{c_1\gamma(d-1)}{T}\right)^\gamma \mathbb{P}(W^{(\mathcal{D}\setminus i)} \ge c_1\gamma(d-1)). && (3.57)
\end{aligned}$$

Bounding Masking Error Probability - Denominator Term

We first derive a bound for $\mathbb{P}(A_i \cap A_j | W^{(\mathcal{D}\setminus i,j)} = w)$. Just for this part, we represent columns of X corresponding to items i and j by lists $\mathcal{T}_i = \{t_{i1}, \dots, t_{i\gamma}\}$ and $\mathcal{T}_j = \{t_{j1}, \dots, t_{j\gamma}\}$. Each list entry is obtained by choosing $t \in \{1, \dots, T\}$ uniformly at random with replacement, so duplicates may occur. Without loss of generality, we assume that the w tests containing items from $\mathcal{D} \setminus \{i, j\}$ are those indexed by $1, \dots, w$. Any given list occurs with probability $1/T^{\gamma}$. Letting $\mathscr{A}_i$ be the set of list pairs $(\mathcal{T}_i, \mathcal{T}_j)$ under which the event A_i occurs, and similarly for $\mathscr{A}_j$, we have

$$\mathbb{P}(A_i \cap A_j | W^{(\mathcal{D} \setminus i,j)} = w) = \frac{N_{ij}}{T^{2\gamma}}, \tag{3.58}$$

where

$$N_{ij} = \sum_{\mathcal{T}_i} \sum_{\mathcal{T}_j} \mathbb{1}\{(\mathcal{T}_i, \mathcal{T}_j) \in \mathscr{A}_i \cap \mathscr{A}_j\}, \tag{3.59}$$

is the number of pairs of lists in $\mathscr{A}_i \cap \mathscr{A}_j$. Here the sets $\mathscr{A}_i$ and $\mathscr{A}_j$ implicitly depend on w. To bound N_{ij}, we separately consider the number of "new positive tests" caused by items i and j; that is, not among the first w. Specifically, letting $N_{ij}(l)$ be defined as above with the summation limited to the case that there are l such new positive tests, we have

$$N_{ij} = \sum_{l=0}^{\gamma} N_{ij}(l), \tag{3.60}$$

where the summation goes up to γ due to the fact that any new positive test containing i must also contain j and vice versa; otherwise, the masking under consideration would not occur.

To bound $N_{ij}(l)$, we consider the following procedure for choosing the lists:

- From $T - w$ tests, choose l of them to be the new defective tests. This is one of $\binom{T-w}{l}$ options.

- For both i and j, assign one list index from $\{1, \dots, \gamma\}$ to each of the l new defective tests. This is at most γ^l options each, for γ^{2l} in total.

- For both i and j, the remaining $\gamma - l$ list entries are chosen arbitrarily from the $w + l$ positive tests. This is $(w + l)^{\gamma - l}$ options each, for $(w + l)^{2(\gamma - l)}$ in total.

Combining these terms gives

$$
\begin{aligned}
N_{ij}(l) &\leq \binom{T-w}{l} \cdot \gamma^{2l} \cdot (w+l)^{2(\gamma-l)} && (3.61)\\
&\leq (T-w)^{l} \cdot \gamma^{2l} \cdot (w+\gamma)^{2(\gamma-l)} && (3.62)\\
&= (w+\gamma)^{2\gamma} \cdot \left(\frac{\gamma^2(T-w)}{(w+\gamma)^2}\right)^{l}. && (3.63)
\end{aligned}
$$

Under the assumption that $w \geq c_1\gamma(d-2)$, the bracketed term $\frac{\gamma^2(T-w)}{(w+\gamma)^2}$ is less than any fixed $\epsilon_1 > 0$ for sufficiently large n. To see this, recall that $T \in \Theta(\gamma d \cdot d^{1/\gamma}) = \Theta(\gamma n^{\theta+\theta/\gamma})$ and $w \in \Omega(\gamma d) = \Omega(\gamma n^{\theta})$. By substituting the scaling regime for T and w into the bracketed term above and taking the log, we get

$$
\begin{aligned}
\log \frac{\gamma^2(T-w)}{(w+\gamma)^2} &= \log \frac{\gamma^2(\Theta(1)\gamma n^{\theta+\theta/\gamma} - \Omega(1)\gamma n^{\theta})}{(\Omega(1)\gamma n^{\theta}+\gamma)^2} && (3.64)\\
&\overset{(a)}{\leq} \log \frac{\Theta(1)\gamma^3 n^{\theta+\theta/\gamma}}{\Omega(1)\gamma^2 n^{2\theta}} && (3.65)\\
&= \log\left(\frac{\Theta(1)}{\Omega(1)}\right) + \log(\gamma n^{\theta/\gamma-\theta}) && (3.66)\\
&= \log\left(\frac{\Theta(1)}{\Omega(1)}\right) + \log\gamma + \left(\frac{\theta}{\gamma}-\theta\right)\log n, && (3.67)
\end{aligned}
$$

where (a) is by using $\Theta(1)\gamma n^{\theta+\theta/\gamma} - \Omega(1)\gamma n^{\theta} \leq \Theta(1)\gamma n^{\theta+\theta/\gamma}$ and $\Omega(1)\gamma n^{\theta} + \gamma \geq \Omega(1)\gamma n^{\theta}$. The term above tends to $-\infty$ when $\gamma > 1$ because $\gamma \in o(\log n)$. This implies that the bracketed term in (3.63), $\frac{\gamma^2(T-w)}{(w+\gamma)^2} \to 0$ as $n \to \infty$. Hence, summing over l gives

$$
\begin{aligned}
N_{ij} &\leq \sum_{l=0}^{\gamma} (w+\gamma)^{2\gamma} \cdot \left(\frac{\gamma^2(T-w)}{(w+\gamma)^2}\right)^{l} && (3.68)\\
&\leq (w+\gamma)^{2\gamma} \cdot \sum_{l=0}^{\infty} \epsilon_1^{l} && (3.69)\\
&= (w+\gamma)^{2\gamma} \cdot \frac{1}{1-\epsilon_1}. && (3.70)
\end{aligned}
$$

Substituting the above result into (3.58), we get

$$
\mathbb{P}(A_i \cap A_j | W^{(\mathcal{D}\setminus i,j)} = w) \leq \left(\frac{w+\gamma}{T}\right)^{2\gamma} \cdot \frac{1}{1-\epsilon_1} \qquad (3.71)
$$

Now for any $c_1, c_2 > 0$, we have

$$\sum_{j\in\mathcal{D}\setminus\{i\}} \mathbb{P}(A_i\cap A_j) = (d-1)\sum_w \mathbb{P}(A_i\cap A_j|W^{(\mathcal{D}\setminus i,j)}=w)\mathbb{P}(W^{(\mathcal{D}\setminus i,j)}=w) \tag{3.72}$$

$$\begin{aligned}&\leq \frac{d-1}{1-\epsilon_1}\sum_{c_1\gamma(d-2)\leq w\leq c_2\gamma(d-2)}\Big(\frac{w+\gamma}{T}\Big)^{2\gamma}\mathbb{P}(W^{(\mathcal{D}\setminus i,j)}=w)\\ &\quad + (d-1)\mathbb{P}(W^{(\mathcal{D}\setminus i,j)}\notin[c_1\gamma(d-2),c_2\gamma(d-2)]) \qquad (3.73)\\ &\leq \frac{d-1}{1-\epsilon_1}\Big(\frac{c_2\gamma(d-2)+\gamma}{T}\Big)^{2\gamma}\mathbb{P}(c_1\gamma(d-2)\leq W^{(\mathcal{D}\setminus i,j)}\leq c_2\gamma(d-2))\\ &\quad + (d-1)\mathbb{P}(W^{(\mathcal{D}\setminus i,j)}<c_1\gamma(d-2)) + (d-1)\mathbb{P}(W^{(\mathcal{D}\setminus i,j)}>c_2\gamma(d-2)). \qquad (3.74)\end{aligned}$$

Combining Both Terms

We choose

$$\begin{aligned} c_1 &= \min\Big\{1-\delta_n^{(1)}-\frac{1}{(\gamma(d-1))^{1/3}}, 1-\delta_n^{(2)}-\frac{1}{(\gamma(d-2))^{1/3}}\Big\} && (3.75)\\ &= 1-\delta_n^{(3)}-\frac{1}{(\gamma(d-2))^{1/3}} && (3.76)\\ c_2 &= 1-\delta_n^{(2)}+\frac{1}{(\gamma(d-2))^{1/3}}, && (3.77)\end{aligned}$$

where $\delta_n^{(3)} = \max\{\delta_n^{(1)},\delta_n^{(2)}\} \in O\big(\frac{\gamma d}{T}\big) \to 0$. With our choices, the concentration results for both $W^{(\mathcal{D}\setminus i)}$ and $W^{(\mathcal{D}\setminus i,j)}$ in Corollary 3.2.2.1 and Corollary 3.2.2.2 respectively will hold. We also introduce

$$c_3 = 1+\Big(\frac{d-2}{d-1}\Big)(-\delta_n^{(2)}+(\gamma(d-2))^{-1/3}), \tag{3.78}$$

and note the useful fact

$$c_3\gamma(d-1) = c_2\gamma(d-2)+\gamma, \tag{3.79}$$

which will be used later.

Applying the concentration results from Corollary 3.2.2.1 and Corollary 3.2.2.2, we get

$$\begin{aligned}\mathbb{P}(A_i) &\overset{(a)}{\geq} \Big(\frac{c_1\gamma(d-1)}{T}\Big)^{\gamma}(1-o(1)) && (3.80)\\ \sum_{j\in\mathcal{D}\setminus\{i\}}\mathbb{P}(A_i\cap A_j) &\overset{(b)}{\leq} \frac{d-1}{1-\epsilon_1}\Big(\frac{c_2\gamma(d-2)+\gamma}{T}\Big)^{2\gamma}(1-o(1))+o(1), && (3.81)\end{aligned}$$

where in (a), we substitute our chosen c_1 into the $\mathbb{P}(\cdot)$ part of (3.57), and then apply the concentration result in Corollary 3.2.2.1. In (b), we substitute our chosen c_1 and c_2 into the $\mathbb{P}(\cdot)$ parts of (3.74), and then apply the concentration results in Corollary 3.2.2.2. Note that we also used the fact that $2(d-1)\exp(-(\gamma(d-2))^{1/3}) \to 0$ since $\gamma \geq 1$ and $d \to \infty$. Also, we have

$$\mathbb{P}(A_i) \leq \left(\frac{\gamma(d-1)}{T}\right)^{\gamma}, \tag{3.82}$$

because the maximum number of positive tests that contain at least one item in $\mathcal{D} \setminus \{i\}$ is at most $\gamma(d-1)$. Combining the results in (3.80), (3.82), and (3.81) into (3.51), we obtain

$$\begin{aligned}
\mathbb{P}^{\mathrm{SSS}}(\mathrm{err}) &\geq \sum_{i\in\mathcal{D}} \frac{\left(\frac{c_1\gamma(d-1)}{T}\right)^{2\gamma}(1-o(1))}{\left(\frac{\gamma(d-1)}{T}\right)^{\gamma} + \frac{d-1}{1-\epsilon_1}\left(\frac{c_2\gamma(d-2)+\gamma}{T}\right)^{2\gamma}(1-o(1)) + o(1)} && (3.83)\\
&\overset{(a)}{=} \frac{d\left(\frac{\gamma(d-1)}{T}\right)^{2\gamma} c_1^{2\gamma}(1-o(1))}{\left(\frac{\gamma(d-1)}{T}\right)^{\gamma} + \frac{d-1}{1-\epsilon_1}\left(\frac{\gamma(d-1)}{T}\right)^{2\gamma} c_3^{2\gamma}(1-o(1)) + o(1)} && (3.84)\\
&= \frac{d\left(\frac{\gamma(d-1)}{T}\right)^{\gamma} c_1^{2\gamma}(1-o(1))}{1 + \frac{d-1}{1-\epsilon_1}\left(\frac{\gamma(d-1)}{T}\right)^{\gamma} c_3^{2\gamma}(1-o(1)) + o(1)} && (3.85)\\
&= \frac{d\left(\frac{c_1^2\gamma(d-1)}{T}\right)^{\gamma}(1-o(1))}{1 + \frac{d-1}{1-\epsilon_1}\left(\frac{c_3^2\gamma(d-1)}{T}\right)^{\gamma}(1-o(1)) + o(1)}, && (3.86)
\end{aligned}$$

where (a) is by applying (3.79) in the denominator. Substituting $T = \gamma d^{1/\gamma}(d-1)(c_3^2)(1-\epsilon_2)$ for some constant $\epsilon_2 > 0$ into (3.86), we get

$$\begin{aligned}
\mathbb{P}^{\mathrm{SSS}}(\mathrm{err}) &\geq \frac{d\left(\frac{c_1^2\gamma(d-1)}{\gamma d^{1/\gamma}(d-1)(c_3^2)(1-\epsilon_2)}\right)^{\gamma}(1-o(1))}{1 + \frac{d-1}{1-\epsilon_1}\left(\frac{c_3^2\gamma(d-1)}{\gamma d^{1/\gamma}(d-1)(c_3^2)(1-\epsilon_2)}\right)^{\gamma}(1-o(1)) + o(1)} && (3.87)\\
&= \frac{\left(\frac{c_1}{c_3}\right)^{2\gamma}\left(\frac{1}{1-\epsilon_2}\right)^{\gamma}(1-o(1))}{1 + \frac{d-1}{d(1-\epsilon_1)}\left(\frac{1}{1-\epsilon_2}\right)^{\gamma}(1-o(1)) + o(1)} && (3.88)\\
&\overset{(a)}{=} \frac{\left(1-\delta_n^{(2)} - (\gamma(d-2))^{-1/3}\right)^{2\gamma}}{\left(1 + \left(\frac{d-2}{d-1}\right)\left(-\delta_n^{(2)} + (\gamma(d-2))^{-1/3}\right)\right)^{2\gamma}} \cdot \frac{1-o(1)}{(1-\epsilon_2)^{\gamma} + \frac{d-1}{d(1-\epsilon_1)}} && (3.89)\\
&\overset{(b)}{=} \frac{1-o(1)}{(1-\epsilon_2)^{\gamma} + \frac{d-1}{d(1-\epsilon_1)}}, && (3.90)
\end{aligned}$$

where we substitute c_1 and c_3 in (a). In (b), we note that both the numerator and denominator are in $\big(1-O\big(\frac{1}{d^{1/\gamma}}\big) \pm O\big(\frac{1}{(\gamma d)^{1/3}}\big)\big)^{2\gamma} = \big(1-O\big(\frac{1}{d^{1/\gamma}}\big)\big)^{2\gamma}$, and then apply Lemma 3.2.1. Consider the scaling regime $\gamma \in \Theta\big((\log n)^c\big)$ for some $c \in [0,1)$. Then, the right hand side above approaches 1 if $c > 0$ (large γ) and is typically close to 1 if $c = 0$ (constant γ).

Bounding COMP Error Probability

For any non-defective item $i \notin \mathcal{D}$, and $w \in [\gamma d(1-\delta_n)-(\gamma d)^{2/3}, \gamma d(1-\delta_n)+(\gamma d)^{2/3}]$, we have

$$\mathbb{P}(i \text{ masked by } \mathcal{D}|W^{(\mathcal{D})}=w) \geq \min_{\substack{w\in[\gamma d(1-\delta_n)-(\gamma d)^{2/3} \\ ,\gamma d(1-\delta_n)+(\gamma d)^{2/3}]}} \mathbb{P}(i \text{ masked by } \mathcal{D}|W^{(\mathcal{D})}=w) \tag{3.91}$$

$$\overset{(a)}{=} \Big(\frac{\gamma d(1-\delta_n^{(4)})}{T}\Big)^{\gamma}, \tag{3.92}$$

where (a) is because we introduced $\delta_n - (\gamma d)^{-1/3} \leq \delta_n^{(4)} \leq \delta_n + (\gamma d)^{-1/3}$ (where w depends on $\delta_n^{(4)}$), which results in $\mathbb{P}(i \text{ masked by } \mathcal{D}|W^{(\mathcal{D})}=w)$ to be minimized. From the inequality of w, we know that $\delta_n^{(4)} \in O\big(\frac{\gamma d}{T}\big) = O\big(\frac{1}{d^{1/\gamma}}\big)$. Using the above equation, we derive an upper bound on $\mathbb{P}^{\mathrm{COMP}}(\mathrm{suc})$:

$$\mathbb{P}^{\mathrm{COMP}}(\mathrm{suc}) = \sum_w \mathbb{P}(W^{(\mathcal{D})}=w)\mathbb{P}^{\mathrm{COMP}}(\mathrm{suc}|W^{(\mathcal{D})}=w) \tag{3.93}$$

$$= \sum_w \mathbb{P}(W^{(\mathcal{D})}=w)\mathbb{P}(\text{all } i \text{ not masked by } \mathcal{D}|W^{(\mathcal{D})}=w) \tag{3.94}$$

$$= \sum_w \mathbb{P}(W^{(\mathcal{D})}=w)\Big(1-\mathbb{P}(i \text{ masked by } \mathcal{D}|W^{(\mathcal{D})}=w)\Big)^{n-d} \tag{3.95}$$

$$= \sum_{\substack{w\in[\gamma d(1-\delta_n)-(\gamma d)^{2/3} \\ ,\gamma d(1-\delta_n)+(\gamma d)^{2/3}]}} \mathbb{P}(W^{(\mathcal{D})}=w)\Big(1-\mathbb{P}(i \text{ masked by } \mathcal{D}|W^{(\mathcal{D})}=w)\Big)^{n-d}$$
$$+ \sum_{\substack{w\notin[\gamma d(1-\delta_n)-(\gamma d)^{2/3} \\ ,\gamma d(1-\delta_n)+(\gamma d)^{2/3}]}} \mathbb{P}(W^{(\mathcal{D})}=w)\Big(1-\mathbb{P}(i \text{ masked by } \mathcal{D}|W^{(\mathcal{D})}=w)\Big)^{n-d} \tag{3.96}$$

$$\overset{(a)}{\leq} \bigg(1-\Big(\frac{\gamma d(1-\delta_n^{(4)})}{T}\Big)^{\gamma}\bigg)^{n-d} + 2\exp(-2(\gamma d)^{\frac{1}{3}}) \tag{3.97}$$

$$\overset{(b)}{=} \bigg(1-\Big(\frac{\gamma d(1-\delta_n^{(4)})}{T}\Big)^{\gamma}\bigg)^{n-d} + o(1), \tag{3.98}$$

where in (a) the first term is obtained by first applying (3.92), then simply upper bounding $\mathbb{P}(W^{(\mathcal{D})} \in [\gamma d(1-\delta_n)-(\gamma d)^{2/3}, \gamma d(1-\delta_n)+(\gamma d)^{2/3}])$ by 1. For the second term, we first upper bound $\big(1-\mathbb{P}(i \text{ masked by } \mathcal{D}|W^{(\mathcal{D})}=w)\big)^{n-d}$ by 1. Next, we apply Lemma 3.2.2 to upper bound $\mathbb{P}(W^{(\mathcal{D})} \notin [\gamma d(1-\delta_n)-(\gamma d)^{2/3}, \gamma d(1-\delta_n)+(\gamma d)^{2/3}])$. (b) is obtained by using the fact that $2\exp(-2(\gamma d)^{\frac{1}{3}}) \to 0$ since $\gamma \geq 1$ and $d \to \infty$. Applying the above results, we provide a lower bound on $\mathbb{P}^{\text{COMP}}(\text{err})$:

$$\begin{aligned}
\mathbb{P}^{\text{COMP}}(\text{err}) &= 1 - \mathbb{P}^{\text{COMP}}(\text{suc}) && (3.99)\\
&\geq 1 - \bigg(1 - \Big(\frac{\gamma d(1-\delta_n^{(4)})}{T}\Big)^{\gamma}\bigg)^{n-d} - o(1) && (3.100)\\
&= 1 - \bigg(1 - \Big(\frac{\gamma d}{T}\Big)^{\gamma}(1-\delta_n^{(4)})^{\gamma}\bigg)^{n-d} - o(1). && (3.101)
\end{aligned}$$

We stick to the same $T = \gamma d^{\frac{1}{\gamma}}(d-1)(c_3^2)(1-\epsilon_2)$ for some constant $\epsilon_2 > 0$. Substituting our T into (3.101), we get

$$\begin{aligned}
\mathbb{P}^{\text{COMP}}(\text{err}) &\geq 1 - \bigg(1 - \Big(\frac{\gamma d}{\gamma d^{1/\gamma}(d-1)(c_3^2)(1-\epsilon_2)}\Big)^{\gamma}(1-\delta_n^{(4)})^{\gamma}\bigg)^{n-d} - o(1) && (3.102)\\
&= 1 - \bigg(1 - \frac{1}{d}\Big(1 + \frac{1}{d-1}\Big)^{\gamma}\frac{(1-\delta_n^{(4)})^{\gamma}}{c_3^{2\gamma}(1-\epsilon)^{\gamma}}\bigg)^{n-d} - o(1) && (3.103)\\
&\overset{(a)}{=} 1 - \bigg(1 - \frac{(1+o(1))}{d(1-\epsilon_2)^{\gamma}}\bigg)^{n-d} - o(1) && (3.104)\\
&\overset{(b)}{=} 1 - \exp\Big(-\frac{(n-d)(1+o(1))}{d(1-\epsilon_2)^{\gamma}}\Big) - o(1) && (3.105)\\
&= 1 - \exp\Big(-\frac{n(1+o(1))}{d(1-\epsilon_2)^{\gamma}}\Big) - o(1), && (3.106)
\end{aligned}$$

where (a) $\big(1+\frac{1}{d-1}\big)^{\gamma} = 1+o(1)$ because $d \gg \gamma$. We also note that $c_3, 1-\delta_n^{(4)} \in 1 - O\big(\frac{1}{d^{1/\gamma}}\big)$ and apply Lemma 3.2.1 to get $c_3^{2\gamma}, (1-\delta_n^{(4)})^{\gamma} \in 1-o(1)$. (b) is by first noting that $\frac{1+o(1)}{d(1-\epsilon_2)^{\gamma}} \in o(1)$, then applying $1-x = e^{-x(1+o(1))}$ when $x \in o(1)$. To see why $\frac{1+o(1)}{d(1-\epsilon_2)^{\gamma}} \in o(1)$, we can take the log of the denominator and substitute the respective scaling regimes to get $\theta \log n + (\log n)^c \log(1-\epsilon_2) \gg 0$. The right hand side approaches 1 as $n \to \infty$ since $\exp\big(-\frac{n}{d(1-\epsilon_2)^{\gamma}}\big) \to 0$ because $d \in \Theta(n^{\theta})$ for some $\theta \in (0,1)$.

3.5.3 Proof of Theorem 3.3.3 (Converse for ρ-sized Tests)

We follow the same steps as Theorem 3.3.1 until equation (3.45). Since $c = \rho T$ for ρ-sized tests, we substitute $c = \rho T$ into (3.42) and further substitute the result into (3.45). Thus, we have

$$\epsilon \geq 1 - \frac{\frac{\rho T d(1+o(1))}{n}\big[\log\big(\frac{n}{\rho d(1+o(1))}\big) + 1 + \frac{1}{\log 2}\big] + \log 2}{\log\binom{n}{d}}, \tag{3.107}$$

which simplifies to

$$T \geq \frac{n}{\rho(1+o(1))} \times \frac{(1-\epsilon)\log(\frac{n}{d}) - \frac{\log 2}{d}}{\log(\frac{n}{\rho d(1+o(1))}) + 1 + \frac{1}{\log 2}} \tag{3.108}$$

$$= \frac{n}{\rho(1+o(1))} \times \frac{(1-\epsilon)\log(\frac{n}{d})}{\log(\frac{n}{\rho d})}(1+o(1)) \tag{3.109}$$

$$\overset{(a)}{=} \frac{1-\epsilon}{1-\beta}\left(\frac{n}{\rho}\right)(1+o(1)), \tag{3.110}$$

where (a) is because $\rho \in \Theta((n/d)^{\beta})$.

3.5.4 Proof of Theorem 3.4.1 (DD Performance)

We observe that first and second steps recover $\mathcal{D}$ correctly when each defective item i is not masked by $\mathcal{PD} \setminus \{i\}$. Hence, we want to derive a bound on T when the probability of each defective item i being masked by $\mathcal{PD} \setminus \{i\}$ is vanishing. Each defective item i is masked by $\mathcal{PD} \setminus \{i\}$ only when the number of collisions between i and $\mathcal{PD}$ is γ. Since $\mathcal{PD}$ can be split into two sets $\mathcal{D}$ and $\mathcal{PD} \setminus \mathcal{D}$, we can consider the number of collisions between i and each of these two sets separately. This motivates the main steps of our proof:

1. We derive a concentration result on the number of non-defective items in $\mathcal{PD}$.

2. We derive a bound on T when the probability of the event, where there exists a defective item i where the number of collisions between i and $\mathcal{D} \setminus \{i\}$ is "close to γ", is vanishing.

3. Conditioning the event where for all $i \in \mathcal{D}$, the number of collisions between defective item i and $\mathcal{D} \setminus \{i\}$ "is small", we derive a bound on T when the probability of the event, where every test that includes defective item i—where it is the only defective item—contains at least one item from $\mathcal{PD} \setminus \mathcal{D}$, is vanishing.

4. Taking the maximum between the two bounds on T gives us the required number of tests.

We analyse the two steps of the DD algorithm separately before combining their results together.

Analysis of the First Step

Let $G = |\mathcal{PD} \setminus \mathcal{D}|$ denote the number of non-defective items in $\mathcal{PD}$, where $G = \sum_{i=1}^{n-d} G_i$ with $G_i \in \{0,1\}$. Setting the number of positive tests as $W^{(\mathcal{D})} = w^{(\mathcal{D})}$, we have

$$(G|W^{(\mathcal{D})} = w^{(\mathcal{D})}) \sim \text{Binomial}\bigg(n-d, \Big(\frac{w^{(\mathcal{D})}}{T}\Big)^{\gamma}\bigg), \tag{3.111}$$

where $\big(\frac{w^{(\mathcal{D})}}{T}\big)^{\gamma} \leq \big(\frac{\gamma d}{T}\big)^{\gamma}$ because each of the d defective items can be tested at most γ times, meaning that the number of positive tests is at most γd. Since $\mathbb{P}[|G_i| \leq 1] = 1$, then for all positive t, we have the following by Bernstein's inequality:

$$\mathbb{P}\Big[\sum_{i=1}^{n-d} G_i > \mathbb{E}[G] + t\Big] \leq \exp\left(\frac{-\frac{1}{2}t^2}{\sum_{i=1}^{n-d}\text{Var}[G_i] + \frac{1}{3}t}\right) \tag{3.112}$$

$$\overset{(a)}{\leq} \exp\left(\frac{-\frac{1}{2}t^2}{\sum_{i=1}^{n-d}\mathbb{E}[G_i^2] + \frac{1}{3}t}\right) \tag{3.113}$$

$$\overset{(b)}{\leq} \exp\left(\frac{-\frac{1}{2}t^2}{\sum_{i=1}^{n-d}\mathbb{E}[G_i] + \frac{1}{3}t}\right) \tag{3.114}$$

$$\overset{(c)}{\leq} \exp\left(\frac{-\frac{1}{2}t^2}{\mathbb{E}[G] + \frac{1}{3}t}\right), \tag{3.115}$$

where (a) is because $\text{Var}[G_i] = \mathbb{E}[G_i^2] - (\mathbb{E}[G_i])^2 \leq \mathbb{E}[G_i^2]$, (b) is because $G_i \in \{0,1\}$ resulting in $\mathbb{E}[G_i^2] = \mathbb{E}[G_i]$, and (c) is due to the linearity of expectation.

Analysis of the Second Step

Firstly, we want to show that the event in which the number of collisions between a chosen defective item i and $\mathcal{D} \setminus \{i\}$ is "close to γ" (to be formalized later) is a rare event. It is easy to see that rearranging the columns of the test matrix does not change the items being involved in each test. Hence, for clarity, we think of the test matrix being rearranged, where the first d columns are for the defective items, as shown in Figure 3.3. Referring to Figure 3.3, let C_i be the number of collisions between a given defective item i (with $i = d$ in Figure 3.3) and $\mathcal{D} \setminus \{i\}$. Recall that $W^{(\mathcal{D}\setminus i)}$ denotes the number of positive tests containing at least one item in $\mathcal{D} \setminus \{i\}$. Given $W^{(\mathcal{D}\setminus i)} = w^{(\mathcal{D}\setminus i)}$, the probability of defective item i occurring in any of those tests equals $\frac{w^{(\mathcal{D}\setminus i)}}{T}$. This gives us the following distribution

$$(C_i|W^{(\mathcal{D}\setminus i)} = w^{(\mathcal{D}\setminus i)}) \sim \text{Binomial}\Big(\gamma, \frac{w^{(\mathcal{D}\setminus i)}}{T}\Big), \tag{3.116}$$

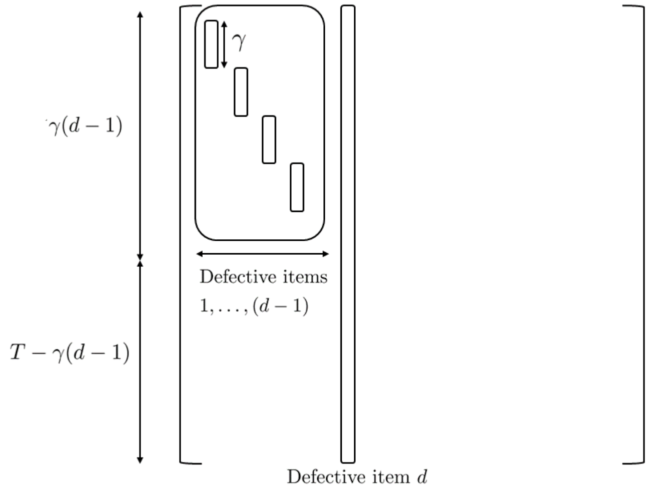

Figure 3.3: Rearranged test matrix to show that collision within defectives are rare.

where $\frac{w^{(\mathcal{D}\setminus i)}}{T} \leq \frac{\gamma(d-1)}{T} < \frac{\gamma d}{T}$ because any $w^{(\mathcal{D}\setminus i)}$ is at most $\gamma(d-1)$. We want to show that $\mathbb{P}[C_i \geq \alpha_2\gamma]$ is small, where $\alpha_2 \in (0,1)$. We first note that

$$\mathbb{P}[C_i = \alpha_2\gamma] = \binom{\gamma}{\alpha_2\gamma}\Big(\frac{w^{(\mathcal{D}\setminus i)}}{T}\Big)^{\alpha_2\gamma}(1-p)^{\gamma-\alpha_2\gamma} \tag{3.117}$$

$$\overset{(a)}{\leq} 2^{\gamma H_2(\alpha_2)}\Big(\frac{\gamma d}{T}\Big)^{\alpha_2\gamma}. \tag{3.118}$$

where (a) is due to $\binom{\gamma}{\alpha_2\gamma} \leq 2^{\gamma H_2(\alpha_2)}$, $\frac{w^{(\mathcal{D}\setminus i)}}{T} < \frac{\gamma d}{T}$, and $(1-p)^{\gamma-\alpha_2\gamma} \leq 1$. We proceed to show that $\mathbb{P}[C_i \geq \alpha_2\gamma]$ behaves similarly to $\mathbb{P}[C_i = \alpha_2\gamma]$:

$$\mathbb{P}[C_i \geq \alpha_2\gamma] = \sum_{k=\alpha_2\gamma}^{\gamma}\mathbb{P}[C_i = k] \tag{3.119}$$

$$\leq \sum_{k=\alpha_2\gamma}^{\gamma} 2^{\gamma H_2(k/\gamma)}\Big(\frac{\gamma d}{T}\Big)^k \tag{3.120}$$

$$\overset{(a)}{\leq} 2^{\gamma H_2(\max\{\alpha_2,\frac{1}{2}\})}\sum_{k=\alpha_2\gamma}^{\infty}\Big(\frac{\gamma d}{T}\Big)^k \tag{3.121}$$

$$\overset{(b)}{=} 2^{\gamma H_2(\max\{\alpha_2,\frac{1}{2}\})}\frac{(\frac{\gamma d}{T})^{\alpha_2\gamma}}{1-(\frac{\gamma d}{T})} \tag{3.122}$$

$$\overset{(c)}{=} 2^{\gamma H_2(\max\{\alpha_2,\frac{1}{2}\})}\Big(\frac{\gamma d}{T}\Big)^{\alpha_2\gamma}(1+o(1)), \tag{3.123}$$

where in (a) we note the fact that $H_2(p)$ is increasing for $p \leq 2$ and decreasing for $p \geq 2$. Hence, α_2 provides a tighter bound when $\alpha_2 > 1/2$. We used the sum to infinity of geometric series in (b), and the fact that $\frac{\gamma d}{T} \in o(1)$ in (c). By the union bound, we have the following:

$$\mathbb{P}\Big[\bigcup_{i=1}^{d}\{C_i \geq \alpha_2\gamma\}\Big] \leq d2^{\gamma H_2(\max\{\alpha_2,\frac{1}{2}\})}\Big(\frac{\gamma d}{T}\Big)^{\alpha_2\gamma}(1+o(1)). \tag{3.124}$$

Hence, for the above union bound to approach 0, we consider the following condition for T, where β_n is a slowly decaying term as $n \to \infty$:

$$2^{\gamma H_2(\max\{\alpha_2,\frac{1}{2}\})}\Big(\frac{\gamma d}{T}\Big)^{\alpha_2\gamma}(1+o(1)) \leq \frac{\beta_n}{d}(1+o(1)), \tag{3.125}$$

which simplifies to

$$T \geq \gamma d 2^{\frac{1}{\alpha_2}H_2(\max\{\alpha_2,\frac{1}{2}\})}\Big(\frac{d}{\beta_n}\Big)^{\frac{1}{\alpha_2\gamma}}. \tag{3.126}$$

Now, we study the probability of defective item i not being in $\widehat{\mathcal{D}}$. We will first condition on the event that for any defective item i, the number of collisions between defective item i and $\mathcal{D}\setminus\{i\}$ "is small" (to be formalized later). After conditioning, we consider the event where every test that includes defective item i—where it is the only defective item—contains at least one item from $\mathcal{PD}\setminus\mathcal{D}$. This is equivalent to the event that defective item i is not in $\widehat{\mathcal{D}}$. We derive a bound on T when the probability of that event is vanishing. We start by claiming that $T \in \Omega\big(\gamma d\big(\frac{n}{d}\big)^{1/\gamma}d^{1/\gamma^2}\big)$.

We condition on the following events with their explanations below:

1. $\bigcap_{i=1}^{d}\{C_i < \alpha_2\gamma\}$ because previously, we derived a bound on T that ensures $\mathbb{P}[\bigcup_{i=1}^{d}\{C_i \geq \alpha_2\gamma\}] \to 0$. Thus, we condition on $\neg[\bigcup_{i=1}^{d}\{C_i \geq \alpha_2\gamma\}] \equiv \bigcap_{i=1}^{d}\{C_i < \alpha_2\gamma\}$ (by De Morgan's law).

2. $W^{(\mathcal{D})} = \gamma d(1-\delta_n^-)$ for $\delta_n^- \in O\big(\frac{1}{n^{(1-\theta)/\gamma}}\big)$ because we want δ_n^- to fall within the concentration result in Lemma 3.2.2. Applying the bound on T from our claim, we get $\delta_n^- \in O\big(\frac{\gamma d}{T}\big) = O\big(\frac{1}{(n/d)^{1/\gamma}d^{1/\gamma^2}}\big) = O\big(\frac{1}{n^{(1-\theta)/\gamma}}\big)$.

For clarity, we repeat the event that we are interested in: given $\bigcap_{i=1}^{d}\{C_i < \alpha_2\gamma\}$ and $W^{(\mathcal{D})} = \gamma d(1-\delta_n^-)$, defective item i is not in $\widehat{\mathcal{D}}$. The main steps to derive a bound on T are as follows: we derive a bound on G ensuring that the event is rare, which gives us our bound on T by further applying the concentration results from the analysis of the first step.

We start by looking at a single defective item. Let $\tilde{\gamma}$ be the number of tests in which defective item i is the only defective item. Recall that we conditioned on $\bigcap_{i=1}^{d}\{C_i < \alpha_2\gamma\}$. Hence, we have $\tilde{\gamma} \geq (1-\alpha_2)\gamma$. We want to find the probability that all $\tilde{\gamma}$ indices occur in tests where at least one non-defective item in $\mathcal{PD}$ is also present. We index the first $\tilde{\gamma}$ tests from 1 to $\tilde{\gamma}$. Let A_i be the event that the positive test indexed i (where i is from 1 to $\tilde{\gamma}$) contains at least one non-defective item in $\mathcal{PD}$. Similar to before, consider a population of $\gamma d(1-\delta_n^-)$ coupons. A collector makes $G\gamma$ uniformly random selections with replacement. We can think of the population as being the total number of positive tests $\gamma d(1-\delta_n^-)$, and the number of coupons collected as $G\gamma$: γ for each column in a total of G columns. From the coupon collection problem, A_i is equivalent to the event that coupon i is collected. We are interested in the following:

$$\mathbb{P}[A_1,\ldots,A_{\tilde{\gamma}}] = \frac{\#\text{ways to collect coupons including all of the first } \tilde{\gamma} \text{ indices}}{\#\text{ways to collect all coupons}}. \tag{3.127}$$

We will bound the numerator and denominator separately. For the denominator, we can think of lining up all the selected $G\gamma$ coupons in a row where each position can be any of the $\gamma d(1-\delta_n^-)$ coupons from the population. This gives us the following:

$$\#\text{ways to collect all coupons} = (\gamma d)^{G\gamma}(1-\delta_n^-)^{G\gamma}. \tag{3.128}$$

For the numerator, we have

$$\begin{aligned}
\begin{matrix}\#\text{ways to collect coupons}\\ \text{including all of the first } \tilde{\gamma} \text{ indices}\end{matrix} &\leq \left[\prod_{i=0}^{\tilde{\gamma}-1}(G\gamma - i)\right](\gamma d)^{G\gamma-\tilde{\gamma}}(1-\delta_n^-)^{G\gamma-\tilde{\gamma}} && (3.129)\\
&\leq (G\gamma)^{\tilde{\gamma}}(\gamma d)^{G\gamma-\tilde{\gamma}}(1-\delta_n^-)^{G\gamma-\tilde{\gamma}} && (3.130)\\
&= \left(\frac{G}{d}\right)^{\tilde{\gamma}}(\gamma d)^{G\gamma}(1-\delta_n^-)^{G\gamma-\tilde{\gamma}}. && (3.131)
\end{aligned}$$

because each index in the set $\{1, 2, \cdots, \tilde{\gamma}\}$ must minimally take one position in the sequence of $G\gamma$ coupons. Each index has at most $G\gamma$ choices. Afterwards, each of the remaining $(G\gamma - \tilde{\gamma})$ positions can take any of the $\gamma d(1-\delta_n^-)$ indices. Combining the bounds on numerator and denominator, we have

$$\mathbb{P}[A_1, \dots, A_{\tilde{\gamma}}] = \frac{\#\text{ways to collect coupons including all of the first } \tilde{\gamma} \text{ indices}}{\#\text{ways to collect all coupons}} \tag{3.132}$$

$$= \Big(\frac{G}{d}\Big)^{\tilde{\gamma}} (1-\delta_n^-)^{-\tilde{\gamma}} \tag{3.133}$$

$$\overset{(a)}{\leq} \Big(\frac{G}{d}\Big)^{(1-\alpha_2)\gamma} (1-\delta_n^-)^{-\tilde{\gamma}} \tag{3.134}$$

$$\overset{(b)}{=} \Big(\frac{G}{d}\Big)^{(1-\alpha_2)\gamma} (1+o(1)), \tag{3.135}$$

where (a) is due to the assumption[1] that $G \leq d$ and $(1-\alpha_2)\gamma \leq \tilde{\gamma}$, and (b) follows by recalling that $\delta_n^- \in O\big(\frac{1}{n^{(1-\theta)/\gamma}}\big)$, where the scaling regime is similar (differing only in the constant multiplicative factor of the power) to $O\big(\frac{1}{d^{1/\gamma}}\big) = O\big(\frac{1}{n^{\theta/\gamma}}\big)$, and applying Lemma 3.2.1. According to the DD algorithm, the event where there exists a defective item not in $\widehat{\mathcal{D}}$ is equivalent to there existing a defective item where all its $\tilde{\gamma}$ indices are collected by the $G\gamma$ coupons. Applying union bound, the bound on the probability is as follows.

$$\mathbb{P}[\exists \text{ defective item not in } \widehat{\mathcal{D}}] \leq \mathbb{P}\Big[\bigcup_{i=1}^{d} \{\text{all } \tilde{\gamma} \text{ indices of } i \text{ are collected}\}\Big] \tag{3.136}$$

$$\leq d\Big(\frac{G}{d}\Big)^{(1-\alpha_2)\gamma}(1+o(1)). \tag{3.137}$$

The bound approaches 0 if $\big(\frac{G}{d}\big)^{(1-\alpha_2)\gamma} \leq \frac{\beta_n}{d}$ where β_n is a slowly decaying term as $n \to \infty$. Rearranging, we get

$$G \leq d\Big(\frac{\beta_n}{d}\Big)^{\frac{1}{(1-\alpha_2)\gamma}}. \tag{3.138}$$

Combining the Analyses of Both Steps

We now combine our bound on G from the analysis of the second step with the concentration result from the analysis of the first step to obtain on a bound on T. We define

$$G_{\max} = d\Big(\frac{\beta_n}{d}\Big)^{\frac{1}{(1-\alpha_2)\gamma}}. \tag{3.139}$$

Recall that $\mathbb{E}[G] = (n-d)\big(\frac{w^{(\mathcal{D})}}{T}\big)^{\gamma}$, and hence, $\mathbb{E}[G] \leq G_{\max}/2$ is guaranteed when

$$(n-d)\Big(\frac{\gamma d}{T}\Big)^{\gamma} \leq \frac{d}{2}\Big(\frac{\beta_n}{d}\Big)^{\frac{1}{(1-\alpha_2)\gamma}}, \tag{3.140}$$

[1] This assumption is later shown to be true when we obtain our bound on G.

because $\big(\frac{w^{(\mathcal{D})}}{T}\big)^{\gamma} \leq \big(\frac{\gamma d}{T}\big)^{\gamma}$. Rearranging, we obtain the condition

$$T \geq 2^{1/\gamma}\gamma d\Big(\frac{n-d}{d}\Big)^{\frac{1}{\gamma}}\Big(\frac{d}{\beta_n}\Big)^{\frac{1}{(1-\alpha_2)\gamma^2}}, \tag{3.141}$$

which satisfies our initial claim. Combining (3.126) and (3.141), we get

$$T \geq \gamma d \max\bigg\{2^{\frac{1}{\alpha_2}H_2(\max\{\alpha_2,\frac{1}{2}\})}\Big(\frac{d}{\beta_n}\Big)^{\frac{1}{\alpha_2\gamma}}, 2^{1/\gamma}\Big(\frac{n-d}{d}\Big)^{\frac{1}{\gamma}}\Big(\frac{d}{\beta_n}\Big)^{\frac{1}{(1-\alpha_2)\gamma^2}}\bigg\}. \tag{3.142}$$

We now provide a bound on the total error probability. Setting $t = G_{\max}/2$ in our inequality in (3.115), we get

$$\mathbb{P}\Big[G > \mathbb{E}[G] + \frac{G_{\max}}{2}\Big] \leq \exp\bigg(\frac{-\frac{1}{2}\big(\frac{G_{\max}}{2}\big)^2}{\mathbb{E}[G] + \frac{1}{3}\big(\frac{G_{\max}}{2}\big)}\bigg). \tag{3.143}$$

Applying that fact that $\mathbb{E}[G] \leq G_{\max}/2$, we get

$$\mathbb{P}[G > G_{\max}] \leq \exp\Big(-\frac{3}{16}G_{\max}\Big) = \exp\bigg(-\frac{3d}{16}\Big(\frac{\beta_n}{d}\Big)^{\frac{1}{(1-\alpha_2)\gamma}}\bigg), \tag{3.144}$$

which approaches 0 as long is β_n is a slowly decaying function (*e.g.*, log factors only). By combining all the error probabilities in (3.144), (3.124), (3.137), and Lemma 3.2.2, we have

$$\begin{aligned}
\mathbb{P}[\text{total error}] &\leq \mathbb{P}[G > G_{\max}] + \mathbb{P}\Big[\bigcup_{i=1}^{d}\{C_i \geq \alpha_2\gamma\}\Big] + \mathbb{P}[\exists \text{ defective item not in } \widehat{\mathcal{D}}] \\
&\quad + \mathbb{P}\Big(\delta_n^- \notin [\delta_n - (\gamma d)^{-1/3}, \delta_n + (\gamma d)^{-1/3}]\Big) & (3.145) \\
&\overset{(a)}{\leq} \exp\bigg(-\frac{3d}{16}\Big(\frac{\beta_n}{d}\Big)^{\frac{1}{(1-\alpha_2)\gamma}}\bigg) + d2^{\gamma H_2(\max\{\alpha_2,\frac{1}{2}\})}\Big(\frac{\gamma d}{T}\Big)^{\alpha_2\gamma}(1+o(1)) \\
&\quad + d\Big(\frac{G}{d}\Big)^{(1-\alpha_2)\gamma}(1+o(1)) + 2\exp(-2(\gamma d)^{1/3}) & (3.146) \\
&\overset{(b)}{\leq} \exp\bigg(-\frac{3d}{16}\Big(\frac{\beta_n}{d}\Big)^{\frac{1}{(1-\alpha_2)\gamma}}\bigg) + 2\beta_n(1+o(1)) + 2\exp(-2(\gamma d)^{1/3}), & (3.147)
\end{aligned}$$

where (a) is because we apply Lemma 3.2.2 for the last term, and (b) is because we have $d2^{\gamma H_2(\max\{\alpha_2,\frac{1}{2}\})}\big(\frac{\gamma d}{T}\big)^{\alpha_2\gamma}(1+o(1)) \leq \beta_n(1+o(1))$ from (3.125) and $d\big(\frac{G}{d}\big)^{(1-\alpha_2)\gamma}(1+o(1)) \leq \beta_n(1+o(1))$ from (3.138).

Chapter 4

Conclusion and Future Work

In Chapter 1, we motivated and set up both the group testing problem and its sparse counterpart.

In Chapter 2, motivated by the fact that relatively little is known in the sparse adaptive setting, we studied the sparse adaptive setting, and provided information theoretic lower bounds for γ-divisible items, and algorithms for both γ-divisible items and ρ-sized tests.

In Chapter 3, we considered the non-adaptive setting. We first provided a generalization for both γ-divisible items constraint and ρ-sized tests constraint, which we used to provided information theoretic lower bounds for both setting. Furthermore, we improved the upper bound on the number of tests for γ-divisible items by considering a more refined algorithm, and analyzing it.

An interesting extension of this work would be to strengthen the converse in Theorem 3.3.2 by extending the arguments presented in [7] to the sparse setting in order to remove the need for a specific test design, and to obtain an error probability that approaches 1 as $n \to \infty$ for test numbers below the converse bound. Another interesting extension to the γ-divisible items constraint is to consider the case where different items have a different divisibility γ_i. This results in the general constraint: the number of ones in the test matrix is less than or equal to $\sum_{i=1}^{n} \gamma_i$. This setup might lead to a more general information theoretic converse and algorithm(s).

References

M. Aldridge. The capacity of Bernoulli nonadaptive group testing. *IEEE Trans. Inf. Theory*, 63(11):7142–7148, 2017.

M. Aldridge, L. Baldassini, and O. Johnson. Group testing algorithms: Bounds and simulations. *IEEE Trans. Inf. Theory*, 60(6):3671–3687, June 2014. ISSN 0018-9448. doi: 10.1109/TIT.2014.2314472.

M. Aldridge, O. Johnson, and J. Scarlett. Group testing: An information theory perspective. *Foundations and Trends in Communications and Information Theory*, 15(3-4):196–392, 2019.

R. B. Ash. *Information Theory.* Dover Publications, Inc., 1990.

L. Baldassini, O. Johnson, and M. Aldridge. The capacity of adaptive group testing. In *IEEE Int. Symp. Inf. Theory*, pages 2676–2680, July 2013. doi: 10.1109/ISIT.2013.6620712.

C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri. Non-adaptive group testing: Explicit bounds and novel algorithms. *IEEE Trans. Inf. Theory*, 60(5):3019–3035, May 2014. ISSN 0018-9448.

A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick. Optimal non-adaptive group testing. https://arxiv.org/abs/1911.02287, 2019.

T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, Inc., 2006.

P. Damaschke and A. S. Muhammad. Competitive group testing and learning hidden vertex covers with minimum adaptivity. *Disc. Math., Algs. and Apps.*, 2(03):291–311, 2010.

D. de Caen. A lower bound on the probability of a union. *Discrete mathematics*, 169(1): 217–220, 1997.

R. Dorfman. The detection of defective members of large populations. *Ann. Math. Stats.*, 14 (4):436–440, 1943.

M. Falahatgar, A. Jafarpour, A. Orlitsky, V. Pichapati, and A. T. Suresh. Estimating the number of defectives with group testing. In *IEEE Int. Symp. Inf. Theory*, pages 1376–1380, 2016.

V. Gandikota, E. Grigorescu, S. Jaggi, and S. Zhou. Nearly optimal sparse group testing. *IEEE Trans. Inf. Theory*, 65(5):2760 – 2773, 2019. doi: 10.1109/TIT.2019.2891651.

J. L. Gastwirth and P. A.Hammick. Rare-allele detection using compressed se(que)nsing. *Journal of Statistical Planning and Inference*, 1989. doi: 10.1016/0378-3758(89)90061-X.

A. Gilbert, M. Iwen, and M. Strauss. Group testing and sparse signal recovery. In *Asilomar Conf. Sig., Sys. and Comp.*, pages 1059–1063, Oct. 2008. doi: 10.1109/ACSSC.2008.5074574.

F. Hwang. A method for detecting all defective members in a population by group testing. *J. Amer. Stats. Assoc.*, 67(339):605–608, 1972.

P. Indyk. Deterministic superimposed coding with applications to pattern matching. *38th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 127–136, 1997. doi: https://doi/10.1137/1.9781611973075.91.

O. Johnson. Strong converses for group testing from finite blocklength results. *IEEE Trans. Inf. Theory*, 63(9):5923–5933, Sept. 2017. ISSN 0018-9448. doi: 10.1109/TIT.2017.2697358.

T. Madej. An application of group testing to the file comparison problem. *9th International Conference on Distributed Computing Systems*, 1989. doi: 10.1109/ICDCS.1989.37952.

C. McDiarmid. *On the method of bounded differences*, page 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

T. Richardson and R. Urbanke. *Modern coding theory.* Cambridge University Press, 2008.

Z. Samuel. *Advanced Calculus, An Introduction To Mathematical Analysis.* World Scientific, 1997.

J. Scarlett and V. Cevher. Phase transitions in group testing. In *Proc. ACM-SIAM Symp. Disc. Alg. (SODA)*, 2016.

N. Shental, A. Amir, and O. Zuk. Rare-allele detection using compressed se(que)nsing. *Nucleic Acids Research*, 2009.

N. Tan and J. Scarlett. Near-optimal sparse adaptive group testing. 2020.

K. H. Thompson. Estimation of the proportion of vectors in a natural population of insects. *Biometrics*, 18(4):568–578, 1962. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2527902.

J. Wang, E. Lo, and M. L. Yiu. Identifying the most connected vertices in hidden bipartite graphs using group testing. *IEEE Transactions on Knowledge and Data Engineering*, 25(10): 2245–2256, 2013.

J. Wolf. Born again group testing: Multiaccess communications. *IEEE Transactions on Information Theory*, 31(2):185–191, March 1985. ISSN 1557-9654. doi: 10.1109/TIT.1985.1057026.