Adaptive Group Testing with a Constrained Number of Positive Responses Improved

Peter Damaschke*

Department of Computer Science and Engineering
Chalmers University, 41296 Göteborg, Sweden
ptr@chalmers.se

Abstract

Group testing aims at identifying the defective elements of a set by testing selected subsets called pools. A test gives a positive response if the tested pool contains some defective elements. Adaptive strategies test the pools one by one. Assuming that only a tiny minority of elements are defective, the main objective of group testing strategies is to minimize the number of tests. De Bonis introduced in COCOA 2014 a problem variant where one also wants to limit the number of positive tests, as they have undesirable side effects in some applications. A strategy was given with asymptotically optimal test complexity, subject to a constant factor. In the present paper we reduce the test complexity, making also the constant factor optimal in the limit. This is accomplished by a routine that searches for a single defective element and uses pools of decreasing sizes even after negative responses. An additional observation is that randomization saves a further considerable fraction of tests compared to the deterministic worst case, if the number of permitted positive responses per defective element is small.

Keywords: group testing, positive test, adaptive strategy, randomized strategy

1 Introduction

Group testing aims at identifying the defective elements of a set by testing selected subsets called pools. That is, some unknown elements are defective, and a test gets a positive response if and only if the tested pool contains at least one defective element. Group testing is one of the most extensively studied combinatorial search problems. It has a history dating back to at least 1943, and has various modern applications including molecular biology [3, 4], fault detection [5, 8], conflict resolution [2], data compression [6], and computer security [10], to mention a few.

Usually the main goal is to find all defectives after a minimum number of tests. Countless variations of group testing differ in the assumptions on the number of defectives, the number

*Tel. 0046-31-772-5405. Fax 0046-31-772-3663.

of rounds of parallel tests, and restrictions on the choice of pools. We must refrain from even a cursory overview. Given the wealth of results it is amazing that only recently [1] a model has been introduced where one also limits the number of positive tests. The motivation is that positive tests can have undesired side effects. As an example, the defective elements could be radioactive sources or toxic substances, hence one wants to limit exposure to them. While, for trivial reasons, some positive tests are necessary to solve the search problem, their number should be kept small.

A test strategy is called adaptive if the tests are performed sequentially, hence the choice of the next test may depend on all earlier responses. Throughout the paper we use the following symbols for the parameters of interest.

n: number of elements
d: number of defective elements
t: number of tests
y: number of positive ("yes") tests

More precisely, n is the given number of elements, and d is a previously known upper bound on the number of defectives to expect (unless said otherwise). Whether t and y denote exact or expected numbers, upper bounds, or worst-case lower bounds will be clear from context or explicitly mentioned. Logarithms are base 2. To avoid heavy notation, rounding brackets and lower-order terms are omitted in expressions, as long as this does not affect their asymptotic behaviour. Symbol e denotes Euler's number $2.718\ldots$

It is folklore in the field that $t > d\log(n/d)$ is the information-theoretic lower bound on the worst-case number of tests, and some simple group testing strategies need essentially this number of tests only, when a bound d is known beforehand. In the model with limited y we focus on the most relevant case where the allowed y is a small fraction of t. (Any limit $y > t/2$ is not a severe restriction, as then we are almost back to the ordinary group testing problem.) To have a predefined limit we may assume that y is smaller than half the trivial information-theoretic lower bound on t.

Due to a simple argument [1], any strategy for d defectives must admit $y \geq d$ positive tests. Next it is shown in [1] that any strategy needs

$$y > \frac{d\log(n/d)}{\log(et/y)}$$

where t is the actual number of tests performed. For any prescribed y we solve this inequality for t and obtain

$$t > \frac{y}{e}\left(\frac{n}{d}\right)^{\frac{d}{y}}.$$

Defining $f := y/d$ (note that $f \geq 1$) this becomes

$$t > \frac{fd}{e}\left(\frac{n}{d}\right)^{\frac{1}{f}}.$$

Moreover, a strategy is provided in [1] that needs

$$t < fd\left(\frac{n}{d}\right)^{\frac{1}{f}} + fd$$

tests, at least for integer-valued f. Notice that there remains, essentially, a multiplicative gap of e. The algorithm in [1] builds upon Li's classic algorithm [7]. It works in f stages of tests that can be performed in parallel. Apart from minor technicalities, every stage partitions the remaining elements (that might still be defective) into the same number of disjoint pools of equal sizes. Only the positive pools in a stage need to be further searched from the next stage on. But when a strategy is adaptive anyway, it does not seem optimal to use equally sized pools within a stage. Intuitively it would be better to let the pool sizes strictly decrease even after negative tests, because the earlier a positive pool is encountered in a sequence, the more tests are still available to find a defective element therein. We will derive pool sizes that make optimal use if this idea. Interestingly, this improvement suffices to close the multiplicative gap, as we will see. This also establishes asymptotic tightness, that is, subject to a factor $1 + o(1)$, of the lower bound from [1].

Our second contribution is a simple randomized version of the strategy. A thorough analysis turns out to be intricate, but we get a few partial results implying that the expected test number is considerably smaller than the worst-case test number in the case of small y/d. We end with a conjecture about the optimal randimized test number.

2 An Asymptotically Optimal Adaptive Strategy

Lemma 1. *Suppose that a given set of n elements is already known to contain some defective elements, and we want to identify one of them, permitting at most y positive responses. This can be accomplished with fewer than $n^{1/y}(y/e)V + y$ tests, where V is some term that goes to 1 as y grows.*

Proof. For $t \geq y$ we define $N(t, y)$ to be the largest n such that it is possible to find one defective out of n elements using at most t test at most y of which may be positive. In the case of $t < y$ we define $N(t, y) := N(t, t)$.

Note that any sequential strategy keeps on testing certain pools until the first positive response is seen. Without loss of generality these pools are pairwise disjoint, since elements in negative pools need not be tested again. Upon a positive test we know that the tested pool has some defective element, and it remains to solve the problem restricted to this pool.

If $y = 1$, we cannot afford testing pools with more than one element, since after a positive response we have already used up the positive answer but have not yet identified a defective element. It follows immediately $N(t, 1) = t + 1$. (If t tests have been negative, we know without testing that the last element is defective.)

Now consider any $y > 1$. The ith pool can have $N(t - i, y - 1)$ elements. Indeed, if the ith pool is positive, we have performed i tests one of which was positive, thus we can still solve the residual problem on the ith pool. The last pool, with $i = t$, has size $N(0, y-1) = 1$. We can append another element that will be known to be defective if all t tests are negative. This shows

$$N(t,y) = 1 + \sum_{i=1}^{t} N(t-i, y-1) = 1 + \sum_{j=0}^{t-1} N(j, y-1).$$

Induction on y yields $N(t,y) \geq (t-y+2)^y/y!$ as follows. For $y = 1$ we have $N(t,1) = t+1 = (t-1+2)^1/1!$, and for $y \geq 2$ we bound the sum of a monotone increasing step function from below by an integral, therefore the argument j is replaced with $x - 1$:

$$N(t,y) > \sum_{j=0}^{t-1} N(j,y-1) > \frac{1}{(y-1)!}\int_{y-2}^{t} ((x-1)-(y-1)+2)^{y-1}\,dx = \frac{(t-y+2)^y}{y!}$$

With $n = N(t,y)$ and Stirling's formula in the upper-bound version of [9] we get

$$t-y+2 \leq n^{1/y}(y!)^{1/y} < n^{1/y}\left(\sqrt{2\pi}\cdot y^{y+1/2}e^{-y}e^{1/(12y)}\right)^{1/y} = n^{1/y}(y/e)\left(\sqrt{2\pi}\cdot y^{1/2}e^{1/(12y)}\right)^{1/y}$$

$$= n^{1/y}(y/e)(2\pi y)^{1/(2y)}e^{1/(12y^2)}.$$

Observe that $V := (2\pi y)^{1/(2y)}e^{1/(12y^2)}$ tends to 1 as y grows. □

We have invoked Stirling's formula and the "vanishing" factor V only for the sake of a simple general analysis that will lead to the optimal factor in the test number when y is large, as we show below in Theorem 2. Actually we improve upon [1] already for small y where Stirling's formula is unsuitable: Consider the problem of identifying one defective element in the smallest "interesting" case $y = 2$. (Case $y = 1$ trivially requires n tests.) Using equal pools of size $\sqrt{n}$ in the first stage would result in the upper bound $2\sqrt{n}$, whereas by using pools of sizes $t, t-1, t-2\ldots, 3, 2, 1$ we get $n = t^2/2$, hence $\sqrt{2}\sqrt{n}$ tests are sufficient; notice that t is here the worst-case test number. This is considerably smaller, by a factor $1/\sqrt{2}$. More generally, for fixed y we reduce the test number by a factor $(y)^{1/y}/y$, which finally goes to $1/e$ for $y \longrightarrow \infty$. This is the key to the main result.

Theorem 2. *In a set of n elements we can find d defective elements while permitting at most y positive responses, with*

$$\frac{fd}{e}Ve^{1/f}\left(\frac{n}{d}\right)^{1/f} + f + d$$

tests, where $f := y/d$ and V is some term that goes to 1 as y grows. Moreover, this bound is asymptotically optimal, provided that also d, f, and $(n/d)^{1/f}$ grow.

Proof. We partition the given set into g disjoint pools of size n/g that we call principal pools. We test the g principal pools, note that at most d of them are positive. Then we extract the at most d defective elements one by one, allowing $f - 1$ positive tests for each. We apply

Lemma 1 d times to n/g elements; note that y in the Lemma becomes $f-1$. Together with the g principal pools, the total number of tests is therefore smaller than

$$\frac{d(f-1)}{e}\left(\frac{n}{g}\right)^{\frac{1}{f-1}}V+g+f+d.$$

This holds regardless which principal pools contain how many defective elements. The d additional tests are used to test, after identification of every defective element, the rest of the principal pool for the presence of further defective elements. Note that for every defective element we do f positive tests: one test to signal that the principal pool must be further investigated, and $f-1$ positive tests in the actual search. Hence we need $y=fd$ positive tests as in [1], which does not depend on g. We set

$$g:=\left(\frac{d}{e}\right)^{\frac{f-1}{f}}n^{\frac{1}{f}}=\left(\frac{1}{e}\right)^{\frac{f-1}{f}}d\left(\frac{n}{d}\right)^{\frac{1}{f}}.$$

This also means

$$\frac{n}{g}=\left(\frac{e}{d}\right)^{\frac{f-1}{f}}n^{\frac{f-1}{f}}=e^{\frac{f-1}{f}}\left(\frac{n}{d}\right)^{\frac{f-1}{f}}.$$

We plug in this g and obtain the test number bound

$$\frac{(f-1)V+1}{e}e^{\frac{1}{f}}d\left(\frac{n}{d}\right)^{\frac{1}{f}}+f+d.$$

The factor V tends to 1 according to Lemma 1. For growing f we note that $e^{1/f}$ tends to 1 as well. Hence, in this case our bound becomes better than the bound in [1] by a factor $1/e$, which also matches the known lower bound. □

In the remainder of the paper we concentrate on the identification of a single element in a principal pool, which is the main new twist in our algorithm. That is, from now on we let $d=1$ and denote by n now the number of elements in a principal pool.

3 On Randomized Adaptive Strategies

Next we reduce the test number further, in a sense: We observe that for small y the expected number of tests is considerably smaller than the worst-case deterministic test number, in a suitably randomized strategy. For $y=1$, by testing the elements in random order, obviously we need only $\frac{1}{2}n$ tests in expectation. Standard arguments (such as Yao's technique) also show that this is optimal.

A natural idea for general y is now to apply a random permutation on the given set and then to run the algorithm from Theorem 2. Then the last stage where only one final positive response is permitted uses, in expectation, only half as many tests as in the worst case. Intuitively, randomization should therefore be most effective for small y, whereas for y approaching $\log n$ the problem converges to classical adaptive group testing where randomization is useless.

We conjecture that the proposed randomized strategy is asymptotically optimal also for general y. Note that it is not clear whether an optimal deterministic strategy applied to a random order is already the optimal randomized strategy, as the reduced test number in the last stage may change the optimal pool sizes to use in previous stages. At least, we can confirm the optimality conjecture in the case $y = 2$, too:

Proposition 3. *In a set of n elements we can find one defective element while permitting at most two positive responses, with fewer than $\frac{2}{3}\sqrt{2}\sqrt{n}$ tests, and this test number is asymptotically optimal.*

Proof. By Yao's minimax principle, some lower bound for randomzied strategies is the expected test number of the best deterministic strategy on an adversarial input distribution. Specifically, let our adversary permute the given set randomly, such that the unknown defective element is at every position with probability $1/n$, and consider any deterministic strategy on it. As argued earlier, the first stage of a strategy is completely characterized by the sizes of disjoint pools tested until a positive response is obtained. Let p denote the number of these pools, and $x_1, \ldots, x_p$ their sizes in the order in which they are tested. Clearly $x_1+\cdots+x_p = n$. Conditional on membership of the defective element in the kth pool, the expected number of tests is $k + \frac{1}{2}x_k$. Note that, due to $y = 2$, the searcher is forced to test single elements sequentially.

Hence the overall expected number of tests in the deterministic strategy amounts to

$$\sum_{k=1}^{p} \frac{x_k}{n} \left(k + \frac{x_k}{2}\right)$$

under the constraint $x_1 + \cdots + x_p = n$. Suppose that we take the best p and $x_1, \ldots, x_p$, that is, the numbers that minimize this expression. Furthermore, let now the searcher permute the set at random and test p pools of these sizes. This strategy obviously has the same number of expected tests until the defective is found, and since this number also equals a lower bound, this search strategy is optimal. (The argument is not tautologic; note that the adversary and the searcher have independent sources of randomness.) It remains to solve the aforementioned optimization problem.

Observe that $p = \Theta(\sqrt{n})$, since p of larger (smaller) order of magnitude would imply more than $O(\sqrt{n})$ tests in the first (second) stage. The only problem is the constant factor. First we fix p to have a fixed number of variables, and we calculate optimal sizes x_k. The Lagrange function of this restricted problem is

$$L = \sum_{k=1}^{p} \frac{x_k}{n} \left(k + \frac{x_k}{2}\right) + \lambda \sum_{k=1}^{p} x_k - \lambda n.$$

with a Lagrange multiplier λ for the constraint. We set the partial derivatives to zero:

$$\frac{\partial L}{\partial x_k} = \frac{k + x_k}{n} + \lambda = 0.$$

It follows that all $k + x_k$ are equal. Let c denote their common value, thus $x_k = c - k$ for all k. This also means $k + \frac{x_k}{2} = \frac{1}{2}(c + k)$. Since $(c - k)(c + k) = c^2 - k^2$, the expected test number simplifies to

$$E[t] = \frac{1}{2n} \sum_{k=1}^{p} (c^2 - k^2) < \frac{c^2 p}{2n} - \frac{p^3}{6n}.$$

We also have $n = \sum_{k=1}^{p} (c - k)$. We set $n = cp - \frac{p^2}{2}$ neglecting a linear term in $p = \Theta(\sqrt{n})$. This greatly simplifies the subsequent calculations and does not affect the asymptotic result for large n. Resolving the last equation for c we get $c = \frac{n}{p} + \frac{p}{2}$, thus

$$c^2 = \frac{n^2}{p^2} + n + \frac{p^2}{4}.$$

We substitute c^2 in $E[t]$ and obtain

$$E[t] < \frac{n}{2p} + \frac{p}{2} + \frac{p^3}{8n} - \frac{p^3}{6n} = \frac{n}{2p} + \frac{p}{2} - \frac{p^3}{24n}.$$

The value $p = \Theta(\sqrt{n})$ that minimizes this expression satisfies

$$-\frac{n}{2p^2} + \frac{1}{2} - \frac{p^2}{8n} = 0,$$

hence $p^2 = 2n$ and $p = \sqrt{2}\sqrt{n}$. We finally obtain

$$E[t] < \left(\frac{1}{2\sqrt{2}} + \frac{\sqrt{2}}{2} - \frac{2\sqrt{2}}{24} \right) \sqrt{n} = \left(\frac{1}{4} + \frac{1}{2} - \frac{1}{12} \right) \sqrt{2}\sqrt{n} = \frac{2}{3}\sqrt{2}\sqrt{n}.$$

This concludes the proof. □

Hence the randomized test number is still smaller than the worst-case test number by a factor of one third. While the above proof takes a general approach, some of the detailed calculations rely on $y = 2$ and do not seem to generalize. We must leave the conjecture as an open question. We argued already that randomization becomes less worthwhile when y is larger (but remains below $\log n$), yet from a theoretical point of view an optimal randomized result would be interesting.

4 Conclusions

We have further invetigated a variant of the group testing problem brought up in [1], where the number y of positive tests needed to identify d defectives is constrained. We got an asymptotically optimal test number t (including the leading constant factor), and we demonstrated the savings by randomization for small y/d. All technical results assumed integer y/d, but, of course, this ratio can also be fractional. However the results give hope that optimal continuous trade-offs between t and y could be achieved both in the deterministic and in the randomized setting by a more sophisticated analysis.

Acknowledgment

I am indebted to the referees for very careful reading.

References

[1] A. De Bonis: Efficient group testing algorithms with a constrained number of positive responses, in: Z. Zhang, L. Wu, W. Xu, D.Z. Du, D.Z. (Eds.), COCOA 2014, LNCS 8881, Springer, Heidelberg, 2014, pp. 506–521.

[2] A. De Bonis, U. Vaccaro: Constructions of generalized superimposed codes with applications to group testing and conflict resolution in multiple access channels, Theor. Comput. Sci. 306 (2003) 223–243.

[3] D.Z. Du, F.K. Hwang: Combinatorial Group Testing and Its Applications. Series on Appl. Math. vol. 3, World Scientific, 2000.

[4] D.Z. Du, F.K. Hwang: Pooling Design and Nonadaptive Group Testing. Series on Appl. Math. vol. 18, World Scientific, 2006.

[5] N.J.A. Harvey, M. Patrascu, Y. Wen, S. Yekhanin, V.W.S. Chan: Non-adaptive fault diagnosis for all-optical networks via combinatorial group testing on graphs, in: INFOCOM 2007, IEEE, 2007, pp. 697–705.

[6] E.S. Hong, R.E. Ladner: Group testing for image compression, IEEE Trans. Image Proc. 11 (2002) 901–911.

[7] C.H. Li: A sequential method for screening experimental variables, J. Amer. Statist. Assoc. 57 (1962) 455–477.

[8] C. Lo, M. Liu, J.P. Lynch, A.C. Gilbert: Efficient sensor fault detection using combinatorial group testing, in: DCOSS 2013, IEEE, 2013, pp. 199–206.

[9] H. Robbins: A remark on Stirling's formula, Amer. Math. Monthly 62 (1955) 26–29.

[10] M.T. Thai: Group Testing Theory in Network Security – An Advanced Solution. Springer, Heidelberg, 2012.