Article:
Aldridge, M, Baldassini, L and Gunderson, K (2017) Almost separable matrices. Journal of Combinatorial Optimization, 33 (1). pp. 215-236. ISSN: 1382-6905

https://doi.org/10.1007/s10878-015-9951-1

Almost Separable Matrices

Matthew Aldridge*[1], Leonardo Baldassini†[2] and Karen Gunderson‡[1]

[1]Heilbronn Institute for Mathematics Research, School of Mathematics, University of Bristol, Bristol, UK
[2]School of Mathematics, University of Bristol, Bristol, UK

July 26, 2018

Abstract

An $m \times n$ matrix A with column supports $\{S_i\}$ is k-separable if the disjunctions $\bigcup_{i \in \mathcal{K}} S_i$ are all distinct over all sets $\mathcal{K}$ of cardinality k. While a simple counting bound shows that $m > k \log_2 n/k$ rows are required for a separable matrix to exist, in fact it is necessary for m to be about a factor of k more than this. In this paper, we consider a weaker definition of 'almost k-separability', which requires that the disjunctions are 'mostly distinct'. We show using a random construction that these matrices exist with $m = O(k \log n)$ rows, which is optimal for $k = O(n^{1-\beta})$. Further, by calculating explicit constants, we show how almost separable matrices give new bounds on the rate of nonadaptive group testing.

1 Introduction

Let $\mathsf{A} \in \{\mathtt{0},\mathtt{1}\}^{m \times n}$ be an $m \times n$ binary matrix, and write S_i for the support of its ith column (that is, the locations of the $\mathtt{1}$s). Then A is said to be *k-separable* if the sets $\bigcup_{i \in \mathcal{K}} S_i$ are all distinct over all sets $\mathcal{K} \in \{1, 2, \dots, n\}$ of cardinality k (see Definition 1, to come).

Separable matrices were first introduced by Erdős and Moser in 1970 [9] and have since been studied in different contexts, including coding theory, combinatorics and, as we discuss later, group testing, where they play a very important role.

Separable matrices are often studied through the slightly stronger concept of *disjunct matrices* (see Definition 3). Disjunct matrices were first introduced by Kautz and Singleton [11] and, just like separable matrices, they have been extensively studied in coding theory, combinatorics and group testing [5, 7, 8, 10, 18].

A central question in the study of both separable and disjunct matrices is the following: Given n and k, how large must m be for there to exist either

*m.aldridge@bristol.ac.uk
†leonardo.baldassini@bristol.ac.uk
‡k.gunderson@bristol.ac.uk

an $m \times n$ k-separable or disjunct matrix? In this paper, we investigate the asymptotics for separability as $n \to \infty$, where k may grow with n.

A simple counting bound (Theorem 2) shows that $m \geq \Omega(k \log n/k)$ rows are required. Disappointingly, when $k = o(n)$ this bound is not tight, and we require roughly a factor of k more than this, as in fact it has been shown [7, 5] that $m \geq \Omega(k^2 \log n/ \log k)$ is needed. This lower bound is motivated by the connection between disjunctness and separability, as we discuss in Section 2. Notice that when k grows linearly with n, taking the identity matrix is order optimal – for this reason, we consider only $k = o(n)$ in this paper.

In order to meet the lower bound $m \geq \Omega(k \log n/k)$, we consider a relaxation of the requirement of k-separability to *almost k-separability*. Roughly speaking, a matrix is *almost k-separable* if the sets $\bigcup_{i \in \mathcal{K}} S_i$ are 'usually' distinct – see Definition 4 for a formal definition.

Our main result shows that it is possible to achieve almost separability with only $O(k \log n)$ rows (Theorem 7). When $k = O(n^{1-\beta})$, for any $\beta \in (0, 1]$, this is order-optimal to the counting bound. However, we also aim to get best possible constants for m - a goal motivated by the study of the rate of group testing algorithms.

Group testing is an old and well-studied search problem, first considered by Dorfman [6], where the goal is to recover a sparse subset of k *defective* elements spread among n otherwise identical items. Instead of testing each item for defectiveness individually, classic group testing algorithms test items in batches. In the noiseless binary model we consider, tests can only reveal whether a given set contains at least one defective (a positive test) or no defectives (a negative test). The connection between separable matrices and nonadaptive group testing is well-known, and we discuss it in Section 5. For the moment, we just observe that a sequence of tests designed a priori (*nonadaptive* group testing) has a natural binary-matrix representation: each length-n row represents a test, with entries being `1` if the corresponding item is being included in the test.

A matrix being k-separable is equivalent to having zero probability of error for nonadaptive group testing, while a matrix being almost k-separable is equivalent to having a small probability of error. The 'arbitrarily small probability of error' criterion we consider here is the same as that in Shannon's theory of channel coding.

With this comparison in mind, we consider the concept of *rate* of group testing (Definition 10) for $k = n^{1-\beta}$ defective items in a population of size n, which can be thought of as the amount of information conveyed by each test. Using a separable matrix with $m = \Omega(k^2 \log n/ \log k)$ rows leads to a group testing rate of 0. However, using an almost separable matrix with $m = O(k \log n)$ rows gives a strictly positive rate, with the rate depending on the contant implied by the big-O. Hence, here we are interested in getting good constants for m, not only in order-wise results.

In Theorem 11, we show that our results meet previous results for the limiting regime where k is fixed as $n \to \infty$, and improves over the previous best known bounds for larger values of the *sparsity parameter* $\beta \in [0, 1]$ in the $k = n^{1-\beta}$ regime frequently considered in the group testing literature.

2 Separable matrices

We begin by recalling the definition of a separable matrix.

Definition 1. Given an $m \times n$ binary matrix $\mathsf{A} = (a_{ij}) \in \{\mathtt{0},\mathtt{1}\}^{m\times n}$, we shall write $S_i := \{j : a_{ij} = \mathtt{1}\}$ for the *support* of column i and for $\mathcal{K} \subseteq \{1, 2, \ldots, n\}$ also write $S(\mathcal{K}) := \bigcup_{i\in\mathcal{K}} S_i$ for the support of a disjunction of columns.

The matrix A is called *k-separable* matrix if the for all sets $\mathcal{K}$ of size k, there is no other set $\mathcal{L}$ also of size k with $S(\mathcal{L}) = S(\mathcal{K})$.

The case $k = 0$ is trivial, so we assume $k \geq 1$ throughout. We shall also assume $k \leq n/2$, which will be no restriction in the limiting regimes we study.

The following counting bound is described by Chen and Hwang as "simple-minded" [5].

Theorem 2. *Let $M(n, k)$ be the smallest m such that an $m \times n$ k-separable matrix exists. Then*

$$M(n,k) \geq \log_2 \binom{n}{k}.$$

Proof. Clearly

$$|\{S(\mathcal{K}) : |\mathcal{K}| = k\}| \leq |\mathcal{P}(\{1, 2, \ldots, m\})| = 2^m,$$

where $\mathcal{P}$ denotes the power set. Hence for A to be k-separable we require $2^m \geq \binom{n}{k}$, and taking logarithms gives the result. □ □

Using the lower bound of

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{\mathrm{e}n}{k}\right)^k \tag{1}$$

(which we shall use many times in this paper), we see that a k-separable matrix must have at least $m \geq k \log_2 n/k = \Omega(k \log n/k)$ rows.

As we anticipated, separable matrices are tightly related to another class of matrices, namely that of disjunct matrices.

Definition 3. With the notation of Definition 1, A is *k-disjunct* if for all sets $\mathcal{K}$ of cardinality $|\mathcal{K}| = k$, there does not exist $i \notin \mathcal{K}$ such that $S_i \subseteq S(\mathcal{K})$.

In the language of set systems, a matrix A being k-seperable is equivalent to the family $\{S_i\}_{i=1}^n$ being k-union-free, and A being k-disjunct is equivalent to $\{S_i\}_{i=1}^n$ being k-cover-free.

It's easy to see that k-disjunctness implies k-separability (see, for example, [11], [7, Section 7.2], or the special case $\epsilon = 0$ of Lemma 6 below). On the other hand, Chen and Hwang [5, Theorem 2] have shown that it is possible to construct a k-disjunct matrix from a $2k$-separable matrix by adding at most one row to it, which means that disjunct and separable matrices share the same order-wise asymptotics. Dyachkov and Rykov have quantified these asymptotics by showing that $m \geq \Omega(k^2 \log n/\log k)$ rows are necessary for a matrix to be k-disjunct [8] – similar results appear elsewhere [18] [10] [7, Theorem 7.2.14]. This means that it is not possible to create a k-separable matrix with $m = O(k \log n)$ rows.

As disjunctness is a stronger (and, in some ways, simpler) property than separability, efforts to derive upper bounds on m for separable matrices have often proceeded via the construction of disjunct matrices. In their seminal paper [11], Kautz and Singleton give a probabilistic existence theorem for k-disjunct matrices with $m = O(k^2 \log n)$ rows. In the group testing literature there exist explicit constructions of testing schemes with $O(k^2 \log n)$ rows, see for example Porat and Rothschild [17].

3 Almost separable matrices

Since separable matrices cannot meet the counting bound, it would be of interest if a matrix could be close to being separable using only $O(k \log n)$ rows. Such a matrix would be order-optimal.

With this in mind, we define the concept of an *almost separable* matrix in a similar manner to Defintion 1.

Definition 4. With the notation of Definition 1, $\mathsf{A} \in \{0,1\}^{m \times n}$ is *ϵ-almost k-separable* if for at most $\epsilon\binom{n}{k}$ sets $\mathcal{K}$ of size k does there exist another set $\mathcal{L}$ of size k with $S(\mathcal{L}) = S(\mathcal{K})$.

An analogous definition is present in for example [22], where almost separable matrices are called *weakly separating designs*. Note that setting $\epsilon = 0$ gives the definition of a separable matrix.

The main result of this paper is to show the existence of ϵ-almost k-separable matrices with $m = O(k \log n)$ rows (see Theorem 7 below). We also examine the implicit constants for the case when $k = n^{1-\beta}$ grows polynomially in n.

Malyutov [14] effectively showed that ϵ-almost k-separable matrices exist with $m = (k + o(1)) \log_2 n$ rows in the regime where k is fixed as $n \to \infty$. This is a special case of a more general result Malyutov proved using an information theoretic argument – this and similar work is reviewed in [15]. Sebő showed effectively the same result [19], again for fixed k, by analysing a concrete bound on the probability that there are two different sets of size k whose disjunctions coincide – we follow a similar route here later. The same result for k fixed and $n \to \infty$ was rediscovered by Zhigljavsky [22, Theorem 5.5]. Although technically different from Sebő's argument, Zhigljavsky's proof is morally similar: given two sets $\mathcal{K}$ and $\mathcal{L}$ of k columns each, Zhigljavsky counts how many rows it is possible to construct that would produce the same value for both $S(\mathcal{K})$ and $S(\mathcal{L})$. He calls this number a Rényi coefficient and only considers designs with fixed- or bounded-size tests.

Our result improves on these by allowing k to vary arbitrarily with n, subject to $k = o(n)$. In our discussion of group testing in Section 5 we show how, in some regimes, this work also improves on recent results on nonadaptive group testing giving bounds of the form $m = O(k \log n)$.

The definition of a disjunct matrix (Defintion 3) can similarly be weakened to give an *almost disjunct matrix*. (This definition also appears in [16] and, previously, in [12].)

Definition 5. With the notation of Definition 1, A is *ϵ-almost k-disjunct* if for at most $\epsilon\binom{n}{k}$ sets $\mathcal{K}$ of size k does there exist a column $i \notin \mathcal{K}$ with $S_i \subseteq S(\mathcal{K})$.

Note again that $\epsilon = 0$ corresponds to a disjunct matrix. Unsurprisingly, almost disjunctness implies almost separability.

Lemma 6. *Let* A *be an* ϵ*-almost* k*-disjunct matrix. Then* A *is* ϵ*-almost* k*-separable (with the same* ϵ *and* k*).*

Proof. We prove the contrapositive. Suppose A is not ϵ-almost k-separable. Then there are more than $\epsilon\binom{n}{k}$ sets of size k breaking separability. Let $\mathcal{K}$ be one of these sets, so there is another set $\mathcal{L}$ of size k with $S(\mathcal{K}) = S(\mathcal{L})$. Letting $i \in \mathcal{K} \setminus \mathcal{L}$, we have $S_i \subseteq S(\mathcal{K})$, breaking disjunctness. Hence there are more than $\epsilon\binom{n}{k}$ sets breaking disjunctness, and A is not ϵ-almost k-disjunct. □ □

Mazumdar [16] shows that there exist almost k-disjunct matrices with $m = O(k^{3/2}\sqrt{\log n})$ rows in the regime $k \sim n^{\delta}$, $\delta > 0$, which is the same as that we consider for group testing. Mazumdar's construction is similar to those of Kautz and Singleton [11] and Porat and Rothschild [17]. In particular, [11] shows how to build fully disjunct matrices with $O(k^2 \log^2_{k \log n} n)$ rows by mapping the symbols of a q-ary Reed-Solomon code to unit-weight binary vectors of length q, while [17] improves on this scheme by replacing the RS code with a linear q-ary code achieving the Gilbert-Varshamov bound. This produces fully disjunct matrices with $O(k^2 \log n)$ rows. This improves on the $\Omega(k^2 \log n / \log k)$ required for full disjunctness or separability, while being less good than the $O(k \log n)$ we achieve for almost separability here.

Our main result is then the following.

Theorem 7. *For any sequence* $k = k(n) = o(n)$ *and* $\epsilon > 0$*, there exist an* ϵ*-almost* k*-separable matrix with* $m = O(k \log n)$ *rows.*

More precisely, for $\alpha \in [\ln 2, 1]$*, define*

$$\begin{aligned} M_1(n,k,\alpha) &= \frac{1}{-\ln(1 - 2\mathrm{e}^{-\alpha} + 2\mathrm{e}^{-2\alpha})} k \ln \frac{n}{k}, \\ M_2(n,k,\alpha) &= \frac{1}{-\ln(1 - 2\mathrm{e}^{-\alpha} + 2\mathrm{e}^{-\alpha(1+1/k)})} \ln nk, \\ M(n,k) &= \min_{\alpha \in [\ln 2, 1]} \max\left\{M_1(n,k,\alpha), M_2(n,k,\alpha)\right\}. \end{aligned} \tag{2}$$

Then for any $\epsilon, \delta > 0$*, for* n *sufficiently large, and* $m > (1+\delta)M(n,k)$*, there exists and* $m \times n$ ϵ*-almost* k*-separable matrix.*

Consider the special case $\alpha = \ln 2$. It is possible to see that M_2 dominates, and hence that there exist almost separable matrices with $m = (1+\delta)k \log_2 nk$ rows. Note that this is sufficient to show the $m = O(k \log n)$ result – and comes with a slightly easier proof than the general case (see below). This bound also meets the Malyutov–Sebő result of $m \sim k \log_2 n$ for k constant. However, it is possible to get slightly better constants for most $k = k(n)$ by allowing different values of α. In particular, M_2 with $\alpha = 1$ gives the best result in many regimes.

In Section 5 we discuss the constants in more detail in the regime $k = n^{1-\beta}$ for $\beta \in (0,1)$. (The reader may wish to skip ahead to Figure 1, to get a feeling for this result.)

Our proof gives a randomised construction where the matrix is chosen to have entries sampled from IID Bernoulli random variables; we discuss this in the next section.

4 Proof of main result

We proceed to prove Theorem 7 as follows. Fix n and k. We will choose A to be an $m \times n$ matrix (where m will be determined later) with each entry independently $\mathtt{1}$ with probability p and $\mathtt{0}$ with probability $q = 1 - p$, for some p also to be chosen later. We aim to show that there is a choice of m and p so that, with positive probability, A is ϵ-almost k-separable, and hence that such a matrix exists.

The following bound will be important, and is fairly well known – see for example Sebő [19], who analyses its asymptotics for fixed k as $n \to \infty$.

Lemma 8. *Let A be a randomly chosen matrix in $\{0,1\}^{m\times n}$ with each entry independently 1 with probability p. For any set $\mathcal{K}$ of size $k \le n/2$, then*

$$\mathbb{P}(\exists\ \mathcal{L} \text{ with } |\mathcal{L}| = k,\ S(\mathcal{L}) = S(\mathcal{K})) \le \sum_{b=0}^{k-1} \binom{k}{b}\binom{n-k}{k-b}\left(1 - 2q^k + 2q^{2k-b}\right)^m. \tag{3}$$

Proof. Say that an *overlap* occurs if there exists $\mathcal{L}$ with $|\mathcal{L}| = k$ and $S(\mathcal{L}) = S(\mathcal{K})$. Take two distinct sets $\mathcal{K}, \mathcal{L}$, both of size k, that have $b = |\mathcal{K} \cap \mathcal{L}|$ elements in common. Then a row j of A could distinguish between $\mathcal{K}$ and $\mathcal{L}$ in two ways: either we have $j \in S(\mathcal{K})$ while $j \notin S(\mathcal{L})$, or the other way round: $j \in S(\mathcal{L})$ while $j \notin S(\mathcal{K})$.

If the entries of the row $\mathbf{a}_j$ are IID Bernoulli(p), these two events each occur with probability $q^k(1 - q^{k-b}) = q^k - q^{2k-b}$. Hence, row j fails to distinguish between $\mathcal{K}$ and $\mathcal{L}$ with probability $1 - 2q^k(1 - q^{k-b}) = 1 - 2q^k + 2q^{2k-b}$.

Since the rows of A are IID, the whole matrix fails to distinguish between $\mathcal{K}$ and $\mathcal{L}$ with probability $(1 - 2q^k + 2q^{k-b})^m$.

The result then follows by a union bound over $\mathcal{L}$, noting that the number of sets of size k sharing b elements with $\mathcal{K}$ is precisely $\binom{k}{b}\binom{n-k}{k-b}$. □ □

The main work in this paper is a careful asymptotic analysis of the overlap probability (3), showing for which m it can be made arbitrarily small.

Lemma 9. *For every sequence $k = k(n) = o(n)$, $\varepsilon, \delta > 0$, there exists n_0 so that if $n > n_0$ and $m > (1 + \delta)M(n, k)$, with $M(n, k)$ as in Theorem 7, then $\mathbb{P}(\text{overlap}) < \varepsilon$.*

Proof. We first prove that it suffices to have $m > (1 + \delta)M_2(n, k, \ln 2)$, with $M_2(n, k, \ln 2) = (1 + o(1))k \log_2 nk$. This is simpler to prove than the full result and illustrates the main techniques.

Here, we take $p = 1 - 2^{-1/k}$, as does Sebő [19], so that $q = 2^{-1/k}$. This is a special case of the general value of p used in the appendix, $p = 1 - \mathrm{e}^{-\alpha/k}$, by taking $\alpha = \ln 2$. Note that, in group testing parlance, this is the value of p that gives a $50 : 50$ chance of a test being positive. The bound (3) then becomes

$$\mathbb{P}(\text{overlap}) \le \sum_{b=0}^{k-1} \binom{k}{b}\binom{n-k}{k-b}\left(\frac{1}{2}2^{b/k}\right)^m.$$

It will be convenient to write $c = k - b$ for the number of nonoverlapping items, to get

$$\begin{aligned}\mathbb{P}(\text{overlap}) &\le \sum_{c=1}^{k} \binom{k}{k-c}\binom{n-k}{c}\left(\frac{1}{2}2^{(k-c)/k}\right)^m \\ &= \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{c}2^{-cm/k}.\end{aligned}$$

When $m > (1+\delta)k\log_2 nk$, then the terms in the above sum are decreasing since

$$\begin{aligned}\frac{\binom{k}{c+1}\binom{n-k}{c+1}2^{-(c+1)m/k}}{\binom{k}{c}\binom{n-k}{c}2^{-cm/k}} &= \frac{(k-c)(n-k-c)2^{-m/k}}{(c+1)^2} \\ &\le \frac{c^2 - nc + k(n-k)}{nk(c^2+2c+1)} \qquad (\text{since } 2^{-m/k} \le 1/nk) \\ &\le \frac{1}{4},\end{aligned}$$

for $n > 2k$ and $k \ge 2$. Thus, the probability of an overlap can be estimated by the largest term with

$$\begin{aligned}\mathbb{P}(\text{overlap}) &\le k(n-k)2^{-m/k}\sum_{c=1}^{k}\left(\frac{1}{4}\right)^{c-1} \\ &\le kn2^{-(1+\delta)\log_2 nk}\frac{4}{3} \\ &= nk(nk)^{-1-\delta}\frac{4}{3} \\ &\le 2(nk)^{-\delta},\end{aligned}$$

which, for fixed $\delta > 0$, can be made arbitrarily small for n sufficiently large.

Further, since $\log_2 nk \le 2\log_2 n$, we see that $m > (1+\delta)k\log_2 nk = O(k\log n)$.

We can get the more general result that it suffices to have $m > (1+\delta)M(n,k)$, with $M(n,k)$ as in (2), by instead taking $p = 1-\mathrm{e}^{-\alpha/k}$, and then optimising over α. The analysis is very similar to that above, but somewhat more longwinded. The interested reader is directed to the appendix for the details. □ □

Proving our main result is now straightforward.

Proof of Theorem 7. Choose the matrix A at random as above, with m and n chosen as in Lemma 9 so that the overlap probability is at most $\epsilon/2$.

Write X for the number of sets $\mathcal{K}$ of size k that experience an overlap. It is clear A will be ϵ-almost k-separable provided that $X \le \epsilon\binom{n}{k}$.

Then we have

$$\mathbb{P}\left(X > \epsilon\binom{n}{k}\right) \le \frac{1}{\epsilon\binom{n}{k}}\mathbb{E}X,$$

by the Markov inequality. But this expectation is, by Lemma 9

$$\mathbb{E}X = \sum_{|\mathcal{K}|=k}\mathbb{P}(\mathcal{K} \text{ has an overlap}) = \binom{n}{k}\mathbb{P}(\text{overlap}) \le \binom{n}{k}\frac{\epsilon}{2}.$$

Hence, our random A is ϵ-almost k-separable with probability at least $1/2$, so such matrices must exist. □ □

5 Rates for nonadaptive group testing

In this section, we show how the use of almost separable matrices can give new results on the rate of nonadaptive group testing.

As we outlined in the introduction, in a nonadaptive group testing procedure we aim to find a subset $\mathcal{K}$ of k defective items within a population of n identical items. We use m pooled tests. Recall that the outcome of a test j is positive if one or more of the defective items is in the test pool, and negative if none of them are. We summarise our testing procedure by a matrix $\mathsf{A} = (a_{ij})$, where $a_{ij} = \mathtt{1}$ denotes that item i is in the pool for test j, and $a_{ij} = \mathtt{0}$ denotes that it is not. Recalling the notation of Definition 1, the set of positive tests for a defective set $\mathcal{K}$ is precisely $S(\mathcal{K})$.

The aim is, given the outcomes $S(\mathcal{K})$ and the matrix A, to identify the defective set $\mathcal{K}$. Clearly if there is no other $\mathcal{L}$ with $S(\mathcal{K}) = S(\mathcal{L})$, then we can find $\mathcal{K}$ (at least theoretically: for study of practical algorithms for this, see, for example, [1, 4, 20, 13, 21]). Conversely, if there is an $\mathcal{L}$ with $S(\mathcal{K}) = S(\mathcal{L})$, then our error probability is at least $1/2$.

A comprehensive survey of combinatorial group testing is given in [7]. Likewise, the study of nondeterministic testing schemes is addressed in the field of probabilistic group testing – see for example [22] and references therein. The derivation of both non-constructive results and practical algorithms has been addressed in different contexts, including combinatorial [7, 14, 16, 19], probabilistic [1, 22] and information-theoretic [2, 3, 15, 17, 20] scenarios.

The connection between separable matrices and nonadaptive group testing is well explored. In particular, if there are known to be exactly k defective items, then a testing matrix will allow us to find the defective set with certainty if and only if it is k-separable. The advantages of using what we call almost separability for group testing in the fixed-k regime have also been discussed in [22].

While separable matrices allow detection with zero probability of error, the study of group testing within the scope of information theory and the need for efficient algorithms generated an interest in nonadaptive group testing with low – but not necessarily zero – probability of error, a situation which has gained considerable attention [1, 4, 20, 14, 13, 2, 21]. Here the probability of error is defined as an average over all possible defective sets of size k; that is,

$$\mathbb{P}(\text{error}) = \frac{1}{\binom{n}{k}} \sum_{|\mathcal{K}|=k} \mathbb{P}(\text{error} \mid \mathcal{K}) \ .$$

Baldassini, Johnson and Aldridge [3] introduced a concept of the *rate* of group testing to quantify how well a group testing design works. (An earlier definition of rate for the fixed k regime had been introduced by Malyutov [15].) The rate is the ratio of the number of tests to the counting bound $\log_2 \binom{n}{k}$. If we interpret the counting bound as a binary labelling of all possible defective sets of size k, the rate can be considered as the number of bits learned per test by the group testing procedure.

Definition 10. Consider a group testing problem with n items of which k are defective. A design with m tests is said to have *rate* $R = m/\log_2 \binom{n}{k}$.

Given a sequence of group testing problems for n items of which $k = k(n)$ are defective, a rate R is said to be *achievable* for a design A if, for any $\epsilon > 0$, the design finds the defective set with error probability at most ϵ with rate at least R for n sufficiently large.

We follow Baldassini et al. [3, 1] and study achievable rates in regimes where $k = k(n) = n^{1-\beta}$ for different values of the sparsity parameter $\beta \in (0, 1]$.

Note from the above that using a k-separable matrix with $m \geq \Omega(k^2 \log n / \log k)$ tests gives rate 0 for all values of $\beta < 1$.

As far as we are aware, the best known rate for nonadaptive group testing until now is achieved by the `DD` algorithm of Aldridge, Baldassini and Johnson [1], which has a lower bound on the maximum achievable rate of

$$R_{\mathtt{DD}}(\beta) = \frac{1}{\mathrm{e} \ln 2} \min\left\{\frac{\beta}{1-\beta}, 1\right\} \approx 0.53 \min\left\{\frac{\beta}{1-\beta}, 1\right\}, \tag{4}$$

together with the Malytuov–Sebő result that $R = 1$ can be achieved in the fixed-k regime.

Baldassini, Johnson and Aldridge [3] also showed that for adaptive group testing, the generalized binary splitting algorithm of Hwang [7] gives a rate of 1 (the best possible) for all $\beta \in (0, 1]$.

From Theorem 7, we know that using an ϵ-almost k-separating matrix will find the defective set with error probability at most ϵ, since the sets $\mathcal{K}$ without overlaps can by definition be recovered with certainty. Hence, the number of rows of the almost separating matrix gives bounds on the rate. Therefore, using our above results, we have the following:

Theorem 11. *For $\beta \in (0, 1]$ and $k = n^{1-\beta}$, the maximum achievable rate of nonadaptive group testing with n items of which k are defective is bounded below by*

$$R \geq \frac{1}{\ln 2} \max_{\alpha \in [\ln 2, 1]} \min\left\{2\alpha \mathrm{e}^{-\alpha} \frac{\beta}{2-\beta}, -\ln\left(1 - 2\mathrm{e}^{-\alpha} + 2\mathrm{e}^{-2\alpha}\right)\right\}. \tag{5}$$

Figure 1 illustrates the result of Theorem 11. Note that our result improves over the best known result for $\beta > 2/3$, and meets the Malyutov–Sebő point as $\beta \to 1$.

Proof. Following directly from Theorem 7 and the definition of rate, we have

$$R \geq \frac{1}{\ln 2} \max_{\alpha \in [\ln 2, 1]} \min\Big\{ -\ln\left(1 - 2\mathrm{e}^{-\alpha} + 2\mathrm{e}^{-\alpha(1+1/k)}\right) k \frac{\beta}{2-\beta}, \\ -\ln\left(1 - 2\mathrm{e}^{-\alpha} + 2\mathrm{e}^{-2\alpha}\right) \Big\},$$

noting that, when $k = n^{1-\beta}$,

$$k \log_2 nk = \frac{2-\beta}{\beta} k \log_2 \frac{n}{k}.$$

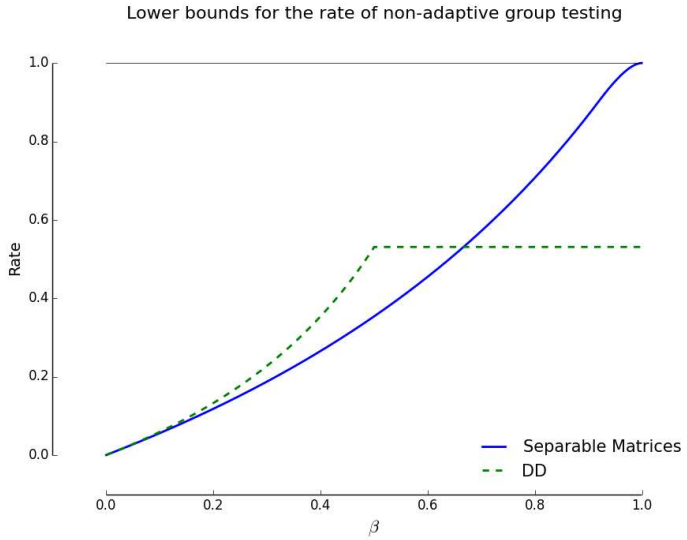

Figure 1: Bounds on rates of group testing, showing the DD bound (4) of Baldassini et al, and our new result Theorem 11.

When $\beta = 1$, the second term is the minimum. When $\beta < 1$, since we have that $k \to \infty$, we can take limits in the first minimand. We have

$$\begin{aligned}
&-\ln\left(1 - 2\mathrm{e}^{-\alpha} + 2\mathrm{e}^{-\alpha(1+1/k)}\right)k \\
&\qquad = -\ln\left(1 - 2\mathrm{e}^{-\alpha} + 2\mathrm{e}^{-\alpha}\mathrm{e}^{-\alpha/k}\right)k \\
&\qquad = -\ln\left(1 - 2\mathrm{e}^{-\alpha} + 2\mathrm{e}^{-\alpha}\left(1 - \frac{\alpha}{k} + o\left(\frac{1}{k}\right)\right)\right)k \\
&\qquad = -\ln\left(1 - 2\mathrm{e}^{-\alpha}\frac{\alpha}{k} + o\left(\frac{1}{k}\right)\right)k \\
&\qquad = \left(2\mathrm{e}^{-\alpha}\frac{\alpha}{k} + o\left(\frac{1}{k}\right)\right)k \\
&\qquad \to 2\alpha\mathrm{e}^{-\alpha}.
\end{aligned}$$

The result follows. □ □

Note that our 'simpler' result with $\alpha = \ln 2$ gives a bound almost as good the general case, namely

$$R(\ln 2) = \frac{\beta}{2 - \beta}.$$

In particular, this choice of $\alpha = \ln 2$ is optimal at $\beta = 1$.

Note also that for all but the sparsest cases, we get the bound by taking $\alpha = 1$. Specifically, for $\beta \le \beta_0$, where

$$\beta_0 = \frac{-2\ln(1 - 2\mathrm{e}^{-1} + 2\mathrm{e}^{-2})}{2\mathrm{e}^{-1} - \ln(1 - 2\mathrm{e}^{-1} + 2\mathrm{e}^{-2})} \approx 0.92,$$

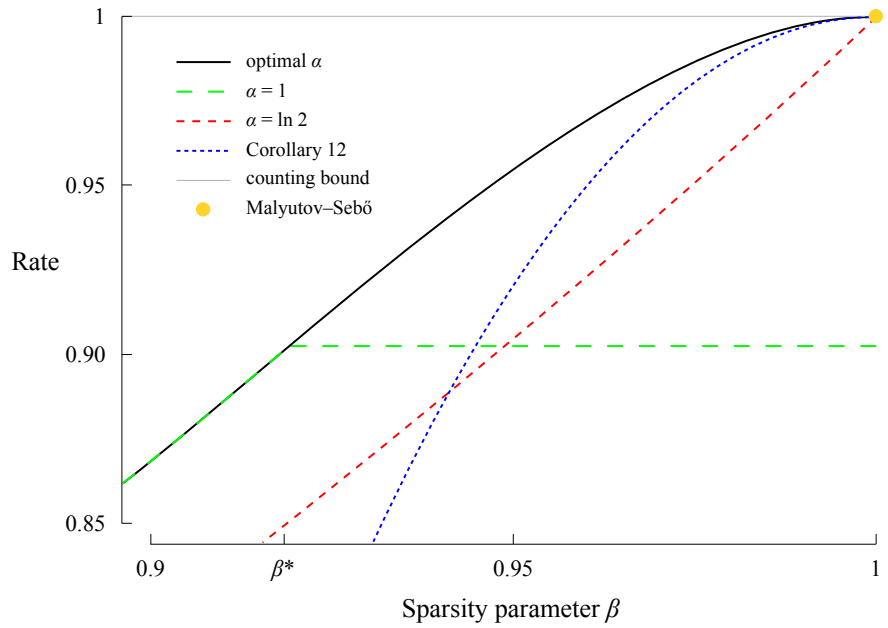

Figure 2: Bounds on rates of group testing for large β, showing Theorem 11 for different values of α and the approximation of Corollary 12.

the best value of the bound is

$$\begin{aligned} R(1) &= \frac{1}{\ln 2} \min\left\{ 2\mathrm{e}^{-1}\frac{\beta}{2-\beta}, -\ln\left(1-2\mathrm{e}^{-1}+2\mathrm{e}^{-2}\right)\right\} \\ &= \frac{1}{\ln 2} 2\mathrm{e}^{-1}\frac{\beta}{2-\beta} \\ &\approx 1.06\frac{\beta}{2-\beta}. \end{aligned}$$

For $\beta \in (\beta_0, 1)$, the optimal rate is given as the maximum in (5), and the optimal α is that which achieves the maximum. It's easy to see for $\beta \geq \beta_0$ that the maximum over α is achieved when the two terms in the minimum are equal, and this is simple to solve numerically. However, here we also provide some closed form approximations to this which could be useful.

Corollary 12. *For $\frac{2\ln 2}{1+\ln 2} < \beta < 1$ and $k = n^{1-\beta}$, the maximum achievable rate of nonadaptive group testing with n items, of which k are defective, is bounded from below by*

$$R \geq 1 - \frac{1}{\ln 2}\ln\left(1+\left(\frac{2(1-\beta)\ln 2}{\beta(1-\ln 2)}\right)^2\right)$$

This is illustrated in Figure 2. From this, we see that the bound of Corollary 12 is very good for $\beta \approx 1$, but that when β is not much above β_0, then the bound of simply $\alpha = 1$ is better. Hence, setting

$$\beta_1 = \frac{2\ln 2}{1-2\mathrm{e}^{-1}+\ln 2+2\mathrm{e}^{-1}\ln 2} \approx 0.94,$$

and taking β_0 as above, we get the following bound:

Corollary 13. *For $\beta \in (0,1)$ and $k = n^{1-\beta}$, the maximum achievable rate of nonadaptive group testing with n items, of which k are defective, is bounded from below by*

$$R \geq \frac{1}{\ln 2} \begin{cases} 2\mathrm{e}^{-1}\dfrac{\beta}{2-\beta} & \text{if } \beta \leq \beta_0 \\ -\ln(1 - 2\mathrm{e}^{-1} + 2\mathrm{e}^{-2}) & \text{if } \beta_0 < \beta \leq \beta_1, \\ \ln 2 - \ln\left(1 + \left(\dfrac{2(1-\beta)\ln 2}{\beta(1-\ln 2)}\right)^2\right) & \text{if } \beta > \beta_1 \end{cases}$$

The proofs of these statements can be found in Appendix B.

6 Conclusions and further work

We have explored the asymptotics of almost separability and we have shown that almost separable matrices exist with $O(k \log n)$ rows. Furthermore, we have proved that the use of almost separable matrice can improve the lower bounds on the rate of nonadaptive group testing in the very sparse regime.

Several interesting questions, however, remain still open, and provide scope for future research. Most notably, while we have given new achievable rates, the maximum rate of nonadative group testing is still unknown. In particular, we know of no upper bounds beyond the trivial counting bound.

As discussed in Section 2, Chen and Hwang [5] have proved that disjunct and separable matrices share the same asymptotics by showing how to construct a k-disjunct matrix out of a $2k$-separable matrices by adding at most one row to it. Unlike its inverse (disjunctness implying separability), this statement doesn't naturally carry through to the case of almost separability/disjunctness.

Another problem is to extend the existing results to other regimes than the $k = n^{1-\beta}$ for $\beta \in (0,1]$ considered here. Of particular interest is the case where $k = cn$ grows like a constant proportion of n, as in recent work by Wadayama [21]. Note that the counting bound now gives a lower bound of $m = O(n)$, while, for coupon-collector reasons, the IID random approach here inevitably leads to the suboptimal $m = \Omega(n \log n)$.

A Asymptotic analysis of the overlap probability

We now show the full result of Lemma 9.

We use the same random construction as the special case described in Section 4, but now take $p = 1 - \mathrm{e}^{-\alpha/k}$, so $q = \mathrm{e}^{-\alpha/k}$, where α is a parameter to be chosen later (simply taking $\alpha = \ln 2$ as in Section 4 gives $p = 1 - 2^{-1/k}$). Within the group testing literature, different values of p have also been considered. For example, the value $p = 1/k$ (which gives an average of one defective per test) has been considered before by many authors [1, 4, 2, 22], while Sejdinovic and Johnson [20] consider the more general α/k for noisy group testing. The same value can be obtained asymptotically in this context, as $p \sim \alpha/k$ if $k \to \infty$ as $n \to \infty$.

Proof of Lemma 9. We wish to find values of m such that $\mathbb{P}(\text{overlap})$ can be made arbitrarily small. It will be convenient to write

$$s = 1 - 2q^k = 1 - 2\mathrm{e}^{-\alpha}, \qquad t = 2q^{2k} = 2\mathrm{e}^{-2\alpha}, \qquad u = \frac{1}{q} = \mathrm{e}^{\alpha/k},$$

allowing us to rewrite the bound (3) as

$$\mathbb{P}(\text{overlap}) \leq \sum_{b=0}^{k-1} \binom{k}{b}\binom{n-k}{k-b}\left(s + tu^b\right)^m.$$

As before, it will be more convenient to deal with $c = b - k$, which gives

$$\mathbb{P}(\text{overlap}) \leq \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{c}\left(s + tu^{k-c}\right)^m. \tag{6}$$

Now, we expand out $(s+tu^b)^m$ in (6) using the binomial theorem and reverse the order of summation to get

$$\begin{aligned}
\mathbb{P}(\text{overlap}) &\leq \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{c} \sum_{j=0}^{m} \binom{m}{j} s^{m-j} t^j u^{(k-c)j} \\
&= \sum_{j=0}^{m} \binom{m}{j} s^{m-j} t^j \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{c} u^{(k-c)j} \\
&= \sum_{j=0}^{m} \binom{m}{j} s^{m-j} t^j u^{jk} \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{c} q^{cj} \qquad (7)
\end{aligned}$$

Consider the inner sum of (7). It is possible to approximate it by its largest term, which will depend on the value of j. To start with, the following bound holds:

$$\binom{k}{c}\binom{n-k}{c} q^{cj} \leq \left(\frac{\mathrm{e}^2 knq^j}{c^2}\right)^c. \tag{8}$$

Note that for any a, the function $(a/x^2)^x$ attains its maximum at $x = \sqrt{a}/\mathrm{e}$; and further is increasing for $x < \sqrt{a}/\mathrm{e}$ and decreasing for $x > \sqrt{a}/\mathrm{e}$. In (8), the maximum corresponds to $c = \sqrt{knq^j}$. Now, $1 < \sqrt{knq^j} < k$ when

$$\frac{1}{-\ln q} \ln \frac{n}{k} < j < -\frac{1}{-\ln q} \ln nk,$$

or, since $q = \mathrm{e}^{-\alpha/k}$,

$$\frac{1}{\alpha} k \ln \frac{n}{k} < j < \frac{1}{\alpha} k \ln nk.$$

Then, in light of the above, we will split between the three cases: first, $j \leq k/\alpha \ln n/k$; second, $k/\alpha \ln n/k < j < k/\alpha \ln nk$; and third, $j \geq k/\alpha \ln nk$.

For the first case, $j \leq k/\alpha \ln n/k$, the maximum of (8) is attained at $c = k$, giving the bound

$$\left(\frac{\mathrm{e}^2 knq^j}{k^2}\right)^k = \mathrm{e}^{2k} \left(\frac{n}{k}\right)^k q^{jk}.$$

Summing over this range for j yields

$$\begin{aligned}\sum_{j=0}^{k/\alpha \ln n/k} \binom{m}{j} s^{m-j} t^j u^{jk} k\mathrm{e}^{2k} \left(\frac{n}{k}\right)^k q^{jk} &= k\mathrm{e}^{2k} \left(\frac{n}{k}\right)^k \sum_{j=0}^{k/\alpha \ln n/k} \binom{m}{j} s^{m-j} t^j \\ &\le k\mathrm{e}^{2k} \left(\frac{n}{k}\right)^k \sum_{j=0}^{m} \binom{m}{j} s^{m-j} t^j \\ &= k\mathrm{e}^{2k} \left(\frac{n}{k}\right)^k (s+t)^m \\ &= k \exp\left(2k + k \ln \frac{n}{k} + m \log(s+t)\right).\end{aligned}$$

Provided that

$$\begin{aligned} m &> (1+\delta) \frac{1}{-\ln(s+t)} k \ln \frac{n}{k} \\ &= (1+\delta) \frac{1}{-\ln(1 - 2\mathrm{e}^{-\alpha} + 2\mathrm{e}^{-2\alpha})} k \ln \frac{n}{k} \\ &= (1+\delta) M_1(n,k,\alpha), \end{aligned} \tag{9}$$

for some $\delta > 0$, then this can be made arbitrarily small for n sufficiently large.

For the second case, $k/\alpha \ln n/k < j < k/\alpha \ln nk$, the maximum is attained at $c = \sqrt{knq^j}$, giving the bound

$$\begin{aligned}\left(\frac{\mathrm{e}^2 knq^j}{knq^j}\right)^{\sqrt{knq^j}} &= \exp(2\sqrt{knq^j}) \le \exp(2\sqrt{knq^{k/\alpha \ln n/k}}) \\ &= \exp\left(2\sqrt{kn\left(\frac{n}{k}\right)^{k/\alpha \ln q}}\right) = \exp\left(2\sqrt{kn\frac{k}{n}}\right) = \exp(2k).\end{aligned}$$

Then we have that

$$\begin{aligned}\sum_{j=k/\alpha \ln n/k}^{k/\alpha \ln nk} \binom{m}{j} s^{m-j} t^j u^{jk} \sum_{c=1}^{k} \binom{k}{c} \binom{n-k}{c} q^{jc} \\ &\le \sum_{j=k/\alpha \ln n/k}^{k/\alpha \ln nk} \binom{m}{j} s^{m-j} (tu^k)^j k\mathrm{e}^{2k} \\ &= k\mathrm{e}^{2k} \mathbb{P}\left(\frac{1}{\alpha} k \ln \frac{n}{k} < X \le \frac{1}{\alpha} k \ln nk\right),\end{aligned}$$

where we called $X \sim \mathrm{Bin}(m, tu^k)$, and we have used that $s = 1 - tu^k$. Then as long as

$$\mathbb{E}X = mtu^k > (1+\delta)\frac{1}{\alpha} k \ln nk, \tag{10}$$

we have by the Azuma–Hoeffding inequality that

$$\begin{aligned} k\mathrm{e}^{2k}\mathbb{P}\left(\frac{1}{\alpha}k\ln\frac{n}{k} < X \le \frac{1}{\alpha}k\ln nk\right) &\le k\mathrm{e}^{2k}\mathbb{P}\left(X \le \frac{1}{\alpha}\ln nk\right) \\ &\le k\mathrm{e}^{2k}\exp\left(-\frac{2}{m}\left(mtu^k - \frac{1}{\alpha}k\ln nk\right)^2\right) \\ &= k\exp\left(2k - 2m\left(tu^k - \frac{k\ln nk}{\alpha m}\right)^2\right). \end{aligned}$$

Given (10), this can be made arbitrarily small for n sufficiently large. We can rewrite (10) as

$$m > (1+\delta)\frac{1}{\alpha t u^k}k\ln nk = (1+\delta)\frac{\mathrm{e}^\alpha}{2\alpha}k\ln nk = (1+\delta)M_2(n,k,\alpha). \tag{11}$$

Now for the final case, when $j \ge k/\alpha\ln nk$. Note that for $j \ge k/\alpha\ln nk$,

$$q^j \le q^{k/\alpha\ln nk} = \mathrm{e}^{-\ln nk} = \frac{1}{nk},$$

hence $nkq^j \le 1$. Then, splitting up $c=1$, $c=2$, and $c\ge 3$, and noting that $\mathrm{e}^2/9 < 1$, we have

$$\begin{aligned} \sum_{c=1}^{k}\left(\frac{\mathrm{e}^2knq^j}{c^2}\right)^c &\le \mathrm{e}^2knq^j\left(1 + \frac{\mathrm{e}^2knq^j}{2^4} + \sum_{c=3}^{k}\frac{1}{c^2}\left(\frac{\mathrm{e}^2knq^j}{c^2}\right)^{c-1}\right) \\ &\le \mathrm{e}^2knq^j\left(1 + \frac{\mathrm{e}^2}{16} + \frac{1}{9}\sum_{c=3}^{\infty}\left(\frac{\mathrm{e}^2}{9}\right)^{c-1}\right) \\ &\le 5\mathrm{e}^2knq^j. \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{j=\alpha\ln nk}^{m}\binom{m}{j}s^{m-j}t^ju^{jk}\sum_{c=1}^{k}\left(\frac{\mathrm{e}^2knq^j}{c^2}\right)^c &\le \sum_{j=\alpha\ln nk}^{m}\binom{m}{j}s^{m-j}t^ju^{jk}5\mathrm{e}^2knq^j \\ &\le 5\mathrm{e}^2kn\sum_{j=0}^{m}\binom{m}{j}s^{m-j}(tu^{k-1})^j \\ &= 5\mathrm{e}^2kn(s+tu^{k-1})^m \\ &= 5\exp\left(\ln nk + m\ln(s+tu^{k-1})\right) \end{aligned}$$

To make this small requires

$$m > (1+\delta)\frac{1}{-\ln(s+tu^{k-1})}\ln nk. \tag{12}$$

In order to compare the condition in (12) to (9) and (11), note that for any $x, y \in (0,1)$,

$$-\ln(1-x(1-\mathrm{e}^{-y})) \le xy.$$

The above inequality can be seen, for example, since for each y, the function $f_y(x) = xy + \ln(1 - x(1 - \mathrm{e}^{-y}))$ is concave for $x \in [0, 1]$ with $f_y(0) = 0 = f_y(1)$. Thus, since $s + tu^{k-1} = 1 - 2\mathrm{e}^{-\alpha}(1 - \mathrm{e}^{-\alpha/k})$, then

$$\frac{-1}{\ln(s + tu^{k-1})} = \frac{-1}{\ln(1 - 2\mathrm{e}^{-\alpha}(1 - \mathrm{e}^{-\alpha/k}))} \geq \frac{k}{2\mathrm{e}^{-\alpha}\alpha}.$$

Thus, condition (12) is always stronger than (9) and one can see that when k tends to infinity, the two conditions are asymptotically equal.

Hence from (9), (11), and (12) our requirements are

$$m > (1 + \delta)M_1(n, k, \alpha) \qquad m > (1 + \delta)M_2(n, k, \alpha).$$

From the above, we can optimise this result over α. Noting that M_1 is minimised at $\alpha = \ln 2$ and M_2 is minimised at $\alpha = 1$, it is sufficient to just consider $\alpha \in [\ln 2, 1]$.

This proves Lemma 9. □ □

B Explicit bounds on rate

Here we give the proofs of Corollaries 12 and 13.

Proof of Corollary 12. The bound on R follows from Theorem 11 by a careful choice of α in terms of β.

In order to simplify some of the expressions that follow, define $y = y(\alpha) = 1 - 2\mathrm{e}^{-\alpha}$ and $t = 1 - \frac{\beta}{2-\beta}$. Then, for $\alpha \in [\ln 2, 1]$ we have $y \in [0, 1 - 2/\mathrm{e}]$ and as β tends to 1, t tends to 0. Further, the expressions in Theorem 11 can be simplified as

$$-\ln(1 - 2\mathrm{e}^{-\alpha} + 2\mathrm{e}^{-2\alpha}) = -\ln\left(\frac{1}{2}(1 + y^2)\right) = \ln 2 - \ln(1 + y^2)$$

and

$$2\alpha\mathrm{e}^{-\alpha}\frac{\beta}{2 - \beta} = (1-y)\left(-\ln\left(\frac{(1 - y)}{2}\right)\right)(1-t) = (1-y)\left(\ln 2 - \ln(1 - y)\right)(1-t).$$

Thus, the result of Theorem 11 can be restated as

$$R \geq \frac{1}{\ln 2}\min_{y\in[0,1-2/\mathrm{e}]}\left\{\ln 2 - \ln(1 + y^2), (1 - t)(1 - y)\left(\ln 2 - \ln(1 - y)\right)\right\} \quad (13)$$

The desired result then follows from equation (13) by choosing

$$y = \frac{\ln 2}{1 - \ln 2} \cdot \frac{t}{1 - t}. \quad (14)$$

Note that, by the definition of t, $\frac{t}{1-t} = \frac{2(1-\beta)}{\beta}$.

What remains is to show that for y given by equation (14),

$$\ln 2 - \ln(1 + y^2) \leq (1 - y)(1 - t)\left(\ln 2 - \ln(1 - y)\right). \quad (15)$$

For y given by equation (14), the right-hand side of equation (15) is

$$\begin{aligned}
&(1-y)(1-t)\left(\ln 2 - \ln(1-y)\right) \\
&\quad = (1-y)(1-t)(\ln 2 + y) - (1-y)(1-t)(y + \ln(1-y)) \\
&\quad = (1-t)\ln 2 + y(1-t)(1-\ln 2) - (1-t)y^2 \\
&\qquad - (1-y)(1-t)(y + \ln(1-y)) \\
&\quad = (1-t)\ln 2 + t\ln 2 - (1-t)y^2 \\
&\qquad - (1-y)(1-t)(y + \ln(1-y)) && \text{(by eq. (14))} \\
&\quad = \ln 2 - (1-t)(y^2 + (1-y)y + (1-y)\ln(1-y)) \\
&\quad = \ln 2 - (1-t)(y + (1-y)\ln(1-y)) \\
&\quad = \ln 2 - \left(\frac{\ln 2}{\ln 2 + y(1-\ln 2)}\right)(y + (1-y)\ln(1-y)) && \text{(by eq. (14))}.
\end{aligned}$$

Thus, in order to show that the inequality in (15) holds, it suffices to show that for all $y \in [0,1]$,

$$y + (1-y)\ln(1-y) \le \left(1 + \frac{y(1-\ln 2)}{\ln 2}\right)\ln(1+y^2) \tag{16}$$

The inequality in (16) is shown by considering separately the cases $y \le 1/2$ and $y > 1/2$.

Consider first the case $y \le 1/2$. Using the fact that $\ln(1-y) < -y$ and

$$\ln(1+y^2) \ge y^2 - y^4/2 \ge y^2 - y^3/4 = y^2(1-y/4).$$

Then,

$$y + (1-y)\ln(1-y) < y^2$$

and for all $y \in [0, 1/2]$,

$$1 \le \left(1 + y\frac{(1-\ln 2)}{\ln 2}\right)\left(1 - \frac{y}{4}\right).$$

Thus, for $y \le 1/2$,

$$\begin{aligned}
y + (1-y)\ln(1-y) \le y^2 &\le y^2\left(1 + y\frac{(1-\ln 2)}{\ln 2}\right)\left(1 - \frac{y}{4}\right) \\
&\le \left(1 + y\frac{(1-\ln 2)}{\ln 2}\right)\ln(1+y^2).
\end{aligned}$$

Consider now the inequality from (16) in the case $y \ge 1/2$. Note that for all $y \in [0,1]$,

$$\ln(1+y^2) \ge \ln 2 - (1-y).$$

The above inequality can be seen to be true since it holds for $y = 0$ and $y = 1$ and $\ln(1+y^2)$ is concave. Thus, in order to prove the inequality in (16), it suffices to show that for $y \in [1/2, 1]$,

$$y + (1-y)\ln(1-y) \le \left(1 + y\frac{(1-\ln 2)}{\ln 2}\right)(\ln 2 - (1-y)). \tag{17}$$

Again, the inequality in equation (17) can be seen to be true since it holds for $y = 1/2$ and $y = 1$ and the function

$$\begin{aligned}
&\left(1+y\frac{(1-\ln 2)}{\ln 2}\right)(\ln 2-(1-y))-y-(1-y)\ln(1-y)\\
&\quad=\ln 2+y(1-\ln 2)-1+y-y\frac{(1-\ln 2)}{\ln 2}+y^2\frac{(1-\ln 2)}{\ln 2}\\
&\qquad-y-(1-y)\ln(1-y)\\
&\quad=(\ln 2-1)+y(1-\ln 2)\left(1-\frac{1}{\ln 2}\right)+y^2\frac{(1-\ln 2)}{\ln 2}-(1-y)\ln(1-y)
\end{aligned}$$

is concave for $y \in [0, 1]$. □ □

Next, is the proof of Corollary 13.

Proof of Corollary 13. For $\beta < \beta_1$, the result follows from Theorem 11 by substituting $\alpha = 1$ and noting that the inequality

$$\frac{2\beta}{\mathrm{e}(2-\beta)} \leq -\ln(1-2/\mathrm{e}+2/\mathrm{e}^2)$$

holds exactly when $\beta < \beta_0$.

For $\beta \geq \beta_1$, the result follows from Corollary 12 by noting that $\beta_1 > \frac{2\ln 2}{1+\ln 2}$. □ □

In Corollaries 12 and 13, a better bound for the case $\beta > \beta_0$ can be obtained by substituting in Theorem 11, α chosen so that

$$1-2\mathrm{e}^{-\alpha} = \frac{-\beta(1-\ln 2)+\sqrt{\beta^2(1-\ln 2)^2+4(1-\beta)(4-3\beta)\ln 2}}{4-3\beta},$$

but the expression obtained does not seem simpler than statement of Theorem 11 itself.

References

[1] M Aldridge, L Baldassini, and O Johnson. Group testing algorithms: bounds and simulations. *IEEE Transactions on Information Theory*, **60**:6, 3671–3687, 2014.

[2] GK Atia and V Saligrama. Boolean compressed sensing and noisy group testing. *IEEE Transactions on Information Theory*, **58**:3, 1880–1901, 2012.

[3] L Baldassini, O Johnson, and M Aldridge. The capacity of adaptive group testing. *2013 IEEE International Symposium on Information Theory Proceedings*, 2676–2680, 2013.

[4] CL Chan, S Jaggi, V Saligrama, and S Agnihotri. Non-adaptive group testing: explicit bounds and novel algorithms. *IEEE Transactions on Information Theory*, **60**:5, 3019–3035, 2014

[5] H-B Chen and FK Hwang. Exploring the missing link among d-separable, $\overline{d}$-separable and d-disjunct matrices. *Discrete Applied Mathematics*, **155**:5, 662—664, 2007.

[6] R Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, **14**:4, 436–440, 1943.

[7] D-Z Du and FK Hwang. *Combinatorial Group Testing and Applications*, second edition. Series on Applied Mathematics, **12**, World Scientific, 2000.

[8] AG D'yachkov and VV Rykov. Bounds on the length of disjunctive codes. *Problems of Information Transmission*, **18**:3, 166—171, 1982.

[9] P Erdős and L Moser. Problem 35. *Proceedings on the Conference of Combinatorial Structures and their Applications*, Gordon and Breach, 1970.

[10] Z Füredi. On r-cover-free families. *Journal of Combinatorial Theory, Series A*, **73**:1, 172–173, 1996.

[11] WH Kautz and RC Singleton. Nonrandom binary superimposed codes. *IEEE Transaction on Information Theory*, **10**:4, 363–377, 1964.

[12] A Macula, V Rykov and S Yekhanin. Trivial two-stage group testing for complexes using almost disjunct matrices. *Discrete Applied Mathematics*, **137**:1, 97–107, 2004.

[13] D Malioutov and M Malyutov. Boolean compressed sensing: Lp relaxation for group testing. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3305—3308, 2012.

[14] MB Malyutov. The separating property of random matrices. *Mathematical Notes of the Academy of Sciences of the USSR*, **23**:1, 84–91, 1978.

[15] M Malyutov. Search for sparse active inputs: a review. In H Aydinian, F Cicalese, and C Deppe (Eds), *Information Theory, Combinatorics and Search Theory* Lecture notes in Computer Science, **7777**, Springer, 609–647, 2013.

[16] A Mazumdar. On almost disjunct matrices for group testing. *Algorithms and Computation*, Lecture Notes in Computer Science, **7676**, 649–658, 2012.

[17] E Porat and A Rothschild. Explicit Non-Adaptive Combinatorial Group Testing Schemes. In L Aceto, I Damgard, LA Goldberg, MM Halldorsson, A Ingolfsdottir and I Walukiewicz (Eds), *ICALP 2008*, Lecture Notes in Computer Science, **5125**, 748–759, 2008.

[18] M Ruszinkó. On the upper bound of the size of r-cover-free families. *Journal of Combinatorial Theory, Series A*, **66**:2, 302–310, 1994.

[19] A Sebő. On two random search problems. *Journal of Statistical Planning and Inference*, **11**:1, 23–31, 1985.

[20] D Sejdinovic and OT Johnson. Note on noisy group testing: asymptotic bounds and belief propagation reconstruction. *Proceedings of the 48th Annual Allerton Conference on Communication, Control and Computing*, 998–1003, 2010.

[21] T Wadayama. An analysis on non-adaptive group testing based on sparse pooling graphs. *2013 IEEE International Symposium on Information Theory*,2681—2685, 2013.

[22] A Zhigljavsky. Probabilistic existence theorems in group testing. *Journal of Statistical Planning and Inference*,**115**:1, 1–43, 2003.