MATH 433

Applied Algebra

Lecture 33:
Linear codes.
Coset leaders and syndromes.

Binary codes

A **binary code** is an injective function $f : \mathbf{B}^m \to \mathbf{B}^n$, where $\mathbf{B}^k$, $k \geq 1$, is regarded as the set of all words of length k in the alphabet $\mathbf{B} = \{0, 1\}$. For any $w \in \mathbf{B}^m$, the word $f(w)$ is called the **codeword** associated to w.

Encoding: The sender splits the message into "blocks", i.e., words of length m: $w_1, w_2, \dots, w_s$. Then they apply f to each of these words and produce a sequence of codewords $f(w_1), f(w_2), \dots, f(w_s)$, which is to be transmitted.

Decoding: The receiver obtains a sequence of words of length n: $w'_1, w'_2, \dots, w'_s$, where w'_i is supposed to be $f(w_i)$ but it may be different due to errors during transmission. Each w'_i is checked for being a codeword. If it is, $w'_i = f(w)$, then w'_i is decoded to w. Otherwise an error (or errors) is detected. In the case of an error-correcting code, the receiver attempts to correct w'_i by applying a correction function $c : \mathbf{B}^n \to \mathbf{B}^n$, then decodes the word $c(w'_i)$.

Error detection and error correction

The distance $d(w_1, w_2)$ between binary words w_1, w_2 of the same length is the number of positions in which they differ. The **weight** of a word w is the number of nonzero digits, which is the distance to the zero word.

The distance between the sent codeword and the received word is equal to the number of errors during transmission.

Theorem Let $f : \mathbf{B}^m \to \mathbf{B}^n$ be a coding function. Then **(i)** f allows detection of k or fewer errors if and only if the minimum distance between distinct codewords is at least $k+1$; **(ii)** f allows correction of k or fewer errors if and only if the minimum distance between distinct codewords is at least $2k+1$.

The correction function c is usually chosen so that $c(w)$ is the codeword closest to w.

Linear codes

The binary alphabet $\mathbf{B} = \{0, 1\}$ is naturally identified with $\mathbb{Z}_2$, the field of 2 elements. Then $\mathbf{B}^n$ can be regarded as an n-dimensional vector space over the field $\mathbb{Z}_2$.

A binary code $f : \mathbf{B}^m \to \mathbf{B}^n$ is called a **group code** (or a **linear code**) if the set of all codewords in $\mathbf{B}^n$ is closed under addition.

Theorem Given a nonempty subset W of $\mathbf{B}^n$, the following conditions are equivalent:

- W is closed under addition;
- W is a subgroup of $\mathbf{B}^n$;
- W is a subspace of $\mathbf{B}^n$.

The Hamming distance on $\mathbf{B}^n$ is **translation invariant**, which means that $d(w_1 + w,\ w_2 + w) = d(w_1, w_2)$ for all $w_1, w_2, w \in \mathbf{B}^n$. In particular, $d(w_1, w_2) = d(w_1 - w_2,\ \mathbf{0})$. Note that $d(w, \mathbf{0})$ is equal to the number of 1's in the word w (called the **weight** of w).

In the case of a linear code, the zero word $\mathbf{0}$ is always a codeword. Moreover, $w_1 - w_2$ (which is the same as $w_1 + w_2$) is a codeword whenever both w_1 and w_2 are. Hence the minimum distance between distinct codewords is equal to the minimum weight of nonzero codewords.

A natural example of a linear code $f : \mathbf{B}^m \to \mathbf{B}^n$ is a linear transformation of vector spaces. Any linear transformation is given by a **generator matrix** G, which is an $m \times n$ matrix with entries from $\mathbb{Z}_2$ such that $f(w) = wG$ (here w is regarded as a row vector). For a systematic code, G is of the form $(I_m|A)$, where I_m is the $m \times m$ identity matrix.

Examples. • Parity bit.

$f : \mathbf{B}^3 \to \mathbf{B}^4$, $f(w) = wx$, where x is the parity bit of w.
This code is linear, given by the generator matrix

$$G = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

Codewords: 0000, 1001, 0101, 0011, 1100, 1010, 0110, 1111.

• Tell three times.

$f : \mathbf{B}^2 \to \mathbf{B}^6$, $f(w) = www$.
This code is also linear, given by the generator matrix

$$G = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

Codewords: 000000, 101010, 010101, 111111.

Coset leaders

A received word w' can be represented as $w_0 + w_1$, where w_0 is the codeword that was sent and w_1 is an **error pattern** (namely, w_1 has 1's in those positions where transmission errors occured).

The set of words received with a particular error pattern w_1 is a coset $w_1 + W$ of the subgroup W of codewords in the group $\mathbf{B}^n$ of all words of length n.

Error correction is based on the following assumption. For every coset C of W, we assume that all words from C are received with the same error pattern w_C. Note that $w_C \in C$ so that $C = w_C + W$. Given a word $w \in C$, the corrected codeword is $w - w_C$ $(= w + w_C)$.

The word w_C is a **coset leader**. It is chosen as the most likely error pattern, namely, the word of smallest weight in the coset C (the choice may not be unique).

Coset decoding table

Using coset leaders, the error correction can be done via the **coset decoding table**, which is a $2^{n-m} \times 2^m$ table containing all words of length m. The table has the following properties:

- the first row consists of codewords,
- the first entry in the first row is the zero word,
- each row is a coset,
- the first column consists of the coset leaders,
- any word is the sum of the first word in its row and the first word in its column.

Once the coset decoding table is build, each word is corrected to the codeword on top of its column.

The coset decoding table can be build simultaneously with choosing coset leaders using a procedure similar to the sieve of Eratosthenes.

Example. Generator matrix: $G = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}$

Coding function: $\begin{cases} 00 \to 00000 \\ 01 \to 01011 \\ 10 \to 10110 \\ 11 \to 11101 \end{cases}$ detects 2 errors, corrects 1 error

Coset decoding table:

00000	01011	10110	11101
00001	01010	10111	11100
00010	01001	10100	11111
00100	01111	10010	11001
01000	00011	11110	10101
10000	11011	00110	01101
00101	01110	10011	11000
10001	11010	00111	01100

Coset decoding table:

00000	01011	10110	11101
00001	01010	10111	11100
00010	01001	10100	11111
00100	01111	10010	11001
01000	00011	11110	10101
10000	11011	00110	01101
00101	01110	10011	11000
10001	11010	00111	01100

- Message: 00 01 01 00 10 11 11 01 00
- After encoding:

00000 01011 01011 00000 10110 11101 11101 01011 00000

- After transmission:

01000 00011 01011 00000 11100 11101 10101 11101 00101

- After correction:

00000 01011 01011 00000 11101 11101 11101 11101 00000

- After decoding: 00 01 01 00 11 11 11 11 00

Parity-check matrix

An alternative way to do error correction is to use the parity-check matrix and syndromes.

Let G be the generator matrix of a linear code $f : \mathbf{B}^m \to \mathbf{B}^n$. We assume the code to be systematic so that G has the form $(I_m|A)$. The **parity-check matrix** of the code is the matrix $P = \begin{pmatrix} A \\ I_{n-m} \end{pmatrix}$. Given a word $w' \in \mathbf{B}^n$, the **syndrome** of w' is the product $w'P$, which is a word of length $n - m$.

Theorem (i) The syndrome of a word w' is the zero word if and only if w' is a codeword. [Hint: $GP = O$]
(ii) The syndromes of two words w' and w'' coincide if and only if these words are in the same coset.

Given a transmitted word w', we compute its syndrome and find a coset leader w_C with the same syndrome. Then the corrected word is $w' - w_C$ $(= w' + w_C)$.

To perform the error correction, we need a two-column table where one column consists of coset leaders and the other consists of the corresponding syndromes.

Example. Generator matrix: $G = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}$.

$$\left(\begin{array}{cc|ccc} 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{array}\right) \to \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \to \left(\begin{array}{ccc} 1 & 1 & 0 \\ 0 & 1 & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right)$$

Coset leaders	Syndromes
00000	000
00001	001
00010	010
00100	100
01000	011
10000	110
00101	101
10001	111