Electronic Colloquium on Computational Complexity, Report No. 141 (2023)

A Tight Lower Bound of $\Omega(\log n)$ for the Estimation of the Number of Defective Items

Nader H. Bshouty
Dept. of Computer Science
Technion
Haifa, Israel
bshouty@cs.technion.ac.il

Gergely Harcos
Number Theory Divison
Alfréd Rényi Institute of Mathematics
Budapest, Hungary
harcos.gergely@renyi.hu

Abstract. Let X be a set of items of size n , which may contain some defective items denoted by I, where $I \subseteq X$. In group testing, a *test* refers to a subset of items $Q \subset X$. The test outcome is 1 (positive) if Q contains at least one defective item, i.e., $Q \cap I \neq \emptyset$, and 0 (negative) otherwise.
We give a novel approach to obtaining tight lower bounds in non-adaptive randomized group testing. Employing this new method, we can prove the following result.
Any non-adaptive randomized algorithm that, for any set of defective items I, with probability at least $2/3$, returns an estimate of the number of defective items $|I|$ to within a constant factor requires at least $\Omega(\log n)$ tests.
Our result matches the upper bound of $O(\log n)$ and solves the open problem posed by Damaschke and Sheikh Muhammad in [8, 9] and by Bshouty in [2].

1 Introduction

Let X be a set of n items, among which are defective items denoted by $I \subseteq X$. In the context of group testing, a *test* is a subset $Q \subseteq X$ of items, and its result is 1 if Q contains at least one defective item (i.e., $Q \cap I \neq \emptyset$), and 0 otherwise.

Although initially devised as a cost-effective way to conduct mass blood testing [10], group testing has since been shown to have a broad range of applications. These include DNA library screening [20], quality control in product testing [22], file searching in storage systems [16], sequential screening of experimental variables [18], efficient contention resolution algorithms for multiple-access communication [16, 26], data compression [14], and computation in the data stream model [7]. Additional information about the history and diverse uses of group testing can be found in [6, 11, 12, 15, 19, 20] and their respective references.

ISSN 1433-8092

Adaptive algorithms in group testing employ tests that rely on the outcomes of previous tests, whereas *non-adaptive* algorithms use tests independent of the outcome of previous tests[1], allowing all tests to be conducted simultaneously in a single step. Non-adaptive algorithms are often preferred in various group testing applications [11, 12].

Estimating the number of defective items $d := |I|$ to within a constant factor of α is the problem of identifying an integer D that satisfies $d \le D < \alpha d$. This problem is widely utilized in a variety of applications [4, 23–25, 17].

Estimating the number of defective items in a set X has been extensively studied, with previous works including [3, 5, 8, 9, 13, 21]. In this paper, we focus specifically on studying this problem in the non-adaptive setting. Bshouty [1] showed that deterministic algorithms require at least $\Omega(n)$ tests to solve this problem. For randomized algorithms, Damaschke and Sheikh Muhammad [9] presented a non-adaptive randomized algorithm that makes $O(\log n)$ tests and, with high probability, returns an integer D such that $D \ge d$ and $\mathbf{E}[D] = O(d)$. Bshouty [1] proposed a polynomial time randomized algorithm that makes $O(\log n)$ tests and, with probability at least $2/3$, returns an estimate of the number of defective items within a constant factor.

As for lower bounds, Damaschke and Sheikh Muhammad [9] gave the lower bound of $\Omega(\log n)$; however, this result holds only for algorithms that select each item in each test uniformly and independently with some fixed probability. They conjectured that any randomized algorithm with a constant failure probability also requires $\Omega(\log n)$ tests. Ron and Tsur [21][2] and independently Bshouty [1] prove this conjecture up to a factor of $\log\log n$. Recently in [2], Bshouty established a lower bound of

$$\Omega\left(\frac{\log n}{(c\log^* n)^{(\log^* n)+1}}\right)$$

tests, where c is a constant and $\log^* n$ is the smallest integer k such that $\log\log \overset{k}{\ldots} \log n < 2$. It follows that the lower bound is

$$\Omega\left(\frac{\log n}{\log\log \overset{k}{\ldots} \log n}\right)$$

for any constant k.

In this paper, we close the gap between the lower and upper bound. We prove

Theorem 1. *Let $\alpha = 1 + \Omega(1)$. Any non-adaptive randomized algorithm that, with probability at least $2/3$, α-estimates the number of defective items must make at least*

$$\Omega\left(\frac{\log n}{\log \alpha}\right)$$

tests.

[1] A test may depend on previous tests but not on the outcomes of the previous tests.

[2] The lower bound in [21] pertains to a different model of non-adaptive algorithms, but their technique implies this lower bound.

In particular, for algorithms that estimate the number of defective items to within a constant factor, the bound is $\Omega(\log n)$.

To prove the Theorem, we first consider any algorithm that makes $m = \log n/(c\log\alpha)$ tests, for a sufficiently large constant c, and α-estimates the number of defective items. Next, we use this algorithm to construct another one that makes $2m$ tests and, when given any pair of sets of defective items where one set is α times the size of the other set, with high probability, can distinguish which set is the larger of the two. We then use Yao's principle to turn the algorithm to a deterministic algorithm that can do the same for a random pair of such sets. The input pairs are generated with a distribution that is uniform over the logarithm of the size d of the smaller set and uniformly distributed over pairs of subsets of X of sizes d and αd.

We then employ a central lemma (Lemma 3) in this paper's analysis. This lemma plays a pivotal role in our proof, requiring an innovative approach for its proof. This Lemma implies that if the number of tests is $2m$ then for an input drawn according to the above distribution, with high probability, the test outcomes for both sets are identical, making them indistinguishable. This leads to a contradiction and, as a result, establishes the lower bound of $m = \Omega(\log n/\log\alpha)$.

The paper is organized as follows: The next section introduces some definitions and notations. In Section 3, we present the main lemma that plays a crucial role in the proof of Theorem 1. Then in Section 4 we prove Theorem 1.

2 Definitions and Notation

In this section, we introduce some definitions and notation.

We will consider the set of *items* $X = [n] = \{1, 2, \ldots, n\}$ and the set of *defective items* $I \subseteq X$. The algorithm is provided with knowledge of n and has access to a test oracle, denoted as $\mathcal{O}_I$. The algorithm uses the oracle $\mathcal{O}_I$ to make a *test* $Q \subseteq X$, and the oracle responds with $\mathcal{O}_I(Q) := 1$ if $Q \cap I \neq \emptyset$, and $\mathcal{O}_I(Q) := 0$ otherwise.

We say that an algorithm $\mathcal{A}$ α*-estimates* the number of defective items with probability at least $1-\delta$ if, for every $I \subseteq X$, $\mathcal{A}$ runs in polynomial time in n, makes tests with the oracle $\mathcal{O}_I$, and with probability at least $1-\delta$, returns an integer $\mathcal{A}(I)$ such that[3] $|I| \le \mathcal{A}(I) < \alpha|I|$. If α is constant, then we say that the algorithm *estimates the number of defective items to within a constant factor*.

The algorithm is called *non-adaptive* if the queries are independent of the answers of previous queries and, therefore, can be executed simultaneously in a single step. Our objective is to develop a non-adaptive algorithm that minimizes the number of tests and provides, with a probability of at least $1-\delta$, an α estimation of the number of defective items.

---

[3] Some papers in the literature provide the following alternative definition: $|I|/\alpha \le \mathcal{A}(I) \le \alpha|I|$. It is worth noting that this alternative definition is equivalent to α^2-estimation, and the results in this paper also hold for this definition.

Throughout this paper, all logarithms are taken to the base 2 unless stated otherwise, and bold letters denote random variables.

In the Appendix, we prove the following lemma:

Lemma 1. *Let $\mathcal{A}$ be an algorithm that makes T tests and, with probability at least $2/3$, α-estimates the number of defective items. Then there is an algorithm $\mathcal{A}'$ that makes $O(T\log(1/\delta))$ tests and, with probability at least $1-\delta$, α-estimates the number of defective items.*

3 Preliminary Results

In this section, we present the main lemma that plays a crucial role in proving Theorem 1.

First, we prove the following lemma:

Lemma 2. *Let n be an integer. Given s integers $1 = q_1 \le q_2 \le \cdots \le q_{s-1} \le q_s = n$, define*

$$\sigma_\ell := \sum_{i=1}^{\ell} q_i \quad \textit{and} \quad \tau_\ell := \sum_{i=\ell+1}^{s} \frac{1}{q_i}.$$

Then,

$$\prod_{\ell=1}^{s-1} \max\left(1, \frac{1}{\sigma_\ell \tau_\ell}\right) > \frac{n}{4^s}.$$

Proof. First, we have

$$\prod_{\ell=1}^{s-1} \left(\frac{q_\ell}{q_{\ell+1}} \frac{\sigma_{\ell+1}}{\sigma_\ell} \frac{\tau_{\ell-1}}{\tau_\ell} \right) = \frac{q_1}{q_s} \cdot \frac{\sigma_s}{\sigma_1} \cdot \frac{\tau_0}{\tau_{s-1}} = \sigma_s \tau_0 > n.$$

On the other hand, the left-hand side satisfies

$$\begin{aligned} \frac{q_\ell}{q_{\ell+1}} \frac{\sigma_{\ell+1}}{\sigma_\ell} \frac{\tau_{\ell-1}}{\tau_\ell} &= \frac{q_\ell}{q_{\ell+1}} \left(1 + \frac{q_{\ell+1}}{\sigma_\ell}\right) \left(1 + \frac{1}{q_\ell \tau_\ell}\right) \\ &= \frac{q_\ell}{q_{\ell+1}} + \frac{q_\ell}{\sigma_\ell} + \frac{1}{q_{\ell+1}\tau_\ell} + \frac{1}{\sigma_\ell \tau_\ell} \\ &\le 3 + \frac{1}{\sigma_\ell \tau_\ell} \le 4 \max\left(1, \frac{1}{\sigma_\ell \tau_\ell}\right). \end{aligned}$$

Hence

$$\prod_{\ell=1}^{s-1} 4 \max\left(1, \frac{1}{\sigma_\ell \tau_\ell}\right) > n,$$

and the result follows. □

We now prove the main Lemma.

Lemma 3. *Let $\alpha \geq 2$ and $s = (\log n)/(2000 \log \alpha)$. Let $1 = q_1 \leq q_2 \leq \cdots \leq q_s = n$. Let*

$$Z = \{2^{\lfloor \log \alpha \rfloor + 1}, 2^{\lfloor \log \alpha \rfloor + 2}, \ldots, 2^{\lfloor \log(n/\alpha) \rfloor}\}.$$

Then:

$$\mathbf{Pr}_{z \in Z}\left[\sum_{q_i \leq z} q_i \leq \frac{z}{100\alpha} \text{ and } \sum_{q_i \geq z} \frac{1}{q_i} \leq \frac{1}{100\alpha z}\right] \geq \frac{99}{100},$$

where z is uniformly drawn from Z.

Proof. Let σ_ℓ and τ_ℓ be as defined in Lemma 2. For each $\ell \in [s-1]$, consider the interval[4] $I_\ell := [100\alpha\sigma_\ell, 1/(100\alpha\tau_\ell)]$. If $z \in I_\ell$, it satisfies $\sigma_\ell \leq z/(100\alpha)$ and $\tau_\ell \leq 1/(100\alpha z)$. Additionally, we have $z \geq 100\alpha\sigma_\ell > q_\ell$ and $z \leq 1/(100\alpha\tau_\ell) < q_{\ell+1}$. Therefore,

$$\sum_{q_i \leq z} q_i = \sigma_\ell \leq \frac{z}{100\alpha} \text{ and } \sum_{q_i \geq z} \frac{1}{q_i} = \tau_\ell \leq \frac{1}{100\alpha z}.$$

Furthermore, $I_\ell \subset (q_\ell, q_{\ell+1}) := \{q | q_\ell < q < q_{\ell+1}\}$. As a result, these sets I_ℓ are disjoint sets and therefore

$$\mathbf{Pr}_{z \in Z}\left[\sum_{q_i \leq z} q_i \leq \frac{z}{100\alpha} \text{ and } \sum_{q_i \geq z} \frac{1}{q_i} \leq \frac{1}{100\alpha z}\right] \geq \frac{\sum_{\ell=1}^{s-1} |Z \cap I_\ell|}{|Z|}. \tag{1}$$

Let Z' be the set of all the powers of 2. We will now show that all the powers of 2 that are in I_ℓ are also in Z. That is, $|Z \cap I_\ell| = |Z' \cap I_\ell|$. This follows from two facts. First, the largest powers of 2 that are in $I := \cup_\ell I_\ell$ are in $I_{s-1} = [100\alpha\sigma_{s-1}, n/(100\alpha)]$, and $\max_{z \in Z} z = 2^{\lfloor \log(n/\alpha) \rfloor} > n/(100\alpha)$. Second, the smallest power of 2 that are in I are in $I_1 = [100\alpha, 1/(100\alpha\tau_\ell)]$, and $\min_{z \in Z} z = 2^{\lfloor \log \alpha \rfloor + 1} < 100\alpha$.

Using Lemma 4 from the Appendix, the number of powers of 2 that are in the interval I_ℓ is

$$|Z' \cap I_\ell| \geq \left\lfloor \log \max\left(1, \frac{1}{10000\alpha^2\sigma_\ell\tau_\ell}\right)\right\rfloor.$$

Therefore, by Lemma 2,

$$\begin{aligned}
\sum_{\ell=1}^{s-1} |Z \cap I_\ell| &= \sum_{\ell=1}^{s-1} |Z' \cap I_\ell| \\
&\geq \sum_{\ell=1}^{s-1} \left\lfloor \log \max\left(1, \frac{1}{10^4\alpha^2\sigma_\ell\tau_\ell}\right)\right\rfloor \\
&\geq \left(\sum_{\ell=1}^{s-1} \log \max\left(1, \frac{1}{10^4\alpha^2\sigma_\ell\tau_\ell}\right)\right) - s
\end{aligned}$$

[4] If $a > b$ then $[a, b] = \emptyset$.

$$
\begin{aligned}
&= \log\left(\prod_{\ell=1}^{s-1} \max\left(1, \frac{1}{10^4\alpha^2\sigma_\ell\tau_\ell}\right)\right) - s \\
&\geq \log\left(\frac{1}{(10^4\alpha^2)^s}\prod_{\ell=1}^{s-1} \max\left(1, \frac{1}{\sigma_\ell\tau_\ell}\right)\right) - s \\
&\geq \log\left(\prod_{\ell=1}^{s-1} \max\left(1, \frac{1}{\sigma_\ell\tau_\ell}\right)\right) - (15+2\log\alpha)s \\
&\geq (\log n - 2s) - (15+2\log\alpha)s \\
&\geq \log n - (17+2\log\alpha)\frac{\log n}{2000\log\alpha} \\
&\geq \log n - \frac{19}{2000}\log n \\
&\geq \frac{99}{100}\log n \geq \frac{99}{100}|Z|. \qquad (2)
\end{aligned}
$$

By (1) and (2) the result follows. □

4 The Lower Bound

In this section, we present the proof of the theorem that establishes the lower bound on the number of tests required for any non-adaptive randomized algorithm to α-estimate the number of defective items, where $\alpha = 1 + \Omega(1)$.

We prove.

Theorem 1. *Let $\alpha = 1 + \Omega(1)$. Any non-adaptive randomized algorithm that, with probability at least $2/3$, α-estimates the number of defective items must make at least*

$$\Omega\left(\frac{\log n}{\log \alpha}\right)$$

tests.

In particular, for algorithms that estimate the number of defective items to within a constant factor, the bound is $\Omega(\log n)$.

Proof. First, it suffices to prove the lower bound for $\alpha \geq 2$, as any α-estimation where $2 > \alpha = 1+\Omega(1)$ also qualifies as a 2-estimation, and the lower bound for 2-estimation is $\Omega(\log n)$, which equates to $\Omega(\log n/\log\alpha)$ when $\alpha = 1 + \Omega(1)$.

Second, without loss of generality, we assume that n and α are both powers of two. This is because the lower bound for $n' = 2^{\lfloor \log n\rfloor}$ and $\alpha' = 2^{\lceil \log\alpha\rceil}$ is also a lower bound for n and α, and $\Omega(\log n'/\log\alpha') = \Omega(\log n/\log\alpha)$.

Furthermore, we will prove the lower bound for algorithms with a success probability of at least $7/8$. To get a success probability of at least $7/8$, just run the algorithm that has a success probability of at least $2/3$ three times and take the median of the outcomes. See the proof of Lemma 1. Therefore, both have the same asymptotic lower bound.

Suppose, to the contrary, that a non-adaptive randomized algorithm $\mathcal{A}$ exists, which makes

$$s := \frac{\log n}{2000 \log \alpha}$$

tests and, with probability at least 7/8, α-estimates the number of defective items. In other words, for any set of defective items $I \subseteq [n]$, the algorithm $\mathcal{A}$ makes s random tests (using the oracle $\mathcal{O}_I$) and, with probability at least 7/8, returns $\mathcal{A}(I)$ satisfying $|I| \le \mathcal{A}(I) < \alpha|I|$.

Now, we construct an algorithm $\mathcal{B}$ that, when given two sets of defective items $\{I_0, I_1\}$ where, for some $\xi \in \{0,1\}$, $I_\xi \supset I_{1-\xi}$ and $|I_\xi| = \alpha|I_{1-\xi}|$, makes $2s$ tests (using the oracles $\mathcal{O}_{I_0}$ and $\mathcal{O}_{I_1}$), and, with probability at least 3/4, can determine which of the two sets is larger, effectively outputting ξ.

Algorithm $\mathcal{B}$ first runs algorithm $\mathcal{A}$ to generate all the tests. This is feasible since algorithm $\mathcal{A}$ is non-adaptive. Then it makes these tests to both I_0 and I_1 using $\mathcal{O}_{I_0}$ and $\mathcal{O}_{I_1}$, respectively. If $\mathcal{A}(I_0) > \mathcal{A}(I_1)$, the algorithm outputs 0; otherwise, it outputs 1. The probability that neither of the following events occurs: $|I_0| \le \mathcal{A}(I_0) < \alpha|I_0|$ or $|I_1| \le \mathcal{A}(I_1) < \alpha|I_1|$, is at most 1/4. Thus, with probability of at least 3/4, $\mathcal{A}(I_\xi) \ge |I_\xi| = \alpha|I_{1-\xi}| > \mathcal{A}(I_{1-\xi})$, and $\mathcal{B}$ provides the correct answer.

We will now define a distribution D over pairs of sets of defective items. Let D_1 be the uniform distribution over $N := \{2^{\log \alpha}, 2^{\log \alpha + 1}, \ldots, 2^{\log(n/\alpha)-1}\}$. Initially, we select $\boldsymbol{d} \in N$ according to the distribution D_1. Next, we randomly and uniformly select $\boldsymbol{\xi}$ from $\{0,1\}$. Finally, we, uniformly at random, draw $\boldsymbol{I_\xi} \subseteq [n]$ of size $\boldsymbol{d}$ and $\boldsymbol{I}_{1-\boldsymbol{\xi}} \subseteq [n]$ such that $\boldsymbol{I}_{1-\boldsymbol{\xi}} \supseteq \boldsymbol{I_\xi}$ of size $\alpha \boldsymbol{d}$.

By applying Yao's Principle, we can conclude the existence of a deterministic, non-adaptive algorithm $\mathcal{C}$ that makes s tests and, when given $\{\boldsymbol{I}_0, \boldsymbol{I}_1\}$ drawn according to the distribution D, with probability of at least 3/4, correctly identifies the largest set.

Let $Q_1, Q_2, \ldots, Q_s \subseteq [n]$ be the tests that $\mathcal{C}$ makes. Note that $\mathcal{C}$ is deterministic, so $Q_1, Q_2, \ldots, Q_s$ are fixed and non-random. Let $q_i = |Q_i|$ for all $i \in [s]$. We can assume, without loss of generality, that $1 = q_1 \le q_2 \le \cdots \le q_{s-1} \le q_s = n$. In case where $q_1 \ne 1$ or $q_n \ne n$, then just add the two tests[5] $Q_0 = \{1\}$ and $Q_{s+1} = [n]$.

If $\boldsymbol{d} \in N$ is drawn according to distribution D_1, then $\boldsymbol{z} = n/\boldsymbol{d}$ is uniformly drawn from $\{2^{\log \alpha + 1}, 2^{\log \alpha + 2}, \ldots, 2^{\log(n/\alpha)}\}$. By Lemma 3, with probability at least 99/100, the chosen $\boldsymbol{z} = z$ ($\boldsymbol{d} = d$) satisfies

$$\sum_{q_i \le z} q_i \le \frac{z}{100\alpha} \quad \text{and} \quad \sum_{q_i \ge z} \frac{1}{q_i} \le \frac{1}{100\alpha z}. \tag{3}$$

Consider $\{\boldsymbol{I}_0, \boldsymbol{I}_1\}$ drawn according to distribution D conditioned on $\boldsymbol{d} = d$ satisfying (3). Without loss of generality, assume that $|\boldsymbol{I}_1| = \alpha d > d = |\boldsymbol{I}_0|$. Now let[6] $q_1 \le q_2 \le \cdots \le q_\ell < z < q_{\ell+1} \le \cdots \le q_s$. Define the event A_0 as the

[5] The lower bound will then be $s-2$.

[6] z cannot be equal to q_ℓ for any $\ell \in [s]$ because, otherwise, $1 = q_\ell \cdot (1/q_\ell) \le (\sum_{q_i \le q_\ell} q_i) \sum_{q_i \ge q_\ell} (1/q_i) \le (z/(100\alpha))(1/(100\alpha z) = 1/(10^4\alpha^2) < 1$.

situation where the outcomes of all the tests $Q_1, Q_2, \ldots, Q_\ell$ in algorithm $\mathcal{C}$ are 0. Then

$$\begin{aligned}
\mathbf{Pr}[\neg A_0 | \boldsymbol{d} = d] &= \mathbf{Pr}_{\boldsymbol{I}_0, \boldsymbol{I}_1, |\boldsymbol{I}_0| = d}[(\exists i \in [\ell])(\mathcal{O}_{\boldsymbol{I}_0}(Q_i) = 1 \vee \mathcal{O}_{\boldsymbol{I}_1}(Q_i) = 1)] \\
&= \mathbf{Pr}_{\boldsymbol{I}_0, \boldsymbol{I}_1, |\boldsymbol{I}_0| = d}\left[\bigvee_{i=1}^{\ell} (\boldsymbol{I}_0 \cap Q_i \neq \emptyset \vee \boldsymbol{I}_1 \cap Q_i \neq \emptyset)\right] \\
&= \mathbf{Pr}_{\boldsymbol{I}_1, |\boldsymbol{I}_1| = \alpha d}\left[\bigvee_{i=1}^{\ell} (\boldsymbol{I}_1 \cap Q_i \neq \emptyset)\right] && (4) \\
&\le \sum_{i=1}^{\ell} \mathbf{Pr}_{\boldsymbol{I}_1, |\boldsymbol{I}_1| = \alpha d}[\boldsymbol{I}_1 \cap Q_i \neq \emptyset] && (5) \\
&= \sum_{i=1}^{\ell} \left(1 - \prod_{j=0}^{\alpha d - 1} \left(1 - \frac{q_i}{n - j}\right)\right) && (6) \\
&\le \sum_{i=1}^{\ell} \left(1 - \left(1 - \frac{2q_i}{n}\right)^{\alpha d}\right) && (7) \\
&\le \sum_{i=1}^{\ell} \frac{2\alpha d q_i}{n} = 2\alpha \frac{1}{z} \sum_{i=1}^{\ell} q_i = \frac{1}{50}. && (8)
\end{aligned}$$

(4) follows from the fact that since $\boldsymbol{I}_0 \subset \boldsymbol{I}_1$ we have $\boldsymbol{I}_0 \cap Q_i \neq \emptyset$ implies $\boldsymbol{I}_1 \cap Q_i \neq \emptyset$. (5) follows from the union-bound rule. (6) follows from the fact that $\boldsymbol{I}_1$ is a random uniform subset of $[n]$ of size αd. Therefore, the probability that $\boldsymbol{I}_1 \cap Q_i \neq \emptyset$ is $1 - \binom{n-q_i}{\alpha d}/\binom{n}{\alpha d}$. Note here that when $n - q_i < \alpha d$ then $1 - \binom{n-q_i}{\alpha d}/\binom{n}{\alpha d} = 1 \le (n - q_i + 1)q_i/n \le \alpha d q_i/n < 2\alpha d q_i/n$ (the term in (8)). In such a case, we can safely disregard the inequality in step (7). Also, for terms where $2q_i/n > 1$ we have $\mathbf{Pr}_{\boldsymbol{I}_1}[\boldsymbol{I}_1 \cap Q_i \neq \emptyset] \le 1 < \alpha d(2q_i/n) = 2\alpha d q_i/n$ and again for those terms you can disregard the inequality in step (7). (7) follows from the fact that $n - j \ge n - \alpha d \ge n - \alpha 2^{\log(n/\alpha) - 1} \ge n/2$. (8) follows from the fact that $(1-x)^y \ge 1 - yx$ for $x \in [0, 1]$ and $y \ge 1$, then from (3) and $z = n/d$.

Now define the event A_1 as the situation where the outcomes of all the tests $Q_{\ell+1}, Q_{\ell+2}, \ldots, Q_s$ in algorithm $\mathcal{C}$ is 1. Then

$$\begin{aligned}
\mathbf{Pr}[\neg A_1 | \boldsymbol{d} = d] &= \mathbf{Pr}_{\boldsymbol{I}_0, \boldsymbol{I}_1, |\boldsymbol{I}_0| = d}[(\exists i \in [\ell])(\mathcal{O}_{\boldsymbol{I}_0}(Q_i) = 0 \vee \mathcal{O}_{\boldsymbol{I}_1}(Q_i) = 0)] \\
&= \mathbf{Pr}_{\boldsymbol{I}_0, \boldsymbol{I}_1, |\boldsymbol{I}_0| = d}\left[\bigvee_{i=\ell+1}^{s} (\boldsymbol{I}_0 \cap Q_i = \emptyset \vee \boldsymbol{I}_1 \cap Q_i = \emptyset)\right] \\
&= \mathbf{Pr}_{\boldsymbol{I}_0, |\boldsymbol{I}_0| = d}\left[\bigvee_{i=\ell+1}^{s} (\boldsymbol{I}_0 \cap Q_i = \emptyset)\right] && (9) \\
&\le \sum_{i=\ell+1}^{s} \mathbf{Pr}_{\boldsymbol{I}_0, |\boldsymbol{I}_0| = d}[\boldsymbol{I}_0 \cap Q_i = \emptyset]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=\ell+1}^{s} \left(\prod_{j=0}^{d-1} \left(1 - \frac{q_i}{n-j} \right) \right) \\
&\le \sum_{i=\ell+1}^{s} \left(1 - \frac{q_i}{n} \right)^d \\
&\le \sum_{i=\ell+1}^{s} \frac{n}{dq_i} = z \sum_{i=\ell+1}^{s} \frac{1}{q_i} \le \frac{1}{100}. \qquad (10)
\end{aligned}
$$

(9) follows from the fact that $\boldsymbol{I}_1 \cap Q_i = \emptyset$ implies that $\boldsymbol{I}_0 \cap Q_i = \emptyset$. (10) follows from the fact that $(1-x)^d \le 1/(dx)$ for any $0 < x \le 1$ and $d > 0$ combined with (3) and $\alpha \ge 2$.

Therefore, when considering $\{\boldsymbol{I}_0, \boldsymbol{I}_1\}$ drawn according to D, with probability at least $97/100$ (since $99/100 - 1/50 - 1/100 = 97/100$), algorithm $\mathcal{C}$ gets the same outcomes for both $\boldsymbol{I}_0$ and $\boldsymbol{I}_1$. Consequently, the success probability in this case is $1/2$ (essentially guessing). As a result, the overall success probability of $\mathcal{C}$ cannot be more than $3/100 + (1/2)(97/100) = 103/200$ which is less than $3/4$. This leads to a contradiction. □

References

1. Nader H. Bshouty. Lower bound for non-adaptive estimation of the number of defective items. In Pinyan Lu and Guochuan Zhang, editors, *30th International Symposium on Algorithms and Computation, ISAAC 2019, December 8-11, 2019, Shanghai University of Finance and Economics, Shanghai, China*, volume 149 of *LIPIcs*, pages 2:1–2:9. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.
2. Nader H. Bshouty. Improved lower bound for estimating the number of defective items. *CoRR*, abs/2308.07721, 2023.
3. Nader H. Bshouty, Vivian E. Bshouty-Hurani, George Haddad, Thomas Hashem, Fadi Khoury, and Omar Sharafy. Adaptive group testing algorithms to estimate the number of defectives. *ALT*, 2017.
4. Chao L. Chen and William H. Swallow. Using group testing to estimate a proportion, and to test the binomial model. *Biometrics.*, 46(4):1035–1046, 1990.
5. Yongxi Cheng and Yinfeng Xu. An efficient FPRAS type group testing procedure to approximate the number of defectives. *J. Comb. Optim.*, 27(2):302–314, 2014.
6. Ferdinando Cicalese. *Fault-Tolerant Search Algorithms - Reliable Computation with Unreliable Information.* Monographs in Theoretical Computer Science. An EATCS Series. Springer, 2013.
7. Graham Cormode and S. Muthukrishnan. What's hot and what's not: tracking most frequent items dynamically. *ACM Trans. Database Syst.*, 30(1):249–278, 2005.
8. Peter Damaschke and Azam Sheikh Muhammad. Bounds for nonadaptive group tests to estimate the amount of defectives. In *Combinatorial Optimization and Applications - 4th International Conference, COCOA 2010, Kailua-Kona, HI, USA, December 18-20, 2010, Proceedings, Part II*, pages 117–130, 2010.
9. Peter Damaschke and Azam Sheikh Muhammad. Competitive group testing and learning hidden vertex covers with minimum adaptivity. *Discrete Math., Alg. and Appl.*, 2(3):291–312, 2010.

10. R. Dorfman. The detection of defective members of large populations. *Ann. Math. Statist.*, pages 436–440, 1943.
11. D. Du and F. K Hwang. Combinatorial group testing and its applications. *World Scientific Publishing Company.*, 2000.
12. D. Du and F. K Hwang. Pooling design and nonadaptive group testing: important tools for dna sequencing. *World Scientific Publishing Company.*, 2006.
13. Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Estimating the number of defectives with group testing. In *IEEE International Symposium on Information Theory, ISIT 2016, Barcelona, Spain, July 10-15, 2016*, pages 1376–1380, 2016.
14. Edwin S. Hong and Richard E. Ladner. Group testing for image compression. *IEEE Trans. Image Processing*, 11(8):901–911, 2002.
15. F. K. Hwang. A method for detecting all defective members in a population by group testing. *Journal of the American Statistical Association*, 67:605—-608, 1972.
16. William H. Kautz and Richard C. Singleton. Nonrandom binary superimposed codes. *IEEE Trans. Information Theory*, 10(4):363–377, 1964.
17. Joseph L.Gastwirth and Patricia A.Hammick. Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: application to estimating the prevalence of aids antibodies in blood donors. *Journal of Statistical Planning and Inference.*, 22(1):15–27, 1989.
18. C. H. Li. A sequential method for screening experimental variables. *J. Amer. Statist. Assoc.*, 57:455–477, 1962.
19. Anthony J. Macula and Leonard J. Popyack. A group testing method for finding patterns in data. *Discrete Applied Mathematics*, 144(1-2):149–157, 2004.
20. Hung Q. Ngo and Ding-Zhu Du. A survey on combinatorial group testing algorithms with applications to DNA library screening. In *Discrete Mathematical Problems with Medical Applications, Proceedings of a DIMACS Workshop, December 8-10, 1999*, pages 171–182, 1999.
21. Dana Ron and Gilad Tsur. The power of an example: Hidden set size approximation using group queries and conditional sampling. *ACM Trans. Comput. Theory*, 8(4):15:1–15:19, 2016.
22. M. Sobel and P. A. Groll. Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Tech. J.*, 38:1179–1252, 1959.
23. William H. Swallow. Group testing for estimating infection rates and probabilities of disease transmission. *Phytopathology*, 1985.
24. Keith H. Thompson. Estimation of the proportion of vectors in a natural population of insects. *Biometrics*, 18(4):568–578, 1962.
25. S. D. Walter, S. W. Hildreth, and B. J. Beaty. Estimation of infection rates in population of organisms using pools of variable size. *Am J Epidemiol.*, 112(1):124–128, 1980.
26. Jack K. Wolf. Born again group testing: Multiaccess communications. *IEEE Trans. Information Theory*, 31(2):185–191, 1985.

Appendix

Lemma 1. *Let $\mathcal{A}$ be an algorithm that makes T tests and, with probability at least $2/3$, α-estimates the number of defective items. Then there is an algorithm $\mathcal{A}'$ that makes $O(T\log(1/\delta))$ tests and, with probability at least $1-\delta$, α-estimates the number of defective items.*

Proof. The algorithm $\mathcal{A}'$ runs $\mathcal{A}$ $m = O(\log(1/\delta))$ times (m is odd) and takes the median of the values it outputs. The probability that the median is not in the interval $[|I|, \alpha|I|]$ is the probability that $\mathcal{A}$ fails at least $\lceil m/2 \rceil$ times. By Chernoff's bound, the result follows. □

Lemma 4. *Let $a, b > 0$. The number of power of 2 that are in the interval $[a, b]$ is at least*

$$\left\lfloor \log \max\left(1, \frac{b}{a}\right) \right\rfloor .$$

Proof. If $b < a$ then $[a, b] = \emptyset$ and the number is 0.

If $b \geq a$ then let i and j be such that $2^i < a \leq 2^{i+1}$ and $2^{i+j+1} > b \geq 2^{i+j}$. Then the power of 2 that are in $[a, b]$ are $\{2^{i+1}, 2^{i+2}, \ldots, 2^{i+j}\}$ and their number is j. Then

$$j = \log \frac{2^{i+j}}{2^i} > \log \frac{b/2}{a} = \log \frac{b}{a} - 1.$$

This implies $j \geq \lfloor \log(b/a) \rfloor$. □

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il