Aalto University
School of Science
and Technology

Coding Theory on Non-Standard Alphabets

Group testing with error correction

Marcus Greferath

Department of Mathematics and Systems Analysis
Aalto University School of Sciences
marcus.greferath@aalto.fi

June 2023

Incidence Structures and Incidence Matrices

For most of what follows, let (P, B) be an incidence structure on the v-element set P of points, and let $b = |B|$ denote the number of blocks of B.

Definition

A binary matrix $M \in \mathbb{B}_2^{b \times v}$ is called an incidence matrix for (P, B), if its rows are labelled by the blocks, while its columns are labelled by the points of (P, B), such that

$$M_{c,p} = \begin{cases} 1 & : \quad p \in c, \\ 0 & : \quad \text{otherwise.} \end{cases}$$

Incidence matrices may thus be considered as indicator functions of their underlying incidence relation.

Partial Linear Spaces

Definition

For natural number s and t, a finite incidence structure (P, L) consisting of points and lines is called a partial linear space of order (s, t) if the following axioms hold:

- Two different points are connected by at most one line.
- Every line is incident with $s + 1$ points, and every point is incident with $t + 1$ lines.

Note: Interchanging the terms “line” and “point” will transform a partial linear space of order (s, t) into a partial linear space of order (t, s).

This comes from the fact, that any two lines in a partial linear space may intersect in at most one point.

Generalized Quadrangles

A well-understood class of partial linear spaces is that of the generalized quadrangles, first introduced by J. Tits.

Definition

A partial linear space (P, L) of order (s, t) is called a generalized quadrangle, denoted by $GQ(s, t)$, if for any non-incident point-line pair (p, ℓ), there exists a unique point q on ℓ that is connected with p by a line.

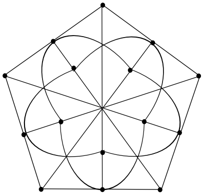

Figure: GQ(2,2) aka W(2)

Remark: A generalized quadrangle of order (s, t) has $(s+1)(st+1)$ points and $(t+1)(st+1)$ lines.

Group Testing Schemes from Partial Linear Spaces

The proof of the following result is rather simple, so we leave it to the interested audience.

Theorem

Let (P, L) be a partial linear space of order (s, t), and let $\ell_1, \dots \ell_m$ denote a collection of m distinct lines in L. If $\ell \in L$ is a line with $\ell \subseteq \ell_1 \cup \cdots \cup \ell_m$ then $\ell = \ell_j$ for some $1 \leq j \leq m$ provided $m \leq s$.

For the incidence matrix of a partial linear space (P, L) we may derive the following immediate conclusion.

Corollary

The group testing scheme resulting from the incidence matrix of a partial linear space of order (s, t) satisfies condition $\mathbf{t}$-rev.

Message Space: Coding vs Group Testing

- In coding theory, the message space is typically of the form $M = \mathbb{F}_2^k$. Due to compression, all messages are of equal probability.
- This implies, that a code

$$C = \{xG \mid x \in M\},$$

with a $k \times n$-generator matrix, will be a k-dimensional subspace of $\mathbb{F}_2^n$ which is endorsed with the uniform distribution.
- This in turn is responsible for the fact that a maximum-likelyhood decoder for the code C is equivalently described by a minimum-distance decoder.
- Understanding group testing as a generalization of coding theory, we need to see where this scenario changes.

Message Space: Coding vs Group Testing (cont'd)

- Messages in $\mathbb{B}_2^n$ represent infection patterns coming with a binomial distribution with parameter σ (prevalence).
- This means, that the message space $M = \mathbb{B}_2^n$ carries the distribution

$$P(x) \;=\; \sigma^{w(x)}(1-\sigma)^{n-w(x)}, \;\; \text{for } x \in \mathbb{B}_2^n.$$

- The group testing scheme $f : \mathbb{B}_2^n \longrightarrow \mathbb{B}_2^k$ takes this distribution to $\mathbb{B}_2^k$, such that

$$P_f(z) \;=\; \sum_{\substack{x \in \mathbb{B}^n \\ f(x)=z}} \sigma^{w(x)}(1-\sigma)^{n-w(x)}, \;\; \text{for } z \in \mathbb{B}_2^k.$$

- Already on message layer, we have a rate $R \leq 1$, namely the Shannon entropy

$$R \;=\; H(\sigma) = -\sigma \log_2(\sigma) - (1-\sigma)\log_2(1-\sigma).$$

Rate and Noise

- The rate of the group testing scheme should apparently be defined as $R_f = H(\sigma) \cdot \frac{n}{k}$.
- In the noiseless case, such schemes should exist, provided $R_f \leq 1$, in other words,

$$H(\sigma) \leq \frac{k}{n}.$$

- **Noise:** Simple antigen tests, say, for CoVid19 are cheap nowadays, however their accuracy has frequently been questioned.

  false pos: The probability p of a single false positive test is generally around 2%, whereas

  false neg: the probability q of a false negative test can easily exceed 20%.

The Binary Asymmetric Channel (BAC)

- This gives rise to what is called a binary asymmetric channel BAC(p, q), described by the channel matrix:

$$\pi(p, q) = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

- The literature provides formulae for the capacity of $\pi(p, q)$.
- The formula is complicated, and for $p = 0.02$ and $q = 0.2$ we obtain the capacity

$$\mathcal{C}(\pi) = 0.5488,$$

which is attained if the prevalence $\sigma = 0.4536$.

- We expect a Shannon-like theorem for the (asymptotic) existence of (error-correcting) group testing schemes if

$$R_f \leq R < \mathcal{C}(\pi).$$

A Non-Probabilistic Approach

- The prevalence σ may also be understood as an upper bound resulting as a ratio $\sigma = \frac{t}{n}$.
- In this case, we wish to identify infection patterns of Hamming weight $\leq t$ out of the n-element population
- We suggest to consider a binary layer code

$$C_t := f(B_n(0,t)) \subseteq \mathbb{B}_2^k.$$

- This code will have size $M_t \leq \sum_{i=0}^{t} \binom{n}{i}$, and a certain minimum distance δ_t, that allows for error correction.
- We will present a few examples of such codes in the sequel. We found them simply by experimentation.

Examples

7 Samples – 7 Tests:

Let $\mathbb{B}_2^7 \longrightarrow \mathbb{B}_2^7$ be the group testing scheme based on the incidence matrix of the binary Fano plane $\mathrm{PG}(2,2)$. By this, we mean $f(x) = xH$ for all $x \in \mathbb{B}_2^7$ where

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}.$$

We observe, that f satisfies **d**-rev for $d \leq 2$ and consider two layer codes induced by this matrix in $\mathbb{B}_2^7$:

Example A1: $A_1 := B_7(0,1) \cdot H$ is a $(7,8,3)$-code with dist. enumerator

$$E(A_1) = 8 + 14\,z^3 + 42\,z^4.$$

This scheme reliably identifies one infected sample out of 7 and correct at the same time one flawed test result.

Examples

Example A2: $A_2 := B_7(0,2) \cdot H$ is a $(7,29,2)$-code with distance enumerator

$$E(A_2) \; = \; 29 + 294\,z^2 + 14\,z^3 + 420\,z^4 + 42\,z^5 + 42\,z^6.$$

This scheme identifies two infected samples out of 7. Correction of a single flawed test result requires the probabilistic approach discussed earlier.

15 Samples – 15 Tests:

Consider $\mathbb{B}_2^{15} \longrightarrow \mathbb{B}_2^{15}$ be the group testing scheme represented by the full circulant matrix J induced by the generating word [000011101100101] of the binary BCH(15,5,7)-code.

Again, this scheme satisfies $\mathbf{d}$-rev for $d \leq 2$ (but not $d = 3$).

Example C1: $C_1 := B_{15}(0,1) \cdot J$ is a $(15, 16, 7)$-code with distance enumerator

$$E(C_1) \; = \; 16 + 30\,z^7 + 210\,z^8.$$

This scheme identifies 1 out of 15 with 15 tests and at the same time recovers 3 test errors.

Example C2 : $C_2 := B_{15}(0,2) \cdot J$ is a $(15, 121, 4)$-code with distance enumerator

$$\begin{aligned} E(C_2) \;&=\; 121 + 3570\,z^4 + 5040\,z^6 + 30\,z^7 \\ &\quad + 5460\,z^8 + 210\,z^{11} + 210\,z^{12}. \end{aligned}$$

This scheme identifies 2 out of 15 and at the same time recover 1 test error. More will be possible using a probabilistic approach.

- **Remark:** Without further justification, we have used the concept of distance enumerator:

$$E(C) \;:=\; \sum\{z^{d(x,y)} \mid x, y \in C\} \;=\; \sum_{i=0}^{k} A_i z^i,$$

where $A_i = |\{(x, y) \in C^2 \mid d(x, y) = i\}|$.
- It is easy to verify that $A_0 = M$ and that the first exponent $i \in \{1, \ldots, k\}$ with $A_i \neq 0$ is the minimum distance of C.
- Moreover, $\sum_{i=0}^{k} A_i = M^2 = A_0^2$.
- The Hamming distance d is not translation invariant, and hence this concept might not share properties commonly known from coding theory involving $\mathbb{F}_2$.