

Improved Bounds and Algorithms for Sparsity-Constrained Group Testing

1

Nelvin Tan and Jonathan Scarlett

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) in which 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 a simple asymptotically optimal adaptive algorithm for ρ -sized tests. In addition, in the non-adaptive setting with γ -divisible items, we use the Definite Defectives (DD) decoding algorithm and study bounds on the required number of tests for vanishing error probability under the random near-constant test-per-item design. We show that the number of tests required can be significantly less than the Combinatorial Orthogonal Matching Pursuit (COMP) decoding algorithm, and is asymptotically optimal in broad scaling regimes.

I. INTRODUCTION

In the group testing problem, the goal is to identify a subset of defective items of size d within a larger set of items of size n based on a number T of tests. We consider the noiseless setting, in which we are guaranteed that the test procedure is perfectly reliable: We get a negative test outcome if all items in the test are non-defective, and a positive outcome if at least one item in the test is defective. This problem is relevant in areas such as medical testing [2], data science [3], communication protocols [4], and many more [5]. One of the defining features of the group testing problem is the distinction between the non-adaptive and adaptive settings. In the non-adaptive

The authors are with the Department of Computer Science, National University of Singapore (NUS). J. Scarlett is also with the Department of Mathematics, NUS. e-mail: nelvintan@u.nus.edu; scarlett@comp.nus.edu.sg

This work will be presented in part at the 2020 IEEE International Symposium on Information Theory (ISIT) [1].

This work was supported by the Singapore National Research Foundation (NRF) under grant number R-252-000-A74-281.

setting, all tests must be designed prior to observing any outcomes, whereas in the adaptive testing, each test can be designed based on previous test outcomes.

While the majority of the group testing literature has allowed for arbitrary (i.e., unconstrained) test designs, these may be unrealistic in many practical scenarios. To address this, a *sparsity-constrained* group testing setting was recently proposed in [6], in which the tests are subject to one of two constraints: (a) items are *finitely divisible* and thus may participate in at most γ tests; or (b) tests are *size-constrained* and thus contain no more than ρ items per test. These constraints are motivated by physical considerations, where each item has a limitation on the number of samples (e.g., blood from a patient) it can be divided into, or the testing equipment has a limitation on the number of items (e.g., volume of blood a machine can hold). The focus in [6] was on non-adaptive testing, and the two main goals of the present paper are the following: (i) provide a detailed treatment of the adaptive setting; (ii) close some notable gaps between the upper and lower bounds derived in [6] in the non-adaptive setting with γ -divisible items (in contrast, the gaps in [6] were less significant for ρ -sized tests).

A. Problem Setup

Let n be the number of items, which we label as $\{1, 2, \dots, n\}$. Let $\mathcal{D} \subset \{1, 2, \dots, n\}$ be the set of defective items, and $d = |\mathcal{D}|$ be the number of defective items. We assume that \mathcal{D} is generated uniformly at random among all sets of size d (also known as the combinatorial prior [5]).¹

We are interested in asymptotic scaling regimes in which n is large and d comparatively is small, and thus assume that $d = o(n)$ throughout. We let $T = T(n)$ be the *number of tests* performed and label the tests $\{1, 2, \dots, T\}$. To keep track of the design of the test pools in the non-adaptive setting, we write $x_{ti} = 1$ to denote that item $i \in \{1, 2, \dots, n\}$ is in the pool for test $t \in \{1, 2, \dots, T\}$, and $x_{ti} = 0$ otherwise. This can be represented by the matrix $\mathbf{X} \in \{0, 1\}^{T \times n}$, known as the *testing matrix* or *test design*.

Similar to most theoretical studies of group testing [5], our study of the non-adaptive setting will focus on random designs. The most commonly considered test design in the unconstrained setting is the Bernoulli design (i.e., \mathbf{X} has i.i.d. Bernoulli entries); however, this creates significant fluctuations in the number of tests per item, which is undesirable in the case of γ -divisible items. Hence, we will pay particular attention to the *near-constant tests-per-item* design [7], [8], in which each item is included in some fixed number L of tests, chosen uniformly at random with replacement. Since we are selecting with replacement, some items may be in fewer than L tests, hence the terminology “near-constant”. This is a mathematical convenience that makes the analysis more tractable.

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

¹Despite this assumption, our adaptive algorithm attains zero error probability (see Theorem 7), thus ensuring success even for the worst case \mathcal{D} of cardinality d .

disjunction) operator \bigvee , we have

$$y_t = \bigvee_{i \in \mathcal{D}} x_{ti}. \quad (1)$$

A *decoding* (or *detection*) *algorithm* is a (possibly randomized) function $\hat{D} : \{0, 1\}^{T \times n} \times \{0, 1\}^T \rightarrow \mathcal{P}(\{1, 2, \dots, n\})$, where the power-set $\mathcal{P}(\{1, 2, \dots, n\})$ is the collection of the subsets of items. Denoting the final estimate by $\hat{\mathcal{D}}$, the error probability is given by

$$P_e = \mathbb{P}(\hat{\mathcal{D}} \neq \mathcal{D}), \quad (2)$$

where the probability is taken over the randomness of the set of defective items, and also over the test design and/or decoding algorithm if they are randomized.

Testing Constraints. As mentioned above, our focus in this paper is on the sparsity-constrained group testing problem introduced in [6], in which 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.

For instance, in the classical application of testing blood samples for a given disease [2], the γ -divisible items constraint may arise when there are limitations on the volume of blood provided by each individual, and the ρ -sized tests constraint may arise when there are limitations on the number of blood samples that the machine can accept.

B. Related Work

There exists extensive literature providing group testing bounds and algorithms for unconstrained group testing [7]–[17]; see [5] for a recent overview. Here we focus our attention on those most relevant to the present paper.

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 [13], [14]. 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 $\epsilon \in (0, 1)$ [18], [19]. A matching upper bound is known for all $d = o(n)$ in the unconstrained adaptive setting [20], whereas matching this lower bound non-adaptively is only possible non-adaptively in certain sparser regimes [8], [15].

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 [7], [13], [14]. 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 for γ -divisible items 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

$$\frac{T}{\gamma d} \rightarrow \infty \quad (3)$$

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

In the non-adaptive setting, Gandikota *et al.* [6] proved the following results for γ -divisible items.

Theorem 1. [6] *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 2. [6] *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 [6].

Theorem 3. [6] *For any sufficiently large n , sufficiently small $\zeta > 0$, $\rho \in \Theta((\frac{n}{d})^\beta)$ (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 4. [6] *For any sufficiently large n , sufficiently small $\epsilon > 0$, $\rho \in \Theta((\frac{n}{d})^\beta)$ (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 $(\frac{1-6\epsilon}{1-\beta}) \frac{n}{\rho}$ 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.² 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 narrowed all the way down to a constant factor in the adaptive setting, and can also be significantly reduced in the non-adaptive setting. See the following subsection for further details.

C. Overview of the Paper

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

- In Section II, we consider the adaptive setting. We present an information-theoretic lower bound for γ -divisible items (Theorem 6), which strengthens the previous information-theoretic lower bound in [6] 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 using a near-optimal number of tests (Theorem

²See also [21] for very recent improvements providing sharp constants.

7 and Section II-C). Informally, under mild assumptions, the optimal number of tests is shown to be within a factor $e^{1+o(1)}$ of $\gamma d (\frac{n}{d})^{1/\gamma}$ for γ -divisible items, and within a $1 + o(1)$ factor of $\frac{n}{\rho}$ for ρ -sized tests.

- In Section III, we consider the non-adaptive setting. To further complement the preceding lower bound, we provide an additional lower bound that can be tighter (Corollary 1), but that is specific to the near-constant tests-per-item design (rather than general designs). In addition, we analyze the performance of the DD algorithm [13] (Theorem 9), and show that the number of tests can be significantly less than that of COMP (considered in [6]). Informally, a special case of our results states that in the scaling regime $d = \Theta(n^\theta)$ and $\gamma = \Theta((\log n)^c)$ with $\theta, c \in (0, 1)$, the optimal number of tests behaves as $\Theta(\gamma d \max\{n^\theta, n^{1-\theta}\}^{\frac{1+o(1)}{\gamma}})$.

In the final stages of preparing this paper, we noticed the concurrent work of [21], whose results are similar to those that we develop for the non-adaptive setting. In particular, the optimal number of tests for the near-constant tests-per-item design are characterized up to a constant factor in [21] whenever $\gamma = \Theta(1)$ (tight bounds are also given for ρ -sized test constraints, but these are more separate from our results). While our results are not quite as strong in this regime (see the discussion following Theorem 9), they have the advantage of also applying in regimes where $\gamma \rightarrow \infty$ (e.g., $\gamma = (\log n)^c$ for $c \in (0, 1)$). In addition, the proof techniques used are complementary, with ours building on [22] and [7], whereas [21] builds on [17] and [8]. Finally, the adaptive setting is not considered in [21].

Notation. Throughout the paper, the function $\log(\cdot)$ has base e , and we make use of Bachmann-Landau asymptotic notation (i.e., O , o , Ω , ω , Θ).

II. THE ADAPTIVE SETTING

In this section, we seek information-theoretic bounds and algorithms for the adaptive setting, considering both the cases of γ -divisible items and ρ -sized tests.

A. 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 $\mathbb{P}(\text{suc}) = 1 - P_e$, following similar proof techniques as [18], with suitable refinements to account for the γ -divisibility constraint. Afterwards, we will use the bound on $\mathbb{P}(\text{suc})$ to prove the converse result (lower bound on T).

Theorem 5. (Counting-Based Bound) *In 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 $\mathbb{P}(\text{suc})$ satisfying*

$$\mathbb{P}(\text{suc}) \leq \frac{\sum_{i=0}^{\gamma d} \binom{T}{i}}{\binom{n}{d}}. \quad (4)$$

Proof. See Section II-D. □

The intuition behind (4) is that the denominator represents the number of defective sets of size d , and the numerator represents the number of possible tests outcomes (since there are always at most γd positive tests). Once (4) is in place, the following converse follows from an asymptotic analysis.

Theorem 6. (General Converse Bound) *Fix $\epsilon \in (0, 1)$, and suppose that $d \in o(n)$, $\gamma \in o(\log n)$, and $\gamma d \rightarrow \infty$ as $n \rightarrow \infty$. 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 $e^{-(1+o(1))} \gamma d (\frac{n}{d})^{1/\gamma}$ tests.*

Proof. See Section II-E. □

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 [18], [19].

Theorem 6 strengthens the previous information-theoretic lower bound in [6] 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 extending its validity to the adaptive setting (whereas [6] used an approach based on Fano's inequality that is specific to the non-adaptive setting).

B. Adaptive 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 .

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.³ Algorithm 1 is reminiscent of Hwang's generalized binary splitting algorithm [20], but the depth of the corresponding tree is controlled by using much more than two branches per split; see Figure 1.

Using Algorithm 1, we have the following theorem, which is proved throughout the remainder of the subsection.

Theorem 7. (Adaptive Algorithm Performance) *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 II-F. □

Comparisons: Referring to Theorem 1, the upper bound for the non-adaptive algorithm of [6] using a randomized test design is $T \leq \lceil e \gamma d (\frac{n}{\epsilon})^{1/\gamma} \rceil$, where ϵ is the target error probability. 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

³Note that we assume $d \in o(n)$ and $\gamma \in o(\log(\frac{n}{d}))$, meaning that $(\frac{n}{d})^{1/\gamma} \rightarrow \infty$. Hence, the effect of rounding is asymptotically negligible, and is accounted for by the $1 + o(1)$ term in the theorem statement.

Algorithm 1 Adaptive algorithm for γ -divisible items

Require: Number of items n , number of defective items d , and divisibility of each item γ

- 1: Initialize $M \leftarrow \left(\frac{n}{d}\right)^{\frac{\gamma-1}{\gamma}}$ and defective set $\mathcal{D} \leftarrow \emptyset$
 - 2: Arbitrarily group the n items into $\frac{n}{M}$ groups of size M
 - 3: Test each group and discard any that return a negative outcome
 - 4: Label the remaining groups incrementally as $G_j^{(0)}$, where $j = 1, 2, \dots$
 - 5: **for** $i = 1$ to $\gamma - 1$ **do**
 - 6: **for** each group $G_j^{(i-1)}$ from the previous stage **do**
 - 7: Arbitrarily group all items in $G_j^{(i-1)}$ into $M^{1/(\gamma-1)}$ sub-groups of size $M^{1-i/(\gamma-1)}$
 - 8: Test each sub-group and discard any that return a negative outcome
 - 9: Label the remaining sub-groups incrementally as $G_j^{(i)}$
 - 10: **end for**
 - 11: **end for**
 - 12: Add the items in all the remaining groups $G_j^{(\gamma-1)}$ to \mathcal{D}
 - 13: **return** \mathcal{D}
-

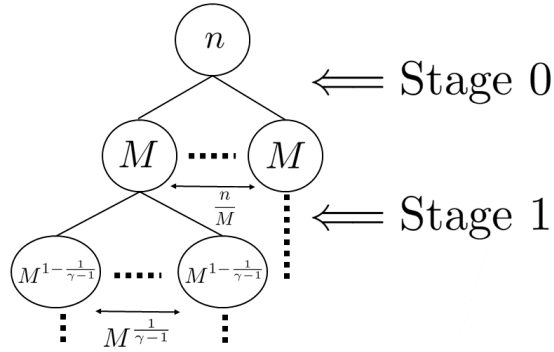


Fig. 1: Visualization of splitting in the adaptive algorithm.

bound of our algorithm matches the information-theoretic lower bound in Theorem 6 up to a constant factor of $e^{1+o(1)}$. This proves that our algorithm is nearly optimal.

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 [23], [24] 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 \hat{d} that is passed to Algorithm 1. Note that we need \hat{d} to be an overestimate for the proof of Theorem 7 to still apply (with \hat{d} in place of d).

Algorithm 2 Estimation of d

Require: Population of items, number of items n , upper bound $d_{\max} \geq d$, and a probability parameter β_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: $\hat{d} \leftarrow \# \text{positive bins} / (1 - \sqrt{\beta_n})$
 - 5: **return** \hat{d}
-

a) *Using 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 $(\frac{n}{d_{\max}})^{(\gamma-1)/\gamma}$. Substituting the updated value of M into (36), we obtain the following:

$$T \leq \frac{n}{(\frac{n}{d_{\max}})^{(\gamma-1)/\gamma}} + (\gamma - 1)d \left[\left(\frac{n}{d_{\max}} \right)^{\frac{\gamma-1}{\gamma}} \right]^{\frac{1}{\gamma-1}}, \quad (5)$$

which simplifies to

$$T \leq (d_{\max} - d + \gamma d) \left(\frac{n}{d_{\max}} \right)^{\frac{1}{\gamma}}. \quad (6)$$

b) *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 $n \rightarrow \infty$. We first run Algorithm 2 to obtain a new input \hat{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 $(\frac{n}{\hat{d}})^{1/\gamma}$ 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 $d \times (\text{bin size}) = d(\frac{\beta_n n}{d_{\max}})$, the (upper bound on the) number of defective items as \hat{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 $n \rightarrow \infty$. 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

$$\frac{d}{B} = \frac{d}{d_{\max}/\beta_n} \leq \frac{d}{d/\beta_n} = \beta_n \rightarrow 0, . \quad (7)$$

Secondly, we show that with high probability as $n \rightarrow \infty$, \hat{d} overestimates d . From (7), we have

$$\mathbb{E}[\# \text{collisions}] \leq d\beta_n, \quad (8)$$

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

$$\mathbb{P}(\text{\#collisions} \geq d\sqrt{\beta_n}) \leq \sqrt{\beta_n}, \quad (9)$$

which implies the following:

$$\mathbb{P}(d - \text{\#collisions} \geq d - d\sqrt{\beta_n}) \geq 1 - \sqrt{\beta_n} \quad (10)$$

$$\implies \mathbb{P}\left(\frac{d - \text{\#collisions}}{1 - \sqrt{\beta_n}} \geq d\right) \geq 1 - \sqrt{\beta_n}. \quad (11)$$

Since $(\text{\#positive bins} \geq d - \text{\#collisions})$ always holds, we have $\mathbb{P}(\hat{d} \geq d) \geq 1 - \sqrt{\beta_n}$, which tends to one because $\beta_n \rightarrow 0$.

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 $(d \times \text{bin size})$ items. To simplify notation, our updated inputs (labeled with subscript “new”) are

$$n_{\text{new}} = \frac{\beta_n dn}{d_{\max}}, \quad d_{\text{new}} = \hat{d}, \quad \gamma_{\text{new}} = \gamma - 1. \quad (12)$$

We can then run Algorithm 1 to recover the defective set. Substituting our updated inputs into (36) and using $M = \left(\frac{\beta_n dn}{d_{\max} \hat{d}}\right)^{\frac{\gamma-2}{\gamma-1}}$, we have the following bound for T :

$$T \leq \frac{d_{\max}}{\beta_n} + \frac{\beta_n dn}{d_{\max} \left(\frac{\beta_n dn}{d_{\max} \hat{d}}\right)^{\frac{\gamma-2}{\gamma-1}}} + (\gamma - 2)d \left(\frac{\beta_n dn}{d_{\max} \hat{d}}\right)^{\frac{1}{\gamma-1}}, \quad (13)$$

which simplifies to

$$T \leq \frac{d_{\max}}{\beta_n} + (\hat{d} - 2d + \gamma d) \left(\frac{\beta_n dn}{d_{\max} \hat{d}}\right)^{\frac{1}{\gamma-1}} \quad (14)$$

$$\stackrel{(a)}{\leq} \frac{d_{\max}}{\beta_n} + \left(\frac{d}{1 - \sqrt{\beta_n}} - 2d + \gamma d\right) \left(\frac{\beta_n n}{d_{\max}}\right)^{\frac{1}{\gamma-1}}, \quad (15)$$

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

Comparisons: By using T satisfying the derived upper bounds, the first approach recovers the defective set with zero error probability, whereas the second approach recovers the defective set with a small error probability determined by the β_n parameter. Referring to (6) and (15), 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, (6) is the better bound; its leading term is $\gamma d \left(\frac{n}{d}\right)^{1/\gamma}$. In particular, we note the following two cases: (i) If $\beta_n \ll \frac{1}{\gamma \left(\frac{n}{d}\right)^{1/\gamma}}$, then the $\frac{d_{\max}}{\beta_n}$ term in (15) is strictly higher than $\gamma d \left(\frac{n}{d}\right)^{1/\gamma}$; (ii) If $\beta_n \gg \frac{1}{\gamma \left(\frac{n}{d}\right)^{1/\gamma}}$, then some simple algebra gives $\frac{\beta_n n}{d} \gg \frac{1}{\gamma} \left(\frac{n}{d}\right)^{(\gamma-1)/\gamma}$, which implies that the $\gamma d \left(\frac{\beta_n n}{d}\right)^{1/(\gamma-1)}$ term from (15) is strictly higher than $\gamma d \left(\frac{n}{d}\right)^{1/\gamma}$ (note that $\left(\frac{1}{\gamma}\right)^{1/(\gamma-1)} = \Theta(1)$).

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 ρ

- 1: Initialize defective set $\mathcal{D} \leftarrow \emptyset$
 - 2: Randomly group n items into n/ρ groups of size ρ
 - 3: **for** each group G_i where $i \in \{1, 2, \dots, n/\rho\}$ **do**
 - 4: **while** testing G_i returns a positive outcome **do**
 - 5: run Algorithm 4 on G_i and add its one defective item output d^* into \mathcal{D}
 - 6: $G_i \leftarrow G_i \setminus \{d^*\}$
 - 7: **end while**
 - 8: **end for**
 - 9: **return** \mathcal{D}
-

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 (6) is $d_{\max} \left(\frac{n}{d_{\max}}\right)^{1/\gamma}$. Since the dominating term $\max \left\{ \frac{d_{\max}}{\beta_n}, \gamma d \left(\frac{\beta_n n}{d_{\max}}\right)^{1/(\gamma-1)} \right\}$ in (15) is not obvious, we consider both possibilities: (i) $d_{\max} \left(\frac{n}{d_{\max}}\right)^{1/\gamma} \gg \frac{d_{\max}}{\beta_n}$ whenever $\beta_n \gg \left(\frac{d_{\max}}{n}\right)^{1/\gamma}$; and (ii) $d_{\max} \left(\frac{n}{d_{\max}}\right)^{1/\gamma} \gg \gamma d \left(\frac{\beta_n n}{d_{\max}}\right)^{\frac{1}{\gamma-1}}$ whenever $\beta_n \ll \left(\frac{d_{\max}}{\gamma d}\right)^{\gamma-1} \left(\frac{d_{\max}}{n}\right)^{1/\gamma}$. Combining these cases, we see that if β_n is in the range $\left(\frac{d_{\max}}{n}\right)^{1/\gamma} \ll \beta_n \ll \left(\frac{d_{\max}}{\gamma d}\right)^{\gamma-1} \left(\frac{d_{\max}}{n}\right)^{1/\gamma}$, the dominating term in (6) is greater than the dominating term in (15).

Since we have assumed β_n to be decaying, we briefly discuss conditions under which the requirement $\left(\frac{d_{\max}}{n}\right)^{1/\gamma} \ll \beta_n$ is consistent with this assumption. While this lower bound on β_n may not always vanish as $n \rightarrow \infty$, 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

$$\lim_{n \rightarrow \infty} \log \left(\frac{d_{\max}}{n} \right)^{\frac{1}{\gamma}} = \lim_{n \rightarrow \infty} (\alpha - 1)(\log n)^{1-c} = -\infty, \quad (16)$$

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

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 .

C. Algorithm for ρ -Sized Tests

While our main focus is on the γ -divisible constraint (motivated by it having larger gaps in the bounds [6]), here we briefly pause to provide a simple adaptive algorithm for the ρ -sized test constraint, shown in Algorithm 3. This is a direct modification of Hwang's generalized binary splitting algorithm [20], in which 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 $d_1, \dots, d_{\frac{n}{\rho}}$ be the number of defective items in each of the initial $\frac{n}{\rho}$ groups. Note that the assumption $\rho \in o(\frac{n}{d})$ (see Section I-B) implies $d \in o(\frac{n}{\rho})$, most groups will not have a defective item. In the binary splitting

Algorithm 4 Binary splitting sub-routine

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 $G_i \leftarrow G'_i$. Otherwise, set $G_i \leftarrow G_i \setminus G'_i$. Return to step 1.
 - 4: **return** \mathcal{D}
-

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 leads to the following upper bound:

$$T \leq \frac{n}{\rho} + d \log_2 \rho + O(d) \quad (17)$$

$$\stackrel{(a)}{=} \frac{n}{\rho}(1 + o(1)) + d \log_2 \rho, \quad (18)$$

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

$$T \leq \frac{n}{\rho}(1 + o(1)). \quad (19)$$

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.

D. Proof of Theorem 5 (Counting-Based 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 [18] as follows: The testing procedure defines a mapping $\theta : \Sigma_{n,d} \rightarrow \{0,1\}^T$. 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 $\mathbf{y} \in \{0,1\}^T$, we write $\mathcal{A}_{\mathbf{y}} \subseteq \Sigma_{n,d}$ for the inverse image of \mathbf{y} under θ :

$$\mathcal{A}_{\mathbf{y}} = \theta^{-1}(\mathbf{y}) = \{S \in \Sigma_{n,d} : \theta(S) = \mathbf{y}\}. \quad (20)$$

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:

$$\mathbb{P}(\text{suc}) \stackrel{(a)}{=} \sum_{S \in \Sigma_{n,d}} \mathbb{P}(\text{suc} | \mathcal{D} = S) \frac{1}{\binom{n}{d}} \quad (21)$$

$$= \frac{1}{\binom{n}{d}} \sum_{S \in \Sigma_{n,d}} \sum_{\mathbf{y} \in \{0,1\}^T} \mathbb{1}(\theta(S) = \mathbf{y}) \mathbb{P}(\text{suc} | \mathcal{D} = S) \quad (22)$$

$$= \frac{1}{\binom{n}{d}} \sum_{S \in \Sigma_{n,d}} \sum_{\mathbf{y} \in \{0,1\}^T : |\mathcal{A}_{\mathbf{y}}| \geq 1} \mathbb{1}(\theta(S) = \mathbf{y}) \frac{1}{|\mathcal{A}_{\mathbf{y}}|} \quad (23)$$

$$= \frac{1}{\binom{n}{d}} \sum_{\mathbf{y} \in \{0,1\}^T : |\mathcal{A}_{\mathbf{y}}| \geq 1} \frac{1}{|\mathcal{A}_{\mathbf{y}}|} \left(\sum_{S \in \Sigma_{n,d}} \mathbb{1}(\theta(S) = \mathbf{y}) \right) \quad (24)$$

$$= \frac{1}{\binom{n}{d}} \sum_{\mathbf{y} \in \{0,1\}^T : |\mathcal{A}_{\mathbf{y}}| \geq 1} \frac{1}{|\mathcal{A}_{\mathbf{y}}|} |\mathcal{A}_{\mathbf{y}}| \quad (25)$$

$$= \frac{|\{\mathbf{y} \in \{0,1\}^T : |\mathcal{A}_{\mathbf{y}}| \geq 1\}|}{\binom{n}{d}} \quad (26)$$

$$\stackrel{(b)}{\leq} \frac{|\{\mathbf{y} \text{ with } \leq \gamma d \text{ ones}\}|}{\binom{n}{d}} = \frac{\sum_{i=0}^{\gamma d} \binom{T}{i}}{\binom{n}{d}}, \quad (27)$$

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.

E. Proof of Theorem 6 (General Converse for γ -Divisible Items)

From the counting bound in (2), we upper bound the sum of binomial coefficients [25, Section 4.7.] to obtain

$$\mathbb{P}(\text{suc}) \leq \frac{e^{TH_2(\frac{\gamma d}{T})}}{\binom{n}{d}} \equiv \delta, \quad (28)$$

where $H_2(\cdot)$ is the binary entropy function in nats. From (28), we have $e^{TH_2(\frac{\gamma d}{T})} / \binom{n}{d} = \delta$, which implies that

$$\log \left(\delta \binom{n}{d} \right) = TH_2 \left(\frac{\gamma d}{T} \right) \quad (29)$$

$$= \gamma d \log \frac{T}{\gamma d} + (T - \gamma d) \log \frac{1}{1 - \frac{\gamma d}{T}} \quad (30)$$

$$\stackrel{(a)}{=} \gamma d \log \frac{T}{\gamma d} + \gamma d(1 + o(1)), \quad (31)$$

where (a) uses a Taylor expansion and the fact that $\frac{\gamma d}{T} \in o(1)$ from (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 (31), we obtain

$$\gamma d \log \frac{T}{\gamma d} = \log \left(\delta \binom{n}{d} \right) - \gamma d(1 + o(1)) \quad (32)$$

$$\implies \log \frac{T}{\gamma d} = \frac{1}{\gamma d} \log \left(\delta \binom{n}{d} \right) - (1 + o(1)), \quad (33)$$

which gives

$$T = e^{-(1+o(1))} \gamma d \left(\delta \binom{n}{d} \right)^{\frac{1}{\gamma d}} \quad (34)$$

$$\stackrel{(a)}{\geq} e^{-(1+o(1))} \gamma d \delta^{\frac{1}{\gamma d}} \left(\frac{n}{d} \right)^{\frac{1}{\gamma}}, \quad (35)$$

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 $\delta = 1 - \epsilon$, $\delta^{1/(\gamma d)} \rightarrow 1$ as $\gamma d \rightarrow \infty$.

F. Proof of Theorem 7 (Adaptive Algorithm Performance)

Similar to Hwang's generalized binary splitting algorithm [20], 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 1). We let M be the group size at the initial splitting stage of the algorithm. The algorithm first tests n/M groups of size M each,⁴ then steadily decrease the sizes of each group down the stages: $M \rightarrow M^{1-1/(\gamma-1)} \rightarrow M^{1-2/(\gamma-1)} \rightarrow \dots \rightarrow 1$ (see Figure 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 $dM^{\frac{1}{\gamma-1}}$, giving the following bound on T :

$$T \leq \frac{n}{M} + (\gamma - 1) d M^{\frac{1}{\gamma-1}}. \quad (36)$$

We optimize with respect to M by differentiating the upper bound and setting it to zero, which gives $M = \left(\frac{n}{d} \right)^{\frac{\gamma-1}{\gamma}}$. Substituting $M = \left(\frac{n}{d} \right)^{\frac{\gamma-1}{\gamma}}$ into the general upper bound in (36), we obtain the following upper bound:

$$T \leq \frac{n}{\left(\frac{n}{d} \right)^{\frac{\gamma-1}{\gamma}}} + (\gamma - 1) d \left[\left(\frac{n}{d} \right)^{\frac{\gamma-1}{\gamma}} \right]^{\frac{1}{\gamma-1}} = \gamma d \left(\frac{n}{d} \right)^{\frac{1}{\gamma}}. \quad (37)$$

III. THE NON-ADAPTIVE SETTING

In this section, we develop bounds and algorithms for the non-adaptive setting with γ -divisible items.

⁴Note that $\frac{n}{M}$ is an integer for our chosen M following (36), which gives $\frac{n}{M} = d \left(\frac{n}{d} \right)^{1/\gamma}$, and $\left(\frac{n}{d} \right)^{1/\gamma}$ was assumed to be an integer earlier.

A. Converse Bound for the Near-Constant Tests-Per-Item Design

In this section, we present an information-theoretic lower bound on the number of tests for the near-constant test-per-item random design with parameter γ . Note that this is in contrast to Theorems 2 and 6, which hold for *arbitrary* non-adaptive test designs. Of course, lower bounds for arbitrary designs are generally preferable; however, the converse specific to the random design will be seen to be significantly tighter in denser scaling regimes. See [7], [13], [27] for similar design-specific converse bounds in other contexts.

We follow the high-level approach of [16], showing that if both the Combinatorial Orthogonal Matching Pursuit (COMP) algorithm [14] and the Smallest Satisfying Set (SSS) algorithm [13] fail, then so does any algorithm. We proceed by introducing these algorithms formally.

Definition 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, yielding $\hat{\mathcal{D}}_{\text{COMP}} = \mathcal{PD}$.

We observe that the COMP algorithm succeeds if and only if every non-defective item is included in at least one negative test.

For the SSS algorithm, we first state a key definition, and then describe the algorithm.

Definition 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. 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 $\hat{\mathcal{D}}_{\text{SSS}}$.

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 $|\hat{\mathcal{D}}_{\text{SSS}}| \leq d$. In addition, we can identify a particular failure event for SSS [13]: If a defective item $i \in \mathcal{D}$ is not the unique defective item in any positive test, then $\mathcal{D} \setminus \{i\}$ will be a smaller satisfying set than \mathcal{D} , so SSS is certain to fail.

Following the above outline, our algorithm-independent converse for the near-constant tests-per-item design will be a simple corollary to the following theorem.

Theorem 8. (Design-Specific Converse for COMP and SSS) *Under the near-constant tests-per-item design, with $d \in \Theta(n^\theta)$ for some positive constant $\theta \in (0, 1)$, and $\gamma \in \Theta((\log n)^c)$ tests per item for some $c \in [0, 1)$, if*

$$T = \gamma d^{1+1/\gamma} (1 - \zeta) \tag{38}$$

for fixed $\zeta \in (0, 1)$, then we have

$$\mathbb{P}(\widehat{\mathcal{D}}_{\text{COMP}} \neq \mathcal{D}) = 1 - o(1) \quad (39)$$

$$\mathbb{P}(|\widehat{\mathcal{D}}_{\text{SSS}}| < d) \geq \frac{1 - o(1)}{(1 - \zeta + o(1))^\gamma + 1}. \quad (40)$$

Proof. See Section III-D. \square

In the scaling regime $\gamma \in \Theta((\log n)^c)$ for some $c \in [0, 1)$, the right-hand side of (40) approaches one if $c > 0$ (large γ), is close to one if $c = 0$ (constant γ) as long as γ is large compared to $\frac{1}{\zeta}$, and is always at least $\frac{1}{2} + o(1)$.

Corollary 1. (Design-Specific Converse for Arbitrary Algorithms) *Under the near-constant tests-per-item design, with $d \in \Theta(n^\theta)$ for some positive constant $\theta \in (0, 1)$, and $\gamma \in \Theta((\log n)^c)$ tests per item for some $c \in [0, 1)$, if*

$$T \leq \gamma d^{1+1/\gamma} (1 - \zeta) \quad (41)$$

for some $\zeta > 0$, then the error probability is bounded away from zero regardless of the decoding algorithm.

Proof. It was proved in [16] that if $\mathbb{P}(\widehat{\mathcal{D}}_{\text{COMP}} \neq \mathcal{D}) + \mathbb{P}(|\widehat{\mathcal{D}}_{\text{SSS}}| < d) \geq 1 + \epsilon$ for some $\epsilon > 0$, then the error probability is at least $\frac{\epsilon}{2}$ for an arbitrary algorithm. Hence, the desired result follows immediately from Theorem 8; it suffices to consider (41) holding with equality, because any decoding algorithm can always choose to ignore some of the tests. \square

We observe that the converse in Corollary 1 is tighter (i.e., has a higher lower bound on T) than that of Theorem 2 when d is “large”, i.e., when θ is above $\frac{1}{2}$, or in particular, close to one.

B. Analysis of the DD Algorithm with γ -Divisible Items

We continue focus on the random near-constant tests-per-item design for the γ -divisible items constraint, where $\gamma \in o(\log(\frac{n}{d}))$ tests are chosen uniformly at random with replacement for each item. However, we now turn our attention to upper bounds.

We will use the Definite Defectives (DD) decoding algorithm [13], which is defined as follows.

Definition 4. 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. The rest of the items, $\mathcal{PD} := \{1, \dots, n\} \setminus \mathcal{ND}$, form 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 the same as COMP; it never makes a mistake in adding to \mathcal{ND} (items are correctly marked as non-defective). Similarly, the second step never makes a 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 $\widehat{\mathcal{D}} \subseteq \mathcal{D}$. 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.

By analyzing the DD algorithm, we obtain the following theorem.

Theorem 9. *For $\gamma \in \Theta((\log n)^c)$ 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, under the near-constant tests-per-item design with parameter γ and a number of tests given by*

$$T = \gamma d \max \left\{ e^{\frac{1}{\alpha_2} H_2(\max\{\alpha_2, \frac{1}{2}\})} \left(\frac{d}{\beta_n} \right)^{\frac{1}{\alpha_2 \gamma}}, 2^{1/\gamma} \left(\frac{n-d}{d} \right)^{\frac{1}{\gamma}} \left(\frac{d}{\beta_n} \right)^{\frac{1}{(1-\alpha_2)\gamma^2}} \right\}, \quad (42)$$

the DD algorithm ensures an error probability of at most

$$P_e \leq \exp \left(-\frac{3d}{16} \left(\frac{\beta_n}{d} \right)^{\frac{1}{(1-\alpha_2)\gamma}} \right) + 2 \exp(-2(\gamma d)^{1/3}) + 2\beta_n(1 + o(1)). \quad (43)$$

Proof. See Section III-E. □

In order to better understand this bound on T , we simplify it in two different scaling regimes:

- 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 $\theta \in (0, 1)$.

In both regimes, we assume that β_n is a slowly decaying term, since we are primarily interested in attaining $P_e \rightarrow 0$ rather than the speed of convergence.

It will be useful to compare the bounds in terms of the following quantity:

$$\eta = \lim_{n \rightarrow \infty} \frac{\log(\frac{n}{d})}{\gamma \log(\frac{T}{\gamma d})}. \quad (44)$$

Observe that for any fixed value of $\eta > 0$, re-arranging gives $T = \gamma d \left(\left(\frac{n}{d} \right)^{\frac{1}{\gamma}} \right)^{\frac{1+o(1)}{\eta}}$. We henceforth use the notation $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to denote the asymptotic behavior of T up to factors that do not impact η , and accordingly omit β_n from such expressions.

For regime 1 (large γ), letting α_2 be a fixed constant close to one, we obtain that $e^{\frac{1}{\alpha_2} H_2(\max\{\alpha_2, \frac{1}{2}\})}$ can be made arbitrarily close to one, and in addition, the assumed scaling on γ and d gives

$$\left(\frac{n-d}{d} \right)^{\frac{1}{\gamma}} d^{\frac{1}{(1-\alpha_2)\gamma^2}} = \left(\frac{n-d}{d^{1-O(1/\gamma)}} \right)^{\frac{1}{\gamma}} \quad (45)$$

$$= \left(\frac{n^{\frac{1}{1-O(1/\gamma)}} (1 - O(d/n))^{\frac{1}{1-O(1/\gamma)}}}{d} \right)^{\frac{1-O(1/\gamma)}{\gamma}} \quad (46)$$

$$= \left(\frac{n}{d} \right)^{\frac{1-o(1)}{\gamma}}. \quad (47)$$

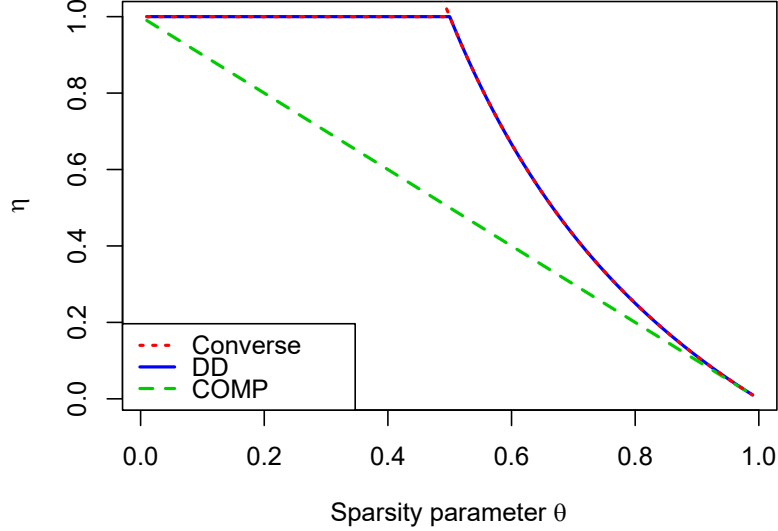


Fig. 2: Plots of the variable η in (44) against the sparsity parameter θ for the converse, DD algorithm, and COMP algorithm, when $n \rightarrow \infty$ and $\gamma = (\log n)^c$ for some $c \in (0, 1)$.

By substituting into (42) and omitting β_n as explained above, we obtain

$$T = \tilde{O}\left(\gamma d \max\{n^\theta, n^{1-\theta}\}^{\frac{1}{\gamma}}\right), \quad (48)$$

which matches the $\Omega(\gamma d \max\{n^\theta, n^{1-\theta}\}^{\frac{1}{\gamma}})$ lower bound obtained by combining Theorem 6 and Corollary 1.

We plot η against $\theta \in (0, 1)$ in Figure 2 to show how the asymptotic bound of the DD algorithm compares to the converse (Theorems 2 and 6) and the COMP bound (Theorem 1). Note that for the COMP algorithm, we have omitted ϵ in our asymptotic bound (similarly to β_n above), giving $T = \tilde{O}(\gamma d n^{1/\gamma})$. From Figure 2, we see that the DD algorithm performs better than the COMP algorithm, and achieves the optimal value of η for all $\theta \in (0, 1)$.

For regime 2 (constant γ), we similarly substitute the scaling laws into (42) (and omit β_n) to get

$$T = \tilde{O}\left(\gamma d \max\left\{n^{\frac{\theta}{\alpha_2 \gamma}}, n^{\frac{1-\theta}{\gamma} + \frac{\theta}{(1-\alpha_2)\gamma^2}}\right\}\right). \quad (49)$$

We numerically optimize with respect to α_2 to obtain our bound on T . Figure 3 shows how the asymptotic bound of the DD algorithm compares to the converse (Theorems 2 and 6) and the COMP bound (Theorem 1), when $\gamma = 10$. We see that DD again significantly outperforms COMP, but falls short of the converse.

This last example provides a useful point of comparison with the concurrent work of [21]. It was shown therein that the converse curve in Figure 3 can in fact be matched exactly. Thus, the proof techniques of [21] appear to

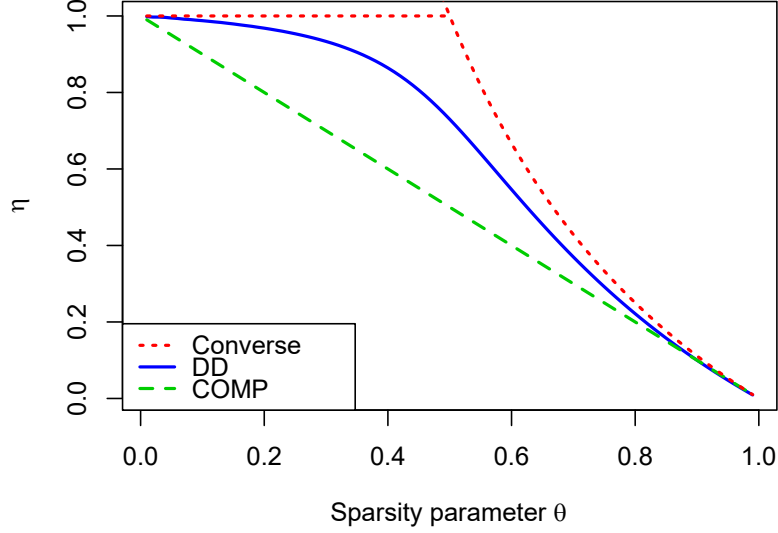


Fig. 3: Plots of the variable η in (44) against the sparsity parameter θ for the converse, DD algorithm, and COMP algorithm, when $n \rightarrow \infty$ and $\gamma = 10$.

be more powerful than ours in the regime $\gamma = \Theta(1)$. On the other hand, regimes where $\gamma \rightarrow \infty$, such as that considered in Figure 2, are not considered in [21].

C. Preliminary Definitions and Results

Before presenting the main proofs, we introduce some useful definitions and auxiliary results.

Definition 5. 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 6. The number of *collisions* between a given item i and a given set of items \mathcal{L} refers to the number of tests selected by the near-constant tests-per-item design for item i (including repetitions in the sampling with replacement) that also include at least one member of \mathcal{L} .

Next, we introduce some auxiliary lemmas that will be used throughout our main proofs.

Lemma 1. If $\gamma \in \Theta((\log n)^c)$ for some $c \in [0, 1)$, and $d \in \Theta(n^\theta)$ for some $\theta \in (0, 1]$, then we have $(1 \pm \frac{1}{d^{\kappa/\gamma}})^\gamma = 1 \pm o(1)$ for any fixed $\kappa > 0$.

Proof. Throughout the proof, we write $g_n \gg f_n$ as a shorthand for $f_n = o(g_n)$. Since $(1 \pm \frac{1}{d^{\kappa/\gamma}})^\gamma \rightarrow 1$ if $d^{\kappa/\gamma} \gg \gamma$, it suffices to show that $d^{\kappa/\gamma} \gg \gamma$. We have

$$\kappa\theta(\log n)^{1-c} \gg c \log \log n \quad (50)$$

$$\implies \frac{\kappa\theta}{(\log n)^c} (\log n) \gg \log(\log n)^c \quad (51)$$

$$\implies \log(n^{\kappa\theta/(\log n)^c}) \gg \log(\log n)^c \quad (52)$$

$$\implies n^{\kappa\theta/(\log n)^c} \gg (\log n)^c. \quad (53)$$

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

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, which follows via the same arguments as the unconstrained setting [7].

Lemma 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*

$$\mathbb{P}(|W^{(\mathcal{D})} - \gamma d(1 - \delta_n)| \geq (\gamma d)^{2/3}) \leq 2 \exp(-2(\gamma d)^{1/3}), \quad (54)$$

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

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

$$\mathbb{E}[W^{(\mathcal{D})}] = \left(1 - \left(1 - \frac{1}{T}\right)^{\gamma d}\right) T \quad (55)$$

$$\stackrel{(a)}{=} \left(1 - \left(1 - \frac{\gamma d}{T} + O\left(\left(\frac{\gamma d}{T}\right)^2\right)\right)\right) T \quad (56)$$

$$= \left(\frac{\gamma d}{T} - O\left(\left(\frac{\gamma d}{T}\right)^2\right)\right) T \quad (57)$$

$$\stackrel{(b)}{=} \gamma d(1 - \delta_n), \quad (58)$$

where (a) applies a second order Taylor expansion, and in (b) we introduce $\delta_n \in O(\frac{\gamma d}{T})$. Let $Z_1, Z_2, \dots, Z_{\gamma d}$ be the labels of the selected coupons and $W(\gamma d) = f(Z_1, \dots, Z_{\gamma d})$ be the number of distinct coupons. We have the bounded property difference property

$$|f(Z_1, \dots, Z_j, \dots, Z_{\gamma d}) - f(Z_1, \dots, \hat{Z}_j, \dots, Z_{\gamma d})| \leq 1 \quad (59)$$

for any $j, Z_1, Z_2, \dots, Z_{\gamma d}$, and Y'_j , since the largest difference we can make is swapping a distinct coupon Z_j for a non-distinct coupon Z'_j , or vice versa. McDiarmid's inequality [26] gives

$$\mathbb{P}(|f(Z_1, Z_2, \dots, Z_{\gamma d}) - \mathbb{E}[f(Z_1, Z_2, \dots, Z_{\gamma d})]| \geq \delta) \leq 2 \exp\left(-\frac{2\delta^2}{\gamma d}\right). \quad (60)$$

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

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 immediately obtain the following two corollaries.

Corollary 2. *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*

$$\mathbb{P}(|W^{(\mathcal{D} \setminus i)} - \gamma(d-1)(1 - \delta_n^{(1)})| \geq (\gamma(d-1))^{2/3}) \leq 2 \exp(-2(\gamma(d-1))^{1/3}), \quad (61)$$

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

Corollary 3. *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*

$$\mathbb{P}(|W^{(\mathcal{D} \setminus i, j)} - \gamma(d-2)(1 - \delta_n^{(2)})| \geq (\gamma(d-2))^{2/3}) \leq 2 \exp(-2(\gamma(d-2))^{1/3}), \quad (62)$$

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

D. Proof of Theorem 8 (Converse for γ -Divisible Items)

Throughout the proof, we condition on a fixed but otherwise arbitrary defective set \mathcal{D} . We consider the event that some defective item $i \in \mathcal{D}$ is masked by the other defective items $\mathcal{D} \setminus \{i\}$, which leads to the event $|\widehat{\mathcal{D}}_{\text{SSS}}| < d$ [13]. Hence, writing A_i for the event that item $i \in \mathcal{D}$ is masked by $\mathcal{D} \setminus \{i\}$, de Caen's lower bound on a union [28] gives

$$\mathbb{P}(|\widehat{\mathcal{D}}_{\text{SSS}}| < d) \geq \mathbb{P}\left(\bigcup_{i \in \mathcal{D}} A_i\right) \geq \sum_{i \in \mathcal{D}} \frac{\mathbb{P}(A_i)^2}{\mathbb{P}(A_i) + \sum_{j \in \mathcal{D} \setminus \{i\}} \mathbb{P}(A_i \cap A_j)}. \quad (63)$$

We proceed by bounding the numerator and denominator separately.

a) *Bounding the Numerator of (63)*: 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

$$\mathbb{P}(A_i) = \sum_w \mathbb{P}(A_i | W^{(\mathcal{D} \setminus i)} = w) \mathbb{P}(W^{(\mathcal{D} \setminus i)} = w) \quad (64)$$

$$= \sum_w \left(\frac{w}{T}\right)^\gamma \mathbb{P}(W^{(\mathcal{D} \setminus i)} = w) \quad (65)$$

$$\geq \sum_{w \geq c_1 \gamma (d-1)} \left(\frac{w}{T}\right)^\gamma \mathbb{P}(W^{(\mathcal{D} \setminus i)} = w) \quad (66)$$

$$\geq \sum_{w \geq c_1 \gamma (d-1)} \left(\frac{c_1 \gamma (d-1)}{T}\right)^\gamma \mathbb{P}(W^{(\mathcal{D} \setminus i)} = w) \quad (67)$$

$$= \left(\frac{c_1 \gamma (d-1)}{T}\right)^\gamma \mathbb{P}(W^{(\mathcal{D} \setminus i)} \geq c_1 \gamma (d-1)). \quad (68)$$

b) *Bounding the Denominator of (63)*: We first derive a bound on $\mathbb{P}(A_i \cap A_j | W^{(\mathcal{D} \setminus i, j)} = w)$ that holds for any given $w = \Theta(\gamma d)$ (an event that holds with high probability by Corollary 3), by suitably adapting the arguments of the unconstrained setting [7].

For this part (and only 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 \mathcal{A}_i be the set of list pairs $(\mathcal{T}_i, \mathcal{T}_j)$ under which the event A_i occurs, and similarly for \mathcal{A}_j , we have

$$\mathbb{P}(A_i \cap A_j | W^{(\mathcal{D} \setminus i, j)} = w) = \frac{N_{ij}}{T^{2\gamma}}, \quad (69)$$

where

$$N_{ij} = \sum_{\mathcal{T}_i} \sum_{\mathcal{T}_j} \mathbb{1}\{(\mathcal{T}_i, \mathcal{T}_j) \in \mathcal{A}_i \cap \mathcal{A}_j\} \quad (70)$$

is the number of pairs of lists in $\mathcal{A}_i \cap \mathcal{A}_j$. Here the sets \mathcal{A}_i and \mathcal{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

$$N_{ij} = \sum_{l=0}^{\gamma} N_{ij}(l), \quad (71)$$

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

$$N_{ij}(l) \leq \binom{T-w}{l} \cdot \gamma^{2l} \cdot (w+l)^{2(\gamma-l)} \quad (72)$$

$$\leq (T-w)^l \cdot \gamma^{2l} \cdot (w+\gamma)^{2(\gamma-l)} \quad (73)$$

$$= (w+\gamma)^{2\gamma} \cdot \left(\frac{\gamma^2(T-w)}{(w+\gamma)^2} \right)^l. \quad (74)$$

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 $w \in \Theta(\gamma d) = \Theta(\gamma n^\theta)$. By substituting the scaling regime for T and w into the bracketed term above and taking the log, we get

$$\log \frac{\gamma^2(T-w)}{(w+\gamma)^2} = \log \frac{\gamma^2(\Theta(1)\gamma n^{\theta+\theta/\gamma} - \Theta(1)\gamma n^\theta)}{(\Theta(1)\gamma n^\theta + \gamma)^2} \quad (75)$$

$$\leq \log \frac{\Theta(1)\gamma^3 n^{\theta+\theta/\gamma}}{\gamma^2 n^{2\theta}} \quad (76)$$

$$= \log \Theta(1) + \log(\gamma n^{\theta/\gamma-\theta}) \quad (77)$$

$$= \log \Theta(1) + \log \gamma + \left(\frac{\theta}{\gamma} - \theta \right) \log n. \quad (78)$$

This expression tends to $-\infty$ for any $\gamma > 1$, since $\gamma \in o(\log n)$. This implies that the bracketed term in (74) satisfies $\frac{\gamma^2(T-w)}{(w+\gamma)^2} \rightarrow 0$ as $n \rightarrow \infty$, and is therefore less than any given $\epsilon_1 > 0$ for sufficiently large n .

Summing over l , we obtain

$$N_{ij} \leq \sum_{l=0}^{\gamma} (w+\gamma)^{2\gamma} \cdot \left(\frac{\gamma^2(T-w)}{(w+\gamma)^2} \right)^l \quad (79)$$

$$\leq (w+\gamma)^{2\gamma} \cdot \sum_{l=0}^{\infty} \epsilon_1^l \quad (80)$$

$$= (w+\gamma)^{2\gamma} \cdot \frac{1}{1-\epsilon_1}, \quad (81)$$

and substituting into (69), we obtain

$$\mathbb{P}(A_i \cap A_j | W^{(\mathcal{D} \setminus \{i,j\})} = w) \leq \left(\frac{w+\gamma}{T} \right)^{2\gamma} \cdot \frac{1}{1-\epsilon_1}. \quad (82)$$

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

$$\begin{aligned} \sum_{j \in \mathcal{D} \setminus \{i\}} \mathbb{P}(A_i \cap A_j) &= (d-1) \sum_w \mathbb{P}(A_i \cap A_j | W^{(\mathcal{D} \setminus \{i,j\})} = w) \mathbb{P}(W^{(\mathcal{D} \setminus \{i,j\})} = w) \\ &\leq \frac{d-1}{1-\epsilon_1} \sum_{c_1 \gamma(d-2) \leq w \leq c_2 \gamma(d-2)} \left(\frac{w+\gamma}{T} \right)^{2\gamma} \mathbb{P}(W^{(\mathcal{D} \setminus \{i,j\})} = w) \end{aligned} \quad (83)$$

$$+ (d-1)\mathbb{P}(W^{(\mathcal{D} \setminus i, j)} \notin [c_1\gamma(d-2), c_2\gamma(d-2)]) \quad (84)$$

$$\leq \frac{d-1}{1-\epsilon_1} \left(\frac{c_2\gamma(d-2)+\gamma}{T} \right)^{2\gamma} \mathbb{P}(c_1\gamma(d-2) \leq W^{(\mathcal{D} \setminus i, j)} \leq c_2\gamma(d-2)) \\ + (d-1)\mathbb{P}(W^{(\mathcal{D} \setminus i, j)} < c_1\gamma(d-2)) + (d-1)\mathbb{P}(W^{(\mathcal{D} \setminus i, j)} > c_2\gamma(d-2)). \quad (85)$$

c) *Combining the Two Terms:* In accordance with Corollaries 2 and 3, we choose c_1 and c_2 in (68) and (85) as follows:

$$c_1 = \min \left\{ 1 - \delta_n^{(1)} - \frac{1}{(\gamma(d-1))^{1/3}}, 1 - \delta_n^{(2)} - \frac{1}{(\gamma(d-2))^{1/3}} \right\} \quad (86)$$

$$\geq 1 - \delta_n^{(3)} \quad (87)$$

$$c_2 = 1 - \delta_n^{(2)} + \frac{1}{(\gamma(d-2))^{1/3}}, \quad (88)$$

where $\delta_n^{(3)} \in O(\frac{\gamma d}{T}) + O(\frac{1}{(\gamma d)^{1/3}}) \rightarrow 0$ (since $\delta_n^{(1)}$ and $\delta_n^{(2)}$ are both $O(\frac{\gamma d}{T})$). We also introduce

$$c_3 = 1 + \left(\frac{d-2}{d-1} \right) (-\delta_n^{(2)} + (\gamma(d-2))^{-1/3}), \quad (89)$$

and note the useful fact

$$c_3\gamma(d-1) = c_2\gamma(d-2) + \gamma, \quad (90)$$

which will be used later.

The concentration results from Corollaries 2 and 3 imply that

$$\mathbb{P}(A_i) \stackrel{(a)}{\geq} \left(\frac{c_1\gamma(d-1)}{T} \right)^\gamma (1 - o(1)) \quad (91)$$

$$\sum_{j \in \mathcal{D} \setminus \{i\}} \mathbb{P}(A_i \cap A_j) \stackrel{(b)}{\leq} \frac{d-1}{1-\epsilon_1} \left(\frac{c_2\gamma(d-2)+\gamma}{T} \right)^{2\gamma} (1 - o(1)) + o(1), \quad (92)$$

where:

- (a) follows by substituting our chosen c_1 into the $\mathbb{P}(\cdot)$ part of (68) and applying the concentration result in Corollary 2;
- (b) follows by substituting our chosen c_1 and c_2 into the $\mathbb{P}(\cdot)$ parts of (85), and then apply the concentration results in Corollary 3.

In these steps, we also used the fact that $2(d-1)\exp(-(\gamma(d-2))^{1/3}) \rightarrow 0$ since $\gamma \geq 1$ and $d \rightarrow \infty$.

In addition to (91), we have the simple upper bound

$$\mathbb{P}(A_i) \leq \left(\frac{\gamma(d-1)}{T} \right)^\gamma, \quad (93)$$

which holds since the number of positive tests that contain at least one item in $\mathcal{D} \setminus \{i\}$ is trivially at most $\gamma(d-1)$.

Substituting (91)–(93) into (63), we obtain

$$\mathbb{P}(|\hat{\mathcal{D}}_{\text{SSS}}| < d) \geq \sum_{i \in \mathcal{D}} \frac{\left(\frac{c_1\gamma(d-1)}{T} \right)^{2\gamma} (1 - o(1))}{\left(\frac{\gamma(d-1)}{T} \right)^\gamma + \frac{d-1}{1-\epsilon_1} \left(\frac{c_2\gamma(d-2)+\gamma}{T} \right)^{2\gamma} (1 - o(1)) + o(1)} \quad (94)$$

$$\stackrel{(a)}{=} \frac{d \left(\frac{\gamma(d-1)}{T} \right)^{2\gamma} c_1^{2\gamma} (1 - o(1))}{\left(\frac{\gamma(d-1)}{T} \right)^\gamma + \frac{d-1}{1-\epsilon_1} \left(\frac{\gamma(d-1)}{T} \right)^{2\gamma} c_3^{2\gamma} (1 - o(1)) + o(1)} \quad (95)$$

$$= \frac{d \left(\frac{\gamma(d-1)}{T} \right)^\gamma c_1^{2\gamma} (1 - o(1))}{1 + \frac{d-1}{1-\epsilon_1} \left(\frac{\gamma(d-1)}{T} \right)^\gamma c_3^{2\gamma} (1 - o(1)) + o(1)} \quad (96)$$

$$= \frac{d \left(\frac{c_1^2 \gamma(d-1)}{T} \right)^\gamma (1 - o(1))}{1 + \frac{d-1}{1-\epsilon_1} \left(\frac{c_3^2 \gamma(d-1)}{T} \right)^\gamma (1 - o(1)) + o(1)}, \quad (97)$$

where (a) is by applying (90) in the denominator.

In the following, it will be convenient to work with the following choice of T :

$$T = \gamma d^{1/\gamma} (d-1) (c_3^2) (1 - \epsilon_2). \quad (98)$$

Since we consider $d \rightarrow \infty$ and $c_2 = 1 - o(1)$, this choice is consistent with (38) for some $\epsilon_2 = \zeta + o(1)$. Substituting (98) into (97), we get

$$\mathbb{P}(|\widehat{\mathcal{D}}_{\text{SSS}}| < d) \geq \frac{d \left(\frac{c_1^2 \gamma(d-1)}{\gamma d^{1/\gamma} (d-1) (c_3^2) (1-\epsilon_2)} \right)^\gamma (1 - o(1))}{1 + \frac{d-1}{1-\epsilon_1} \left(\frac{c_3^2 \gamma(d-1)}{\gamma d^{1/\gamma} (d-1) (c_3^2) (1-\epsilon_2)} \right)^\gamma (1 - o(1)) + o(1)} \quad (99)$$

$$= \frac{\left(\frac{c_1}{c_3} \right)^{2\gamma} \left(\frac{1}{1-\epsilon_2} \right)^\gamma (1 - o(1))}{1 + \frac{d-1}{d(1-\epsilon_1)} \left(\frac{1}{1-\epsilon_2} \right)^\gamma (1 - o(1)) + o(1)} \quad (100)$$

$$\stackrel{(a)}{=} \frac{(1 - \delta_n^{(2)} - (\gamma(d-2))^{-1/3})^{2\gamma}}{\left(1 + \left(\frac{d-2}{d-1} \right) (-\delta_n^{(2)} + (\gamma(d-2))^{-1/3}) \right)^{2\gamma}} \cdot \frac{1 - o(1)}{(1 - \epsilon_2)^\gamma + \frac{d-1}{d(1-\epsilon_1)}} \quad (101)$$

$$\stackrel{(b)}{=} \frac{1 - o(1)}{(1 - \epsilon_2)^\gamma + \frac{d-1}{d(1-\epsilon_1)}}, \quad (102)$$

where (a) follows by substituting c_1 and c_3 , and in (b), we note that both the numerator and denominator are in $(1 - O(\frac{1}{d^{1/\gamma}}) \pm O(\frac{1}{(\gamma d)^{1/3}}))^{2\gamma} = (1 - O(\frac{1}{d^{1/\gamma}}))^{2\gamma}$, and then apply Lemma 1. The proof is concluded by recalling that ϵ_1 may be arbitrarily small, and $\epsilon_2 = \zeta + o(1)$.

d) Bounding the COMP Error Probability: We use the fact that COMP fails if and only if at least one non-defective item is masked by \mathcal{D} , and denote the associated error probability by $\mathbb{P}^{\text{COMP}}(\text{err})$.

Recall from Lemma 2 that the number $W^{(\mathcal{D})}$ of positive tests lies in $[\gamma d(1 - \delta_n) - (\gamma d)^{2/3}, \gamma d(1 - \delta_n) + (\gamma d)^{2/3}]$ with high probability, where $\delta_n \in O(\frac{\gamma d}{T}) \rightarrow 0$. For any w in this range and any non-defective item i , we have

$$\mathbb{P}(i \text{ masked by } \mathcal{D} | W^{(\mathcal{D})} = w) \geq \left(\frac{\gamma d(1 - \delta_n^{(4)})}{T} \right)^\gamma, \quad (103)$$

where $\delta_n^{(4)} \in O(\frac{\gamma d}{T}) = O(\frac{1}{d^{1/\gamma}}) \rightarrow 0$, since i is masked if and only if all of its γ tests are those among the w positive ones.

Using (103), we derive an upper bound on $\mathbb{P}^{\text{COMP}}(\text{suc}) = 1 - \mathbb{P}^{\text{COMP}}(\text{err})$:

$$\mathbb{P}^{\text{COMP}}(\text{suc}) = \sum_w \mathbb{P}(W^{(\mathcal{D})} = w) \mathbb{P}^{\text{COMP}}(\text{suc} | W^{(\mathcal{D})} = w) \quad (104)$$

$$= \sum_w \mathbb{P}(W^{(\mathcal{D})} = w) \mathbb{P}(\text{all } i \text{ not masked by } \mathcal{D} | W^{(\mathcal{D})} = w) \quad (105)$$

$$= \sum_w \mathbb{P}(W^{(\mathcal{D})} = w) \left(1 - \mathbb{P}(i \text{ masked by } \mathcal{D} | W^{(\mathcal{D})} = w)\right)^{n-d} \quad (106)$$

$$\stackrel{(a)}{\leq} \left(1 - \left(\frac{\gamma d(1 - \delta_n^{(4)})}{T}\right)^\gamma\right)^{n-d} + 2 \exp(-2(\gamma d)^{\frac{1}{3}}) \quad (107)$$

$$\stackrel{(b)}{=} \left(1 - \left(\frac{\gamma d(1 - \delta_n^{(4)})}{T}\right)^\gamma\right)^{n-d} + o(1), \quad (108)$$

where (a) follows from Lemma 2 and (103), and (b) uses $2 \exp(-2(\gamma d)^{\frac{1}{3}}) \rightarrow 0$ since $\gamma \geq 1$ and $d \rightarrow \infty$. It follows from (108) that

$$\mathbb{P}^{\text{COMP}}(\text{err}) = 1 - \mathbb{P}^{\text{COMP}}(\text{suc}) \quad (109)$$

$$\geq 1 - \left(1 - \left(\frac{\gamma d(1 - \delta_n^{(4)})}{T}\right)^\gamma\right)^{n-d} - o(1) \quad (110)$$

$$= 1 - \left(1 - \left(\frac{\gamma d}{T}\right)^\gamma (1 - \delta_n^{(4)})^\gamma\right)^{n-d} - o(1). \quad (111)$$

Recall that we choose $T = \gamma d^{\frac{1}{\gamma}}(d-1)(c_3^2)(1 - \epsilon_2)$ for some constant $\epsilon_2 > 0$ (see (98)); substituting into (111), we obtain

$$\mathbb{P}^{\text{COMP}}(\text{err}) \geq 1 - \left(1 - \left(\frac{\gamma d}{\gamma d^{1/\gamma}(d-1)(c_3^2)(1 - \epsilon_2)}\right)^\gamma (1 - \delta_n^{(4)})^\gamma\right)^{n-d} - o(1) \quad (112)$$

$$= 1 - \left(1 - \frac{1}{d} \left(1 + \frac{1}{d-1}\right)^\gamma \frac{(1 - \delta_n^{(4)})^\gamma}{c_3^{2\gamma}(1 - \epsilon_2)^\gamma}\right)^{n-d} - o(1) \quad (113)$$

$$\stackrel{(a)}{=} 1 - \left(1 - \frac{(1 + o(1))}{d(1 - \epsilon_2)^\gamma}\right)^{n-d} - o(1) \quad (114)$$

$$\stackrel{(b)}{=} 1 - \exp\left(-\frac{(n-d)(1 + o(1))}{d(1 - \epsilon_2)^\gamma}\right) - o(1) \quad (115)$$

$$= 1 - \exp\left(-\frac{n(1 + o(1))}{d(1 - \epsilon_2)^\gamma}\right) - o(1), \quad (116)$$

where:

- (a) follows by applying $(1 + \frac{1}{d-1})^\gamma = 1 + o(1)$ (since $d \gg \gamma$ under our considered scaling laws), as well as noting that both c_3 and $1 - \delta_n^{(4)}$ behave as $1 - O(\frac{1}{d^{1/\gamma}})$, and applying Lemma 1 to get $c_3^{2\gamma} = 1 - o(1)$ and $(1 - \delta_n^{(4)})^\gamma \in 1 - o(1)$.
- (b) follows by first noting that $\frac{1+o(1)}{d(1-\epsilon_2)^\gamma} \in o(1)$ (proved shortly), and then applying $1 - z = e^{-z(1+o(1))}$ when $z \in o(1)$. 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 $\Theta(\theta \log n) + (\log n)^c \log(1 - \epsilon_2) = \Theta(\log n)$.

The right-hand side of (116) approaches one as $n \rightarrow \infty$, since $\exp\left(-\frac{n}{d(1-\epsilon_2)^\gamma}\right) \rightarrow 0$ by the assumption that $d \in \Theta(n^\theta)$ with $\theta \in (0, 1)$.

E. Proof of Theorem 9 (DD Performance)

We again condition on a fixed but otherwise arbitrary defective set \mathcal{D} . We observe that first and second steps recover \mathcal{D} correctly when each defective item i is not masked (see Definition 5) by $\mathcal{PD} \setminus \{i\}$. Hence, we want to derive a bound on T ensuring that 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 (see Definition 6) 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 such that there is a low probability of any defective item i incurring “too many” collisions with $\mathcal{D} \setminus \{i\}$.
- 3) Conditioned on not having too many collisions between defectives in the sense of the previous item, we derive a bound on T such that there is also a low probability of any defective item i having every one of its “collision-free” tests contain at least one item from $\mathcal{PD} \setminus \mathcal{D}$.
- 4) Taking the maximum between the two bounds on T gives us the required number of tests.

We proceed by analysing the two steps of the DD algorithm separately.

a) 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 $G_i \in \{0, 1\}$ being the corresponding indicator variable for a single non-defective. Conditioned on the number of positive tests being $W^{(\mathcal{D})} = w^{(\mathcal{D})}$, we have

$$(G|W^{(\mathcal{D})} = w^{(\mathcal{D})}) \sim \text{Binomial}\left(n - d, \left(\frac{w^{(\mathcal{D})}}{T}\right)^\gamma\right), \quad (117)$$

where $\left(\frac{w^{(\mathcal{D})}}{T}\right)^\gamma \leq \left(\frac{\gamma d}{T}\right)^\gamma$ because the number of positive tests is at most γd . Since $G_i \in \{0, 1\}$, Bernstein’s inequality gives the following for any $t > 0$:

$$\mathbb{P}\left(\sum_{i=1}^{n-d} G_i > \mathbb{E}[G] + t\right) \leq \exp\left(\frac{-\frac{1}{2}t^2}{\sum_{i=1}^{n-d} \text{Var}[G_i] + \frac{1}{3}t}\right) \quad (118)$$

$$\stackrel{(a)}{\leq} \exp\left(\frac{-\frac{1}{2}t^2}{\sum_{i=1}^{n-d} \mathbb{E}[G_i^2] + \frac{1}{3}t}\right) \quad (119)$$

$$\stackrel{(b)}{\leq} \exp\left(\frac{-\frac{1}{2}t^2}{\sum_{i=1}^{n-d} \mathbb{E}[G_i] + \frac{1}{3}t}\right) \quad (120)$$

$$\stackrel{(c)}{\leq} \exp\left(\frac{-\frac{1}{2}t^2}{\mathbb{E}[G] + \frac{1}{3}t}\right), \quad (121)$$

where (a) uses $\text{Var}[G_i] = \mathbb{E}[G_i^2] - (\mathbb{E}[G_i])^2 \leq \mathbb{E}[G_i^2]$, (b) follows since $G_i \in \{0, 1\}$ and hence $\mathbb{E}[G_i^2] = \mathbb{E}[G_i]$, and (c) is due to the linearity of expectation. We will return to (121) and select t later in the analysis.

b) 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 only amounts to re-labeling items. Hence, for clarity, we think

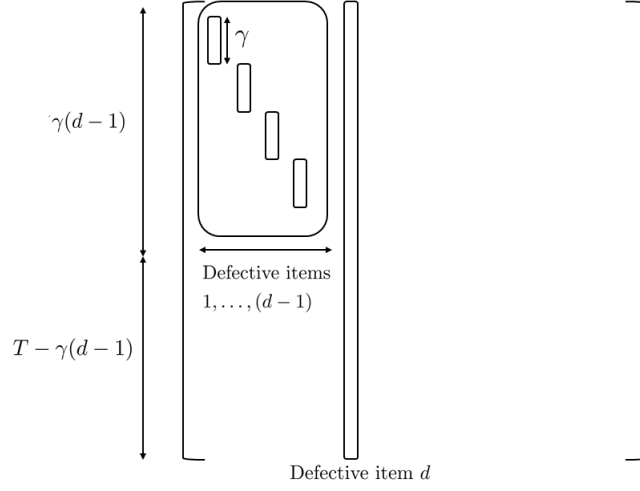


Fig. 4: Rearranged test matrix to illustrate that collision within defectives are rare. This illustration shows the extreme case that the first $d - 1$ defective items are each in a distinct set of γ tests.

of the test matrix as being rearranged such that the first d columns are for the defective items, as shown in Figure 4. Referring to Figure 4, let C_i be the number of collisions between a given defective item i (with $i = d$ in Figure 4) 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 $W^{(\mathcal{D} \setminus i)} = w^{(\mathcal{D} \setminus i)}$, we have

$$(C_i | W^{(\mathcal{D} \setminus i)} = w^{(\mathcal{D} \setminus i)}) \sim \text{Binomial}\left(\gamma, \frac{w^{(\mathcal{D} \setminus i)}}{T}\right), \quad (122)$$

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

$$\mathbb{P}(C_i = \alpha_2 \gamma) = \binom{\gamma}{\alpha_2 \gamma} \left(\frac{w^{(\mathcal{D} \setminus i)}}{T}\right)^{\alpha_2 \gamma} (1-p)^{\gamma - \alpha_2 \gamma} \quad (123)$$

$$\stackrel{(a)}{\leq} e^{\gamma H_2(\alpha_2)} \left(\frac{\gamma d}{T}\right)^{\alpha_2 \gamma}. \quad (124)$$

where (a) is due to $\binom{\gamma}{\alpha_2 \gamma} \leq e^{\gamma H_2(\alpha_2)}$ (where $H_2(\cdot)$ is the binary entropy function in nats), $\frac{w^{(\mathcal{D} \setminus i)}}{T} < \frac{\gamma d}{T}$, 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 $\mathbb{P}(C_i = \alpha_2 \gamma)$:

$$\mathbb{P}(C_i \geq \alpha_2 \gamma) = \sum_{k=\alpha_2 \gamma}^{\gamma} \mathbb{P}(C_i = k) \quad (125)$$

$$\leq \sum_{k=\alpha_2 \gamma}^{\gamma} e^{\gamma H_2(k/\gamma)} \left(\frac{\gamma d}{T}\right)^k \quad (126)$$

$$\stackrel{(a)}{\leq} e^{\gamma H_2(\max\{\alpha_2, \frac{1}{2}\})} \sum_{k=\alpha_2 \gamma}^{\infty} \left(\frac{\gamma d}{T}\right)^k \quad (127)$$

$$\stackrel{(b)}{=} e^{\gamma H_2(\max\{\alpha_2, \frac{1}{2}\})} \frac{(\frac{\gamma d}{T})^{\alpha_2 \gamma}}{1 - (\frac{\gamma d}{T})} \quad (128)$$

$$\stackrel{(c)}{=} e^{\gamma H_2(\max\{\alpha_2, \frac{1}{2}\})} \left(\frac{\gamma d}{T}\right)^{\alpha_2 \gamma} (1 + o(1)), \quad (129)$$

where (a) uses the fact that $H_2(p)$ is increasing for $p \leq \frac{1}{2}$ and decreasing for $p \geq \frac{1}{2}$, (b) uses the sum to infinity of a geometric series, and (c) uses the fact that $\frac{\gamma d}{T} \in o(1)$. By the union bound, we have the following:

$$\mathbb{P}\left(\bigcup_{i=1}^d \{C_i \geq \alpha_2 \gamma\}\right) \leq d e^{\gamma H_2(\max\{\alpha_2, \frac{1}{2}\})} \left(\frac{\gamma d}{T}\right)^{\alpha_2 \gamma} (1 + o(1)). \quad (130)$$

For (130) to approach zero, we consider the following condition for T , where β_n is a slowly decaying term as $n \rightarrow \infty$:

$$e^{\gamma H_2(\max\{\alpha_2, \frac{1}{2}\})} \left(\frac{\gamma d}{T}\right)^{\alpha_2 \gamma} \leq \frac{\beta_n}{d}, \quad (131)$$

which simplifies to

$$T \geq \gamma d e^{\frac{1}{\alpha_2} H_2(\max\{\alpha_2, \frac{1}{2}\})} \left(\frac{d}{\beta_n}\right)^{\frac{1}{\alpha_2 \gamma}}. \quad (132)$$

Now, we study the probability of defective item i not being in $\hat{\mathcal{D}}$. 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 not too high (to be formalized later). After conditioning, we consider the event where every test that includes defective item i , and no other defective item, contains at least one item from $\mathcal{PD} \setminus \mathcal{D}$. This is equivalent to the event that defective item $i \notin \hat{\mathcal{D}}$. We derive a bound on T such that the probability of this event is vanishing.

We condition on the following events:

- 1) $\bigcap_{i=1}^d \{C_i < \alpha_2 \gamma\}$. This occurs with high probability since we already ensured $\mathbb{P}(\bigcup_{i=1}^d \{C_i \geq \alpha_2 \gamma\}) \rightarrow 0$.
- 2) $W^{(\mathcal{D})} = \gamma d(1 - \delta_n^-)$, where $\delta_n^- \in [\delta_n - (\gamma d)^{-1/3}, \delta_n + (\gamma d)^{-1/3}]$ in accordance with Lemma 2, and hence $\delta_n^- \in O(\frac{\gamma d}{T} + (\gamma d)^{-1/3})$. By the choice of T in (42), this scaling on δ_n^- can be simplified to $\delta_n^- \in O(\frac{1}{(n/d)^{1/\gamma} d^{1/\gamma^2}}) = O(\frac{1}{n^{(1-\theta)/\gamma}})$, as the $(\gamma d)^{-1/3}$ term is comparatively negligible.

In addition, we condition on a fixed value of $G \leq d$; this condition will be seen to hold with high probability once we choose t in (121).

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. Since we conditioned on $\bigcap_{i=1}^d \{C_i < \alpha_2 \gamma\}$, we have $\tilde{\gamma} \geq (1 - \alpha_2)\gamma$. We want to find the probability that all $\tilde{\gamma}$ indices correspond to tests where at least one non-defective item in \mathcal{PD} is also present. Without loss of generality, we assume that the $\tilde{\gamma}$ tests of interest are those labeled 1 to $\tilde{\gamma}$. We let A_i be the event that the positive test indexed by i contains at least one non-defective item in \mathcal{PD} .

To study the A_i events, we first note that the non-defective test placements are independent of the defective ones, and recall that we condition on a fixed value of $G = |\mathcal{PD} \setminus \mathcal{D}|$ and a fixed number $\gamma d(1 - \delta_n^-)$ of positive tests. We consider the process of collecting $G\gamma$ ‘‘coupons’’ (placements into tests) corresponding to the non-defective items

in $\mathcal{PD} \setminus \mathcal{D}$. Each coupon collected must correspond to a positive test, since otherwise the item would not be in $\mathcal{PD} \setminus \mathcal{D}$. In addition, since the prior test placement distribution was uniform, the conditional distribution remains uniform, but is now only over the $\gamma d(1 - \delta_n^-)$ positive tests.

Putting the above observations together, we consider a population of $\gamma d(1 - \delta_n^-)$ coupons, the first $\tilde{\gamma}$ of which correspond to i being the unique defective item. We consider collecting $G\gamma$ coupons chosen uniformly with replacement. Then, the event A_i is equivalent to the event that coupon i is collected, yielding

$$\mathbb{P}(A_1, \dots, A_{\tilde{\gamma}}) = \frac{\text{\#ways to collect coupons including all of the first } \tilde{\gamma} \text{ indices}}{\text{\#ways to collect all coupons}}, \quad (133)$$

where the probability is implicitly conditioned on the events described above.

We will bound the numerator and denominator of (133) separately. For the denominator, we can think of listing the selected $G\gamma$ coupons in a vector, where each entry can be any of the $\gamma d(1 - \delta_n^-)$ coupons from the population. This gives the following:

$$(\text{\#ways to collect all coupons}) = (\gamma d)^{G\gamma} (1 - \delta_n^-)^{G\gamma}. \quad (134)$$

For the numerator in (133), we have

$$\begin{aligned} & (\text{\#ways to collect coupons} \\ & \text{including all of the first } \tilde{\gamma} \text{ indices}) \end{aligned} \stackrel{(a)}{\leq} (G\gamma)^{\tilde{\gamma}} (\gamma d)^{G\gamma - \tilde{\gamma}} (1 - \delta_n^-)^{G\gamma - \tilde{\gamma}} \quad (135)$$

$$= \left(\frac{G}{d}\right)^{\tilde{\gamma}} (\gamma d)^{G\gamma} (1 - \delta_n^-)^{G\gamma - \tilde{\gamma}}, \quad (136)$$

where (a) follows since each index in the set $\{1, 2, \dots, \tilde{\gamma}\}$ must occupy at least one position in the sequence of $G\gamma$ coupons; after assigning one such position to each index (in one of at most $(G\gamma)^{\tilde{\gamma}}$ ways), 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

$$\mathbb{P}(A_1, \dots, A_{\tilde{\gamma}}) = \frac{\text{\#ways to collect coupons including all of the first } \tilde{\gamma} \text{ indices}}{\text{\#ways to collect all coupons}} \quad (137)$$

$$= \left(\frac{G}{d}\right)^{\tilde{\gamma}} (1 - \delta_n^-)^{-\tilde{\gamma}} \quad (138)$$

$$\stackrel{(a)}{\leq} \left(\frac{G}{d}\right)^{(1-\alpha_2)\gamma} (1 - \delta_n^-)^{-\tilde{\gamma}} \quad (139)$$

$$\stackrel{(b)}{=} \left(\frac{G}{d}\right)^{(1-\alpha_2)\gamma} (1 + o(1)), \quad (140)$$

where (a) holds since $G \leq d$ and $(1 - \alpha_2)\gamma \leq \tilde{\gamma}$, and (b) follows by recalling that $\delta_n^- \in O\left(\frac{1}{n^{(1-\theta)/\gamma}}\right)$ and applying Lemma 1.

According to the DD algorithm, the event where there exists a defective item not in $\hat{\mathcal{D}}$ 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.

$$\mathbb{P}(\exists \text{ defective item not in } \hat{\mathcal{D}}) \leq \mathbb{P}\left(\bigcup_{i=1}^d \{\text{all } \tilde{\gamma} \text{ indices of } i \text{ are collected}\}\right) \quad (141)$$

$$\leq d\left(\frac{G}{d}\right)^{(1-\alpha_2)\gamma}(1+o(1)). \quad (142)$$

The bound approaches zero if $\left(\frac{G}{d}\right)^{(1-\alpha_2)\gamma} \leq \frac{\beta_n}{d}$ where β_n is a slowly decaying term as $n \rightarrow \infty$. Rearranging, we obtain the following sufficient condition on G to ensure that the right-hand side of (142) vanishes:

$$G \leq d\left(\frac{\beta_n}{d}\right)^{\frac{1}{(1-\alpha_2)\gamma}}. \quad (143)$$

c) Combining the two Steps: We now combine the two steps to obtain our final bound on T , as well as studying the overall error probability. In accordance with (143), we define

$$G_{\max} = d\left(\frac{\beta_n}{d}\right)^{\frac{1}{(1-\alpha_2)\gamma}}. \quad (144)$$

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

$$(n-d)\left(\frac{\gamma d}{T}\right)^\gamma \leq \frac{d}{2}\left(\frac{\beta_n}{d}\right)^{\frac{1}{(1-\alpha_2)\gamma}}, \quad (145)$$

which we will shortly re-arrange to deduce a condition on T . Then, setting $t = G_{\max}/2$ in our inequality in (121) gives

$$\mathbb{P}\left(G > \mathbb{E}[G] + \frac{G_{\max}}{2}\right) \leq \exp\left(\frac{-\frac{1}{2}\left(\frac{G_{\max}}{2}\right)^2}{\mathbb{E}[G] + \frac{1}{3}\left(\frac{G_{\max}}{2}\right)}\right). \quad (146)$$

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

$$\mathbb{P}(G > G_{\max}) \leq \exp\left(-\frac{3}{16}G_{\max}\right) = \exp\left(-\frac{3d}{16}\left(\frac{\beta_n}{d}\right)^{\frac{1}{(1-\alpha_2)\gamma}}\right), \quad (147)$$

which approaches zero as long as β_n does not decay too rapidly. By combining all the error probabilities in (147), (130), (142), and Lemma 2 (with δ_n defined therein), we have

$$\begin{aligned} P_e &\leq \mathbb{P}(G > G_{\max}) + \mathbb{P}\left(\bigcup_{i=1}^d \{C_i \geq \alpha_2 \gamma\}\right) + \mathbb{P}(\exists \text{ defective item not in } \hat{\mathcal{D}}) \\ &\quad + \mathbb{P}\left(\delta_n^- \notin [\delta_n - (\gamma d)^{-1/3}, \delta_n + (\gamma d)^{-1/3}]\right) \end{aligned} \quad (148)$$

$$\begin{aligned} &\leq \exp\left(-\frac{3d}{16}\left(\frac{\beta_n}{d}\right)^{\frac{1}{(1-\alpha_2)\gamma}}\right) + de^{\gamma H_2(\max\{\alpha_2, \frac{1}{2}\})}\left(\frac{\gamma d}{T}\right)^{\alpha_2 \gamma}(1+o(1)) \\ &\quad + d\left(\frac{G}{d}\right)^{(1-\alpha_2)\gamma}(1+o(1)) + 2\exp(-2(\gamma d)^{1/3}) \end{aligned} \quad (149)$$

$$\stackrel{(a)}{\leq} \exp\left(-\frac{3d}{16}\left(\frac{\beta_n}{d}\right)^{\frac{1}{(1-\alpha_2)\gamma}}\right) + 2\beta_n(1+o(1)) + 2\exp(-2(\gamma d)^{1/3}), \quad (150)$$

where (a) applies $de^{\gamma H_2(\max\{\alpha_2, \frac{1}{2}\})}\left(\frac{\gamma d}{T}\right)^{\alpha_2 \gamma}(1+o(1)) \leq \beta_n(1+o(1))$ from (131), as well as $d\left(\frac{G}{d}\right)^{(1-\alpha_2)\gamma}(1+o(1)) \leq \beta_n(1+o(1))$ from (143).

As for the number of tests, rearranging (145) gives

$$T \geq 2^{1/\gamma} \gamma d \left(\frac{n-d}{d}\right)^{\frac{1}{\gamma}} \left(\frac{d}{\beta_n}\right)^{\frac{1}{(1-\alpha_2)\gamma^2}}. \quad (151)$$

Combining (132) and (151) gives

$$T \geq \gamma d \max \left\{ e^{\frac{1}{\alpha_2} H_2(\max\{\alpha_2, \frac{1}{2}\})} \left(\frac{d}{\beta_n} \right)^{\frac{1}{\alpha_2 \gamma}}, 2^{1/\gamma} \left(\frac{n-d}{d} \right)^{\frac{1}{\gamma}} \left(\frac{d}{\beta_n} \right)^{\frac{1}{(1-\alpha_2)\gamma^2}} \right\}, \quad (152)$$

which coincides with (42) in the theorem statement.

IV. CONCLUSION

We have studied the problem of group testing with γ -divisible items (and, more briefly, ρ -sized tests). In the adaptive setting, we characterized the optimal number of tests up to a multiplicative factor of $e^{1+o(1)}$ in broad scaling regimes, via both a strengthened counting-based converse and a novel adaptive splitting algorithm. In the non-adaptive setting, we provided an algorithm-independent converse the near-constant tests-per-item design, and gave a strengthened achievability bound (essentially matching the converse in broad scaling regimes) via the DD algorithm. An open challenge for future work would be to pursue optimal or near-optimal constant factors, as opposed to only optimality with respect to η defined in (44).

REFERENCES

- [1] N. Tan and J. Scarlett, “Near-optimal sparse adaptive group testing,” 2020, <https://arxiv.org/abs/2004.03119>.
- [2] R. Dorfman, “The detection of defective members of large populations,” *Ann. Math. Stats.*, vol. 14, no. 4, pp. 436–440, 1943.
- [3] A. Gilbert, M. Iwen, and M. Strauss, “Group testing and sparse signal recovery,” in *Asilomar Conf. Sig., Sys. and Comp.*, Oct. 2008, pp. 1059–1063.
- [4] A. Fernández Anta, M. A. Mosteiro, and J. Ramón Muñoz, “Unbounded contention resolution in multiple-access channels,” in *Distributed Computing*. Springer Berlin Heidelberg, 2011, vol. 6950, pp. 225–236.
- [5] M. Aldridge, O. Johnson, and J. Scarlett, “Group testing: An information theory perspective,” *Found. Trend. Comms. Inf. Theory*, vol. 15, no. 3–4, pp. 196–392, 2019.
- [6] V. Gandikota, E. Grigorescu, S. Jaggi, and S. Zhou, “Nearly optimal sparse group testing,” *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2760 – 2773, 2019.
- [7] O. Johnson, M. Aldridge, and J. Scarlett, “Performance of group testing algorithms with near-constant tests-per-item,” *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 707–723, Feb. 2019.
- [8] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick, “Optimal non-adaptive group testing,” 2019, <https://arxiv.org/abs/1911.02287>.
- [9] M. Malyutov, “The separating property of random matrices,” *Math. Notes Acad. Sci. USSR*, vol. 23, no. 1, pp. 84–91, 1978.
- [10] M. B. Malyutov and P. S. Mateev, “Screening designs for non-symmetric response function,” *Mat. Zametki*, vol. 29, pp. 109–127, 1980.
- [11] D. Du and F. K. Hwang, *Combinatorial group testing and its applications*. World Scientific, 2000, vol. 12.
- [12] G. Atia and V. Saligrama, “Boolean compressed sensing and noisy group testing,” *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1880–1901, March 2012.
- [13] M. Aldridge, L. Baldassini, and O. Johnson, “Group testing algorithms: Bounds and simulations,” *IEEE Trans. Inf. Theory*, vol. 60, no. 6, pp. 3671–3687, June 2014.
- [14] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri, “Non-adaptive group testing: Explicit bounds and novel algorithms,” *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 3019–3035, May 2014.
- [15] J. Scarlett and V. Cevher, “Phase transitions in group testing,” in *Proc. ACM-SIAM Symp. Disc. Alg. (SODA)*, 2016.
- [16] M. Aldridge, “The capacity of Bernoulli nonadaptive group testing,” *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7142–7148, 2017.

- [17] A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick, “Information-theoretic and algorithmic thresholds for group testing,” in *Int. Colloq. Aut., Lang. and Prog. (ICALP)*, 2019.
- [18] L. Baldassini, O. Johnson, and M. Aldridge, “The capacity of adaptive group testing,” in *IEEE Int. Symp. Inf. Theory*, July 2013, pp. 2676–2680.
- [19] O. Johnson, “Strong converses for group testing from finite blocklength results,” *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5923–5933, Sept. 2017.
- [20] F. Hwang, “A method for detecting all defective members in a population by group testing,” *J. Amer. Stats. Assoc.*, vol. 67, no. 339, pp. 605–608, 1972.
- [21] O. Gebhard, M. Hahn-Klimroth, O. Parczyk, M. Penschuck, and M. Rolvien, “Optimal group testing under real world restrictions,” 2019, <https://arxiv.org/abs/2004.11860>.
- [22] M. Aldridge, L. Baldassini, and K. Gunderson, “Almost separable matrices,” *J. Comb. Opt.*, pp. 1–22, 2015.
- [23] P. Damaschke and A. S. Muhammad, “Competitive group testing and learning hidden vertex covers with minimum adaptivity,” *Disc. Math., Algs. and Apps.*, vol. 2, no. 03, pp. 291–311, 2010.
- [24] 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*, 2016, pp. 1376–1380.
- [25] R. B. Ash, *Information Theory*. Dover Publications Inc., New York, 1990.
- [26] C. McDiarmid, *On the method of bounded differences*, ser. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989, p. 148–188.
- [27] Z. Li, M. Fresacher, and J. Scarlett, “Learning Erdős-Rényi random graphs via edge detecting queries,” in *Conf. Neur. Inf. Proc. Sys (NeurIPS)*, 2019.
- [28] D. de Caen, “A lower bound on the probability of a union,” *Discrete mathematics*, vol. 169, no. 1, pp. 217–220, 1997.