The optimal number of tests

This lecture focuses on the most important objective for group testing strategies: minimizing the number of tests. Given $1 \leq d \leq N-1$, let $t(d,N)$ denote the minimum t for which a d-disjunct matrix with t rows and N columns exists. We study the behavior of the function $t(d,N)$. We shall also briefly introduce basic concepts and results from coding theory which will be used to construct good disjunct matrices.

1 Lower bounds

Exercise 1.1. Show that $t(1,N) \geq \min\{3,N\}$, for all $N \geq 1$.

1.1 Large d

The following 1975 result was attributed to Bassalygo by Dyachkov and Rykov [3].

Proposition 1.2 (Bassalygo – 1975)**.** *For the d-disjunct matrices, we have the following bound*

$$t(d,N) \geq \min\left\{\binom{d+2}{2}, N\right\}. \tag{1}$$

Proof. Exercise 1.1 proves the base case. For the induction step, consider $d \geq 2$ and a d-disjunct matrix $\mathbf{M}$ with $t = t(d,N)$ rows and N columns. Let $N(w)$ denote the number of columns of $\mathbf{M}$ with weight w. A row $i \in [t]$ is said to be *private* for a column j if j is the only column in the matrix having a 1 on row i. If column $\mathbf{M}^j$ has weight at most d, then it must have at least one private element. The total number of private elements of all columns is at most t. Hence,

$$\sum_{w=1}^{d} N(w) \leq t.$$

Let $w_{\max}$ denote the maximum column weight of $\mathbf{M}$. If $w_{\max} \leq d$ then $N = \sum_w N(w) \leq t$. Now, suppose $w_{\max} \geq d+1$ and consider a column $\mathbf{M}^j$ with weight equal to $w_{\max}$. If we remove column $\mathbf{M}^j$ and all rows i for which $m_{ij} = 1$, we are left with a $(d-1)$-disjunct matrix with $t - w_{\max}$ rows and $N-1$ columns. Thus, $t - w_{\max} \geq t(d-1, N-1)$ which along with the induction hypothesis implies

$$t - (d+1) \geq \min\left\{\binom{d+1}{2}, N-1\right\}.$$

The bound is thus proved. □

Note that $t(d,N) \leq N$ is a trivial upper bound: the $N \times N$ identity matrix is d-disjunct. If $\binom{d+2}{2} \geq N$ then $t(d,N) \geq N$, by the Bassalygo bound. Hence, when $d \geq \sqrt{2N}$ we cannot do better than testing items individually: $t(d,N) = N$.

1.2 Small d

Consider a $t \times N$ binary matrix $\mathbf{M}$. Its columns can naturally be viewed as a family of subsets of $[t]$. The collection of columns of a 1-disjunct matrix satisfies the property that no set in the family is contained in another set in the family. Such a family is called an *anti-chain* is partially order set theory [1]. A classic (topology) lemma by Sperner in 1928 [4, 14] states that the maximum size of such an anti-chain is $\binom{t}{\lfloor t/2 \rfloor}$. Since the proof of Sperner's lemma is short and illustrates a nice (probabilistic) technique, we reproduce it here.

Lemma 1.3 (Sperner's Lemma)**.** *Let $\mathcal{F}$ be a collection of subsets of $[t]$ such that no member of $\mathcal{F}$ is contained in another member of $\mathcal{F}$. Then, $|\mathcal{F}| \leq \binom{t}{\lfloor t/2 \rfloor}$. Equality can be reached by picking $\mathcal{F}$ to be the collection of all $\lfloor t/2 \rfloor$-subsets of $[t]$.*

Proof. Pick a random permutation π of $[t]$, uniformly. For each member $F \in \mathcal{F}$, let A_F be the event that F is a prefix of π. For example, if $\pi = 3, 4, 1, 5, 2$ then $\{1, 3, 4\}$ is a prefix a π. If $|F| = k$ then The probability that A_F holds is

$$\text{Prob}[A_F] = \frac{k!(t-k)!}{t!} = \frac{1}{\binom{t}{k}} \geq \frac{1}{\binom{t}{\lfloor t/2 \rfloor}}.$$

Because no member of $\mathcal{F}$ is contained in another, the events A_F are all mutually exclusive. Thus,

$$1 \geq \sum_{F \in \mathcal{F}} \text{Prob}[A_F] \geq |\mathcal{F}| \cdot \frac{1}{\binom{t}{\lfloor t/2 \rfloor}}.$$

□

A subset $F \subseteq [t]$ is called a *private subset* of column $\mathbf{M}^j$ if $F \subseteq \mathbf{M}^j$ and $F \not\subseteq \mathbf{M}^{j'}$ for any $j' \neq j$. In order to prepare for a bound for the small d case (say $d < \sqrt{2N}$), we need the following lemma (Lemma 9.1 from Erdős-Frankl-Füredi [5]).

Lemma 1.4. *Let $\mathbf{M}$ be a $t \times N$ d-disjunct matrix. Fix a positive integer $w \leq t$. Let $\mathcal{C}$ denote the set of all columns of $\mathbf{M}$. Let C be any column in $\mathcal{C}$ which has no private w-subset. Consider any $k \geq 0$ other columns $C_1, \cdots, C_k \in \mathcal{C}$. We have*

$$\left| C \setminus \bigcup_{j=1}^{k} C_j \right| \geq (d-k)w + 1. \tag{2}$$

In particular, if $\mathbf{M}$ has at least $d+1$ columns $C_1, \ldots, C_{d+1}$ none of which have any private w-subset, then

$$\left| \bigcup_{j=1}^{d+1} C_j \right| \geq \frac{1}{2}(d+1)(dw+2). \tag{3}$$

Proof. If $\left| C \setminus \bigcup_{j=1}^{k} C_j \right| \leq (d-k)w$ then C can be covered by the k sets C_j, $j \in [k]$, plus $(d-k)$ other sets because C has no private w-subset. This contradicts the fact that C cannot be covered by the union of

any d sets. To see (3), we apply (2) as follows.

$$\begin{aligned}\left|\bigcup_{j=1}^{d+1} C_j\right| &= |C_1| + |C_2 \setminus C_1| + \cdots + |C_{d+1} \setminus C_1 \cup \cdots \cup C_d| \\ &\geq (dw+1) + ((d-1)w+1) + \cdots + (w+1) + 1 \\ &= \frac{d}{2}(d+1)w + (d+1) \\ &= \frac{1}{2}(d+1)(dw+2).\end{aligned}$$

□

Theorem 1.5. *For $N \geq d \geq 2$ and any d-disjunct matrix $\mathbf{M}$ with t rows and N columns, we have*

$$N \leq d + \binom{t}{\left\lceil \frac{t-d}{\binom{d+1}{2}} \right\rceil}.$$

Proof. Let $\mathcal{C}_w$ be the sub-collection of columns of $\mathbf{M}$ each of which has a private w-subset, and $\mathcal{C}_{<w}$ be the sub-collection of columns of $\mathbf{M}$ each of which has weight $< w$. Then, the same technique used in the proof of Sperner's lemma above can be used to show that, for any $w \leq t/2$, $|\mathcal{C}_w| + |\mathcal{C}_{<w}| \leq \binom{t}{w}$. Now, if there were at least $d+1$ columns **not** in $\mathcal{C}_w \cup \mathcal{C}_{<w}$, then by Lemma 1.4 the union of columns not in $\mathcal{C}_w \cup \mathcal{C}_{<w}$ is at least $\frac{1}{2}(d+1)(dw+2)$. Suppose we choose w such that

$$\frac{1}{2}(d+1)(dw+2) \geq t+1, \tag{4}$$

then we reach a contradiction and thus we can conclude that $n \leq d + \binom{t}{w}$. The minimum w for which (4) holds is $w = \left\lceil \frac{t+1-(d+1)}{\binom{d+1}{2}} \right\rceil$, which is at most $t/2$ when $d \geq 2$. □

Exercise 1.6. Show the missing piece in the above proof that, for any $w \leq t/2$, $|\mathcal{C}_w| + |\mathcal{C}_{<w}| \leq \binom{t}{w}$.

Corollary 1.7. *When $\binom{d+2}{2} < N$, we have*

$$t(d,N) \geq \frac{(d+1)^2}{24 \log d} \log N = \Omega\left(\frac{d^2}{\log d} \log N\right).$$

2 Codes

2.1 Preliminaries

Let Σ be a finite set, $|\Sigma| \geq 2$. We will refer to elements of Σ as *symbols* or *letters*, and Σ as an *alphabet*. A *code* C over alphabet Σ is a subset of Σ^n, where the positive integer n is called the *length* (or *block length*) and $|C|$ is the *size* of the code. Each member of C is called a *codeword*. Thus, a codeword is a vector of dimension n, each of whose coordinates is also called a *position*.

The *Hamming distance* between two codewords $\mathbf{c}$ and $\mathbf{c}'$, denoted by $\Delta(\mathbf{c}, \mathbf{c}')$ is the number of positions where $\mathbf{c}$ and $\mathbf{c}'$ are different. The *minimum distance* of a code C, denoted by $\Delta(C)$, is the minimum Hamming distance between two different codewords of C. The *dimension* of a code C on alphabet Σ is defined to be $\dim(C) := \log_{|\Sigma|} |C|$. A code with length n and dimension k on an alphabet of size q is called an $(n, k)_q$-code. An $(n, k)_q$-code with minimum distance Δ is called an $(n, k, \Delta)_q$-code. Sometimes, to emphasize a specific alphabet in use, we use the notations $(n, k)_\Sigma$ and $(n, k, \Delta)_\Sigma$.

Proposition 2.1 (Singleton Bound [13])**.** *For any $(n, k, \Delta)_q$-code, $k \leq n - \Delta + 1$.*

A code achieving equality in the Singleton bound is called a *Maximum distance separable* code, or MDS code. A very widely used MDS code is the celebrated *Reed-Solomon code*, named after its two inventors Irving Reed and Gustave Solomon [12][1].

Exercise 2.2. Prove the Singleton bound. (**Hint**: consider any code C of minimum distance Δ and length n. Let C' be the projection of C on to the first $n - (\Delta - 1)$ coordinates. Note that $|C'| = |C|$. Bound $|C'|$.)

It is often the case that the alphabet Σ is a finite field $\mathbb{F}_q$[2], because then we are able to take advantage of the underlying (linear) algebraic structures for designing the codes, analyzing its parameters, and discovering good encoding and decoding algorithms. In this case, when C is a linear subspace of $\mathbb{F}_q^n$ we call C a *linear code*. To emphasize the fact that C is linear, we replace $(n, k)_q$ and $(n, k, \Delta)_q$ by $[n, k]_q$ and $[n, k, \Delta]_q$. Note that the dimension k of the code is now precisely the dimension of the subspace C.

2.2 Reed-Solomon Codes

Definition 2.3 (Reed-Solomon code)**.** Let $k \leq n \leq q$ be positive integers where q is a prime power. The *Reed-Solomon code* is an $[n, k, n - k + 1]_q$-code (i.e. a linear MDS code) defined as follows. Let $\{\alpha_1, \cdots, \alpha_n\}$ be any n distinct members of $\mathbb{F}_q$. These are called *evaluation points* of the code. For each vector $\mathbf{m} = (m_0, \ldots, m_{k-1}) \in \mathbb{F}_q^k$, define a polynomial

$$f_{\mathbf{m}}(x) = \sum_{i=0}^{k-1} m_i x^i$$

which is of degree at most $k-1$. Then, for each $\mathbf{m} \in \mathbb{F}_q^k$ there is a corresponding codeword $RS(\mathbf{m})$ defined by

$$RS(\mathbf{m}) = \langle f_{\mathbf{m}}(\alpha_1), \cdots, f_{\mathbf{m}}(\alpha_n) \rangle.$$

Exercise 2.4 (Optional)**.** Prove the following

1. If $\mathbf{m} \neq \mathbf{m}'$ then $RS(\mathbf{m}) \neq RS(\mathbf{m}')$. Thus, the RS code defined above has precisely q^k codewords.
2. For any $\mathbf{m}, \mathbf{m}' \in \mathbb{F}_q^k$, and any scalar $a \in \mathbb{F}_q$,

$$\begin{aligned} RS(\mathbf{m} + \mathbf{m}') &= RS(\mathbf{m}) + RS(\mathbf{m}') \\ RS(a\mathbf{m}) &= a \cdot RS(\mathbf{m}). \end{aligned}$$

Thus, the RS code is a linear code.

[1] This paper and the likes of Shannon and Hamming's papers are perfect examples illustrating that we don't have to write huge papers to be influential.

[2] See, `http://www.cs.cmu.edu/~venkatg/teaching/codingtheory/notes/algebra-brief-notes.pdf` for a brief introduction to finite fields

3. Use that fact that any polynomial of degree at most $k-1$ over $\mathbb{F}_q$ has at most $k-1$ roots to show that, for any $\mathbf{m} \neq \mathbf{m}'$ the Hamming distance between $RS(\mathbf{m})$ and $RS(\mathbf{m}')$ is at least $n-k+1$.

4. Lastly, consider the distance between the all-zero codeword and the codeword corresponding to the polynomial $\prod_{i=1}^{k-1}(x-\alpha_i)$, prove that the above RS code is an $[n, k, n-k+1]_q$-code. (We could also use the Singleton bound to show this fact.)

Last but not least, to see that the $[n, k]_q$-RS code is strongly explicit, note that

$$RS(\mathbf{m}) = (m_0, \ldots, m_{k-1}) \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_n \\ \alpha_1^2 & \alpha_2^2 & \cdots & \alpha_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_1^{k-1} & \alpha_2^{k-1} & \cdots & \alpha_n^{k-1} \end{pmatrix}$$

The matrix is called the $k \times n$ *Vandermonde matrix*, which occurs in many other applications such as in computing the FFT.

2.3 Code concatenation

Let q, n, m, N be integers such that $N \leq q^n$ and $2^m \geq q$. Let C_{out} be a code of length n and size N over an alphabet Σ of size q. Without loss of generality (up to isomorphism) we might as well set $\Sigma = [q]$. Let C_{in} be a binary code (i.e. alphabet $\{0, 1\}$) of length m and size q. A *concatenation* $C = C_{\text{out}} \circ C_{\text{in}}$ of C_{out} and C_{in} is a code C of length mn and size N constructed by replacing each symbol a of a codeword in C by the ath codeword in C_{in}. Here, we order the codewords in C_{in} in an arbitrary manner.

For example, consider the case when $n = q = 3$, $m = 2$

$$C_{\text{out}} = \left\{ \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ 2 \\ 3 \end{bmatrix} \right\}, \quad C_{\text{in}} = \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}.$$

Then,

$$C_{\text{out}} \circ C_{\text{in}} = \left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}.$$

Abusing notation, we often also state that a code is a matrix which is constructed by putting all codewords of the code as columns of the matrix in any order. For example, the matrix $\mathbf{M} = C_{\text{out}} \circ C_{\text{in}}$ above is

$$\mathbf{M} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

In the concatenation $C_{\text{out}} \circ C_{\text{in}}$, C_{out} is called the *outer code* and C_{in} the inner code. By instantiating the outer and inner codes with carefully chosen codes, we obtain good group testing matrices. Some of the constructions are illustrated in this lecture. We will have more examples of code concatenation in later lectures.

One of the most basic inner code is the trivial *identity code*, ID_q, which is the binary code of length q and size q whose ith codeword is the ith standard basis vector. The corresponding matrix is the identity matrix of order q.

2.4 Gilbert-Varshamov Bound

Let $A_q(n, \Delta)$ denote the maximum size of a q-ary code of length n and minimum distance Δ. Determining $A_q(n, \Delta)$ is a major (open) problem in coding theory. Define

$$\text{Vol}_q(n, l) = \sum_{j=0}^{l} \binom{n}{j} (q-1)^j$$

to be the "volume" of the Hamming ball of radius l around any codeword, i.e. the number of vectors of distance at most l from a given vector in $\mathbb{F}_q^n$. Gilbert [6] and Varshamov [15] proposed a simple greedy algorithm which constructs a linear code with size at least $q^n/\text{Vol}_q(n, \Delta - 1)$. Actually, Gilbert's algorithm does not produce a linear code; Varshamov's does. However, their algorithms are very similar and achieves similar bounds.

Theorem 2.5 (Gilbert-Varshamov Bound)**.** *The maximum size of a code of length n, alphabet size q, and distance Δ satisfies*

$$A_q(n, \Delta) \geq \frac{q^n}{\text{Vol}_q(n, \Delta - 1)} = \frac{q^n}{\sum_{j=0}^{\Delta-1} \binom{n}{j}(q-1)^j}.$$

There also exists linear codes achieving the bound.

Exercise 2.6 (Gilbert algorithm)**.** Consider the following algorithm for code construction. Let Σ be an alphabet of size q. Initially let $C = \emptyset$. While there still exists a vector $\mathbf{c}\Sigma^n$ which is of distance at least Δ from all the codewords in C, add $\mathbf{c}$ into C. Prove that when the algorithm stops, we obtain a code C whose size is at least $q^n/\text{Vol}_q(n, \Delta - 1)$.

To show that there exist linear codes attaining the Gilbert-Varshamov (GV) bound, we use the probabilistic method. In fact, we will prove the asymptotic form of the GV bound for linear codes.

Definition 2.7 (q-ary entropy)**.** Let $q \geq 2$ be an integer. The *q-ary entropy function* $H_q : [0, 1] \to \mathbb{R}$ is defined by

$$H_q(\delta) := \delta \log_q \frac{q-1}{\delta} + (1-\delta) \log_q \frac{1}{1-\delta}. \tag{5}$$

When $q = 2$, we drop the subscript q and write the famous (Shannon) *binary entropy function* as

$$H(\delta) = \delta \log \frac{1}{\delta} + (1-\delta) \log \frac{1}{1-\delta}.$$

Sometimes, it might be easier grasp if we re-write (5) by

$$H_q(\delta) = \delta \log_q(q-1) - \delta \log_q \delta - (1-\delta) \log_q(1-\delta).$$

We define $H_q(0) = 0$. The function $H_q(x)$ is continuous in the interval $[0, 1]$, is increasing from 0 to $1-1/q$, and decreasing from $1 - 1/q$ to 1. The following lemma can be shown with careful analysis using Stirling approximation.

Lemma 2.8. *For any positive integers $n, q \geq 2$ and real number $0 \leq \delta \leq 1 - 1/q$,*

$$q^{n(H_q(\delta)-o(1))} \leq \text{Vol}_q(n, \delta n) \leq q^{nH_q(\delta)}.$$

A central problem in coding theory is to characterize the tradeoff between the distance and the rate of a code. The *relative distance* $\delta(C)$ of a code C of length n is $\Delta(C)/n$. If C has dimension k then its *rate* is defined to be $R(C) = k/n$.

Theorem 2.9 (Asymptotic form of GV bound)**.** *Let $q \geq 2$ be an integer. For any $0 \leq \delta \leq 1 - 1/q$, there exists an infinite family of q-ary codes with rate $R \geq 1 - H_q(\delta) - o(1)$. In fact, such code exists for all sufficiently large length n.*

We will in fact prove the linear code version of the above bound.

Exercise 2.10. Show that for a linear code the minimum distance is equal to the minimum weight of a non-zero codeword. (The *weight* of a codeword is the number of non-zero entries.)

Exercise 2.11. For positive integers $k < n$, let $\mathbf{G}$ be a random $k \times n$ matrix chosen by picking each of its entries from $\mathbb{F}_q$ uniformly and independently. Fix a vector $\mathbf{y} \in \mathbb{F}_q^k$. Prove that the vector $\mathbf{yG}$ is a uniformly random vector in $\mathbb{F}_q^n$.

Theorem 2.12 (Linear code version of the asymptotic form of the GV bound)**.** *Let $q \geq 2$ be any prime power. Let $0 \leq \delta < 1 - 1/q$, and $0 < \epsilon < H_q(\delta)$ be arbitrary reals, and n be any sufficiently large integer. For any integer $k \leq (1 - H_q(\delta))n$ there exists an $[n, k, \delta n]_q$-code.*

Proof. We want a k-dimensional linear subspace C of $\mathbb{F}_q^n$ where the minimum weight of non-zero codewords is at least $\Delta = \delta n$. The subspace can be generated by a $k \times n$ matrix $\mathbf{G}$ of rank k, called the *generator matrix* for the code. The rows of $\mathbf{G}$ form a basis for the subspace. We pick a random generator matrix $\mathbf{G}$ and show that it satisfies two properties with positive probability:

(a) $\mathbf{G}$ has full row rank, and

(b) for every non-zero vector $\mathbf{y} \in \mathbb{F}_q^k$, $\mathbf{yG}$ has weight at least Δ.

Let $\text{wt}(\mathbf{x})$ denote the weight of vector $\mathbf{x}$. Actually, property (b) implies property (a) because if the rows of $\mathbf{G}$ are linearly dependent then there is some non-zero vector $\mathbf{y}$ for which $\mathbf{yG} = \mathbf{0}$.

To pick the random matrix $\mathbf{G}$, we simply pick each of its entry from $\mathbb{F}_q$ uniformly and independently. For any fixed non-zero vector $\mathbf{y} \in \mathbb{F}_q^k$, $\mathbf{yG}$ is a uniformly random vector in $\mathbb{F}_q^n$. Hence, by Lemma 2.8

$$\text{Prob}\left[\text{wt}(\mathbf{yG}) \leq \delta n\right] = \frac{\text{Vol}_q(n, \delta n)}{q^n} \leq q^{(H_q(\delta)-1)n}.$$

Now, taking a union bound over all non-zero vectors $\mathbf{y} \in \mathbb{F}_q^k$, the probability that $\text{wt}(\mathbf{yG}) \leq \delta n$ for some $\mathbf{y}$ is at most

$$(q^k - 1)q^{(H_q(\delta)-1)n} < q^{(1-H_q(\delta))n} q^{(H_q(\delta)-1)n} = 1.$$

□

3 Upper bounds and constructions

3.1 A greedy algorithm

Let S be the collection of all binary row vectors of length N, each with weight w (i.e. each vector has w non-zero entries), where $w \leq N - d$. We construct a d-disjunct matrix $\mathbf{M}$ by picking members of S to be rows of $\mathbf{M}$. For $\mathbf{M}$ to be d-disjunct, for each $j \in [N]$ and each d-subset $A \in \binom{[N]}{d}$, $j \notin A$, we want to pick a row $\mathbf{s} \in S$ such that $s_j = 1$ and $\mathbf{s}|_A = \mathbf{0}$, in which case we say that $\mathbf{s}$ "covers" (j, A). The main objective is to pick a small subset of S which covers all $(d+1)\binom{N}{d+1}$ possible pairs (j, A). This is a special case of the SET COVER problem where each "set" is a member of S and the universe consists of pairs (j, A) as described.

A natural algorithm for the SET COVER problem is to pick a set which covers as many uncovered elements as possible, then remove all covered elements and repeat until all elements are covered. This is the greedy algorithm for SET COVER. A classic result by Lovasz [9] (and independently by Chvatal [2]) implies that the greedy algorithm finds a set cover for all the (j, A) of size at most

$$\begin{aligned}
t &\leq \frac{\binom{N}{w}}{\binom{N-d-1}{w-1}} \left(1 + \ln\left(w\binom{N-w}{d}\right)\right) \\
&= \frac{N-d}{w} \cdot \frac{N}{N-d} \cdot \frac{N-1}{N-d-1} \cdots \frac{N-w+1}{N-d-w+1} \left(1 + \ln\left(w\binom{N-w}{d}\right)\right) \\
&< \frac{N-d}{w} \left(\frac{N-w+1}{N-d-w+1}\right)^w \left(1 + \ln w + d\ln\left(\frac{e(N-w)}{d}\right)\right) \\
&= \frac{N-d}{w} \left(1 + \frac{d}{N-d-w+1}\right)^w \left(1 + d + \ln w + d\ln\left(\frac{(N-w)}{d}\right)\right) \\
&< \frac{N-d}{w} e^{\frac{dw}{N-d-w+1}} \left(1 + d + \ln w + d\ln\left(\frac{(N-w)}{d}\right)\right).
\end{aligned}$$

This fact can also be seen from the *dual-fitting* analysis of the greedy algorithm for SET COVER [16]. This set cover is exactly the set of rows of the d-disjunct matrix we are looking for. The final expression might seem a little unwieldy. Note, however, that for most meaningful ranges of w and d, the factor $\left(1 + d + \ln w + d\ln\left(\frac{(N-w)}{d}\right)\right)$ can safely be thought of as $O(d\ln(N/d))$. Last but not least, if $dw = O(N)$ then $e^{\frac{dw}{N-d-w+1}} = O(1)$ and the number of rows l is not exponential. Also, when $dw = \Theta(N)$ the overall cost is $t = O(d^2 \log(N/d))$, matching the best known bound for disjunct matrices. This optimality only applies when we are free to choose w in terms of N and d; in particular, when we have this freedom we will pick $w = \Theta(N/d)$.

Exercise 3.1. Suppose instead of applying the greedy algorithm, we simply pick independently each round a random member $\mathbf{s}$ of S to use as a row of $\mathbf{M}$. In expectation, how many rounds must be performed so that $\mathbf{M}$ is d-disjunct. You should set $w = N/d$, and assume $\binom{d+2}{2} < N$ as usual.

Hwang and Sós [7] gave a different greedy algorithm achieving asymptotically the same number of tests. There algorithms have running time $\Omega(N^d)$, and thus are not practical unless d is a small constant.

3.2 Concatenating a random code with the identity code

Let q, N, d, n be integers such that $q > d$, and $q^n \geq N$. Let C_{in} be the identity code ID_q. Let C_{out} be a random code of length n, size N, alphabet $[q]$ constructed as follows. We randomly select each codeword

$\mathbf{c}$ of C_{out} by picking uniformly a random symbol from $[q]$ for each position of $\mathbf{c}$ independently. Let $\mathbf{M} = C_{\text{out}} \circ C_{\text{in}}$. We bound the probability that $\mathbf{M}$ is not d-disjunct.

Let $j_0, \cdots, j_d$ be a fixed set of $d+1$ columns of $\mathbf{M}$. Then,

$$\text{Prob}[\text{codeword } \mathbf{M}^{j_0} \text{ is covered by } \mathbf{M}^{j_1}, \ldots, \mathbf{M}^{j_d}] \leq (d/q)^n.$$

Thus, by the union bound

$$\text{Prob}[\mathbf{M} \text{ is not } d\text{-disjunct}] \leq (d+1)\binom{N}{d+1}(d/q)^n.$$

We just proved the following.

Proposition 3.2. *Let q, N, d, n be integers such that $q > d$ and $q^n \geq N$. If*

$$(d+1)\binom{N}{d+1}(d/q)^n < 1$$

then there exists a d-disjunct matrix with qn rows and N columns.

Corollary 3.3. *When $\binom{d+2}{2} \leq N$, we have*

$$t(d, N) = O\left(d^2 \log(N/d)\right).$$

Proof. Set $q = 2d$, and $n = 10(d+1)\log_2(N/(d+1))$ in the previous proposition. Since $\binom{d+2}{2} \leq N$, we have $(d+1)e \leq N$ and thus $(N/(d+1))^2 \geq Ne/(d+1)$. Also, $(d+1) \leq (N/(d+1))^2$. Thus,

$$(d+1)\binom{N}{d+1} \leq (d+1)(Ne/(d+1))^{d+1} \leq (N/(d+1))^{2+2(d+1)} = 2^{2(d+2)\log_2(N/(d+1))} < 2^n = (q/d)^n.$$

□

Open Problem 3.4. The upperbound $O(d^2 \log(N/d))$ is only slightly larger than the best known lower bound $\Omega(d^2 \log N / \log d)$ of Corollary 1.7. Closing this gap is the major open question in group testing theory.

3.3 Concatenating the Reed-Solomon code with the identity code

Code concatenation seems like a neat little trick to construct disjunct matrices. However, how do we choose the inner and outer codes? What are the necessary and/or sufficient conditions required on the properties of the codes so that the concatenation is d-disjunct? We will derive a simple sufficient condition due to Kautz and Singleton [8] (this is the same Richard Collom Singleton of the Singleton bound fame mentioned above). Kautz and Singleton studied and constructed the so called *superimposed codes* which turn out to be equivalent to disjunct matrices. Their influential 1964 paper was also the first to give a strongly explicit construction of disjunct matrices with $t = O(d^2 \log^2 N)$, which we present in this section.

We first need a simple lemma which relates the weights of the codewords and their pairwise intersections to disjunctiveness. Two columns of a binary matrix "intersects" at a row if both columns contain a 1 on that row.

Lemma 3.5. *Let $\mathbf{M}$ be a binary matrix such that each column has weight at least w and every two different columns intersect at at most λ rows. Then, $\mathbf{M}$ is a d-disjunct matrix for any $d \leq (w-1)/\lambda$.*

Proof. Consider arbitrary columns $\mathbf{M}^{j_0}, \cdots, \mathbf{M}^{j_d}$ of $\mathbf{M}$. When $w \geq 1 + d\lambda$, there must be at least one row on which $\mathbf{M}^{j_0}$ has a 1 and the other d columns have 0. □

An analogous proof leads to the following lemma regarding concatenated codes.

Lemma 3.6. *Suppose C_{out} is an $(n, k, \Delta)_q$ code, and the matrix corresponding to C_{in} is d-disjunct, then $\mathbf{M} = C_{\text{out}} \circ C_{\text{in}}$ is d-disjunct if $n > d(n - \Delta)$.*

Proof. Consider arbitrary columns $\mathbf{M}^{j_0}, \cdots, \mathbf{M}^{j_d}$ of $\mathbf{M}$. Every two codewords of the outer code share symbols in at most $(n - \Delta)$ positions. Hence, there is a position $p \in [n]$ such that $\mathbf{M}^{j_0}$ has a symbol different from all the symbols of the $\mathbf{M}^{j_i}$, $i \in [d]$. Due to the fact that the inner-code matrix is d-disjunct, there is a row in $\mathbf{M}$ belonging to this position which $\mathbf{M}^{j_0}$ has a 1 and the other $\mathbf{M}^{j_i}$ all have 0. □

Open Problem 3.7. The above sufficient conditions might be a little too strong for many applications. Find a more relaxed condition.

Corollary 3.8. *Let $k \leq n \leq q$ be positive integers with q a prime power. Let C_{out} be the $[n, k]_q$-RS code, and C_{in} be the ID$_q$ code. Then, $\mathbf{M} = C_{\text{out}} \circ C_{\text{in}}$ is a strongly explicit d-disjunct matrix for any $d \leq (n-1)/(k-1)$.*

Proof. Note that ID$_q$ is d-disjunct when $q \geq d$, and that $n \geq 1 + d(n - (n - k + 1))$ is equivalent to $d \leq (n-1)/(k-1)$. □

Corollary 3.9. *Given $1 \leq d < N$, there is a strongly explicit d-disjunct matrix with N columns and $t = O(d^2 \log^2 N)$ rows.*

Proof. We want to pick parameters $k \leq n \leq q$ such that $d \leq (n-1)/(k-1)$ and $N \leq q^k$, and then apply the previous corollary. To make the calculation simpler, we replace the constraint $d \leq (n-1)/(k-1)$ by $d \leq n/k$. This replacement is OK because $n/k \leq (n-1)/(k-1)$.

Let us ignore the integrality issue for the moment. Suppose we pick $n = q$, and $\log N = k \log q$. Then, we need $d \log N / \log q \leq q$. Hence, we should pick q to be the smallest number such that $q \log q \geq d \log N$. Let's pick $n = q \approx \frac{2d \log N}{\log(d \log N)}$, and then set $k \approx \log N / \log q$. The overall number of tests is $qn \approx \Theta\left(\frac{d^2 \log^2 N}{\log^2(d \log N)}\right)$.

With the integrality issue taken into account, it is not hard to see that $t = \Theta\left(\frac{d^2 \log^2 N}{\log^2(d \log N)}\right)$ suffices. □

Nguyen and Zeisel [10] used Lemma 3.6 and a result by Zinoviev [17] to prove an interesting upper bound on $t(d, N)$. The main idea is to recursively apply Lemma 3.6 many times with suitably chosen parameters.

3.4 Porat-Rothschild's derandomization

Porat and Rothschild [11] derandomized the code construction in Theorem 2.12, and concatenated the resulting code with the identity code to obtain a polynomial time construction of d-disjunct matrices with $\min(d^2 \log N, N)$ rows. We will not describe the derandomization here. We will, however, briefly specify how such a construction can lead to $\min(d^2 \log N, N)$ rows.

First, if $d^2 \ln N \geq N$, then we can use the identity matrix. Hence, we can assume $d^2 \ln N < N$. Set $\delta = 1 - 1/(d+1)$. Let $q \in [2d, 4d)$ be a prime power, $k = \log_q N$ and $n = \frac{k}{(1 - H_q(\delta))} = \Theta(kd \ln d) = \Theta(d \log N)$. Now, use Theorem 2.12 to construct an $[n, k, \delta n]_q$-code and concatenate it (as an outer code)

with the identity inner code. Because $d(n - \delta n) < n$, by Lemma 3.6, the concatenated code is d-disjunct. The overall number of rows is $nq = O(d^2 \log N)$.

Exercise 3.10. Show that with $q \in [2d, 4d)$, and $\delta = 1 - 1/(d+1)$, we have $1 - H_q(\delta) = \Theta(\frac{1}{d \ln d})$.

Open Problem 3.11. We do not know of a strongly explicit construction of d-disjunct matrices with $t = O(d^2 \log N)$ tests.

References

[1] B. BOLLOBÁS, *Combinatorics*, Cambridge University Press, Cambridge, 1986. Set systems, hypergraphs, families of vectors and combinatorial probability.

[2] V. CHVÁTAL, *A greedy heuristic for the set-covering problem*, Math. Oper. Res., 4 (1979), pp. 233–235.

[3] A. G. D'YACHKOV AND V. V. RYKOV, *A survey of superimposed code theory*, Problems Control Inform. Theory/Problemy Upravlen. Teor. Inform., 12 (1983), pp. 229–242.

[4] K. ENGEL, *Sperner theory*, vol. 65 of Encyclopedia of Mathematics and its Applications, Cambridge University Press, Cambridge, 1997.

[5] P. ERDŐS, P. FRANKL, AND Z. FÜREDI, *Families of finite sets in which no set is covered by the union of r others*, Israel J. Math., 51 (1985), pp. 79–89.

[6] E. N. GILBERT, *A comparison of signalling alphabets*, The Bell system technical journal, 31 (1952), pp. 504–522.

[7] F. K. HWANG AND V. T. SÓS, *Nonadaptive hypergeometric group testing*, Studia Sci. Math. Hungar., 22 (1987), pp. 257–263.

[8] W. H. KAUTZ AND R. C. SINGLETON, *Nonrandom binary superimposed codes*, IEEE Trans. Inf. Theory, 10 (1964), pp. 363–377.

[9] L. LOVÁSZ, *On the ratio of optimal integral and fractional covers*, Discrete Math., 13 (1975), pp. 383–390.

[10] A. Q. NGUYEN AND T. ZEISEL, *Bounds on constant weight binary superimposed codes*, Problems Control Inform. Theory/Problemy Upravlen. Teor. Inform., 17 (1988), pp. 223–230.

[11] E. PORAT AND A. ROTHSCHILD, *Explicit non-adaptive combinatorial group testing schemes*, in ICALP (1), 2008, pp. 748–759.

[12] I. REED AND G. SOLOMON, *Polynomial codes over certain finite fields*, Journal of the Society for Industrial and Applied Mathematics, 8 (1960), pp. 300–304.

[13] R. C. SINGLETON, *Maximum distance q-nary codes*, IEEE Trans. Inf. Theory, 10 (1964), pp. 116–118.

[14] E. SPERNER, *Ein satz uber untermengen einer endlichen Menge*, Math. Z., 27 (1928), pp. 544–548.

[15] R. R. VARSHAMOV, *Estimate of the number of signals in error correcting codes*, Dokl. Akad. Nauk. SSSR, 117 (1957), pp. 739–741.

[16] V. V. VAZIRANI, *Approximation algorithms*, Springer-Verlag, Berlin, 2001.

[17] V. ZINOVIEV, *Cascade equal weight codes and maximal packing*, Problems Control Inform. Theory/Problemy Upravlen. Teor. Inform., 12 (1983), pp. 3–10.