Near-Optimal Sparse Adaptive Group Testing

Nelvin Tan
National University of Singapore
nelvintan@u.nus.edu

Jonathan Scarlett
National University of Singapore
scarlett@comp.nus.edu.sg

***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 items are finitely divisible and thus may participate in at most γ tests (or alternatively, each test may contain at most ρ items). While information-theoretic limits and algorithms are known for the non-adaptive setting, relatively little is known in the adaptive setting. In this paper, 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. In broad scaling regimes, our upper and lower bounds on the number of tests asymptotically match up to a factor of e.**

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 outcome if at least one item in the test is defective. This problem is relevant in areas such as medical testing [1], data science [2], communication protocols [3], and many more [4].

One of the defining features of the group testing problem is the distinction between the non-adaptive and adaptive settings. In the non-adaptive 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.

A sparse group testing setting was recently proposed in [5] 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 [5] was on non-adaptive testing, and the main goal of this paper is to handle the adaptive setting.

A. Related Work

In the standard group-testing setting, in the absence of testing constraints, $T > (1-\epsilon)(d\log(\frac{n}{d}))$ tests are necessary to identify all defectives with error probability at most ϵ [6], [7]. 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 $\epsilon \in (0,1)$ [8], [9]. A matching upper bound is known for all $d \in o(n)$ in the unconstrained adaptive setting [10], whereas matching this lower bound non-adaptively is only possible in certain sparser regimes [11]–[13].

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 asymptotically vanishing error probability [7], [11], [14], [15]. Hence, the parameter regime of primary interest in the size-constrained setting is $\rho \in o(\frac{n}{d})$. By a similar argument, the parameter regime of primary interest in the finitely divisible setting is $\gamma \in o(\log(\frac{n}{d}))$. Combined with the condition $T \in \Omega(d\log(\frac{n}{d}))$, the latter scaling regime implies that

$$\frac{T}{\gamma d} \to \infty \tag{1}$$

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

For the non-adaptive setting, Gandikota *et al.* [5] proved the following results for the γ-divisible setting.

Theorem 1. [5] *For any sufficiently large n, sufficiently small $\epsilon > 0$, $\gamma \in o(\log n)$, and $d \in \Theta(n^\alpha)$ for some positive constant $\alpha \in [0,1)$, 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. [5] *For any sufficiently large n, sufficiently small $\epsilon > 0$, $\gamma \in o(\log n)$, and $d \in \Theta(n^\alpha)$ for some positive constant $\alpha \in [0,1)$, 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 [5].

Theorem 3. [5] *For any sufficiently large n, sufficiently small $\zeta > 0$, $\rho \in \Theta\big((\frac{n}{d})^\beta\big)$ (for some constant $\beta \in [0,1)$), and $d \in \Theta(n^\alpha)$ for some positive constant $\alpha \in [0,1)$, there exists a randomized non-adaptive group testing design that includes at most ρ items per test, using at most $\big\lceil\frac{1+\zeta}{(1-\alpha)(1-\beta)}\big\rceil\big\lceil\frac{n}{\rho}\big\rceil$ tests and ensuring a reconstruction error of at most $\epsilon = n^{-\zeta}$.*

Theorem 4. [5] *For any sufficiently large n, sufficiently small $\epsilon > 0$, $\rho \in \Theta\big((\frac{n}{d})^\beta\big)$ (for some constant $\beta \in [0,1)$), and $d \in \Theta(n^\alpha)$ for some positive constant $\alpha \in [0,1)$, any non-adaptive group testing algorithm that includes at most ρ items per test and has a probability of error of at most ϵ requires at least $\big(\frac{1-6\epsilon}{1-\beta}\big)\frac{n}{\rho}$ tests.*

We observe that for ρ-sized tests, both the lower and upper bounds have the same leading order term $\frac{n}{\rho}$. Hence, there is little gap between the lower and upper bounds. However, for γ-divisible items, the lower bound contains $(\frac{n}{d})^{(1-5\epsilon)/\gamma}$, while the upper bound contains $(\frac{n}{\epsilon})^{1/\gamma}$. Hence, there is significant gap between the lower and upper bounds; we will see that the gap can be made much smaller in the adaptive setting.

B. Contributions

For γ-divisible items, we provide a converse bound that holds even in the adaptive setting. In addition, we provide a noiseless adaptive algorithm and study the number of tests for reliable recovery with zero error probability. We establish that our algorithm is near-optimal by showing that the upper and lower bounds match up to factor of $e^{1+o(1)}$. In the appendix, we provide a noiseless adaptive algorithm for ρ-sized tests and study the number of tests for reliable recovery.

II. Problem Setup

We consider a population of n items $\{1, 2, \ldots, n\}$, among which a small subset $\mathcal{D} \subset \{1, 2, \ldots, n\}$ of size $d = |\mathcal{D}|$ is defective. Here we assume that the set $\mathcal{D}$ of defective items is chosen uniformly at random among all sets of size d (also known as the combinatorial prior [4]).[1] Formally, we represent the population by the binary vector $\mathbf{x} \in \{0, 1\}^n$ with weight d, where one represents a defective item and zero represents a non-defective item. We focus on the sparse regime $d \in o(n)$.

We wish to identify the defective items through a series of tests. At the i-th stage, we pick a subset (group) $\mathcal{S}_i$ of the population, represented by a binary vector $\mathbf{v}_i = (v_{i1}, \ldots, v_{in}) \in \{0, 1\}^n$, where $v_{ij} = 1$ denotes that item j is in the subset (group) for test i, and $v_{ij} = 0$ otherwise. We then test the items in $\mathcal{S}_i$ together and observe the test outcome $y_i = \bigvee_{j \in \mathcal{D}} v_{ij}$ (*i.e.*, whether there exists at least one defective item in the subset). We seek to minimize the number T of tests. We allow the tests to be adaptive, and focus primarily on the γ-divisibility constraint described in the introduction.

Given the tests and their outcomes, the decoding algorithm outputs an estimate vector $\widehat{\mathbf{x}} \in \{0, 1\}^n$, representing an estimate of $\mathbf{x}$. We seek to successfully identify $\mathbf{x}$ with a small probability of error. Concretely, we target the probability of error being bounded by some $\epsilon > 0$:

$$P_e = \mathbb{P}[\widehat{\mathbf{x}} \neq \mathbf{x}] \leq \epsilon, \tag{2}$$

where the probability is taken over the randomness of the set of defective items.

III. Converse

We first prove a counting-based upper bound on the success probability $\mathbb{P}(\mathrm{suc}) = 1 - P_e$, following similar proof techniques as [8], with suitable refinements to account for the γ-divisibility constraint. Afterwards, we will use the bound on $\mathbb{P}(\mathrm{suc})$ to prove our main converse bound, providing a lower bound on T for attaining a given target error probability.

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

Theorem 5. *Consider the case of n items with d defectives where each item can be tested at most γ times. Any algorithm (possibly adaptive) to recover the defective set $\mathcal{D}$ with T tests has success probability $\mathbb{P}(\mathrm{suc})$ satisfying*

$$\mathbb{P}(\mathrm{suc}) \leq \frac{\sum_{i=0}^{\gamma d} \binom{T}{i}}{\binom{n}{d}}. \tag{3}$$

Proof. Given a population of n objects, we write $\Sigma_{n,d}$ for the collection of subsets of size d from the population.

We follow the steps of [8] as follows: The testing procedure defines a mapping $\theta : \Sigma_{n,d} \to \{0, 1\}^T$. Given a putative defective set $S \in \Sigma_{n,d}$, $\theta(S)$ is the vector of test outcomes, with positive tests represented as 1s and negative tests represented as 0s. For each vector $\mathbf{y} \in \{0, 1\}^T$, we write $\mathcal{A}_\mathbf{y} \subseteq \Sigma_{n,d}$ for the inverse image of $\mathbf{y}$ under θ,

$$\mathcal{A}_\mathbf{y} = \theta^{-1}(\mathbf{y}) = \{S \in \Sigma_{n,d} : \theta(S) = \mathbf{y}\}. \tag{4}$$

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, the conditional probability of success given $\mathcal{D} = S$ 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:

$$\mathbb{P}(\mathrm{suc}) \stackrel{(a)}{=} \sum_{S \in \Sigma_{n,d}} \mathbb{P}(\mathrm{suc}|\mathcal{D} = S) \frac{1}{\binom{n}{d}} \tag{5}$$

$$= \frac{1}{\binom{n}{d}} \sum_{S \in \Sigma_{n,d}} \sum_{\mathbf{y} \in \{0,1\}^T} \mathbb{1}(\theta(S) = \mathbf{y}) \mathbb{P}(\mathrm{suc}|\mathcal{D} = S) \tag{6}$$

$$= \frac{1}{\binom{n}{d}} \sum_{S \in \Sigma_{n,d}} \sum_{\mathbf{y} \in \{0,1\}^T : |\mathcal{A}_\mathbf{y}| \geq 1} \mathbb{1}(\theta(S) = \mathbf{y}) \frac{1}{|\mathcal{A}_\mathbf{y}|} \tag{7}$$

$$= \frac{1}{\binom{n}{d}} \sum_{\mathbf{y} \in \{0,1\}^T : |\mathcal{A}_\mathbf{y}| \geq 1} \frac{1}{|\mathcal{A}_\mathbf{y}|} \Big(\sum_{S \in \Sigma_{n,d}} \mathbb{1}(\theta(S) = \mathbf{y}) \Big) \tag{8}$$

$$= \frac{1}{\binom{n}{d}} \sum_{\mathbf{y} \in \{0,1\}^T : |\mathcal{A}_\mathbf{y}| \geq 1} \frac{1}{|\mathcal{A}_\mathbf{y}|} |\mathcal{A}_\mathbf{y}| \tag{9}$$

$$= \frac{|\{\mathbf{y} \in \{0,1\}^T : |\mathcal{A}_\mathbf{y}| \geq 1\}|}{\binom{n}{d}} \tag{10}$$

$$\stackrel{(b)}{\leq} \frac{|\{\mathbf{y} \text{ with} \leq \gamma d \text{ ones}\}|}{\binom{n}{d}} = \frac{\sum_{i=0}^{\gamma d} \binom{T}{i}}{\binom{n}{d}}, \tag{11}$$

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. $\square$

We now use the result in (3) to prove the following converse.

Theorem 6. *Fix $\epsilon \in (0,1)$, and suppose that $d \in o(n)$, $\gamma \in o(\log n)$, and $\gamma d \to \infty$ as $n \to \infty$. Then any non-adaptive or adaptive group testing algorithm that tests each item at most γ times and has a probability of error of at most ϵ requires at least $e^{-(1+o(1))}\gamma d\big(\frac{n}{d}\big)^{1/\gamma}$ tests.*

Proof. From the counting bound in (2), we upper bound the sum of binomial coefficients [16, Section 4.7.] to obtain

$$\mathbb{P}(\text{suc}) \le \frac{e^{Th(\frac{\gamma d}{T})}}{\binom{n}{d}} \equiv \delta, \tag{12}$$

where $h(\cdot)$ is the binary entropy function in nats. From (12), we have $e^{Th(\frac{\gamma d}{T})}/\binom{n}{d} = \delta$, which implies that

$$\log\left(\delta\binom{n}{d}\right) = Th\Big(\frac{\gamma d}{T}\Big) \tag{13}$$

$$= \gamma d \log\frac{T}{\gamma d} + (T-\gamma d)\log\frac{1}{1-\frac{\gamma d}{T}} \tag{14}$$

$$\overset{(a)}{=} \gamma d \log\frac{T}{\gamma d} + \gamma d(1+o(1)), \tag{15}$$

where (a) uses a Taylor expansion and the fact that $\frac{\gamma d}{T} \in o(1)$ from (1). 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 (15), we obtain

$$\gamma d \log\frac{T}{\gamma d} = \log\left(\delta\binom{n}{d}\right) - \gamma d(1+o(1)) \tag{16}$$

$$\implies \log\frac{T}{\gamma d} = \frac{1}{\gamma d}\log\left(\delta\binom{n}{d}\right) - (1+o(1)), \tag{17}$$

which gives

$$T = e^{-(1+o(1))}\gamma d\left(\delta\binom{n}{d}\right)^{\frac{1}{\gamma d}} \tag{18}$$

$$\overset{(a)}{\ge} e^{-(1+o(1))}\gamma d\delta^{\frac{1}{\gamma d}}\left(\frac{n}{d}\right)^{\frac{1}{\gamma}}, \tag{19}$$

where (a) follows from the fact that $\binom{n}{d} \ge \big(\frac{n}{d}\big)^d$.

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

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 [8], [9].

Theorem 6 strengthens the previous information-theoretic lower bound in [5] for γ-divisible items (stating that $T \ge \gamma d(\frac{n}{d})^{(1-5\epsilon)/\gamma}$) by improving the dependence on ϵ, as well as extending its validity to the adaptive setting (whereas [5] used an approach based on Fano's inequality that is specific to the non-adaptive setting).

IV. Algorithm

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.

Algorithm 1 Adaptive algorithm for γ-divisible items

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

Initialize $M \leftarrow (\frac{n}{d})^{\frac{\gamma-1}{\gamma}}$ and the estimate $\hat{\mathcal{D}} \leftarrow \emptyset$

Arbitrarily group the n items into $\frac{n}{M}$ groups of size M

Test each group and discard any that return negative

Label the remaining groups incrementally as $G_j^{(0)}$, where $j = 1, 2, \ldots$

for $i = 1$ to $\gamma - 1$ **do**

    for each group $G_j^{(i-1)}$ from the previous stage **do**

        Arbitrarily group all items in $G_j^{(i-1)}$ into $M^{1/(\gamma-1)}$ sub-groups of size $M^{1-i/(\gamma-1)}$

        Test each sub-group and discard any that return a negative outcome

        Label the remaining sub-groups incrementally as $G_j^{(i)}$

    end for

end for

Add the items from all of the remaining singleton groups $G_j^{(\gamma-1)}$ to $\hat{\mathcal{D}}$

return $\hat{\mathcal{D}}$

A. 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.[2] Using Algorithm 1, we have the following theorem, which is proved throughout the remainder of the subsection.

Theorem 7. *For $\gamma \in o(\log n)$ and $d \in o(n)$, the adaptive algorithm in Algorithm 1 tests each item at most γ times and uses at most $\gamma d(\frac{n}{d})^{1/\gamma}(1+o(1))$ tests to recover the defective set exactly with zero error probability given knowledge of d.*

Proof. Similar to Hwang's generalized binary splitting algorithm [10], 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 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 (e.g., see 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,[3] then steadily decrease the sizes of each group down the stages: $M \to M^{1-1/(\gamma-1)} \to M^{1-2/(\gamma-1)} \to \cdots \to 1$ (see Figure 1). 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

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

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

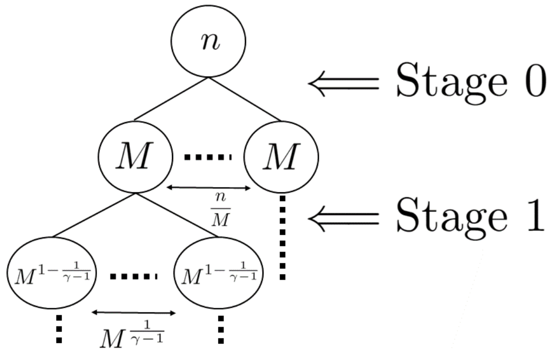

Figure 1: Visualization of splitting in the adaptive algorithm.

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:

$$T \leq \frac{n}{M} + (\gamma-1)dM^{\frac{1}{\gamma-1}}. \tag{20}$$

We optimize with respect to M by differentiating the upper bound and setting it to zero. This gives $M = (\frac{n}{d})^{\frac{\gamma-1}{\gamma}}$, and substituting $M = (\frac{n}{d})^{\frac{\gamma-1}{\gamma}}$ into the general upper bound in (20) gives the following upper bound:

$$T \leq \frac{n}{(\frac{n}{d})^{\frac{\gamma-1}{\gamma}}} + (\gamma-1)d\left[\left(\frac{n}{d}\right)^{\frac{\gamma-1}{\gamma}}\right]^{\frac{1}{\gamma-1}} = \gamma d\left(\frac{n}{d}\right)^{\frac{1}{\gamma}}. \tag{21}$$

□

Comparisons: Referring to Theorem 1, the upper bound for the non-adaptive algorithm of [5] using a randomized test matrix design is $T \leq \lceil e\gamma d(\frac{n}{\epsilon})^{1/\gamma}\rceil$. The non-adaptive algorithm has a $(\frac{n}{\epsilon})^{1/\gamma}$ term in the upper bound, while our adaptive algorithm has a $(\frac{n}{d})^{1/\gamma}$ term. Since ϵ is small but d is large, we see that our adaptive algorithm gives a significantly improved bound on the number of tests. Furthermore, the upper bound of our algorithm matches the information-theoretic lower bound in Theorem 6 up to a constant factor of $e^{1+o(1)}$. This proves that our algorithm is nearly optimal.

B. 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 [17], [18] are not suitable when $\gamma \ll \log\log d$, and may be wasteful of the budget γ even when $\gamma \gg \log\log d$.

To overcome this limitation, we introduce and evaluate two approaches to obtain a suitable input for d in Algorithm 1 given knowledge of an upper bound $d_{\max} \geq d$. The first approach uses $d_{\max}$ directly in Algorithm 1, while the second approach refines $d_{\max}$ by deriving an estimate $\widehat{d}$ that is passed to Algorithm 1. Note that we need $\widehat{d}$ to be an overestimate for the proof of Theorem 7 to still apply (with $\widehat{d}$ 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: $\widehat{d} \leftarrow$ (#positive bins)$/(1-\sqrt{\beta_n})$
5: **return** $\widehat{d}$

1) 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 (20), we obtain the following:

$$T \leq \frac{n}{(\frac{n}{d_{\max}})^{(\gamma-1)/\gamma}} + (\gamma-1)d\left[\left(\frac{n}{d_{\max}}\right)^{\frac{\gamma-1}{\gamma}}\right]^{\frac{1}{\gamma-1}}, \tag{22}$$

which simplifies to

$$T \leq (d_{\max} - d + \gamma d)\left(\frac{n}{d_{\max}}\right)^{\frac{1}{\gamma}}. \tag{23}$$

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

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

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

$$\frac{d}{B} = \frac{d}{d_{\max}/\beta_n} \leq \frac{d}{d/\beta_n} = \beta_n \to 0,. \tag{24}$$

Next, we show that with high probability as $n \to \infty$, $\widehat{d}$ overestimates d. From (24), we have

$$\mathbb{E}[\#\text{collisions}] \leq d\beta_n, \tag{25}$$

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

$$\mathbb{P}[\#\text{collisions} \geq d\sqrt{\beta_n}] \leq \sqrt{\beta_n}, \tag{26}$$

which implies the following:

$$\mathbb{P}[d - \#\text{collisions} \geq d - d\sqrt{\beta_n}] \geq 1 - \sqrt{\beta_n} \tag{27}$$

$$\implies \mathbb{P}\Big[\frac{d - \#\text{collisions}}{1-\sqrt{\beta_n}} \geq d\Big] \geq 1 - \sqrt{\beta_n}. \tag{28}$$

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

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

$$n_{\text{new}} = \frac{\beta_n dn}{d_{\max}},\ d_{\text{new}} = \widehat{d},\ \gamma_{\text{new}} = \gamma - 1. \tag{29}$$

We can then run Algorithm 1 to recover the defective set. Substituting our updated inputs into (20) and using $M = \big(\frac{\beta_n dn}{d_{\max}\widehat{d}}\big)^{\frac{\gamma-2}{\gamma-1}}$, we have the following bound for T:

$$T \leq \frac{d_{\max}}{\beta_n} + \frac{\beta_n dn}{d_{\max}\big(\frac{\beta_n dn}{d_{\max}\widehat{d}}\big)^{\frac{\gamma-2}{\gamma-1}}} + (\gamma-2)d\Big(\frac{\beta_n dn}{d_{\max}\widehat{d}}\Big)^{\frac{1}{\gamma-1}}, \tag{30}$$

which simplifies to

$$T \leq \frac{d_{\max}}{\beta_n} + (\widehat{d} - 2d + \gamma d)\Big(\frac{\beta_n dn}{d_{\max}\widehat{d}}\Big)^{\frac{1}{\gamma-1}} \tag{31}$$

$$\stackrel{(a)}{\leq} \frac{d_{\max}}{\beta_n} + \Big(\frac{d}{1-\sqrt{\beta_n}} - 2d + \gamma d\Big)\Big(\frac{\beta_n n}{d_{\max}}\Big)^{\frac{1}{\gamma-1}}, \tag{32}$$

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

Comparisons: By using T within the derived upper bounds, the first approach recovers the defective set with zero error probability while the second approach recovers the defective set with a small error probability determined by the β_n parameter. Referring to (23) and (32), 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, (23) is the better bound; its leading term is $\gamma d\big(\frac{n}{d}\big)^{1/\gamma}$. In particular, we note the following two cases: (i) If $\beta_n \ll \frac{1}{\gamma(\frac{n}{d})^{1/\gamma}}$, then the $\frac{d_{\max}}{\beta_n}$ term in (32) is strictly higher than $\gamma d\big(\frac{n}{d}\big)^{1/\gamma}$; (ii) If $\beta_n \gg \frac{1}{\gamma(\frac{n}{d})^{1/\gamma}}$, then some simple algebra gives $\frac{\beta_n n}{d} \gg \frac{1}{\gamma}\big(\frac{n}{d}\big)^{(\gamma-1)/\gamma}$, which implies that the $\gamma d\big(\frac{\beta_n n}{d}\big)^{1/(\gamma-1)}$ term from (32) is strictly higher than $\gamma d\big(\frac{n}{d}\big)^{1/\gamma}$ (note that $\big(\frac{1}{\gamma}\big)^{1/(\gamma-1)} = \Theta(1)$).

For $\gamma d \ll d_{\max} \ll n$, the choice of β_n can impact which bound is smaller. First note that the dominating term in (23) is $d_{\max}\big(\frac{n}{d_{\max}}\big)^{1/\gamma}$. Since the dominating term $\max\big\{\frac{d_{\max}}{\beta_n}, \gamma d\big(\frac{\beta_n n}{d_{\max}}\big)^{1/(\gamma-1)}\big\}$ in (32) is not obvious, we consider both possibilities: (i) $d_{\max}\big(\frac{n}{d_{\max}}\big)^{1/\gamma} \gg \frac{d_{\max}}{\beta_n}$ whenever $\beta_n \gg \big(\frac{d_{\max}}{n}\big)^{1/\gamma}$; and (ii) $d_{\max}\big(\frac{n}{d_{\max}}\big)^{1/\gamma} \gg \gamma d\big(\frac{\beta_n n}{d_{\max}}\big)^{\frac{1}{\gamma-1}}$ whenever $\beta_n \ll \big(\frac{d_{\max}}{\gamma d}\big)^{\gamma-1}\big(\frac{d_{\max}}{n}\big)^{1/\gamma}$. Combining these cases, we see that if β_n is in the range $\big(\frac{d_{\max}}{n}\big)^{1/\gamma} \ll \beta_n \ll \big(\frac{d_{\max}}{\gamma d}\big)^{\gamma-1}\big(\frac{d_{\max}}{n}\big)^{1/\gamma}$, the dominating term in (23) is greater than the dominating term in (32).

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

$$\lim_{n\to\infty} \log\Big(\frac{d_{\max}}{n}\Big)^{\frac{1}{\gamma}} = \lim_{n\to\infty} (\alpha-1)(\log n)^{1-c} = -\infty, \tag{33}$$

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.

V. Appendix

Our adaptive algorithm under the ρ-sized test constraint is a modification of Hwang's generalized binary splitting algorithm [10] where we divide the n items into $\frac{n}{\rho}$ groups of size ρ, instead of d groups of size $\frac{n}{d}$ as in the original algorithm.

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

$$T \leq \frac{n}{\rho} + d\log_2 \rho + O(d) \stackrel{(a)}{=} \frac{n}{\rho}(1+o(1)) + d\log_2 \rho, \tag{34}$$

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

$$T \leq \frac{n}{\rho}(1+o(1)). \tag{35}$$

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.

Acknowledgment. This work was funded by an NUS Early Career Research Award (grant R-252-000-A36-133).

REFERENCES

[1] R. Dorfman, "The detection of defective members of large populations," *Ann. Math. Stats.*, vol. 14, no. 4, pp. 436–440, 1943.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] M. Malyutov, "The separating property of random matrices," *Math. Notes Acad. Sci. USSR*, vol. 23, no. 1, pp. 84–91, 1978.

[7] 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.

[8] L. Baldassini, O. Johnson, and M. Aldridge, "The capacity of adaptive group testing," in *IEEE Int. Symp. Inf. Theory*, July 2013, pp. 2676–2680.

[9] O. Johnson, "Strong converses for group testing from finite blocklength results," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5923–5933, Sept. 2017.

[10] 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.

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

[12] 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.

[13] ——, "Optimal non-adaptive group testing," 2019, https://arxiv.org/abs/1911.02287.

[14] 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.

[15] 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.

[16] R. B. Ash, *Information Theory*. Dover Publications Inc., New York, 1990.

[17] 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.

[18] 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.