Constraining the Number of Positive Responses in Adaptive, Non-Adaptive, and Two-Stage Group Testing

Annalisa De Bonis
Dipartimento di Informatica, Università di Salerno,
Fisciano (SA), Italy.
email: debonis@dia.unisa.it

Abstract

Group testing is a well known search problem that consists in detecting the defective members of a set of objects O by performing tests on properly chosen subsets (*pools*) of the given set O. In classical group testing the goal is to find all defectives by using as few tests as possible. We consider a variant of classical group testing in which one is concerned not only with minimizing the total number of tests but aims also at reducing the number of tests involving defective elements. The rationale behind this search model is that in many practical applications the devices used for the tests are subject to deterioration due to exposure to or interaction with the defective elements. In this paper we consider adaptive, non-adaptive and two-stage group testing. For all three considered scenarios, we derive upper and lower bounds on the number of "yes" responses that must be admitted by any strategy performing at most a certain number t of tests. In particular, for the adaptive case we provide an algorithm that uses a number of "yes" responses that exceeds the given lower bound by a small constant. Interestingly, this bound can be asymptotically attained also by our two-stage algorithm, which is a phenomenon analogous to the one occurring in classical group testing. For the non-adaptive scenario we give almost matching upper and lower bounds on the number of "yes" responses. In particular, we give two constructions both achieving the same asymptotic bound. An interesting feature of one of these constructions is that it is an explicit construction. The bounds for the non-adaptive and the two-stage cases follow from the bounds on the optimal sizes of new variants of d-cover free families and (p, d)-cover free families introduced in this paper, which we believe may be of interest also in other contexts.

1 Introduction

Group testing is a well known search paradigm that consists in detecting the defective members of a set of objects O by performing tests on properly chosen subsets (*pools*) of the given set O. A test yields a "yes" response if the tested pool contains one or more defective elements, and a "no" response otherwise. The goal is to find all defectives by using as few tests as possible. Group testing origins date back to World War II when it was introduced as a possible technique for mass blood testing [14]. Since then group testing has found applications in a wide variety of situations ranging from conflict resolution algorithms for multiple-access systems [12], [34], fault diagnosis in optical networks [20], quality control in product testing [30], failure detection in wireless sensor networks [27], data compression [21], and many others. Among the modern applications of group testing, some of the most important are related to the field of molecular biology, where group testing is especially employed in the design of screening experiments. Du and Hwang [16] provide an extensive coverage of the most relevant applications of group testing in this area.

The different contexts to which group testing applies often call for variations of the classical model that best adapt to the characteristics of the problems. These variants concern the test model [3], [4], [7], [12], [13], the number of pursued defective elements [1], [8], as well as the structure of the test groups [5], [8], [33].

In this paper, we consider a variant of the classical model in which one is concerned not only with minimizing the total number of tests but aims also at reducing the number of tests involving defective elements. Therefore, the test groups should be structured so as to reduce the number of groups intersecting the set of defectives. The rationale behind this search model is that in many practical applications the devices used for the tests are subject to deterioration due to exposure to or interaction with the defective elements. In some contexts, the positive groups may even represent a risk for the safety of the persons that perform the tests. An example of such applications are leak testing procedures aimed at guaranteeing the safety of sealed radioactive sources [31, 32]. Radioactive sources are widely used in medical, industrial and agricultural applications, as well as in scientific research. Sealed sources are small metal containers in which radioactive material is sealed. As long as the sealed sources are handled correctly and the enclosing capsules are intact, they do not represent a health hazard. According to the radiation safety standards, sealed radioactive sources should be tested at regular intervals in order to verify the integrity of the capsules. Leak testing procedures are crucial in preventing contamination of facilities and personnel due to the escape of radioactive material. However, these procedures put the safety personnel at the risk of being exposed to radiation whenever a leak in the tested sources is present. Commonly, when not used, the sources are stored in lead-shielded drawers. In order to be tested for leakage, sources are removed one at time from the storage area and wiped with absorbent paper or a cotton swab held by a long pair of forceps. The wipe sample is then analyzed for radioactive contamination. An alternative procedure consists in testing the sources in groups. To this aim, the sources are not removed from the shielded storage drawer and a wipe sample is taken from the upper surface of the storage drawer. If the sample is contaminated then at least one source in the tested storage drawer is leaking; otherwise all sources in the drawer are intact. This idea suggests the use of group testing in leak testing procedures. Since leak testing procedures expose to risk the personnel that perform the tests on contaminated wipe samples, the number of positive tests admitted by the group testing procedure should depend on the dose of radiation which is judged to be of no danger for the health. Obviously, the total number of tests should also be taken into account in order to reduce the costs and the work load of the safety personnel. Trivially, the procedure that tests all elements individually attains the minimum number of positive responses, which is equal to the number of defectives in the input set. While this procedure may be an option when the danger implied by testing positive samples is extremely high, many practical applications call for procedures that can be tuned to obtain the desired tradeoff between the number of "admissible" positive responses and the total number of tests.

1.1 Summary of results

We consider *adaptive*, *non-adaptive* and *two-stage* group testing procedures. In adaptive group testing, at each step the algorithm decides which group to test by observing the responses of the previous tests. For classical group testing, there exist adaptive strategies that achieve the information theoretic lower bound $\Omega(d\log(n/d))$, where n is the total number of elements and d is the upper bound on the number of defectives. We will prove that in our model any adaptive algorithm must tolerate a number y of positive responses of order $\Omega\left(\frac{d\log(n/d)}{\log(et/y)}\right)$, where t is the total number of tests (i.e., the total of positive and negative tests), and give an adaptive algorithm that attains this lower bound. In fact, the exact values of the two bounds differ by a little constant. Therefore, if we require that $y = O(t^{1-c})$, for any positive constant $c < 1$, then the number of positive responses admitted by our optimal algorithm grows as $O\left(\frac{d\log(n/d))}{\log t}\right)$.

In many practical scenarios adaptive strategies are useless due to the fact that assembling the groups for the tests may be very time consuming and that some kind of group tests may take long time to give a response. In such applications, it is preferable to use non-adaptive strategies, i.e., strategies in which all tests are decided in advance and can be performed in parallel. Non-adaptive group testing strategies are much more costly than adaptive algorithms. Indeed, the minimum number of tests used by these procedures is equal to the minimum length of certain combinatorial structures known under the name of *d-superimposed codes* (or equivalently, the minimum size of the ground set of *d-cover free families*) [17], [18], [24]. The known bounds for these combinatorial structures imply that the number of tests of any non-adaptive group testing algorithm is lower bounded by $\Omega((d^2/\log d)\log n)$ and that there exist non-adaptive group testing algorithms that use $O(d^2\log n)$ tests. In order to study the non-adaptive case under our model, we will introduce a new variant of d-cover free families and derive upper and lower bounds on the size of these combinatorial structures. In particular, we will show that any non-adaptive algorithm for our group testing problem must admit a number of positive responses y of order $\Omega\left(\frac{d^2}{\log\left(\frac{etd^2}{y}\right)}\log n\right)$ and give two almost optimal algorithms that must tolerate a number of positive responses y of order $O\left(\frac{d^2}{\log\left(\frac{et}{y}\right)}\log n\right)$. An interesting feature of one of these constructions consists in being an *explicit* construction, in that there exists an efficient algorithm to design the underlying combinatorial structure. Interestingly, the gap between the above upper and lower bounds decreases as the ratio between the total number t of tests and the number y of positive responses admitted by the algorithm increases. Fot $y = t$, i.e., for algorithms that admit an unlimited number of positive responses, this gap is equal to that existing between the best upper and lower bounds on the minimum number of tests for classical group testing. Closing this gap is considered a major open problem in extremal combinatorics.

In [11] it has been proved that by allowing a little adaptiveness it is possible to dramatically improve on the number of tests used by non-adaptive group testing procedures. Indeed, the authors of that paper gave a trivial two-stage algorithm for classical group testing that uses the same number of tests of the best adaptive procedures. A trivial two-stage group testing algorithm consists of two non-adaptive stages. In the first stage the algorithm performs parallel tests on certain pools of elements with the aim of selecting a "small" subset of elements that are candidates to be the defective elements. In the second stage the elements selected by the first stage are tested individually so as to determine those that are really defective. In many contexts, such as molecular biology experiments involving the screening of library of clones [25], two-stage algorithms are considered as practical as non-adaptive algorithms. Indeed, in those applications, an element must undergo an individual test in order to be confirmed as defective, even though the responses to previous group tests indicate with no doubt that it is defective. Therefore, the tests carried out in the second stage are not considered an additional cost since the confirmatory tests should be performed anyway.

The problem of designing efficient group testing strategies consisting in a constant number of non-adaptive stages has been extended to different settings and variants of group testing and has received much attention in the recent literature [8], [9]. In the present paper we prove that a phenomenon similar to the one exhibited by classical group testing occurs also in our model. Indeed, we give a trivial two-stage group testing strategy that must tolerate the same number of "yes" responses admitted by the optimal adaptive algorithm. This result relies on an existential result proved for a new variant of the well known (p, d)-cover free families [17].

In Section 2, we present the lower bound for the adaptive case and give an algorithm that asymptotically achieves this bound. In Section 3, we first recall the definitions of d-separable families, d-cover free families and (p, d)-cover free families, and describe the existing relationship between these combinatorial structures and classical group testing. Then, in Section 3.1, we introduce our variants of these families which represent our main combinatorial tools. In Section 4, we consider the non-adaptive scenario and derive a lower bound on the number of "yes" responses that must be tolerated by any non-adaptive algorithm that uses at most a certain number t of tests. This lower bound is a consequence of an upper bound we prove in Section 4.1 on the size of our variant of (p, d)-cover free families. In Section 4.2 we give an existential result for these families based on the probabilistic method. For $p = 1$, this result shows that there exist non-adaptive algorithms achieving bounds which are very close to the lower bound. In Section 4.3, we give an explicit construction for our variant of d-cover free families which achieves the same asymptotic bound of the construction of Section 4.2. In Section 5, we consider trivial two-stage group testing and give an algorithm that admits the same asymptotic number of positive responses of the optimal adaptive algorithm of Section 2. This result is based on the existential result for our variant of (p, d)-cover free families of Section 4.2.

2 Adaptive group testing

In this section we deal with the case when tests are performed adaptively by looking at the feedbacks of already performed tests. For the purpose of our analysis, we need to introduce the following definition.

Definition 1 *Let t, n, d be positive integers with $n \geq d \geq 1$, and let O be a set of n elements containing at most d defective elements. Moreover, let $\mathcal{A}$ be a group testing strategy that finds all defective items in O by at most t tests. We denote by $y_{\mathcal{A}}(d,n,t)$ the maximum number of positive responses that occurs during the search process performed by $\mathcal{A}$, where the maximum is taken over all possible subsets of up to d defectives. The minimum value of $y_{\mathcal{A}}(d,n,t)$ is denoted by $y(d,n,t)$, where the minimum is taken over all group testing algorithms that use at most t tests to find all defectives in O.*

Notice that $y(d,n,t)$ represents the minimum number of positive responses that must be admitted in order to find up to d defectives in a set of n elements by at most t tests. The following lemma is quite straightforward.

Lemma 1 *Let t, n, d be positive integers with $n \geq d \geq 1$. Then, $y(d,n,t) \geq d$.*

Proof. Suppose by contradiction that $y(d,n,t) < d$. Then, in the case when the number of defectives is exactly d, there would be at least one defective element which either is never tested or appears only in groups that contain also other defective elements. In both cases, the algorithm could not decide whether this element is defective or not. This is due to the fact that the algorithm does not know a priori whether the number of defectives is d or it is smaller than d. □

In order to derive a lower bound on $y(d,n,t)$, we describe the search process by a binary tree where each internal node corresponds to a test and each leaf to one of the possible outcomes of the algorithm. For each internal node, its left branch is labelled with 0 and corresponds to a negative response, while its right branch is labelled with 1 and corresponds to a positive response. A path from the root to a leaf x represents the sequence of tests performed by the algorithm when the set of defective items is the one associated with x. Obviously, for an input set of size n that contains d defective elements, a group testing strategy is successful if and only if the corresponding tree has $\binom{n}{d}$ leaves. Let us denote by y the maximum number of "yes" responses in the whole sequence of test responses. Each root-to-leaf path can be represented by the binary vector whose entries are the labels of the branches along the path taken in the order they are encountered starting from the root. Since each path that starts from the root and ends in a leaf must contain at most y branches labelled with 1, the number of such binary vectors is smaller than or equal to $\sum_{i=0}^{y}\binom{t}{i}$. Since the number of leaves cannot be larger than the upper bound on the number of root-to-leaf paths, it holds

$$\sum_{i=0}^{y}\binom{t}{i} \geq \binom{n}{d}. \tag{1}$$

The above bound obviously holds also in the case when d is an upper bound on the number of defective elements.

Inequality (1) allows to derive a lower bound on $y(d,n,t)$. In order to obtain the desired bound, we make use of the following lemma which establishes an upper bound on the binary entropy $H(\frac{a}{b}) = -\frac{a}{b}\log\frac{a}{b} - (1-\frac{a}{b})\log(1-\frac{a}{b})$, for any a and b such that $0 < a < b$. In the following, unless specified differently, all logarithms are in base 2.

Lemma 2 *Let a and b such that $0 < a < b$. It holds*

$$H\left(\frac{a}{b}\right) \leq \frac{a}{b}\log\left(\frac{eb}{a}\right).$$

Proof. By the definition of binary entropy, one has that

$$\begin{aligned} H\left(\frac{a}{b}\right) &= \frac{a}{b}\log\frac{b}{a} + \left(\frac{b-a}{b}\right)\log\left(\frac{b}{b-a}\right) \\ &= \frac{a}{b}\log\frac{b}{a} + \frac{1}{b}\cdot\log\left(1+\frac{a}{b-a}\right)^{b-a} \\ &\leq \frac{a}{b}\log\frac{b}{a} + \frac{1}{b}\cdot\log e^{a}, \end{aligned} \tag{2}$$

from which the upper bound in the statement of the lemma follows. ◻

Below we will often resort to the following well known inequalities on the binomial coefficient

$$\binom{N}{m} \geq \left(\frac{N}{m}\right)^{m}, \tag{3}$$

$$\binom{N}{m} \leq \left(\frac{eN}{m}\right)^{m}, \tag{4}$$

where e denotes the Neper's constant $e = 2,71828\ldots$.

Theorem 1 *Let t, n, d be positive integers with $n \geq d \geq 1$. It holds that*

$$y(d,n,t) > \max\left\{d\,,\ \frac{d\log\binom{n}{d}}{\log\alpha}\right\},$$

where $\alpha = 4$ if $y(d,n,t) > t/2$, and $\alpha = \frac{et}{y(d,n,t)} \leq \frac{et\log\left(\frac{et}{d}\right)}{d\log\binom{n}{d}}$ if $y(d,n,t) \leq t/2$.

Proof. Let y denote the maximum number of positive responses admitted by an adaptive group testing algorithm that uses at most t tests to find up to d defectives. By inequality (1) we have that $\sum_{i=0}^{y}\binom{t}{i} \geq \binom{n}{d}$.

First we consider the case $y \leq t/2$. Stirling approximation implies the following well known inequality [19]

$$\sum_{i=0}^{\ell}\binom{m}{i} \leq 2^{mH(\ell/m)}. \tag{5}$$

where $\ell/m \leq 1/2$. By setting $m = t$ and $\ell = y$ in inequality (5), we get

$$\sum_{i=0}^{y} \binom{t}{i} \leq 2^{tH(y/t)}. \tag{6}$$

Lemma 2 implies that $H(\frac{y}{t}) \leq \frac{y}{t} \log \frac{et}{y}$, from which one has that

$$\sum_{i=0}^{y} \binom{t}{i} \leq 2^{y \log \frac{et}{y}}. \tag{7}$$

The lower bound on the binomial coefficients in (3) implies that

$$\binom{n}{d} \geq \left(\frac{n}{d}\right)^{d}. \tag{8}$$

Therefore, inequalities (1), (7), and (8) imply that, for $y \leq t/2$,

$$2^{y \log\left(\frac{et}{y}\right)} \geq \left(\frac{n}{d}\right)^{d}, \tag{9}$$

from which one has that

$$y \geq \frac{d \log\left(\frac{n}{d}\right)}{\log\left(\frac{et}{y}\right)}. \tag{10}$$

Now let us turn our attention to the case when $y > t/2$. In this case the bound follows from the information theoretic lower bound. One has that

$$y > t/2 \geq \frac{1}{2}\left\lceil \log \binom{n}{d} \right\rceil. \tag{11}$$

Inequalities (8) and (11) imply that

$$y \geq \frac{d}{2} \log\left(\frac{n}{d}\right). \tag{12}$$

The lower bound in the statement of the theorem is obtained by taking the maximum between the lower bound in Lemma 1 and either lower bound (10) or lower bound (12), according to whether $y \leq t/2$ or $y > t/2$. The term α in the bound of the theorem is equal to 4 when $y > t/2$, and is equal to $\frac{et}{y}$ when $y \leq t/2$. In this latter case we limit from above α by applying lower bound (10) to y in the expression of α, thus getting $\alpha = \frac{et}{y} \leq \frac{et \log\left(\frac{et}{y}\right)}{d \log\left(\frac{n}{d}\right)}$, which by the lower bound in Lemma 1 is at most $\frac{et \log\left(\frac{et}{d}\right)}{d \log\left(\frac{n}{d}\right)}$. $\square$

2.1 An asymptotically optimal algorithm

Now we present an algorithm that almost attains the lower bound of Theorem 1.

The algorithm is designed after Li's stage group testing algorithm [26]. While Li's analysis aims at minimizing the total number of tests, our algorithm performs a number of tests that depends on the number of positive responses admitted by the algorithm.

The algorithm works as follows. The tests are organized in stages in such a way that each stage tests a collection of disjoint subsets that form a partition of the search space. At stage i the search space is partitioned into $g_i \geq d$ groups, $g_i - 1$ of which have size k_i, while the remaining one might have size smaller than k_i. The elements in the subsets that test negative are discarded, while those in the subsets that test positive are grouped together to form the new search space. Notice that the tests in each stage can be performed in parallel. Let f denote the total number of stages. Notice that in stage i, $i = 1, \ldots, f$, the defective elements are contained in at most d of the g_i groups and therefore, after this stage, the search space consists of at most dk_i elements. The algorithm is successful if and only if after stage f the search space contains only the defective elements. This is insured by setting $k_f = 1$.

Let us ignore for the moment the integral constraints. The total number of tests performed by the algorithm is

$$t = \sum_{i=1}^{f} g_i \leq \frac{n}{k_1} + \frac{dk_1}{k_2} + \frac{dk_2}{k_3} + \ldots + \frac{dk_{f-2}}{k_{f-1}} + dk_{f-1}. \tag{13}$$

As observed before, in each stage at most d groups test positive and consequently, the total number of positive responses is upper bounded by fd. Obviously, the minimum is attained for $f = 1$, i.e, in the case when the algorithm consists in a single stage that tests each element individually. Therefore, it trivially holds

$$y(d, n, n) = d. \tag{14}$$

If we fix the number of stages f, the values of the k_i's do not affect the upper bound on the number of positive responses (as far as $g_i = \frac{dk_{i-1}}{k_i} \geq d$, i.e., $k_{i-1} \geq k_i$). Therefore, we choose the values of $k_1, \ldots, k_{f-1}$ which minimize the upper bound on t. As shown in [26], the minimum value of the right-hand side of (13) is attained for $k_i^* = \left(\frac{n}{d}\right)^{\frac{f-i}{f}}$, $i = 1, \ldots, f-1$. As a consequence, we have $g_1 = \left\lceil \frac{n}{k_1^*} \right\rceil$ and $g_i = d \left\lceil \frac{k_i^*}{k_{i+1}^*} \right\rceil$, for $i = 2, \ldots, f$. In each stage, the number of tests is at most $d\left\lceil (\frac{n}{d})^{\frac{1}{f}} \right\rceil$, and consequently, the total number of tests is

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

The above upper bound on t implies

$$f \leq \frac{\log(\frac{n}{d})}{\log(\frac{t}{fd} - 1 + \frac{1}{fd})} = \frac{\log(\frac{n}{d})}{\log(\frac{t+1}{fd} - 1)}. \tag{15}$$

Since the maximum number of positive responses is fd, we set $y_{\mathcal{A}}(d,n,t) = fd$ and have that inequality (15) implies that

$$y_{\mathcal{A}}(d,n,t) \leq \frac{d\log(\frac{n}{d})}{\log(\frac{t+1}{y_{\mathcal{A}}(d,n,t)}-1)}. \tag{16}$$

If the number $y_{\mathcal{A}}(d,n,t)$ of "yes" responses tolerated by the algorithm is larger than $\frac{t}{3}$ and $t < n$, then, in place of the above described algorithm, we use Hwang's algorithm [23] for classical group testing. This algorithm performs at most $d-1$ more tests than the information theoretic lower bound and therefore we have

$$y_{\mathcal{A}}(d,n,t) \leq t \leq \left\lceil \log\binom{n}{d} \right\rceil + d - 1. \tag{17}$$

The bounds in the statement of the following theorem follow from (17), (16), and (14). The lower bound on $\gamma = \frac{t+1}{y_{\mathcal{A}}(d,n,t)} - 1$ in the statement of the theorem is obtained by observing that, by upper bound (16), it holds

$$\gamma = \frac{t+1}{y_{\mathcal{A}}(d,n,t)} - 1 \geq \frac{(t+1)\log\left(\frac{t+1}{y_{\mathcal{A}}(d,n,t)}-1\right)}{d\log(\frac{n}{d})} - 1 > \frac{(t+1)}{d\log(\frac{n}{d})} - 1,$$

where the last inequality is a consequence of $y_{\mathcal{A}}(d,n,t)$ being at most $\frac{t}{3}$, from which it follows that $\log\left(\frac{t+1}{y_{\mathcal{A}}(d,n,t)}-1\right) > 1$.

Theorem 2 *Let t, n, d be positive integers with $n \geq d \geq 1$. There exists a group testing strategy $\mathcal{A}$ for which it holds that*

$$y_{\mathcal{A}}(d,n,t) \leq \begin{cases} d & \text{if } t = n, \\ \left\lceil \log\binom{n}{d} \right\rceil + d & \text{if } t < n \text{ and } y_{\mathcal{A}}(d,n,t) > t/3, \\ \frac{d\log(\frac{n}{d})}{\log\gamma} & \text{if } t < n \text{ and } y_{\mathcal{A}}(d,n,t) \leq t/3, \end{cases}$$

where $\gamma = \frac{t+1}{y_{\mathcal{A}}(d,n,t)} - 1 > \frac{(t+1)}{d\log(\frac{n}{d})} - 1$.

If we consider the case when more than 1/3 of the tests may receive a "yes" response, then it is immediate to see that the algorithm of Theorem 2 asymptotically attains the lower bound of Theorem 1.

Let us consider the case when at most 1/3 of the total number of tests are allowed to receive a "yes" response. Notice that the upper bounds of Theorem 2 translate into upper bounds on the number of tests that suffice to find up to d defective elements by a group testing algorithm that admits at most $y = y_{\mathcal{A}}(d,n,t)$ "yes" responses. Seen in this way, Theorem 2 implies that there exists an algorithm that uses

$$t \leq y2^{\frac{d\log\left(\frac{n}{d}\right)}{y}} + y - 1 \tag{18}$$

tests, where $y \leq \frac{t}{3}$ is the maximum number of positive responses admitted by the algorithm. Similarly, the lower bounds stated by Theorem 1 translate into lower bounds on the number of tests performed by any group testing algorithm that admits at most a certain number y of positive responses. If we consider algorithms that allow at most 1/3 of the tests to yield a "yes" response, Theorem 1 implies that any such algorithm performs at least

$$t \geq \frac{1}{e} y 2^{\frac{d \log\left(\frac{n}{d}\right)}{y}} \tag{19}$$

tests. The ratio between the upper bound (18) and the lower bound (19) is a constant, and as a consequence, the algorithm of Theorem 2 is asymptotically optimal.

3 Cover-free families and group testing

In this section, we describe the existing relationship between non-adaptive group testing and well known combinatorial structures such as d-separable families, d-cover free families and (p, d)-cover free families. We recall that a group testing algorithm is said to be *non-adaptive* if all tests must be decided beforehand without looking at the responses of previous tests.

In the following, for any positive integer m, we denote by $[m]$ the set of integers $\{1, \ldots, m\}$ and by $[m]_k$, $1 \leq k \leq m$, the set of all k-element subsets of $[m]$.

There exists a correspondence between non-adaptive group testing algorithms for input sets of size n and families of n subsets. Indeed, given a family $\mathcal{F} = \{F_1, \ldots, F_n\}$ with $F_i \subseteq [t]$, we design a non-adaptive group testing strategy as follows. We denote the elements in the input set by the integers in $[n] = \{1, \ldots, n\}$ and for $i = 1, \ldots, t$, define the group $T_i = \{j : i \in F_j\}$. Obviously, $T_1, \ldots, T_t$ can be tested in parallel and therefore the resulting algorithm is non-adaptive. Conversely, given a non-adaptive group testing strategy for an input set of size n that tests $T_1, \ldots, T_t$, we define a family $\mathcal{F} = \{F_1, \ldots, F_n\}$ by setting $F_j = \{i \in [t] : j \in T_i\}$, for $j = 1, \ldots, n$. Equivalently, any non-adaptive group testing algorithm for an input set of size n that performs t tests corresponds to a binary code of length t and size n. This is due to the fact that any family of size n on the ground set $[t]$ can be represented by the binary code of length t whose codewords are the characteristic vectors of the members of the family. Given such a binary code $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$, one has that j belongs to pool T_i if and only if the i-th entry $\mathbf{c}_j(i)$ of $\mathbf{c}_j$ is equal to 1.

A non-adaptive group testing strategy is successful if and only if the corresponding family is a *$\bar{d}$-separable* family, i.e., a family in which the unions of up to d members are pairwise distinct [15, 16]. To see this, let us represent the test responses by a binary vector whose i-th entry is equal to 1 if and only if T_i tests positive. We call this vector the *response vector*. Notice that the response vector is the characteristic vector of the union of the members of the family associated with the defective elements. In the binary code representation, this is equivalent to saying that the response vector is the OR of the codewords associated with the defective elements. Therefore, the set of the defective elements is univocally identified if and only if the union of up to d members of the family are pairwise distinct, that is, if and only if the family is $\bar{d}$-separable. The reader is referred to [15, 16] for a detailed account on these issues.

In spite of the equivalence between separable families and non adaptive group testing strategies, typically in the literature the design of non-adaptive algorithms is based on families satisfying a slightly stronger property that allows for a more efficient decoding algorithm to obtain the set of defectives from the test responses. These families satisfy the property that no member of the family is contained in the union of any other d members. Families with this property are called *d-cover free* families [18], whereas the corresponding binary codes are said to be *d-superimposed* or *d-disjunct* [15], [16], [17], [24]. Such codes have the property that for each codeword $\mathbf{c}$ and any other d codewords $\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}$ there exists an index i such that $\mathbf{c}$ has the i-th entry equal to 1, whereas all of $\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}$ have the i-th entry equal to 0. Given two binary vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ of length t, we say that $\mathbf{c}_2$ *covers* $\mathbf{c}_1$ if for any $i \in [t]$, $\mathbf{c}_1(i) = 1$ implies that $\mathbf{c}_2(i) = 1$. By using this terminology, we say that a code is d-superimposed (or d-disjunct) if and only if no codeword is covered by the Boolean OR of any other d columns. A consequence of this property is that any codeword associated with a regular (e.g., non defective) element is not covered by the response vector. Therefore, it is possible to recover the set of the defective elements by simply comparing the response vector with each codeword. On the other hand, if we use an algorithm based on a $\bar{d}$-separable family then, in order to obtain the set of the defective elements, we need to examine all subsets of up to d codewords.

The d-cover free families are a particular case of the *(p, d)-cover free families* introduced by D'yachkov and Rykov in [17] under the name of superimposed (d, n, p)-codes, where n denotes the size of the family. A (p, d)-cover free family is a family such that the union of any p members of the family is not contained in the union of any other d members of the family. For $p = 1$, (p, d)-cover free families are equivalent to d-cover free families. Analogously to what happens with d-cover free families, (p, d)-cover free families can be associated with non-adaptive group testing algorithms. However, these algorithms do not guarantee to determine exactly all defectives but allow only to obtain a subset of at most $p + d - 1$ elements containing all defective elements. Indeed, given a response vector $\mathbf{z}$, there might be up to $p + d - 1$ members of the families whose characteristic vectors are covered by $\mathbf{z}$. This is due to the fact that for any possible subset of up to d defective elements there are at most $p - 1$ other elements such that the members of the families corresponding to these elements are contained in the union of the members associated with the defective elements. The authors of [11] used a (d, d)-cover free family to design the first stage of their two-stage algorithm. This stage allows to determine a subset of up to $2d - 1$ elements including all defective elements. The elements in this subset are individually tested during the second stage in order to find out which ones of them are defective.

As a matter of fact, the authors of [11] based their algorithms on (k, m, n)-selectors, a combinatorial structure satisfying a slighter stronger property than that of (p, d)-cover free families. Their existential result for this combinatorial structure implies that there exists a (p, d)-cover free family of size n on a ground set of size

$$t < \frac{e(p+d)^2}{p} \ln \frac{n}{p+d} + \frac{e(p+d)(2(p+d)-1)}{p}. \qquad (20)$$

3.1 New variants of separable and cover-free families

In this section we introduce variants of separable and cover-free families that can be used to derive upper and lower bounds for the group testing problem we are considering.

Let $\mathcal{F} = \{F_1, \ldots, F_n\}$ be a family of subsets of $[t] = \{1, \ldots, t\}$. We will refer to the set $[t]$ as the *ground set* of the family. For a positive integer $k \leq t$, a family $\mathcal{F} = \{F_1, \ldots, F_n\}$ of subsets of $[t]$ is said to be *k-uniform*, if $|F_i| = k$, for $i = 1, \ldots, n$.

Given a family $\mathcal{F} = \{F_1, \ldots, F_n\}$, the corresponding group testing algorithm must admit a number of positive responses which is as large as the size of the largest union of up to d members of the family. Indeed, let $j_1, \ldots, j_m$, with $m \leq d$, be the defective elements. A group T_i intersects $\{j_1, \ldots, j_m\}$ if and only if $i \in F_{j_1} \cup \ldots \cup F_{j_m}$. Therefore, the number of positive responses is equal to $|F_{j_1} \cup \ldots \cup F_{j_m}|$. By the above argument, a non-adaptive group testing strategy that uses t tests and admits at most s positive responses is equivalent to the following notion of $\cup_{\leq s}$ $\bar{d}$-separable family.

Definition 2 *Let d, s, and t, $s \leq t$, be positive integers. We say that a family $\mathcal{F}$ on the ground set $[t]$ is a $\cup_{\leq s}$* $\bar{d}$-separable *family if the unions of up to d members of $\mathcal{F}$ are all distinct, and the union of any d members of $\mathcal{F}$ has size at most s. The maximum cardinality of a $\cup_{\leq s}$ $\bar{d}$-separable family on the ground set $[t]$ is denoted by $n_{sep}(d, \cup_{\leq s}, t)$.*

Analogously to what happens in classical group testing, cover free families allow to decode the response vector much more efficiently. Therefore, we introduce the following definition.

Definition 3 *Let d, s, and t, $s \leq t$, be positive integers. We say that a family $\mathcal{F}$ on the ground set $[t]$ is a $\cup_{\leq s}$* d-cover free *family if no member of $\mathcal{F}$ is contained in the union of other d members of $\mathcal{F}$, and the union of any d members of $\mathcal{F}$ has size at most s. The maximum cardinality of a $\cup_{\leq s}$ d-cover free family on the ground set $[t]$ is denoted by $n_{cf}(d, \cup_{\leq s}, t)$.*

It is immediate to see that $\cup_{\leq s}$ d-cover free families are $\cup_{\leq s}$ $\bar{d}$-separable families, and consequently, existential results for the former families apply also to the latter families. The following theorem shows that upper bounds on the maximum cardinality of $\cup_{\leq s}$ $(d-1)$-cover free families can be used to derive upper bounds on the maximum size of $\cup_{\leq s}$ $\bar{d}$-separable families.

Theorem 3 *Let d, s, and t, $s \leq t$, be positive integers. Any $\cup_{\leq s}$ $\bar{d}$-separable family is $\cup_{\leq s}$ $(d-1)$-cover free.*

Proof. First we show that any $\bar{d}$-separable family is a $(d-1)$-cover free family. This relation was noted by Kautz and Slingleton [24] and is quite simple to see. Indeed, suppose by contradiction that a $\bar{d}$-separable family is not $(d-1)$-cover free. As a consequence, there exist d members of the family $F_1, F_2, \ldots, F_d$ such that $F_d \subseteq F_1 \cup \ldots \cup F_{d-1}$, and therefore, it holds $\bigcup_{i=1}^{d} F_i = \bigcup_{i=1}^{d-1} F_i$ thus contradicting the fact that the family is $\bar{d}$-separable. Moreover, for any d members $F_1, F_2, \ldots, F_d$, it holds $|\bigcup_{i=1}^{d-1} F_i| < |\bigcup_{i=1}^{d} F_i| \leq s$, thus proving that the family is $\cup_{\leq s}$ $(d-1)$-cover free. $\square$

If we are not interested in determining exactly which elements are defective but only in confining the defective elements inside a reasonably small subset, then the following definition provides an useful combinatorial tool.

Definition 4 *Let p, d, s, and t, $s \leq t$, be positive integers. We say that a family $\mathcal{F}$ on the ground set $[t]$ is a $\cup_{\leq s}$ (p, d)-*cover free *family if the union of any p members of $\mathcal{F}$ is not contained in the union of other d members of $\mathcal{F}$, and the union of any d members of $\mathcal{F}$ has size at most s, The maximum cardinality of a $\cup_{\leq s}$ (p, d)-cover free family on the ground set $[t]$ will be denoted by $n_{cf}(p, d, \cup_{\leq s}, t)$*

The non-adaptive algorithm designed after a $\cup_{\leq s}$ (p, d)-cover free family has the property that at most s pools test positive and that at most $p-1$ non defective elements cannot be classified as such. Indeed, there are at most $p-1$ non defective elements that appear only in pools containing one or more defective elements. In other words, the response vector has weight at most s and covers at most $p+d-1$ codewords of the binary code associated with the family, that is, at most $p-1$ codewords in addition to those associated with the defective elements.

In Section 5, a $\cup_{\leq s}$ (p, d)-cover free family is employed to design the pools tested in the first stage of our trivial two-stage algorithm so that at most $d+p-1$ elements are candidates to be the defective elements after the first stage and should be individually probed during the second stage.

4 Non-adaptive group testing

In this section we present almost matching upper and lower bounds on the number of positive responses that should be admitted by a non-adaptive algorithm that uses at most t tests to find up to d defective elements in a given set of n elements. These bounds are obtained by establishing upper and lower bounds on the maximum size of $\cup_{\leq s}$ d-cover free families on the ground set $[t]$. Indeed, these bounds translate, respectively, into lower and upper bounds on the number s of positive responses that might be given to the tests. Our upper bound as well as one of our two constructions are given for the more general case of $\cup_{\leq s}$ (p, d)-cover free families. This existential result is proved by the probabilistic method and for $p = 1$ it achieves the same asymptotic bound of the construction for $\cup_{\leq s}$ d-cover free families given in [10], while improving on the estimate of the constant hidden in the asymptotic notation. The construction for $\cup_{\leq s}$ (p, d)-cover free families will be also employed to design the pools tested in the first stage of the two-stage algorithm of Section 5. Our second existential result is proved directly for $\cup_{\leq s}$ d-cover free families. This construction exhibits the interesting feature of being an explicit construction while attaining the same bound as the probabilistic construction.

In the following, given a non-adaptive algorithm $\mathcal{A}$ that finds up to d defective elements in an input set of size n by at most t tests, we denote by $\tilde{y}_{\mathcal{A}}(n, d, t)$ the maximum number of positive responses that may occur during the search process performed by $\mathcal{A}$, where the maximum is taken over all possible subsets of up to d defectives. Moreover, we denote by

$\tilde{y}(n,d,t)$ the minimum value of $\tilde{y}_{\mathcal{A}}(n,d,t)$ over all non-adaptive strategies $\mathcal{A}$ that find up to d defective elements in an input set of size n by at most t tests.

4.1 Negative Result

Theorem 4 *Let d and p be positive integers and let s and t be integers such that $s \le t$. The maximum size of a $\cup_{\le s}$ (p,d)-cover free family on the ground set $[t]$ is*

$$n_{cf}(p,d,\cup_{\le s},t) \le \begin{cases} \binom{t}{\lceil t/2 \rceil} & \text{if } d=1,\ p=1,\ \text{and } t<2s, \\ \binom{t}{s} & \text{if } d=1,\ p=1,\ \text{and } t\ge 2s, \\ (p+d-1)2^{\frac{t}{d}} & \text{if } d=1<p \text{ or } 2\le d<2p,\ \text{and } t<2s, \\ (p+d-1)\left(\frac{et}{s}\right)^{\frac{s}{d}} & \text{if } d=1<p \text{ or } 2\le d<2p,\ \text{and } t\ge 2s, \\ p\left(\frac{etd(d+2)}{4ps}\right)^{\left\lceil \frac{s}{p\lfloor d/(2p)\rfloor^2+\lfloor d/(2p)\rfloor}\right\rceil}+\frac{d}{2}+2p-2 & \text{if } d\ge 2p. \end{cases}$$

Proof. The first bound for the case $d=1$ and $p=1$ follows from the upper bound $\mathcal{F} \le \binom{t}{\lceil t/2\rceil}$ on the size of a Sperner family $\mathcal{F}$ on the ground set $[t]$ with members of unlimited size, while the second bound for the case $d=1$ and $p=1$ follows from the upper bound $\mathcal{F} \le \binom{t}{s}$ on the size of a Sperner family $\mathcal{F}$ on the ground set $[t]$ and with members of size at most $s \le t/2$.

Let us prove the bound for $2 \le d < 2p$. In this case the bound is a consequence of Proposition 2 in [17]. The authors of [17] noticed that for any subfamily Q, with $|Q| \le d$, of a (p,d)-cover free family, there are at most $\binom{d+p-1}{d}$ subfamilies of d members of the family such that the union of the d members in each of these subfamilies is equal to the union of the d members of Q. This implies that for a (p,d)-cover free family of size n, there are at least $\frac{\binom{n}{d}}{\binom{d+p-1}{d}}$ distinct sets that can be obtained from the union of d members of the family. Since our (p,d)-cover free families have the additional property that the union of any d members of the family has size at most s, the following condition must be satisfied.

$$\sum_{i=0}^{s}\binom{t}{i} \ge \frac{\binom{n}{d}}{\binom{d+p-1}{d}}, \tag{21}$$

where the sum in the left-hand side represents the maximum number of subsets of $[t]$ of size less than or equal to s.

For $t \ge 2s$, we bound $\sum_{i=0}^{s}\binom{t}{i}$ by exploiting inequality (7) in Section 2, whereas for $t<2s$, we bound from above $\sum_{i=0}^{s}\binom{t}{i}$ by $\sum_{i=0}^{t}\binom{t}{i}$, and therefore, we have that

$$\sum_{i=0}^{s}\binom{t}{i} \le \begin{cases} 2^{s\log\frac{et}{s}} & \text{if } t\ge 2s, \\ 2^t & \text{if } t<2s. \end{cases} \tag{22}$$

By inequality (22) and inequality (21), one has that for $t \ge 2s$,

$$2^{s\log\frac{et}{s}} \ge \frac{\binom{n}{d}}{\binom{d+p-1}{d}}, \tag{23}$$

whereas for $t < 2s$, it holds that

$$2^t \geq \frac{\binom{n}{d}}{\binom{d+p-1}{d}}. \tag{24}$$

The bound (24) is the same bound obtained by [11] for the case $d < 2p$.

The right-hand side of (21) is equal to

$$\frac{n!}{(n-d)!d!} \cdot \frac{d!(p-1)!}{(d+p-1)!} = \frac{n(n-1)\cdots(n-d+1)}{(d+p-1)(d+p-2)\cdots p} \geq \left(\frac{n}{d+p-1}\right)^d.$$

Therefore, we can lower bound the right-hand sides of (23) and (24) by $\left(\frac{n}{d+p-1}\right)^d$, thus getting

$$2^{s \log \frac{et}{s}} \geq \left(\frac{n}{d+p-1}\right)^d, \quad \text{for } t \geq 2s, \tag{25}$$

whereas for $t < 2s$, it holds that

$$2^t \geq \left(\frac{n}{d+p-1}\right)^d, \quad \text{for } t < 2s. \tag{26}$$

The bounds for $2 \leq d < 2p$ in the statement of the theorem follow immediately from (25) and (26).

Now let us turn our attention to the case $d \geq 2p$. We assume for the moment that d be a multiple of $2p$ and drop this assumption later on. Let $\mathcal{F}$ be a $\cup_{\leq s}$ (p,d)-cover free family on the ground set $[t]$ and let us define the sets $G_1, \ldots, G_{d/2}$ as follows. We set G_1 to be the largest member of $\mathcal{F}$ and, for each $i = 2, \ldots, d/2$, G_i to be the largest set in $\{F \setminus \bigcup_{j=1}^{i-1} G_j : F \in \mathcal{F} \setminus \{G_1, \ldots, G_{i-1}\}\}$. In other words, after choosing G_1 as the largest member of the family, we remove the elements of G_1 from all members of $\mathcal{F} \setminus \{G_1\}$ and set G_2 to be the largest of the resulting sets. Then, we remove the elements of G_2 from all unselected sets and set G_3 to be the largest of the sets of the form $F \setminus (G_1 \cup G_2)$, for $F \in \mathcal{F} \setminus \{G_1, G_2\}$, and so on until $d/2$ sets are selected. Let $\mathcal{F}'$ be the family obtained by removing the elements of $G_1, \ldots, G_{d/2}$ from all members of $\mathcal{F} \setminus \{G_1, \ldots, G_{d/2}\}$, i.e., $\mathcal{F}' = \{F \setminus \bigcup_{j=1}^{d/2} G_j : F \in \mathcal{F} \setminus \{G_1, \ldots, G_{d/2}\}\}$. We show that the the union of any p members of $\mathcal{F}'$ is not contained in the union of any other $d/2$ members of the family. Suppose by contradiction that there are $d/2 + p$ sets $F'_1, F'_2, \ldots, F'_{d/2+p} \in \mathcal{F}'$ such that $F'_1 \cup \ldots \cup F'_p \subseteq F'_{p+1} \cup \ldots \cup F'_{d/2+p}$. Since for $i = 1, \ldots d/2+p$, it is $F'_i = F_i \setminus \bigcup_{j=1}^{d/2} G_j$ for some set $F_i \in \mathcal{F} \setminus \{G_1, \ldots, G_{d/2}\}$, it holds $F_1 \cup \ldots \cup F_p \subseteq F_{p+1} \cup \ldots \cup F_{d/2+p} \cup G_1 \cup \ldots \cup G_{d/2}$, thus contradicting the fact that $\mathcal{F}$ is (p,d)-cover free. Notice that it might be that the members of $\mathcal{F}'$ are not pairwise distinct and that some members of $\mathcal{F}'$ are empty. By the same argument as above one can prove that there exist at most $p-1$ sets $B_i \in \mathcal{F}' \cup \{G_1, \ldots, G_{d/2}\}$ such that $B_i = \emptyset$ or $B_i = B_j$ for some other member of $B_j \in \mathcal{F}' \cup \{G_1, \ldots, G_{d/2}\}$. If we remove these up to $p-1$ sets from $\mathcal{F}'$, we obtain a collection whose members are non-empty and pairwise distinct. Let us denote by $\mathcal{F}''$ this collection. By construction, $\mathcal{F}''$ is a $\cup_{\leq s}$ $(p, d/2)$-cover free family of cardinality larger than or equal to $|\mathcal{F}| - d/2 - p + 1$. In the following, we derive an upper bound on the cardinality of $\mathcal{F}''$. To this aim, we exploit the fact that the members of $\mathcal{F}''$ are non-empty and pairwise distinct and that $\mathcal{F}''$ is $\cup_{\leq s}$ $(p, d/2)$-cover free.

Notice that $G_1, \ldots, G_{d/2}$ are pairwise disjoint and that $|G_1| \geq |G_2| \geq \ldots \geq |G_{d/2}|$. Moreover, it holds $G_i \cap F'' = \emptyset$ and $|G_i| \geq |F''|$, for any $i = 1, \ldots, d/2$ and $F'' \in \mathcal{F}''$. Therefore, for any member $F'' \in \mathcal{F}''$, one has that

$$|\bigcup_{i=1}^{d/2} G_i \cup F''| = \sum_{i=1}^{d/2} |G_i| + |F''| \geq (d/2+1)|F''|. \tag{27}$$

Since $G_1, \ldots, G_{d/2}$ are members of $\mathcal{F}$ and F'' is subset of some member of $\mathcal{F}$, one has that $|\bigcup_{i=1}^{d/2} G_i \cup F''| \leq s$, which, along with (27), implies $|F''| \leq \lfloor \frac{2s}{d+2} \rfloor$. Since F'' is an arbitrary member of $\mathcal{F}''$, inequality (27) holds for any member F'' of $\mathcal{F}''$.

Observe that if d is a multiple of p then for any p members $F_1, \ldots, F_p$ of size at most m of a (p, d)-cover free family, there exists a subset A of at most $\lceil mp/d \rceil$ elements such that $A \subseteq F_j$ for some $F_j \in \{F_1, \ldots, F_p\}$ and $A \not\subseteq F$ for any member F of the family such that $F \notin \{F_1, \ldots, F_p\}$. Indeed, if otherwise it would be possible to partition each of $F_1, \ldots, F_p$ into d/p subsets of size at most $\lceil mp/d \rceil$ each of which is contained in a member of the family different from $F_1, \ldots, F_p$. This would imply that there exist $\leq d$ members of the family that contain all elements of $F_1 \cup \ldots \cup F_p$, thus contradicting the hypothesis of the family being a (p, d)-cover free family. Since, by assumption, $d/2$ is a multiple of p, we can apply this observation to our $(p, d/2)$-cover free family $\mathcal{F}''$. We proved that all members of $\mathcal{F}''$ have size at most $\lfloor \frac{2s}{d+2} \rfloor$, therefore the above observation implies that, for any p members $F''_1, \ldots F''_p$ of $\mathcal{F}''$, there exists a set A of size at most $\lceil 4sp/(d(d+2)) \rceil$ such that $A \subseteq F''_j$ for some $F''_j \in \{F''_1, \ldots, F''_p\}$ and $A \not\subseteq F''$ for any member F'' of the family different from $F''_1, \ldots, F''_p$. Now let us form $\lfloor |\mathcal{F}''|/p \rfloor$ pairwise disjoint subfamilies $\mathcal{F}''_1, \ldots, \mathcal{F}''_{\lfloor |\mathcal{F}''|/p \rfloor}$ of $\mathcal{F}''$ each consisting of p members of $\mathcal{F}''$. By the above argument, for each such a subfamily $\mathcal{F}''_i$ there exists a subset A_i of at most $\lceil 4sp/(d(d+2)) \rceil$ elements such that A_i is entirely contained in some member of $\mathcal{F}''_i$ and is not contained in any member of $\mathcal{F}''_j$, for $j \neq i$. It follows that the family $\{A_1, \ldots, A_{\lfloor |\mathcal{F}''|/p \rfloor}\}$ is a Sperner family, i.e., an antichain. The following celebrated inequality, known under the name of LYM inequality, establishes a relationship between the cardinalities of the members of a Sperner family $\mathcal{G}$ and the size m of the ground set of the family.

$$\sum_{G \in \mathcal{G}} \frac{1}{\binom{m}{|G|}} \leq 1. \tag{28}$$

Since $\{A_1, \ldots, A_{\lfloor |\mathcal{F}''|/p \rfloor}\}$ is a Sperner family on the ground set $[t]$, LYM inequality implies

$$\sum_{i=1}^{\lfloor |\mathcal{F}''|/p \rfloor} \frac{1}{\binom{t}{|A_i|}} \leq 1. \tag{29}$$

Moreover, $A_1, \ldots, A_{\lfloor |\mathcal{F}''|/p \rfloor}$ have size at most $\lceil 4sp/(d(d+2)) \rceil$ which, by the assumption $d \geq 2p \geq 2$, is at most $\lceil s/2 \rceil \leq \lceil t/2 \rceil$. Therefore, one has that $\binom{t}{|A_i|} \leq \binom{t}{\lceil 4sp/(d(d+2)) \rceil}$, for $i = 1, \ldots, \lfloor |\mathcal{F}''|/p \rfloor$. It follows that the left-hand side of (29) is larger than or equal to $\frac{\lfloor |\mathcal{F}''|/p \rfloor}{\binom{t}{\lceil 4sp/(d(d+2)) \rceil}}$ thus implying $\lfloor |\mathcal{F}''|/p \rfloor \leq \binom{t}{\lceil 4sp/(d(d+2)) \rceil}$, from which

$$|\mathcal{F}''| \leq p \binom{t}{\lceil 4sp/(d(d+2)) \rceil} + p - 1. \tag{30}$$

Since $|\mathcal{F}| \leq |\mathcal{F}''| + d/2 + p - 1$, inequality (30) implies

$$|\mathcal{F}| \leq p \binom{t}{\lceil 4sp/(d(d+2)) \rceil} + d/2 + 2p - 2. \tag{31}$$

Now let us drop the assumption that d is a multiple of $2p$. Observe that $d \geq 2p\lfloor d/(2p) \rfloor$ and therefore, one has that

$$n_{cf}(p, d, \cup_{\leq s}, t) \leq n_{cf}(p, 2p\lfloor d/(2p) \rfloor, \cup_{\leq s}, t).$$

We upper bound $n_{cf}(p, 2p\lfloor d/(2p) \rfloor, \cup_{\leq s}, t)$ by using (31) with d replaced by $2p\lfloor d/(2p) \rfloor$, thus obtaining

$$n_{cf}(p, d, \cup_{\leq s}, t) \leq p \binom{t}{\lceil \frac{s}{p\lfloor d/(2p) \rfloor^2 + \lfloor d/(2p) \rfloor} \rceil} + p\lfloor d/(2p) \rfloor + 2p - 2.$$

The bound for $d \geq 2p$ in the statement of the theorem follows from applying the upper bound in (4) to the binomial coefficient in the above inequality. $\square$

By setting $p = 1$ in the bound of Theorem 4, we obtain the following upper bound on the maximum size of $\cup_{\leq s}$ d-cover free families.

Corollary 1 *Let $d \geq 1$, s and t, $s \leq t$, be positive integers. The maximum size of a $\cup_{\leq s}$ d-cover free family on the ground set $[t]$ is*

$$n_{cf}(d, \cup_{\leq s}, t) \leq \begin{cases} \binom{t}{\lceil t/2 \rceil} & \text{if } d = 1 \text{ and } t < 2s, \\ \binom{t}{s} & \text{if } d = 1 \text{ and } t \geq 2s, \\ \left(\frac{etd(d+2)}{4s}\right)^{\left\lceil \frac{s}{\lfloor d/2 \rfloor^2 + \lfloor d/2 \rfloor} \right\rceil} + \frac{d}{2} & \text{if } d \geq 2. \end{cases}$$

The following theorem establishes an upper bound on the maximum size of $\cup_{\leq s}$ $\bar{d}$-separable families on the ground set $[t]$.

Theorem 5 *Let $d \geq 1$, s and t, $s \leq t$, be positive integers. The maximum size of a $\cup_{\leq s}$ $\bar{d}$-separable family on the ground set $[t]$ is*

$$n_{sep}(d, \cup_{\leq s}, t) \leq \begin{cases} 2^{2s-1} & \text{if } d = 1 \text{ and } t < 2s, \\ 2^{s \log(et/s)} & \text{if } d = 1 \text{ and } t \geq 2s, \\ 2^{(t+1)/2} + 1 & \text{if } d = 2 \text{ and } t < 2s, \\ 2^{\frac{s}{2}\log\left(\frac{et}{s}\right) + \frac{1}{2}} + 1 & \text{if } d = 2 \text{ and } t \geq 2s, \\ \left(\frac{et(d^2-1)}{4s}\right)^{\left\lceil \frac{s}{\lfloor d-1/2 \rfloor^2 + \lfloor d-1/2 \rfloor} \right\rceil} + \frac{d-1}{2} & \text{if } d \geq 3. \end{cases}$$

Proof. The bounds for $d = 1$ follow from the fact that the members of a $\cup_{\le s}$ $\bar{d}$-separable family are pairwise distinct and have size at most s. As a consequence, it holds $n_{sep}(d, \cup_{\le s}, t) \le \sum_{i=1}^{s} \binom{t}{i}$. For $d = 1$ and $t < 2s$, we bound $\sum_{i=1}^{s} \binom{t}{i}$ by 2^t , thus obtaining $n_{sep}(d, \cup_{\le s}, t) \le 2^t \le 2^{2s-1}$. For $d = 1$ and $t \ge 2s$, we bound $\sum_{i=1}^{s} \binom{t}{i}$ by exploiting inequality (7) which implies $\sum_{i=1}^{s} \binom{t}{i} \le 2^{s \log(et/s)}$, and consequently, $n_{sep}(d, \cup_{\le s}, t) \le 2^{s \log(et/s)}$.

The bound for $d = 2$ and $t < 2s$ follows directly from Lindstorm's bound [15] which limits from above the size of $\bar{2}$-separable families on the ground set $[t]$ by $1 + 2^{(t+1)/2}$. For $d = 2$ and $t \ge 2s$, the stated bound follows from observing that the unions of any two members of a $\cup_{\le s}$ $\bar{2}$-separable family are distinct and have size smaller than or equal to s. Therefore, it must be $\binom{n_{sep}(d, \cup_{\le s}, t)}{2} \le \sum_{i=1}^{s} \binom{t}{i}$. Then, the stated bound for $d = 2$ and $t \ge 2s$ follows from inequality (7).

The bound for $d \ge 3$ follows immediately from Theorem 3 and from the upper bound stated by Corollary 1 for $d \ge 2$. $\square$

Lemma 1 and Theorem 5 imply the following lower bound on $\tilde{y}(d, n, t)$.

Theorem 6 *Let t, n, d be positive integers with $n \ge d \ge 1$. It holds*

$$\tilde{y}(d, n, t) \ge \max\{d\, , \, \beta\},$$

where

$$\beta \ge \begin{cases} \frac{\log(n+1)}{2} & \text{if } d = 1 \text{ and } \tilde{y}(d,n,t) > t/2, \\[2ex] \frac{\log n}{\log\left(\frac{et}{\tilde{y}(d,n,t)}\right)} \ge \frac{\log n}{\log\left(\frac{et\log(et)}{\log n}\right)} & \text{if } d = 1 \text{ and } \tilde{y}(d,n,t) \le t/2, \\[4ex] \log(n-1) & \text{if } d = 2 \text{ and } \tilde{y}(d,n,t) > t/2, \\[2ex] \frac{2\log(n-1)-1}{\log\left(\frac{et}{\tilde{y}(d,n,t)}\right)} \ge \frac{2\log(n-1)-1}{\log\left(\frac{et\log(et/2)}{2\log(n-1)-1}\right)} & \text{if } d = 2 \text{ and } \tilde{y}(d,n,t) \le t/2, \\[4ex] \left(\left\lfloor\frac{d-1}{2}\right\rfloor^2 + \left\lfloor\frac{d-1}{2}\right\rfloor\right)\left(\frac{\log(n-\frac{d}{2}+\frac{1}{2})}{\log\left(\frac{et(d^2-1)}{4\tilde{y}(d,n,t)}\right)} - 1\right) \ge \left(\left\lfloor\frac{d-1}{2}\right\rfloor^2 + \left\lfloor\frac{d-1}{2}\right\rfloor\right)\left(\frac{\log\left(n-\frac{d}{2}+\frac{1}{2}\right)}{\log \eta} - 1\right) & \text{if } d \ge 3, \end{cases}$$

with $\eta = \frac{e(d-1)^2}{2}$ *if* $\tilde{y}(d,n,t) > t/2$, *and* $\eta = \frac{2et\log(\frac{etd}{4})}{\log\left(n-\frac{d}{2}+\frac{1}{2}\right)-\log(\frac{etd}{4})}$ *if* $\tilde{y}(d,n,t) \le t/2$.

Proof. Lemma 1 implies that $\tilde{y}(d, n, t) \ge d$. The lower bounds on β follow from the corresponding upper bounds of Theorem 5 on the maximum size of a $\cup_{\le s}$ $\bar{d}$-separable family on the ground set $[t]$. The bounds holding for the case when $\tilde{y}(d, n, t) > t/2$ and $d \le 2$, as well as those on the lefthand sides for the remaining cases, are an immediate consequence of Theorem 5. For the case when $\tilde{y}(d, n, t) > t/2$ and $d \ge 3$, the bound on the right-hand side follows from the bound on the left-hand side by simply upper bounding

$\tilde{y}(d,n,t)$ by $\frac{t}{2}$. For $\tilde{y}(d,n,t) \le t/2$, the lower bounds on the right-hand sides are obtained as follows. Observe, that for $\tilde{y}(d,n,t) \le t/2$, the lower bounds on the left-hand sides are

$$\begin{cases} \frac{\log n}{\log\left(\frac{et}{\tilde{y}(d,n,t)}\right)} & \text{if } d = 1, \\ \frac{2\log(n-1)-1}{\log\left(\frac{et}{\tilde{y}(d,n,t)}\right)} & \text{if } d = 2, \\ \left(\left\lfloor\frac{d-1}{2}\right\rfloor^2 + \left\lfloor\frac{d-1}{2}\right\rfloor\right)\left(\frac{\log(n-\frac{d}{2}+\frac{1}{2})}{\log\left(\frac{et(d^2-1)}{4\tilde{y}(d,n,t)}\right)} - 1\right) & \text{if } d \ge 3. \end{cases} \tag{32}$$

By Lemma 1, it holds $\tilde{y}(d,n,t) \ge d$, and consequently, the above lower bounds are at least

$$\begin{cases} \frac{\log n}{\log(et)} & \text{if } d = 1, \\ \frac{2\log(n-1)-1}{\log\left(\frac{et}{2}\right)} & \text{if } d = 2, \\ \left(\left\lfloor\frac{d-1}{2}\right\rfloor^2 + \left\lfloor\frac{d-1}{2}\right\rfloor\right)\left(\frac{\log(n-\frac{d}{2}+\frac{1}{2})}{\log\left(\frac{et(d^2-1)}{4d}\right)} - 1\right) & \text{if } d \ge 3. \end{cases} \tag{33}$$

The lower bounds on the right-hand sides for the case $\tilde{y}(d,n,t) \le t/2$ are obtained by applying lower bounds (33) to $\tilde{y}(d,n,t)$ in lower bounds (32). In order to derive the bound for the case $\tilde{y}(d,n,t) \le t/2$ and $d \ge 3$, one needs also to observe that $\frac{d^2-1}{\left\lfloor\frac{d-1}{2}\right\rfloor^2+\left\lfloor\frac{d-1}{2}\right\rfloor} \le 8$.
□

4.2 Almost optimal $\cup_{\le s}$ (p,d)-cover free families

The following theorem proves the existence of $\cup_{\le s}$ (p,d)-cover free families with size very close to the upper bound implied by Theorem 4.

Theorem 7 *Let d and p be positive integers and let s and t be integers such that $s \le t$. There exists a $\cup_{\le s}$ (p,d)-cover free family on the ground set $[t]$ with size*

$$n \ge \begin{cases} \frac{1}{e}(p+d)2^{\left(\frac{p}{d(d+p)}\left(s-d\log\left(\frac{e(d+p)}{p}\right)-\frac{d}{p}\right)\right)} & \textit{if } t < 2s, \\ \frac{1}{e}(p+d)2^{\left(\frac{p}{d(d+p)}\left(s\log\left(\frac{et}{s}\right)-d\log\left(\frac{e(d+p)}{p}\right)-\frac{d}{p}\right)\right)} & \textit{if } t \ge 2s. \end{cases}$$

Proof. We will prove the theorem by the probabilistic method. In the following, we will conveniently represent a family $\mathcal{F}$ of n subsets of $[t]$ by the $t \times n$ binary matrix having as columns the characteristic vectors $\mathbf{c}_1, \dots, \mathbf{c}_n$ of the subsets belonging to $\mathcal{F}$, i.e., for each $i = 1, \dots, t$ and $j = 1, \dots, n$, the matrix has entry (i,j) set to 1 if and only if the member of $\mathcal{F}$ associated with the j-th column contains i. The number of 1-entries of a column $\mathbf{c}$ will be called the *weight* of $\mathbf{c}$. Given m columns $\mathbf{c}_{j_1}, \dots, \mathbf{c}_{j_m}$, we will denote by $\mathbf{c}_{j_1} \vee \dots \vee \mathbf{c}_{j_m}$ the Boolean *OR* of columns $\mathbf{c}_{j_1}, \dots, \mathbf{c}_{j_m}$.

Let us consider a $t \times n$ random binary matrix $\mathcal{M}$ where each entry is 0 with probability z and 1 with probability $1-z$, with $z = \left(1 - \left(\frac{s}{et}\right)^{s(\frac{p}{d}+1)}\right)^{\frac{1}{d}}$. In order for $\mathcal{M}$ to represent a $\cup_{\leq s}$ (p,d)-cover free family, it must hold that for any choice of d columns $\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}$ the following two events E_1 and E_2 occur.

E_1: The weight of $\mathbf{c}_{j_1} \vee \ldots \vee \mathbf{c}_{j_d}$ is at most s, i.e., there is a number a of rows, $a \leq s$, such that in correspondence of each of these a rows at least one of $\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}$ has an entry equal to 1, whereas in correspondence of the remaining $t-a$ rows, all entries of $\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}$ are equal to 0.

E_2: For any choice of p other columns $\mathbf{c}_{k_1}, \ldots, \mathbf{c}_{k_p}$, the column $\mathbf{c}_{j_1} \vee \ldots \vee \mathbf{c}_{j_d}$ does not cover the column $\mathbf{c}_{k_1} \vee \ldots \vee \mathbf{c}_{k_p}$, i.e., there exists a row index i such that at least one of $\mathbf{c}_{k_1}, \ldots, \mathbf{c}_{k_p}$ has the i-th entry equal to 1 whereas all columns $\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}$ have the i-th entry equal to 0.

We say that a set of d columns $\{\mathbf{c}_{j_1} \vee \ldots \vee \mathbf{c}_{j_d}\}$ is *good* if both events E_1 and E_2 occur. We will prove that the probability that $\mathcal{M}$ contains a set of d columns which is not good is smaller than 1, thus proving that $\mathcal{M}$ has a positive probability of representing a $\cup_{\leq s}$ (p,d)-cover free family.

For a given set of d columns $\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}$ of $\mathcal{M}$, we want to estimate probability

$$\begin{aligned} Pr\{\{\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}\} \text{ is good}\} &= Pr\{E_1 \cap E_2\} = Pr\{E_2|E_1\} Pr\{E_1\} \\ &= \left(1 - Pr\{\overline{E_2}|E_1\}\right) Pr\{E_1\}. \end{aligned} \tag{34}$$

Let us estimate the probability $Pr\{\overline{E_2}|E_1\}$. Notice that event E_1 implies that there are at most s entries equal to 1 in $\mathbf{c}_{j_1} \vee \ldots \vee \mathbf{c}_{j_d}$. Let $0 \leq a \leq s$ be an integer and let $i_1, \ldots, i_a$ be a row indices of $\mathcal{M}$. We denote by $E_{i_1,\ldots,i_a}$ the event that the vector $\mathbf{c}_{j_1} \vee \ldots \vee \mathbf{c}_{j_d}$ has all entries with indices in $\{i_1, \ldots, i_a\}$ equal to 1 and all other entries equal to 0. For the given set of row indices $\{i_1, \ldots, i_a\}$, let us estimate the probability $Pr\{\overline{E_2} \cap E_{i_1,\ldots,i_a}|E_1\}$.

$$\begin{aligned} & Pr\{\overline{E_2} \cap E_{i_1,\ldots,i_a}|E_1\} \\ = \; & Pr\{\overline{E_2}|E_{i_1,\ldots,i_a} \cap E_1\} \cdot Pr\{E_{i_1,\ldots,i_a}|E_1\} \\ = \; & Pr\left\{\exists\, \mathbf{c}_{k_1}, \ldots, \mathbf{c}_{k_p} \notin \{\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}\} \text{ such that } \mathbf{c}_{k_1} \vee \ldots \vee \mathbf{c}_{k_p} \text{ is covered by } \right. \\ & \left. \quad \mathbf{c}_{j_1} \vee \ldots \vee \mathbf{c}_{j_d} | E_{i_1,\ldots,i_a} \cap E_1\right\} \cdot Pr\{E_{i_1,\ldots,i_a}|E_1\} \\ = \; & Pr\left\{\exists\, \mathbf{c}_{k_1}, \ldots, \mathbf{c}_{k_p} \notin \{\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}\} \text{ such that } (\mathbf{c}_{k_1} \vee \ldots \vee \mathbf{c}_{k_p})(i) = 0, \right. \\ & \left. \quad \text{for all } i \in [t] \setminus \{i_1, \ldots, i_a\}\right\} \cdot Pr\{E_{i_1,\ldots,i_a}|E_1\} \qquad (35) \\ \leq \; & \left[\binom{n-d}{p} z^{p(t-a)}\right] \cdot \left[(1-z^d)^a z^{d(t-a)}\right]. \qquad (36) \end{aligned}$$

The second term in (36) has been obtained by observing that $Pr\{E_{i_1,\ldots,i_a}|E_1\} = Pr\{E_{i_1,\ldots,i_a}\}$.

Notice that for $\{i_1, \ldots i_a\} \neq \{i'_1, \ldots, i'_{a'}\}$, with $0 \leq a \leq s$ and $0 \leq a' \leq s$, it is $E_{i_1,\ldots i_a} \cap E_{i'_1,\ldots,i'_{a'}} = \emptyset$. By the law of total probability and upper bound (36), we have that

$$\begin{aligned} Pr\{\overline{E_2}|E_1\} &= \sum_{a=0}^{s} \sum_{(i_1,\ldots,i_a)\in[t]_a} Pr\{\overline{E_2} \cap E_{i_1,\ldots,i_a}|E_1\} \\ &\leq \sum_{a=0}^{s} \binom{t}{a} \left[\binom{n-d}{p} z^{(p+d)(t-a)}(1-z^d)^a\right] \\ &\leq \binom{n-d}{p}(1-z^d)z^{(p+d)(t-s)} \sum_{a=0}^{s} \binom{t}{a}. \end{aligned} \tag{37}$$

By upper bound (37) and by (34), we have that

$$Pr\{\{\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}\} \text{ is good}\} \geq \left(1 - \binom{n-d}{p}(1-z^d)z^{(p+d)(t-s)} \sum_{a=0}^{s} \binom{t}{a}\right) \cdot Pr\left\{E_1\right\}. \tag{38}$$

Now let us estimate $Pr\{E_1\}$, that is the probability that $\mathbf{c}_{j_1} \vee \ldots \vee \mathbf{c}_{j_d}$ has weight at most s. For a fixed row index i, the probability that $\mathbf{c}_{j_1} \vee \ldots \vee \mathbf{c}_{j_d}$ has the i-th entry equal to 1 is $(1 - z^d)$. For $i = 1, \ldots, t$, let X_i be the Bernoulli random variable which is 1 if and only if at least one of $\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}$ has the i-th entry equal to 1. Therefore, the random variable $\sum_{i=1}^{t} X_i$ has a binomial distribution with probability of success equal to $(1 - z^d)$. By Markov inequality, the probability that $\sum_{i=1}^{t} X_i > s$ is at most $\frac{E[\sum_{i=1}^{t} X_i]}{s+1} = \frac{t(1-z^d)}{s+1}$, thus implying that $Pr\{E_1\} \geq \left(1 - \frac{t(1-z^d)}{s+1}\right)$. It follows that

$$\begin{aligned} & Pr\{\{\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}\} \text{ is good}\} \\ \geq\ & \left(1 - \binom{n-d}{p}(1-z^d)z^{(p+d)(t-s)} \sum_{a=0}^{s} \binom{t}{a}\right) \left(1 - \frac{t(1-z^d)}{s+1}\right) \\ =\ & 1 - \frac{t(1-z^d)}{s+1} - \left(1 - \frac{t(1-z^d)}{s+1}\right)\binom{n-d}{p}(1-z^d)z^{(p+d)(t-s)} \sum_{a=0}^{s} \binom{t}{a} \\ \geq\ & 1 - \frac{t(1-z^d)}{s+1} - \binom{n-d}{p}(1-z^d) \sum_{a=0}^{s} \binom{t}{a}. \end{aligned} \tag{39}$$

Now we are ready to estimate the probability that $\mathcal{M}$ does not represent a $\cup_{\leq s}$ (p, d)-cover free family. Inequality (39) allows to upper bound the probability that a given set of d columns is not good. Therefore, we have that

$$\begin{aligned} & Pr\left\{\mathcal{M} \text{ does not represent a } \cup_{\leq s} (p,d)\text{-cover free family}\right\} \\ =\ & Pr\left\{\text{there exists a set } \{\mathbf{c}_{j_1}, \ldots, \mathbf{c}_{j_d}\} \text{ which is not good }\right\} \end{aligned}$$

$$\leq \binom{n}{d}\left(\frac{t(1-z^d)}{s+1}+\binom{n-d}{p}(1-z^d)\sum_{a=0}^{s}\binom{t}{a}\right). \tag{40}$$

By setting $z=\left(1-\left(\frac{s}{et}\right)^{s(\frac{p}{d}+1)}\right)^{\frac{1}{d}}$ in (40) we obtain that

$$\begin{aligned}
& Pr\left\{\mathcal{M} \text{ does not represent a } \cup_{\leq s} (p,d)\text{-cover free family}\right\} \\
\leq\ & \binom{n}{d}\frac{t}{s+1}\left(\frac{s}{et}\right)^{s(\frac{p}{d}+1)}+\binom{n}{d}\binom{n-d}{p}\left(\frac{s}{et}\right)^{s(\frac{p}{d}+1)}\sum_{a=0}^{s}\binom{t}{a} \\
<\ & 2\binom{n}{d}\binom{n-d}{p}\left(\frac{s}{et}\right)^{s(\frac{p}{d}+1)}\sum_{a=0}^{s}\binom{t}{a} \\
=\ & 2\binom{n}{d+p}\binom{d+p}{p}\left(\frac{s}{et}\right)^{s(\frac{p}{d}+1)}\sum_{a=0}^{s}\binom{t}{a}.
\end{aligned} \tag{41}$$

Let P denote $Pr\left\{\mathcal{M} \text{ does not represent a } \cup_{\leq s} (p,d)\text{-cover free family}\right\}$. By (41), we have that

$$P<2\binom{n}{d+p}\binom{d+p}{p}\left(\frac{s}{et}\right)^{s(\frac{p}{d}+1)}\sum_{a=0}^{s}\binom{t}{a}. \tag{42}$$

In order for a $\cup_{\leq s}$ (p,d)-cover free family of size n on the ground set $[t]$ to exist, it is sufficient that $P<1$.

We first consider the case $t\geq 2s$ and then the case $t<2s$.

For $t\geq 2s$, inequality (7) implies that $\sum_{a=0}^{s}\binom{t}{a}\leq 2^{s\log\frac{et}{s}}$, and consequently, by (42) we have that

$$P<2\binom{n}{d+p}\binom{d+p}{p}2^{-s\log\frac{et}{s}(\frac{p}{d}+1)}2^{s\log\frac{et}{s}}. \tag{43}$$

By (43), one has that $P<1$ holds if

$$\frac{sp}{d}\log(et/s)>\log\left(2\binom{n}{d+p}\binom{d+p}{p}\right).$$

Therefore, one has that $P<1$ if

$$s\geq\frac{d}{p}\cdot\frac{\log\left(2\binom{n}{d+p}\binom{d+p}{p}\right)}{\log\left(\frac{et}{s}\right)}. \tag{44}$$

By the upper bound in (4), we can limit from above the binomial coefficients in the right-hand side of (44), and obtain that there exists a $\cup_{\leq s}$ (p,d)-cover free family of size n on the ground set $[t]$ if

$$s\geq\frac{\frac{d(d+p)}{p}\log\left(\frac{en}{d+p}\right)+d\log\left(\frac{e(d+p)}{p}\right)+\frac{d}{p}}{\log\left(\frac{et}{s}\right)},$$

which is satisfied for any $n \leq \frac{1}{e}(p+d)2^{\left(\frac{p}{d(d+p)}\left(s\log\left(\frac{et}{s}\right)-d\log\left(\frac{e(d+p)}{p}\right)-\frac{d}{p}\right)\right)}$. Therefore, we have that, for $t \geq 2s$, there exists a $\cup_{\leq s}$ (p,d)-cover free family of size n on the ground set $[t]$ that satisfies the second bound in the statement of the theorem.

Now, let us consider the case $t < 2s$. In this case, we observe that $\left(\frac{s}{et}\right)^{s(\frac{p}{d}+1)}$ decreases with s and therefore, we can limit it from above by $\left(\frac{1}{2e}\right)^{\frac{t}{2}(\frac{p}{d}+1)}$ in the right-hand side of (42). Moreover, we upper bound $\sum_{a=0}^{s}\binom{t}{a}$ by 2^t. Consequently, one has

$$\begin{aligned}
P &< 2\binom{n}{d+p}\binom{d+p}{p}2^t\left(\frac{1}{2e}\right)^{\frac{t}{2}(\frac{p}{d}+1)} \\
&< 2\binom{n}{d+p}\binom{d+p}{p}2^t\left(\frac{1}{2}\right)^{t(\frac{p}{d}+1)} \\
&= 2\binom{n}{d+p}\binom{d+p}{p}\left(\frac{1}{2}\right)^{t(\frac{p}{d})} \\
&\leq 2\binom{n}{d+p}\binom{d+p}{p}\left(\frac{1}{2}\right)^{s(\frac{p}{d})}, \qquad (45)
\end{aligned}$$

where the last inequality follows from s being smaller than or equal to t.

Therefore, one has that $P < 1$ if

$$s \geq \frac{d}{p}\cdot\log\left(2\binom{n}{d+p}\binom{d+p}{p}\right). \qquad (46)$$

By the upper bound in (4), we can limit from above the binomial coefficients in the right-hand side of (46), and obtain that there exists a $\cup_{\leq s}$ (p,d)-cover free family of size n on the ground set $[t]$ if

$$s \geq \frac{d(d+p)}{p}\log\left(\frac{en}{d+p}\right)+d\log\left(\frac{e(d+p)}{p}\right)+\frac{d}{p},$$

which is satisfied for any $n \leq \frac{1}{e}(p+d)2^{\left(\frac{p}{d(d+p)}\left(s-d\log\left(\frac{e(d+p)}{p}\right)-\frac{d}{p}\right)\right)}$. It follows that, for $t < 2s$, there exists a $\cup_{\leq s}$ (p,d)-cover free family of size n on the ground set $[t]$ that satisfies the first bound in the statement of the theorem. ∎

In the following, we compare the lower bounds of Theorem 7 with the upper bounds of Theorem 4. In fact, we will estimate the gap between the upper and lower bounds on $\log(n_{cf}(p,d,\cup_{\leq s},t))$, thus showing that this gap is not larger than that existing between the best upper and lower bounds on the logarithm of the maximum size of classical (p,d)-cover free families. For the case $t \geq 2s$, Theorem 7 implies an $\Omega\left(\frac{sp}{d(p+d)}\log\left(\frac{et}{s}\right)\right)$ lower bound on $\log(n_{cf}(p,d,\cup_{\leq s},t))$. Theorem 4 implies that $\log(n_{cf}(p,d,\cup_{\leq s},t))$ is upper bounded by $O\left(\frac{s}{d}\log\left(\frac{et}{s}\right)\right)$ for $d < 2p$, and by $O\left(\frac{sp}{d^2}\log\left(\frac{etd^2}{4ps}\right)\right)$ for $d \geq 2p$. Therefore, for $d < 2p$, the gap between the upper and lower bounds on $\log(n_{cf}(p,d,\cup_{\leq s},t))$ is $O\left(\frac{\frac{s}{d}\log\left(\frac{et}{s}\right)}{\frac{sp}{d(p+d)}\log\left(\frac{et}{s}\right)}\right) =$

$O\left(\frac{\frac{s}{d}\log\left(\frac{et}{s}\right)}{\frac{s}{d}\log\left(\frac{et}{s}\right)}\right) = O(1)$. For $d \geq 2p$, the gap is limited from above by

$$O\left(\frac{\frac{sp}{d^2}\log\left(\frac{etd^2}{4ps}\right)}{\frac{sp}{d(p+d)}\log\left(\frac{et}{s}\right)}\right) = O\left(\frac{\frac{sp}{d^2}\log\left(\frac{etd^2}{4ps}\right)}{\frac{sp}{d^2}\log\left(\frac{et}{s}\right)}\right) = O\left(1+\frac{\log\left(\frac{d^2}{4p}\right)}{\log\left(\frac{et}{s}\right)}\right).$$

Interestingly, the above bound decreases as the ratio between the size of the ground set t and the bound on the number of elements in the union of any d members of the family increases. If we set $s = t$ in the above bound, we obtain the same asymptotic gap existing between the best upper and lower bounds on the logarithm of the maximum size of classical (p, d)-cover free families.

For the case $t < 2s$, Theorem 7 implies that $\log(n_{cf}(p, d, \cup_{\leq s}, t))$ is $\Omega\left(\frac{sp}{d(d+p)}\right)$. Theorem 4 implies that $\log(n_{cf}(p, d, \cup_{\leq s}, t))$ is upper bounded by $O\left(\frac{s}{d}\right)$ for $d < 2p$, and by $O\left(\frac{sp}{d^2}\log\left(\frac{etd^2}{4ps}\right)\right) = O\left(\frac{sp}{d^2}\log\left(\frac{d^2}{p}\right)\right)$ for $d \geq 2p$. For $d < 2p$, one has $\Omega\left(\frac{sp}{d(d+p)}\right) = \Omega\left(\frac{s}{d}\right)$, and consequently, the lower bound on $\log(n_{cf}(p, d, \cup_{\leq s}, t))$ asymptotically matches the upper bound. For $d \geq 2p$, one has $\Omega\left(\frac{sp}{d(d+p)}\right) = \Omega\left(\frac{sp}{d^2}\right)$ and the ratio between the upper and lower bounds on $\log(n_{cf}(p, d, \cup_{\leq s}, t))$ is

$$O\left(\frac{\frac{sp}{d^2}\log\left(\frac{d^2}{p}\right)}{\frac{sp}{d^2}}\right) = O\left(\log\left(\frac{d^2}{p}\right)\right),$$

which is the same gap existing between the best upper and lower bounds on the logarithm of the maximum size of classical (p, d)-cover free families.

By setting $p = 1$ in the bound of Theorem 7, we obtain the following lower bound on the maximum size of $\cup_{\leq s}$ d-cover free families on the ground set $[t]$.

Theorem 8 *Let d be a positive integer and let s and t be integers such that $s \leq t$. There exists a $\cup_{\leq s}$ d-cover free family on the ground set $[t]$ with size*

$$n \geq \begin{cases} \frac{1}{e}(d+1)2^{\left(\frac{1}{d(d+1)}(s-d\log(e(d+1))-d)\right)} & \text{if } t < 2s, \\ \frac{1}{e}(d+1)2^{\left(\frac{1}{d(d+1)}\left(s\log\left(\frac{et}{s}\right)-d\log(e(d+1))-d\right)\right)} & \text{if } t \geq 2s. \end{cases}$$

The above theorem implies the following upper bound on the number of "yes" responses admitted by a non-adaptive group testing algorithm that uses at most t tests.

Theorem 9 *Let t, n, d be positive integers with $d \geq 1$ and $n \geq d$. There exists a non-adaptive group testing strategy $\mathcal{A}$ for which $\tilde{y}_{\mathcal{A}}(d, n, t)$ is at most*

$$\begin{cases} d(d+1)\log\left(\frac{en}{d+1}\right) + d\log\left(e(d+1)\right) + d & \text{if } t < 2\tilde{y}_{\mathcal{A}}(d,n,t), \\ \frac{d(d+1)}{\log\left(\frac{et}{\tilde{y}_{\mathcal{A}}(d,n,t)}\right)}\left(\log\left(\frac{en}{d+1}\right) + \frac{\log(e(d+1))+1}{d+1}\right) \leq \frac{d(d+1)}{\log\mu}\left(\log\left(\frac{en}{d+1}\right) + \frac{\log(e(d+1))+1}{d+1}\right) & \text{if } t \geq 2\tilde{y}_{\mathcal{A}}(d,n,t), \end{cases}$$

where $\mu = \log\left(\frac{et\log(2e)}{d(d+1)\left(\log\left(\frac{en}{d+1}\right)+\log(2\sqrt{e})\right)}\right)$.

Proof. The upper bound for $t < 2\tilde{y}_{\mathcal{A}}(d,n,t)$ follows immediately from the lower bound in Theorem 8. For $t \geq 2\tilde{y}_{\mathcal{A}}(d,n,t)$, Theorem 8 implies

$$\tilde{y}_{\mathcal{A}}(d,n,t) \leq \frac{d(d+1)}{\log\left(\frac{et}{\tilde{y}_{\mathcal{A}}(d,n,t)}\right)}\left(\log\left(\frac{en}{d+1}\right)+\frac{\log(e(d+1))+1}{d+1}\right). \tag{47}$$

Since $\frac{\log(e(d+1))+1}{d+1}$ decreases with d, we upper bound it by $\log(2\sqrt{e})$ in (47) and obtain

$$\tilde{y}_{\mathcal{A}}(d,n,t) \leq \frac{d(d+1)}{\log\left(\frac{et}{\tilde{y}_{\mathcal{A}}(d,n,t)}\right)}\left(\log\left(\frac{en}{d+1}\right)+\log(2\sqrt{e})\right). \tag{48}$$

In order to derive an upper bound on $\tilde{y}_{\mathcal{A}}(d,n,t)$, expressed in terms of d, n, and t only, we first exploit upper bound (48) to limit from above $\tilde{y}_{\mathcal{A}}(d,n,t)$ in upper bound (47), thus obtaining

$$\tilde{y}_{\mathcal{A}}(d,n,t) \leq \frac{d(d+1)\left(\log\left(\frac{en}{d+1}\right)+\frac{\log(e(d+1))+1}{d+1}\right)}{\log\left(\frac{et\log\left(\frac{et}{\tilde{y}_{\mathcal{A}}(d,n,t)}\right)}{d(d+1)\left(\log\left(\frac{en}{d+1}\right)+\log(2\sqrt{e})\right)}\right)}. \tag{49}$$

Then, we upper bound $\tilde{y}_{\mathcal{A}}(d,n,t)$ in (49) by $t/2$ thus obtaining $\log\left(\frac{et}{\tilde{y}_{\mathcal{A}}(d,n,t)}\right) \geq \log(2e)$, and consequently, the upper bound that appears on the right-hand side of case $t \geq 2\tilde{y}_{\mathcal{A}}(d,n,t)$. ▯

The result of Theorem 7 will be exploited in Section 5 to prove the existence of a trivial two-stage algorithm that admits the same number of positive responses of the best adaptive procedures.

4.3 An almost optimal explicit non-adaptive algorithm

In this section we present another non-adaptive algorithm that gets very close to the lower bound of Theorem 6.

We remark that this result translates into a lower bound on the size of $\cup_{\leq s}$ d-cover free families which is very close to the upper bound of Corollary 1. The underlying combinatorial structures of the algorithm consist of families in which any two members share at most a certain number λ of elements. The following simple lemma will be used in the analysis of both algorithms.

Lemma 3 *Let d and λ be two positive integers and let $\mathcal{F}$ be a family of sets with $|\mathcal{F}| \geq d$ and such that any two members $F_1, F_2 \in \mathcal{F}$ intersect in at most λ elements. Then, for any d members $F_1, \ldots, F_d$ of $\mathcal{F}$, it holds $|\bigcup_{i=1}^{d} F_i| \geq \sum_{i=1}^{d} |F_i| - \frac{1}{2}d(d-1)\lambda$.*

Proof. Observe that

$$|\bigcup_{i=1}^{d} F_i| \geq |\bigcup_{i=1}^{d}(F_i \setminus \bigcup_{j=1}^{i-1}(F_i \cap F_j))|. \tag{50}$$

Since for $i \neq \ell$, it holds $\left(F_i \setminus \bigcup_{j=1}^{i-1}(F_i \cap F_j)\right) \cap \left(F_\ell \setminus \bigcup_{j=1}^{\ell-1}(F_\ell \cap F_j)\right) = \emptyset$, one has that the right-hand side of (50) is equal to

$$\sum_{i=1}^{d} |F_i \setminus \bigcup_{j=1}^{i-1}(F_i \cap F_j)| \tag{51}$$

Notice that for any two sets A and B, one has that $|A \setminus B| \geq |A| - |B|$, with equality holding if and only if $B \subseteq A$. Therefore, it holds $|F_i \setminus \bigcup_{j=1}^{i-1}(F_i \cap F_j)| \geq |F_i| - |\bigcup_{j=1}^{i-1}(F_i \cap F_j)|$, and consequently, expression (51) is larger than or equal to

$$\sum_{i=1}^{d} |F_i| - |\bigcup_{j=1}^{i-1}(F_i \cap F_j)| = \sum_{i=1}^{d} |F_i| - \sum_{i=1}^{d} |\bigcup_{j=1}^{i-1}(F_i \cap F_j)| \geq \sum_{i=1}^{d} |F_i| - \sum_{i=1}^{d} \sum_{j=1}^{i-1} |F_i \cap F_j|.$$

Since $\sum_{i=1}^{d} \sum_{j=1}^{i-1} |F_i \cap F_j| \leq \sum_{i=1}^{d} \sum_{j=1}^{i-1} \lambda = \binom{d}{2}\lambda$, the lemma follows. $\square$

An interesting feature of the construction presented in this section is that it is an explicit construction. It is based on a breakthrough result by Porat and Rothschild [28] which provides the first deterministic explicit construction of error correcting codes meeting the Gilbert-Varshamov bound. In fact, the result in [28] provides a construction for $[m, k, \delta m]_q$-linear codes. We recall that an $[m, k, \delta m]_q$-linear code is a q-ary code over the alphabet $\mathbb{F}_q$ with length m, size $n = q^k$ and Hamming distance equal to δm. In the following, we denote by $H_q(p)$ the q-ary entropy function

$$H_q(p) = p \log_q \frac{q-1}{p} + (1-p) \log_q \frac{1}{1-p},$$

which, with respect to the Hamming distance over q-ary alphabets plays a role analogue to that played by binary entropy with respect to the binary alphabet. Porat and Rothschild proved the following

Theorem 10 *[28] Let q be a prime power, m and k positive integers, and $\delta \in [0, 1]$. If $k \leq (1 - H_q(\delta))m$, then it is possible to construct an $[m, k, \delta m]_q$-linear code in time $\Theta(mq^k)$.*

In [28], Porat and Rothschild show how to construct an (n, r)-strongly selective family [6] from a linear code with properly chosen parameters and then exploit the above mentioned theorem to construct in time $\Theta(rn \ln n)$ a linear code that can be reduced to an (n, r)-*strongly selective* family of size $\Theta(r^2 \ln n)$. We just mention that an (n, r)-strongly selective family is a combinatorial structure which is essentially equivalent to an $(r-1)$-cover free family. The following theorem rephrases the result in [28] in terms of cover free families.

Theorem 11 *If there exists an $[m, k, \delta m]_q$-linear code then it is possible to construct an m-uniform $(\lceil \frac{1}{1-\delta} \rceil - 1)$-cover free family of size $n = q^k$ on the ground set $[mq]$, with the property that any two members of the family intersect in at most $m - \delta m$ elements.*

Proof. Given an $[m,k,\delta m]_q$-linear code $\mathcal{C} = \{\mathbf{c}_1,\ldots,\mathbf{c}_n\}$, let us define the family $\mathcal{F}$ as $\mathcal{F} = \{F(\mathbf{c}_1),\ldots,F(\mathbf{c}_n)\}$, where $F(\mathbf{c}_j) = \{f(i,a) : (i,a) \in [m]\times[q],\ \mathbf{c}_j[i] = a\}$, with f being an injection from $[m]\times[q]$ to $[mq]$. It is immediate to see that $\mathcal{F}$ is m-uniform in that for each index $i \in [m]$ there is a unique pair $(i,a) \in [m]\times[q]$ such that $\mathbf{c}_j[i] = a$. Moreover, any two members of $\mathcal{F}$ intersect in at most $m-\delta m$ elements. Indeed, for any two distinct words $\mathbf{c}_j, \mathbf{c}_\ell \in \mathcal{C}$ there are at least δm indices $i \in [m]$ such that $\mathbf{c}_\ell(i) \neq \mathbf{c}_j(i)$. This implies that there are at least δm pairs $(i,a) \in [m]\times[q]$ such that $f(i,a) \in F(\mathbf{c}_j)$ and $f(i,a) \notin F(\mathbf{c}_\ell)$, and consequently, $F(\mathbf{c}_j)$ and $F(\mathbf{c}_\ell)$ share at most $m-\delta m$ elements. It follows that the union of any $\lceil \frac{1}{1-\delta} \rceil - 1$ members of $\mathcal{F}$ shares at most $(\lceil \frac{1}{1-\delta} \rceil - 1)(m-\delta m) \leq m-1$ elements with any other member of the family, implying that $\mathcal{F}$ is $(\lceil \frac{1}{1-\delta} \rceil - 1)$-cover free. $\square$

Theorem 12 *Let t, n, d be positive integers with $n \geq d \geq 1$. There exists a non-adaptive group testing strategy $\mathcal{A}$ for which*

$$\tilde{y}_{\mathcal{A}}(d,n,t) = \begin{cases} \Theta(d^2 \ln n) & \text{if } \tilde{y}_{\mathcal{A}}(d,n,t) \geq \frac{(t+1)d}{2(d+1)}, \\ \Theta\left(\frac{d^2 \ln n}{\ln(\frac{t}{\tilde{y}_{\mathcal{A}}(d,n,t)})}\right) = O\left(\frac{d^2 \ln n}{\ln\left(\frac{t}{d^2 \ln n}\right)}\right) & \text{if } \tilde{y}_{\mathcal{A}}(d,n,t) < \frac{(t+1)d}{2(d+1)} \end{cases}.$$

The underlying family can be constructed in time $\Theta(dn \ln n)$ if $\tilde{y}_{\mathcal{A}}(d,n,t) \geq \frac{(t+1)d}{2(d+1)}$, and in time $\Theta\left(\frac{dn \ln n}{\ln(\frac{t}{\tilde{y}_{\mathcal{A}}(d,n,t)})}\right) = O\left(\frac{dn \ln n}{\ln(\frac{t}{d^2 \ln n})}\right)$ otherwise.

Proof. For $\tilde{y}_{\mathcal{A}}(d,n,t) \geq \frac{(t+1)d}{2(d+1)}$, the stated bound follows from Theorem 1 of [28] which implies that there exists a non-adaptive group testing algorithm that uses $t = \Theta(d^2 \ln n)$ and is such that the underlying family can be constructed in time $\Theta(dn \ln n)$. Since in the case we are considering it is $\tilde{y}_{\mathcal{A}}(d,n,t) = \Theta(t)$, we have $\tilde{y}_{\mathcal{A}}(d,n,t) = \Theta(d^2 \ln n)$.

Let us consider the case when $\tilde{y}_{\mathcal{A}}(d,n,t) < \frac{(t+1)d}{2(d+1)}$. By Theorem 10 it is possible to construct an $[m,k,\delta m]_q$ linear code in time $\Theta(mq^k)$, where q is a prime power, m a positive integer, $\delta \in [0,1]$ and $k = (1 - H_q(\delta))m$. Theorem 11 then implies that such a code can be transformed into an m-uniform $(\lceil \frac{1}{1-\delta} \rceil - 1)$-cover free family $\mathcal{F}$ of size $n = q^k$ on the ground set $[mq]$. Let us set $\delta = \frac{d}{d+1}$, and let $q \geq 2d+2$. It holds

$$\begin{aligned} 1 - H_q(\delta) &= 1 - \left[\frac{d}{d+1} \log_q\left(\frac{(d+1)(q-1)}{d}\right) + \frac{1}{d+1}\log_q(d+1)\right] \\ &= \frac{1}{(d+1)\ln q}\left[(d+1)\ln q - d\ln\left(\frac{(d+1)(q-1)}{d}\right) - \ln(d+1)\right] \\ &= \frac{1}{(d+1)\ln q}\left[d\ln q + \ln q - d\ln\left(\frac{d+1}{d}\right) - d\ln(q-1) - \ln(d+1)\right] \\ &= \frac{1}{(d+1)\ln q}\left[d\ln\left(\frac{q}{q-1}\right) - d\ln\left(\frac{d+1}{d}\right) + \ln\left(\frac{q}{d+1}\right)\right]. \end{aligned} \tag{52}$$

We can exploit the well known relation $\ln \frac{z}{z-1} = \frac{1}{z} + o(\frac{1}{z})$, to estimate (52). Therefore, we get

$$1 - H_q(\delta) = \frac{1}{(d+1)\ln q}\left[\frac{d}{q} - \frac{d}{d+1} + \ln\left(\frac{q}{d+1}\right)\right] + o\left(\frac{1}{(d+1)\ln q}\right). \qquad (53)$$

We will prove that

$$c \cdot \ln\left(\frac{q}{d+1}\right) \leq \left[\frac{d}{q} - \frac{d}{d+1} + \ln\left(\frac{q}{d+1}\right)\right] < \ln\left(\frac{q}{d+1}\right),$$

for any constant $c \leq 1/6$. Indeed, we are assuming $q \geq 2d+2$ and therefore, we have that

$$\frac{d}{q} - \frac{d}{d+1} + \ln\left(\frac{q}{d+1}\right) \leq \frac{d}{2d+2} - \frac{d}{d+1} + \ln\left(\frac{q}{d+1}\right) < \ln\left(\frac{q}{d+1}\right).$$

Now, let us prove that

$$\frac{d}{q} - \frac{d}{d+1} + \ln\left(\frac{q}{d+1}\right) \geq c\ln\left(\frac{q}{d+1}\right), \qquad (54)$$

for any positive constant $c \leq \frac{1}{6}$. Notice that inequality (54) holds if and only if

$$1 - c \geq \frac{\frac{d}{d+1} - \frac{d}{q}}{\ln\left(\frac{q}{d+1}\right)}. \qquad (55)$$

Since $q \geq 2d+2$, the right-hand side of inequality (55) is smaller than

$$\frac{\frac{q-d-1}{q}}{\ln\left(\frac{q}{d+1}\right)} = \frac{\frac{q-d-1}{q}}{-\ln\left(1 - \frac{q-d-1}{q}\right)} = \frac{1}{-\ln\left(1 - \frac{q-d-1}{q}\right)^{\frac{q}{q-d-1}}} \leq \frac{1}{2\ln 2},$$

where the last inequality follows from setting $f = \frac{q}{q-d-1}$ and observing that $-f\ln(1-1/f)$ decreases with f. Since $q \geq 2d+2$ implies $f \leq 2$, it holds $-f\ln(1-1/f) \geq 2\ln 2$. Therefore, one has that inequality (55) holds for any c such that $1 - c \geq \frac{1}{2\ln 2}$. Since $1 - \frac{1}{2\ln 2} \geq 1 - \frac{1}{1.2} = \frac{1}{6}$, it follows that inequality (54) holds for any $c \leq \frac{1}{6}$. Therefore,

$$1 - H_q(\delta) = \Theta\left(\frac{1}{(d+1)\ln q}\ln\left(\frac{q}{d+1}\right)\right). \qquad (56)$$

It follows that

$$\log_q n = k = m(1 - H_q(\delta)) = \Theta\left(\frac{m}{(d+1)\ln q}\ln\left(\frac{q}{d+1}\right)\right). \qquad (57)$$

By setting $s = dm$ and $t = mq$ in (57), we get

$$\ln n = \log_q n \ln q = \Theta\left(\frac{s}{d(d+1)}\ln\left(\frac{td}{s(d+1)}\right)\right) = \Theta\left(\frac{s}{d^2}\ln\left(\frac{t}{s}\right)\right). \qquad (58)$$

The maximum number $\tilde{y}_{\mathcal{A}}(d,n,t)$ of positive responses admitted by the algorithm is equal to the maximum number of elements contained in the union of d members of the family. Since s is an upper bound on the size of the union of any d members of the family, one has that $\tilde{y}_{\mathcal{A}}(d,n,t) \leq s$. By Theorem 11, any two members of the family intersect in at most $m - \delta m$ elements. Hence, Lemma 3 implies $|\bigcup_{i=1}^{d} F_{j_i}| \geq md - \frac{1}{2}(m-\delta m)d(d-1) = s - s(d-1)/(2d+2) \geq s/2$, for any d members $F_{j_1},\ldots,F_{j_d}$ of the family. Therefore, it holds $s/2 \leq \tilde{y}_{\mathcal{A}}(d,n,t) \leq s$, from which the first bound for $\tilde{y}_{\mathcal{A}}(d,n,t) < \frac{(t+1)d}{2(d+1)}$ in the statement of the theorem follows.

In order to obtain the bound expressed only in terms of d, t and n, we apply recursively the first bound to limit $\tilde{y}_{\mathcal{A}}(d,n,t)$ in its expression, thus obtaining

$$\tilde{y}_{\mathcal{A}}(d,n,t) = \Theta\left(\frac{d^2 \ln n}{\ln\left(\frac{t}{d^2 \ln n} \ln\left(\frac{t}{\tilde{y}_{\mathcal{A}}(d,n,t)}\right)\right)}\right). \tag{59}$$

Since $\tilde{y}_{\mathcal{A}}(d,n,t) \leq \frac{(t+1)d}{2(d+1)}$, we have that the right-hand side of (59) is $O\left(\frac{d^2 \ln n}{\log\left(\frac{t}{d^2 \log n}\right)}\right)$, thus obtaining the second bound in the statement of the theorem.

The time needed to construct the family is $\Theta(q^k m) = \Theta\left(n\frac{\tilde{y}_{\mathcal{A}}(d,n,t)}{d}\right)$. By applying the bound $\tilde{y}_{\mathcal{A}}(d,n,t) = \Theta\left(\frac{d^2 \ln n}{\ln(\frac{t}{\tilde{y}_{\mathcal{A}}(d,n,t)})}\right)$, we obtain $\Theta\left(n\frac{\tilde{y}_{\mathcal{A}}(d,n,t)}{d}\right) = \Theta\left(\frac{dn \ln n}{\ln(\frac{t}{\tilde{y}_{\mathcal{A}}(d,n,t)})}\right)$, whereas by applying the right-hand side bound $\tilde{y}_{\mathcal{A}}(d,n,t) = O\left(\frac{d^2 \ln n}{\ln\left(\frac{t}{d^2 \ln n}\right)}\right)$, we obtain $\Theta\left(n\frac{\tilde{y}_{\mathcal{A}}(d,n,t)}{d}\right) = O\left(\frac{dn \ln n}{\ln(\frac{t}{d^2 \ln n})}\right)$. $\square$

5 Optimal two-stage group testing

We consider trivial two-stage algorithms, i.e., algorithms that consist of two non-adaptive stages, with the first stage performing parallel tests on pools of elements, and the second stage performing individual tests on certain selected elements. More precisely, in the first stage a non-adaptive group testing algorithm is used to determine a "small" number of potential defective elements, i.e., a subset of elements that contains all defectives; in the second stage the subset of elements selected by the first stage are individually tested so as to find those that are really defective. In this section we give a trivial two-stage algorithm that admits the same maximum number of "yes" responses as the optimal adaptive algorithm, thus showing that by allowing just a little adaptiveness, one can an achieve the same performance as the best adaptive algorithms.

In the following, given a trivial two-stage algorithm $\mathcal{A}$ that finds up to d defective elements in an input set of size n by at most t tests, we denote by $\hat{y}_{\mathcal{A}}(n,d,t)$ the maximum number of positive responses that may occur during the search process performed by $\mathcal{A}$, where the maximum is taken over all possible subsets of up to d defectives. Moreover, we denote by $\hat{y}(n,d,t)$ the minimum value of $\hat{y}_{\mathcal{A}}(n,d,t)$ over all trivial two-stage strategies $\mathcal{A}$ that find up to d defective elements in an input set of size n by at most t tests.

As observed in Section 3.1, a $\cup_{\leq s}$ (p,d)-cover free family can be used to design a non-adaptive algorithm that selects a subset of up to $p+d-1$ elements containing all defective elements and admits at most s "yes" responses. Therefore, such an algorithm can be employed in the first stage of a trivial two-stage algorithm to select the elements that will undergo individual tests during the second stage. Notice that the total number of positive responses admitted by the two-stage algorithm is at most $s+d$, since at most d individual probes yield a positive response in the second stage.

The following theorem follows from the above discussion.

Theorem 13 *Let t, n, d, p be positive integers with $t \geq d+p$ and $d+p \leq n \leq n_{cf}(p,d,\cup_{\leq s},t-d-p+1)$. There exists a two-stage group testing strategy $\mathcal{A}$ for which*

$$\hat{y}_{\mathcal{A}}(d,n,t) \leq s+d.$$

The following theorem is a consequence of Theorem 13 and Theorem 7.

Theorem 14 *Let t, n, d, p be positive integers with $t \geq d+p$ and $n \geq d+p$. There exists a two-stage group testing strategy $\mathcal{A}$ for which $\hat{y}_{\mathcal{A}}(d,n,t)$ is at most*

$$\begin{cases} \frac{d(d+p)}{p}\log\left(\frac{en}{d+p}\right)+d\log\left(\frac{e(d+p)}{p}\right)+\frac{d}{p}+d & \text{if } \hat{y}_{\mathcal{A}}(d,n,t) > (t+d-p+1)/2, \\ \frac{\frac{d(d+p)}{p}\log\left(\frac{en}{d+p}\right)+d\log\left(\frac{e(d+p)}{p}\right)+\frac{d}{p}}{\log\left(\frac{e(t-d-p+1)}{\hat{y}_{\mathcal{A}}(d,n,t)-d}\right)}+d \leq \frac{\frac{d(d+p)}{p}\log\left(\frac{en}{d+p}\right)+d\log\left(\frac{e(d+p)}{p}\right)+\frac{d}{p}}{\log\chi}+d & \text{if } \hat{y}_{\mathcal{A}}(d,n,t) \leq (t+d-p+1)/2, \end{cases}$$

where $\chi = \frac{e(t-d-p+1)\log(2e)}{\frac{d(d+p)}{p}\left(\log\left(\frac{en}{d+p}\right)+\log(e\sqrt{2})\right)}$.

Proof. The two-stage algorithm consists in a first stage in which the pools corresponding to the rows of a $\cup_{\leq s}$ (p,d)-cover free family are tested in parallel, and in a second stage that performs individual probes on the up to $d+p-1$ elements selected by the first stage. The bound in the statement of the theorem follows from the lower bound of Theorem 7 on the maximum size of a $\cup_{\leq s}$ (p,d)-cover free family on the ground set $[t-d-p+1]$. The lower bound of Theorem 7 implies that the number of positive responses in the first stage is

$$s \leq \begin{cases} \frac{d(d+p)}{p}\log\left(\frac{en}{d+p}\right)+d\log\left(\frac{e(d+p)}{p}\right)+\frac{d}{p} & \text{if } s > (t-d-p+1)/2, \\ \frac{\frac{d(d+p)}{p}\left(\log\left(\frac{en}{d+p}\right)+\frac{p}{d+p}\log\left(\frac{e(d+p)}{p}\right)+\frac{1}{d+p}\right)}{\log\left(\frac{e(t-d-p+1)}{s}\right)} & \text{if } s \leq (t-d-p+1)/2. \end{cases}$$

Since up to d individual probes yield a positive response in the second stage, we set $s = \hat{y}_{\mathcal{A}}(d,n,t)-d$ so that the algorithm is guaranteed to receive no more that $\hat{y}_{\mathcal{A}}(d,n,t)$ "yes" responses in total. By setting $s = \hat{y}_{\mathcal{A}}(d,n,t)-d$ in the above bounds, we get the bound for $\hat{y}_{\mathcal{A}}(d,n,t) > (t+d-p+1)/2$ in the statement of the theorem and the first of the two bounds stated for $\hat{y}_{\mathcal{A}}(d,n,t) \le (t+d-p+1)/2$. In order to obtain the second bound for $\hat{y}_{\mathcal{A}}(d,n,t) \le (t+d-p+1)/2$, we first observe that $\frac{p}{d+p}\log\left(\frac{e(d+p)}{p}\right)+\frac{1}{d+p}$ decreases with d and consequently is smaller than $\frac{p}{1+p}\log\left(\frac{e(1+p)}{p}\right)+\frac{1}{1+p} \le \log(e\sqrt{2})$. Therefore, we have that

$$\hat{y}_A(d,n,t) \le \frac{\frac{d(d+p)}{p}\left(\log\left(\frac{en}{d+p}\right)+\log(e\sqrt{2})\right)}{\log\left(\frac{e(t-d-p+1)}{\hat{y}_A(d,n,t)-d}\right)} + d.$$

Then, we bound $\hat{y}_{\mathcal{A}}(d,n,t)$ by $(t+d-p+1)/2$ in the above upper bound, thus obtaining

$$\hat{y}_A(d,n,t) \le \frac{\frac{d(d+p)}{p}\left(\log\left(\frac{en}{d+p}\right)+\log(e\sqrt{2})\right)}{\log(2e)} + d. \tag{60}$$

We exploit upper bound (60) to limit from above $\hat{y}_{\mathcal{A}}(d,n,t)$ in the first of the two bounds stated for $\hat{y}_{\mathcal{A}}(d,n,t) \le (t+d-p+1)/2$, thus getting the second bound for $\hat{y}_{\mathcal{A}}(d,n,t) \le (t+d-p+1)/2$ in the statement of the theorem. $\square$

By setting $p = d$ in the bound of Theorem 14, we obtain the following corollary that states the existence of a trivial two-stage algorithm which asymptotically attains the same bound of the optimal adaptive algorithm.

Corollary 2 *Let t, n, d be positive integers with $t \ge 2d$ and $n \ge 2d$. There exists a two-stage group testing strategy $\mathcal{A}$ for which*

$$\hat{y}_{\mathcal{A}}(d,n,t) \le \begin{cases} 2d\log\left(\frac{en}{2d}\right)+d\log(2e)+d+1 & \text{if } \hat{y}_{\mathcal{A}}(d,n,t) > t/2, \\ \frac{2d\log\left(\frac{en}{2d}\right)+d\log(2e)+1}{\log\left(\frac{e(t-2d+1)}{\hat{y}_A(d,n,t)-d}\right)} + d \le \frac{2d\log\left(\frac{en}{2d}\right)+d\log(2e)+1}{\log\chi'} + d & \text{if } \hat{y}_{\mathcal{A}}(d,n,t) \le t/2, \end{cases}$$

where $\chi' = \frac{e(t-2d+1)\log(2e)}{2d\left(\log\left(\frac{en}{2d}\right)+\log(e\sqrt{2})\right)}$.

References

[1] Ahlswede, R., Deppe, C., Lebedev, V.: Threshold and majority group testing. In: Aydinian, H., Cicalese, F., Deppe, C. (eds.), Ahlswede Festschrift. LNCS, vol. 7777, pp. 488–508, Springer, Heidelberg (2013)

[2] Alon, N., Asodi, V.: Learning a hidden subgraph. SIAM J. Discrete Math. 18, no. 4, pp. 697–712 (2005)

[3] Chen, H.B., De Bonis, A.: An Almost Optimal Algorithm for Generalized Threshold Group Testing with Inhibitors. J. of Comp. Biology 18, pp. 851–864 (2011)

[4] Chin, F.Y.L., Leung, H.C.M., Yiu, S.M.: Non-adaptive complex group testing with multiple positive sets. Theoretical Computer Science 505, pp. 11–18 (2013)

[5] Cicalese, F., Damaschke, P., Vaccaro, U.: Optimal group testing strategies with interval queries and their application to splice site detection. Int. Journal of Bioinformatics Research and Application 1, no. 4, pp. 363–388 (2015)

[6] Clementi, A. E. F., Monti, A., Silvestri, R.: Selective families, superimposed codes, and broadcasting on unknown radio networks. In: Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 709–718 (2001)

[7] Damaschke, P.: Randomized group testing for mutually obscuring defectives. Inf. Process. Lett. 67, pp. 131–135 (1998)

[8] Damaschke, P., Sheikh Muhammad, A., Triesch, E.: Two New Perspectives on Multi-Stage Group Testing. Algorithmica 67, no. 3, pp. 324–354 (2013)

[9] Damaschke, P., Sheikh Muhammad: Randomized Group Testing Both Query-Optimal and Minimal Adaptive. SOFSEM 2012, pp. 214–225 (2012).

[10] De Bonis, A., Efficient Group Testing Algorithms with a Constrained Number of Positive Responses. In: Proceedings of The 8th Annual International Conference on Combinatorial Optimization and Applications-COCOA'14. LNCS, Springer–Verlag (Germania), vol. 8881, pp. 506–521 (2014).

[11] De Bonis, A, Gąsieniec, L, Vaccaro, U.: Optimal two-stage algorithms for group testing problems. SIAM J. Comput. 34, no. 5, pp.1253–1270 (2005)

[12] De Bonis, A., Vaccaro, U.: Constructions of generalized superimposed codes with applications to group testing and conflict resolution in multiple access channels. Theoretical Computer Science, 306, pp. 223–243 (2003)

[13] De Bonis, A., Vaccaro, U.: Optimal Algorithms for Two Group Testing Problems and New Bounds on Generalized Superimposed Codes. IEEE Trans Inf. Theory 10, pp. 4673–4680 (2006)

[14] Dorfman, R.: The detection of defective members of large populations. Ann. Math. Statist. 14, pp. 436–440 (1943)

[15] Du, D.Z., Hwang, F. K.: Combinatorial Group Testing and Its Applications. World Scientific, River Edge, NJ (2000)

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

[17] D'yachkov, A.G., Rykov, V.V.: A survey of superimposed code theory. Problems Control Inform. Theory 12, pp. 229–242 (1983)

[18] Erdös, P., Frankl, P., Füredi, Z.: Families of finite sets in which no set is covered by the union of r others. Israel J. Math. 51, pp. 75–89 (1985)

[19] Flum, J., Grohe, M.: Parameterized Complexity Theory. Springer. (2006)

[20] Harvey, N.J.A., Patrascu, M., Wen, Y., Yekhanin, S., Chan, V.W.S.: Non-Adaptive Fault Diagnosis for All-Optical Networks via Combinatorial Group Testing on Graphs. In: 26th IEEE Int. Conf. on Comp. Communications, pp. 697–705 (2007)

[21] Hong, E.S., Ladner, R.E.: Group Testing for Image Compression. IEEE Transactions on Image Processing 11, no. 8, pp. 901–911 (2002)

[22] Hwang, F.K., Sós, V.T.: Non adaptive hypergeometric group testing. Studia Sc. Math. Hungarica 22, pp. 257–263 (1987)

[23] Hwang, F.K.: A method for detecting all defective members in a population by group testing. Journal of the American Statistical Association 67, no. 339, pp. 605–608 (1972)

[24] Kautz, W.H., Singleton, R.C.: Nonrandom binary superimposed codes. IEEE Trans Inf. Theory 10, pp. 363–377 (1964)

[25] Knill, E.: Lower bounds for identifying subset members with subset queries. In: Symposium on Discrete Algorithms 1995 pp. 369–377 (1995)

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

[27] Lo, C., Liu, M., Lynch, J.P., Gilbert, A.C.: Efficient Sensor Fault Detection Using Combinatorial Group Testing. In: 2013 IEEE International Conference on Distributed Computing in Sensor Systems, pp. 199–206 (2013)

[28] Porat, E., Rothschild, A.: Explicit non adaptive combinatorial group testing schemes. IEEE Trans Inf. Theory 57, no. 12, pp. 7982–7989 (2011)

[29] Ruszinkó, M.: On the upper bound of the size of the r-cover-free families. J. Combin. Theory Ser. A 66, pp. 302–310 (1994)

[30] Sobel M., Groll, P.A.: Group testing to eliminate efficiently all defectives in a binomial sample. Bell System Tech. J. 38, pp. 1179–1252 (1959)

[31] Pasternack, B.S., Bohnin, D.E., Thomas, J.: Group-Sequential Leak-Testing of Sealed Radium Sources. Technometrics 18 no. 1, pp. 59–66 (1975)

[32] Thomas, J., Pasternack, B.S., Vacirca, S.J., Thompson, D.L.: Application of Group Testing Procedures in Radiological Health. Health Physics 25, pp. 259–266 (1973)

[33] Wang, F., Du, H.D., Jia, X, Deng, P., Wu, W., MacCallum, D.: Non-unique probe selection and group testing. Theoretical Computer Science 381, pp. 29–32 (2007)

[34] Wolf, J.: Born again group testing: multiaccess communications. IEEE Trans. Inf. Theory 31, pp. 185–191 (1985)