



Quaternary splitting algorithm in group testing

Jinn Lu¹ · Hung-Lin Fu¹

Accepted: 14 October 2020 / Published online: 4 November 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

In Classical group testing, one is given a population of n items N which contains some defective d items inside. A group test (pool) is a test on a subset of N . Under the circumstance of no errors, a test is negative if the testing pool contains no defective items and the test is positive if the testing pool contains at least one defective item but we don't know which one. The goal is to find all defectives by using as less tests as possible, mainly to minimize the number of tests (in the worst case situation). Let $M(d, n)$ denote the minimum number of tests in the worst case situation where $|N| = n$ and d is the number of defectives. In this paper, we focus on estimating $M(d, n)$ and obtain a better result than known ones in various cases of d and n .

Keywords Group testing · Adaptive algorithm · Quaternary splitting

1 Introduction

The group testing problem has been developed since 1943 during the World War II (Dorfman 1943). The motivation was to detect syphilis among large number of blood samples via blood tests. Even the idea was not actually carried out, nowadays, it has been applied to many fields such as chemical leaking testing, codes, multi-access channel communication (Wolf 1985), data compression (Hong and Ladner 2002), compressed sensing (Donoho 2006), network security (Thai 2012), multimedia fingerprinting (Cheng and Miao 2011), and DNA library screening (Bruno et al. 1995; Farach et al. 1997). Various revised models are also considered in Du and Hwang (2000), Hu et al. (1981), Torney (1999) and Weng and Hwang (1993) to fit the real situation of applications, such as complex models, inhibitor models, and threshold models.

✉ Jinn Lu
jinnlu.am05g@g2.nctu.edu.tw

Hung-Lin Fu
hlifu@math.nctu.edu.tw

¹ Department of Applied Mathematics, National Chiao Tung University, 30010 HsinChu, Taiwan

Given a set N of n items and each item can be either defective or pure. A group test (or test in short) will be on a set $S \subset N$, denoted by S for convenience. Under the circumstance of no errors, the outcome of S is positive (denoted by P) if it contains at least one defective item and the outcome is negative (N) if it contains all pure items. The classical group testing problem is to use less tests as possible to find all defective items. Mainly, for all algorithms, the goal is to minimize the number of tests in the worst case situation.

Let $M_G(d, n)$ denote the minimum number of tests which are used in algorithm “G” on a set N of n items and there are d defective items. Moreover, let $M(d, n) = \min\{M_G(d, n) | G \text{ is an algorithm of group testing}\}$. So, as a consequence, $M_G(d, n) \geq M(d, n)$.

In general, there are two basic algorithms to tackle the problem: adaptive and non-adaptive algorithms. The tests designed in an adaptive algorithm may depend on the outcome of previous tests. But, in a non-adaptive algorithm all tests are carried out simultaneously. Therefore, an adaptive algorithm takes less tests in determining all the defectives in most of the cases. In this paper, we shall focus on applying adaptive algorithms.

The following facts are well-known.

Proposition 1.1 $M(d, n) \geq \lceil \log_2 \binom{n}{d} \rceil$.

This is known as the information lower bound. By using binary search, we have

Proposition 1.2 $M(1, n) = \lceil \log_2 n \rceil$.

Surprisingly, determining $M(2, n)$ for general n remains unsettled at this moment (Chang and Hwang 1981; Chang et al. 1989; Du and Hwang 2000). Based on the generalized binary splitting algorithm developed by in Hwang (1972), we can prove the following result.

Proposition 1.3 (Hwang 1972) *Let H be the generalized binary splitting algorithm. Then, $M_H(d, n) - \lceil \log_2 \binom{n}{d} \rceil \leq d - 1$ for $d \geq 2$.*

Indeed, this result provides a good estimation of $M(d, n)$ by way of the information lower bound.

In this paper, we propose an algorithm to estimate $M(d, n)$. Mainly, we prove that $M(d, n) - \lceil \log_2 \binom{n}{d} \rceil \leq d - 2$ for $3 \leq d \leq 5$. For larger d , a similar idea can be utilized, but it is getting more tedious even it can be done.

2 Results

We shall use $M_L(d, n)$ to denote the number of tests needed in algorithm “L” (depending on d) for finding d defectives among n items. Since we need $d = 2$ case for finding small cases, $M_L(2, n)$ will be introduced first.

The following results are easy to see.

Lemma 2.1 $M(2, 5) = M_L(2, 5) = 4$ and $M(2, 7) = M_L(2, 7) = 5$.

Proposition 2.2 $M_L(2, 10 \times 2^k) = 2k + 6$ and $M_L(2, 14 \times 2^k) = 2k + 7$, $k \geq 0$.

Proof We shall prove these two facts by induction alternatively on $k \geq -1$. By Lemma 2.1, the assertion is true for $k = -1$. Now, consider $k \geq 0$ and assume the inductive hypothesis holds for small k , i.e., $M_L(2, 10 \times 2^{k-1}) = 2k + 4$ and $M_L(2, 14 \times 2^{k-1}) = 2k + 5$.

First, we partition N into 4 subsets respectively, $N = S_0 \cup S_1 \cup S_2 \cup S_3$. If $N = 10 \times 2^k$, then let $|S_0| = 2^k$, $|S_1| = |S_2| = 2 \cdot 2^k$ and $|S_3| = 5 \cdot 2^k$, and if $N = 14 \times 2^k$, then $|S_0| = 2^k$, $|S_1| = |S_2| = 3 \cdot 2^k$ and $|S_3| = 7 \cdot 2^k$. Let T_i denote the i th test and $L(T_i)$ be its outcome.

Let $T_1 = S_0 \cup S_1$ be the first test for $N = 10 \times 2^k$. If $L(T_1) = N$, then the defectives are in $S_2 \cup S_3$. Since $|S_2 \cup S_3| = 7 \cdot 2^k = 14 \cdot 2^{k-1}$, the proof follows by induction, $1 + 2(k-1) + 7 = 2k + 6$. On the other hand, if $L(T_1) = P$, then let $T_2 = S_0 \cup S_2$. Now, if $L(T_2) = N$, then all the defectives are in $S_1 \cup S_3$. By finding one defective in S_1 (using binary splitting test) and one defective in $S_1 \cup S_3$ ($k + 3$ tests) we use at most $2k + 6$ tests. So, consider $L(T_2) = P$. Let $T_3 = S_1$. If $L(T_3) = N$, then find one defective in S_0 first by using k tests and find the other one in $S_0 \cup S_2 \cup S_3$ by at most $k + 4$ tests. In total, at most $2k + 6$ tests are used. On the other hand, if $L(T_3) = P$, then find one defective in S_1 by using at most $k + 1$ tests and on defective in $S_0 \cup S_2$ with at most $k + 3$ tests. Hence, at most $2k + 6$ tests are needed.

Consider $N = 14 \times 2^k$. Now, the process can be carried out by a similar argument. First if $L(S_0 \cup S_1) = N$, then we use $2k + 6$ tests for finding two defectives in $S_2 \cup S_3$. If $L(S_0 \cup S_1) = P$, then test $S_0 \cup S_2$. Now, in case that $L(S_0 \cup S_2) = N$, then all defectives are in $S_1 \cup S_3$, moreover $L(S_1) = P$. Hence, let $S_1 = S'_0 \cup S'_1$ and $S_3 = S'_2 \cup S'_3$ such that $|S'_0| = 2^k$, $|S'_1| = |S'_2| = 2 \times 2^k$ and $|S'_3| = 5 \times 2^k$. By previous argument, we start from testing $|S'_0 \cup S'_2|$ to find all the defectives in at most $2k + 5$ tests. If $L(S_0 \cup S_2) = P$, then let $T_3 = S_1$. Again, if $L(T_3) = N$, then all defectives are in $S_0 \cup S_2 \cup S_3$. By finding one in S_0 and one in $S_0 \cup S_2 \cup S_3$, we have the proof. On the other hand, if $L(T_3) = P$, then, one in S_1 and the other one in $S_0 \cup S_2$. This concludes the proof. $\square$

In summary, we have

Proposition 2.3 $M(2, n) \leq \begin{cases} 2k + 6, & \text{if } 14 \times 2^{k-1} < n \leq 10 \times 2^k, \text{ and} \\ 2k + 7, & \text{if } 10 \times 2^k < n \leq 14 \times 2^k. \end{cases}$

So, if $14 \times 2^{k-1} < n \leq 10 \times 2^k$, $\lceil \log_2 \binom{n}{2} \rceil \geq \lceil \log_2 (14 \times 2^{k-1} + 1) \rceil \geq \lceil \log_2 (7 \times 2^{k-1} \cdot (14 \times 2^{k-1} + 1)) \rceil \geq \lceil \log_2 98 \cdot 2^{2k-2} \rceil = \lceil \log_2 98 \rceil + (2k - 2) = 2k + 5$. On the other hand, if $10 \times 2^k < n \leq 14 \times 2^k$, then $\lceil \log_2 \binom{n}{2} \rceil \geq \lceil \log_2 (10 \times 2^k + 1) \rceil \geq \lceil \log_2 50 \rceil + 2k = 2k + 6$. In any case, $\lceil \log_2 \binom{n}{2} \rceil \leq |M(2, n)| \leq \lceil \log_2 \binom{n}{2} \rceil + 1$.

Now, consider $d = 3$. The following results are the initial cases of our algorithm.

Lemma 2.4 $M_L(3, 12) = 9$ and $M_L(3, 15) = 10$.

Proof For $n = 12$, test two items first (T_1). If $L(T_1) = P$, then it takes one more test to find one defective, and seven other tests to find the others since $M_L(2, 11) \leq 7$. On

Proof For $d = 5$, $n = 46 \times 2^k$, 53×2^k , 61×2^k , 70×2^k and 80×2^k , we partition n items into four subsets S_0 , S_1 , S_2 and S_3 with ratio $1 : 5 : 5 : 35$, $1 : 6 : 6 : 40$, $1 : 7 : 7 : 46$, $1 : 8 : 8 : 53$ and $1 : 9 : 9 : 61$, respectively. We take $T_1 = S_0 \cup S_1$ and $T_2 = S_0 \cup S_2$ if $L(T_1) = P$. Moreover, if $L(T_2) = P$, then take $T_3 = S_0$.

For $n = 46 \times 2^k$ and $L(T_3) = P$, we need $3 + k + M_L(4, 46 \times 2^k - 1) \leq 5k + 23$. If $L(T_3) = N$, we need $3 + (2k + 5) + M_L(3, 45 \times 2^k - 2) \leq 5k + 23$.

For $n = 53 \times 2^k$ and $L(T_3) = P$, we need $3 + k + M_L(4, 53 \times 2^k - 1) \leq 5k + 24$. If $L(T_3) = N$, we need $3 + 2 \times (k + 3) + M_L(3, (52 - 4) \times 2^k - 2) = 3 + 2 \times (k + 3) + M_L(3, 48 \times 2^k - 2) \leq 5k + 24$.

For $n = 61 \times 2^k$ and $L(T_3) = P$, we need $3 + k + M_L(4, 61 \times 2^k - 1) \leq 5k + 24$. If $L(T_3) = N$, we need $3 + 2 \times (k + 3) + M_L(3, (60 - 2) \times 2^k - 2) = 3 + 2 \times (k + 3) + M_L(3, 58 \times 2^k - 2) \leq 5k + 25$.

For $n = 70 \times 2^k$ and $L(T_3) = P$, we need $3 + k + M_L(4, 70 \times 2^k - 1) \leq 5k + 25$. If $L(T_3) = N$, we need $3 + 2 \times (k + 3) + M_L(3, 69 \times 2^k - 2) \leq 5k + 26$.

For $n = 80 \times 2^k$ and $L(T_3) = P$, we need $3 + k + M_L(4, 80 \times 2^k - 1) \leq 5k + 26$. If $L(T_3) = N$, we need $3 + (2k + 7) + M_L(3, (79 - 4) \times 2^k - 2) = 3 + (2k + 7) + M_L(3, 75 \times 2^k - 2) \leq 5k + 27$ or $3 + (2k + 6) + M_L(3, 79 \times 2^k - 2) \leq 5k + 27$. $\square$

Now, by calculating $\lceil \log_2 \binom{n}{4} \rceil$ and $\lceil \log_2 \binom{n}{5} \rceil$, we have the following

Corollary 2.12 $\lceil \log_2 \binom{n}{d} \rceil \leq M(d, n) \leq \lceil \log_2 \binom{n}{d} \rceil + d - 2$ for $3 \leq d \leq 5$.

Finally, we note that if $d > 5$ is considered, then d algorithms are needed to determine $M_L(d, n)$. Moreover, the sizes of four sets S_0 , S_1 , S_2 and S_3 are of sizes 2^k , $(d + t) \cdot 2^k$, $(d + t) \cdot 2^k$, and $(\frac{(3d-1)d}{2} + td + \frac{t(t-1)}{2}) \cdot 2^k$ respectively provided $n = (1 + 2(d + t) + \frac{(3d-1)d}{2} + td + \frac{t(t-1)}{2}) \cdot 2^k$ where $t \in \{0, 1, 2, \dots, d-1\}$.

3 Conclusions

By using quaternary splitting on the testing set N , we are able to obtain a more accurate estimation of $M(d, n)$. Though not match the information lower bound $\lceil \log_2 \binom{n}{d} \rceil$, we believe that the results obtained in this paper are very close to the exact answer of $M(d, n)$. This is due to the fact that for larger d , binary informatic partition is in general not possible.

Acknowledgements The authors are grateful to the reviewers for carefully reading the manuscripts and making helpful comments. Also, we appreciate the suggestions of Kuo-Ching Hwang and Wu-Hsiung Lin in writing this article

References

- Bruno WJ, Knill E, Bulding DJ, Bruce DC, Doggett NA, Sawhill WW, Stallings RL, Whittaker CC, Torney DC (1995) Efficient pooling designs for library screening. *Genomics* 26:21–30
- Chang G, Hwang FK (1981) A group testing problem on two disjoint sets. *SIAM J Algebraic Discrete Methods* 2:35–38

- Chang XM, Hwang FK, Weng JF (1989) Group testing with two and three defectives. *Ann N Y Acad Sci* 576:86–96
- Cheng M, Miao Y (2011) On anti-collision codes and detection algorithms for multimedia fingerprinting. *IEEE Trans Inf Theory* 57:4843–4851
- Donoho DL (2006) Compressed sensing. *IEEE Trans Inf Theory* 52:1289–1306
- Dorfman R (1943) The detection of defective members of large populations. *Ann Math Stat* 14:436–440
- Du DZ, Hwang FK (2000) Combinatorial group testing and its applications, 2nd edn. World Scientific, Singapore
- Parach M, Kannan S, Knill E, Muthukrishnan S (1997) Compression and complexity of sequences In: *Proceedings*, pp 357–367
- Hong ES, Ladner RE (2002) Group testing for image compression. *IEEE Trans Image Process* 11:901–911
- Hu MC, Hwang FK, Wang JK (1981) A boundary problem for group testing. *SIAM J Algebraic Discrete Methods* 2:81–87
- Hwang FK (1972) A method for detecting all defective members in a population by group testing. *J Am Stat Assoc* 67:605–608
- Thai MT (2012) *Group testing theory in network security: an advanced solution*. Springer, New York
- Torrey DC (1999) Sets pooling designs. *Ann Comb* 3:95–101
- Weng JF, Hwang FK (1993) An optimal group testing algorithm on k disjoint sets. *Op Res Lett* 13:43–44
- Wolf J (1985) Born again group testing: multiaccess communications. *IEEE Trans Inf Theory* 31:185–191

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.