Algorithms of Information Security: Error-correcting codes I

Faculty of Information Technology
Czech Technical University in Prague

October 5, 2022

Basic definitions

Definition

Let $A = \{a_1, \ldots, a_q\}$ be an alphabet; we call the a_i values symbols. A block code C of length n over A is a subset of A^n. A vector $c \in C$ is called a codeword. The number of elements in C, denoted by $|C|$, is called the size of the code. A code of length n and size M is called an (n, M)-code.

Example. A code over $A = \{0, 1\}$ is called a binary code and a code over $A = \{0, 1, 2\}$ is called a ternary code.
Example. The set $\{(0, 0, 0), (1, 1, 1)\}$ is the binary (3,2)-code.

Basic definitions

Definition

The Hamming distance between two strings x and y of the same length over a finite alphabet A is defined as the number of positions at which the two strings differ. Let $x = x_1, \ldots, x_n$ and $y = y_1, \ldots, y_n$, then for every i defined

$$\delta(x_i, y_i) = \begin{cases} 1, & x_i \neq y_i, \\ 0, & x_i = y_i \end{cases}$$

Hamming distance is defined by

$$d(x, y) = \sum_{i=1}^{n} \delta(x_i, y_i).$$

Example. In the space F_2^5 the Hamming distance satisfies $d(10111, 11001) = 3$ and in F_3^4 we have $d(1122, 1220) = 2$.

Note. Hamming distance d defines a metric on A^n. That is, for every $x, y, z \in A^n$:

1. $0 \leq d(x, y) \leq n$
2. $d(x, y) = 0$ if and only if $x = y$
3. $d(x, y) = d(y, x)$
4. (triangle inequality) $d(x, z) \leq d(x, y) + d(y, z)$.

Note. We stress that the Hamming distance is not dependent on the actual values of x_i and y_i but only if they are equal to each other or not equal.

Definition

Let C be a code of length n over an alphabet A. The *nearest neighbor* decoding rule states that every $x \in A^n$ is decoded to $c_x \in C$ that is closest to x. That is, $D(x) = c_x$ where c_x is such that $d(x, c_x) = \min_{c \in C} d(x, c)$.

Definition

Let C be a code. The distance of the code denoted by $d(C)$ is defined by

$$d(C) = \min \{d(c_1, c_2) \mid c_1, c_2 \in C, c_1 \neq c_2\}$$

An (n, M)-code of distance d is called an (n, M, d)-code. The values n, M, d are called the parameters of the code.

Restating what we have discussed above, the aim of coding theory is to construct a code with a short n, and large M and d. We now show a connection between the distance of a code and the possibility of detecting and correcting errors.

Definition

Let C be a code of length n over alphabet A.

- C detects u errors if for every codeword $c \in C$ and every $x \in A^n$ with $x \neq c$, it holds that if $d(x, c) \leq u$ then $x \notin C$.
- C corrects v errors if for every codeword $c \in C$ and every $x \in A^n$ it holds that if $d(x, c) \leq v$ then nearest neighbor decoding of x outputs c.

Theorem

- *A code C detects u errors if and only if $d(C) > u$.*
- *A code C corrects v errors if and only if $d(C) \geq 2v + 1$.*

Linear code

We denote by F_q a finite field of size q. Recall that there exists such a finite field for any q that is a power of a prime. In this course, we will just assume that we are given such a field. In linear codes, the alphabet of the code consists of the elements of some finite field F_q.

Definition

A linear code with length n over F_q is a vector subspace of F_q^n.

Example. The repetition code $C = \{\underbrace{(x, \ldots, x)}_{n} \mid x \in F_q\}$ is a linear code.

Notation. A linear code of length n and dimension k is denoted by $[n, k]$-code (or an $[n, k, d]_q$-code when the distance d and the size of the alphabet q are specified).

Note. Dimension k is not M, i.e., the size of the code.

Linear code

Definition

Let C be a linear $[n,k]_q$ code over F_q^n. Then:

1. The *dual code* of C is $C^\perp$ (the orthogonal complement of C in F_q^n, $C^\perp = \{x \in F_q^n \mid \langle x, c\rangle = 0 \text{ for all } c \in C\}$). Notice that $C^\perp$ is an $[n, n-k]_q$ code.
2. The *dimension* of C is the dimension of C as a vector subspace of F_q^n, denoted by dim(C).

Theorem

Let C be a linear code of length n over F_q. Then

1. $\mid C \mid = q^{\mathit{dim}(C)}$ *(dim$(C) = k$, i.e., dimension of a code).*
2. $C^\perp$ *is a linear code, and dim(C) + dim$(C^\perp) = n$.*
3. $(C^\perp)^\perp = C$.

Linear code

Definition

Let C be a linear code. Then:

1. C is *self orthogonal* if $C \subseteq C^{\perp}$.
2. C is *self dual* if $C = C^{\perp}$.

The following theorem is an immediate corollary of the fact that $\dim(C) + \dim(C^{\perp}) = n$.

Theorem

1. *Let C be a self-orthogonal code of length n. Then $\dim(C) \leq \frac{n}{2}$.*
2. *Let C be a self-dual code of length n. Then $\dim(C) = \frac{n}{2}$.*

Definitions

Definition

Let $x \in F_q^n$. The Hamming weight of x denoted by wt(x) is defined to be the number of coordinates that are not zero. That is, wt$(x) = d(x, 0)$.

Definition

Let C be a code (not necessarily linear). The weight of C denoted by wt(C) is defined by

$$\text{wt}(C) = \min_{c \in C; c \neq 0} \{\text{wt}(c)\}.$$

The following theorem holds only for linear codes:

Theorem

Let C be a linear code over F_q^n. Then $d(C) = wt(C)$.

Generator and Parity-Check Matrices

Definition

1. *A generator matrix* G for a linear code C is a matrix whose rows form a basis for C.
2. *A parity check matrix* H for C is a generator matrix for the dual code $C^{\perp}$.

Remarks:

1. If C is a linear $[n, k]$-code then $G \in F_q^{k \times n}$ (recall that k denotes the number of rows and n the number of columns), and $H \in F_q^{(n-k) \times n}$.
2. The rows of a generator matrix are linearly independent.
3. To show that a $k-$ by $-n$ matrix G is a generator matrix of a code C it suffices to show that the rows of G are codewords in C and that they are linearly independent.

Definition

1. A generator matrix is said to be in standard form if it is of the form $(I_k \mid X)$, where I_k denotes the $k \times k$ identity matrix.
2. A parity check matrix is said to be in standard form if it is of the form $(Y \mid I_{n-k})$.

Lemma

Let C be a linear $[n,k]$-code with generator matrix G. Then for every $v \in F_q^n$ it holds that $v \in C^\perp$ if and only if $v \cdot G^T = 0$. In particular, a matrix $H \in F_q^{(n-k)\times n}$ is a parity check matrix if and only if its rows are linearly independent and $H \cdot G^T = 0$.

An equivalent formulation: Let C be a linear $[n,k]$-code with a parity check matrix H. Then $v \in C$ if and only if $v \cdot H^T = 0$.

Theorem

Let C be a linear code and let H the parity check matrix for C. Then

1. $d(C) \geq d$ *if and only if every subset of* $d-1$ *columns of* H *are linearly independent.*
2. $d(C) \leq d$ *if and only if there exists a subset of* d *columns of* H *that are linearly dependent.*

Corollary. Let C be a linear code and let H be a parity check matrix for C. Then $d(C) = d$ if and only if every subset of $d-1$ columns in H are linearly independent and there exists a subset of d columns that are dependent in H.

Theorem

If $G = (I_k \mid X)$ *is the generator matrix in standard form for a linear* $[n, k]$*-code* C*, then* $H = (-X^T \mid I_{n-k})$ *is a parity check matrix for* C.

Equivalence of Codes

Definition

Two (n, M)-codes are equivalent if one can be derived from the other by a permutation of the coordinates and multiplication of any specific coordinate by a non-zero scalar.

Theorem

Every linear code C is equivalent to a linear code C' with a generator matrix in standard form.

Polynomial code

Fix a finite field F_q. For the purpose of constructing polynomial codes, we identify a word of n elements $c = (c_0, \ldots, c_{n-1})$ with its representing polynomial $c(x) = \sum_{i=0}^{n-1} c_i x^i$.

Definition

Fix some integer n and let $g(x)$ be some fixed polynomial of degree $m \leq n-1$. The polynomial code generated by $g(x)$ is the code whose codewords are the polynomials of degree less than n that are divisible (without remainder) by $g(x)$.

Example. Let $n = 5, m = 2$ and consider the polynomial $g(x) = x^2 + x$ over F_2. Using $g(x)$, we generate the polynomials of degree ≤ 4, i.e., polynomials in the form $p(x) \cdot g(x)$, where $p(x) \in \{0, 1, x, (x+1), x^2, (x^2+1), (x^2+x), (x^2+x+1)\}$. Written explicitly:
$0, x^2+x, x^3+x^2, x^3+x, x^4+x^3, x^4+x^3+x^2+x, x^4+x^2, x^4+x$.
And we can represent them as strings of binary digits:
00000, 00110, 01100, 01010, 11000, 11110, 10100, 10010.

Reed–Muller codes

- Reed-Muller codes are named after David E. Muller, who developed the codes in 1954, and Irving S. Reed, who designed the first efficient decoding algorithm.
- Reed-Muller codes are error-correcting codes that are used in wireless communication applications, especially in space communication.
- Reed-Muller codes with parameters r and m are denoted by $R(r, m)$, where r and m are integers such that $0 \leq r \leq m$.
- Reed-Muller codes can be considered as a generalization of Reed-Solomon codes.
- Reed-Muller codes are linear codes defined by evaluating polynomials of several variables. In the lecture, we consider mainly binary Reed-Muller codes.

Basic Definitions

Definition

The Boolean function of m variables is a map $F_2^m \to F_2$.

Definition

Polynomial $f(x_1, \ldots, x_m)$ in m variables over F_2 is *boolean polynomial*, if in each member of the sum

$$f(x_1, \ldots, x_m) = \sum_{(i_1, \ldots, i_m)} a_{i_1 \ldots i_m} x_1^{i_1} \ldots x_m^{i_m}$$

all exponents $i_1, \ldots, i_m$ are equal to 0 or 1.

Basic Definitions

- Boolean polynomial $f(x_1, \ldots, x_m)$ is thus the sum of monomials in a form
$$x_{j_1} x_{j_2} \ldots x_{j_k}$$
where $1 \leq j_1 < \ldots < j_k \leq m$.
- Each set $I \subset \{1, \ldots, m\}$ corresponds to a monomial
$$x_I = \prod_{i \in I} x_i.$$
- Monomial $x_\emptyset$ is denoted by the symbol 1.
- Polynomial 0 denotes the sum of an empty set of monomials.
- The total degree of the polynomial $f \in F_q[x_1, \ldots, x_m]$ is the value $\max \sum_{j=1}^{m} i_j$, where the maximum is over all members $x_1^{i_1} \ldots x_m^{i_m}$, which have a non-zero coefficient.

Basic Definitions

- Since in the field F_2 holds that $0^2 = 0$ and $1^2 = 1$, then for $i = 1, \ldots, m$ the following equality holds:

$$x_i{}^2 = x_i.$$

- Using this property, we can (uniquely) modify the product of two Boolean polynomials into a polynomial, which is again Boolean. For example:

$$x_1x_3 \cdot (x_1 + x_2) = x_1x_3 + x_1x_2x_3.$$

- Each Boolean polynomial f determines the Boolean function $\hat{f}$: if we substitute for individual variables, the resulting value is uniquely determined.
- The number of Boolean functions of m variables is the same as the number of Boolean polynomials in the variables $x_1, \ldots, x_m$.

Basic Definitions

Theorem

For every Boolean function h with m variables, there is a Boolean polynomial $f \in F_2[x_1, \ldots, x_m]$ having the property that $h = \hat{f}$.

Note. The above theorem allows us to identify a Boolean function with a uniquely determined Boolean polynomial.
Notation. If $b = (b_1, \ldots, b_m)$ is an ordered m-tuple of elements of the field F_q, then the symbol $f(b)$ denotes the value $f(b_1, \ldots, b_m)$.

Definition

Let $B_0, \ldots, B_{q^m-1}$ be the numbering of all ordered m-tuples over F_q. Reed-Muller code $R_q(r, m)$ consists of the words in a form:

$$(f(B_0), f(B_1), \ldots, f(B_{q^m-1}))$$

where words are obtained from all polynomials f in $F_q[x_1, \ldots, x_m]$, whose total degree is at most r. The lenght of the code $R_q(r, m)$ is therefore q^m.

Binary Reed–Muller codes

Notation. For any polynomial $f \in F_2[x_1, \ldots, x_m]$ let's denote

$$N(f) = \{(i_1, \ldots, i_m) \in F_2^m : f(i_1, \ldots, i_m) = 1\}.$$

The lower bound on the size of the set $N(f)$ implies an estimate of the minimum distance of the (binary) Reed-Muller codes.

Theorem

Let $f \in F_2[x_1, \ldots, x_m]$ be nonzero Boolean polynomial of total degree at most r. Then

$$|N(f)| \geq 2^{m-r}.$$

Consequence. A set $B_r \subset R(r, m)$, consisting of the evaluations of all monomials of the total degree at most r is the base of the code $R(r, m)$.
Consequence. Reed–Muller code $R(r, m)$ has length 2^m, dimension $\binom{m}{0} + \ldots + \binom{m}{r}$ and minimal weight 2^{m-r}.

Theorem

The codes $R(r, m)$ and $R(m - r - 1, m)$ are dual to each other.

Binary Reed–Muller codes

Example. Let $r = 1$ and $m = 3$, then the length of $R_2(1,3)$ code is $n = 8$. Monomials in $F_2[x_1, x_2, x_3]$ of degree atmost 1 are $\{1, x_1, x_2, x_3\}$. When evaluating, consider the elements of the set F_2^3 in the order:

$$(x_3x_2x_1) : 000, 001, 010, 011, 100, 101, 110, 111.$$

Vectors over F_2^8 associated with these monomials are:

$$\begin{array}{rcl}
(\text{evaluation of})\ 1 \rightarrow & (1 & 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1) \\
(\text{evaluation of})\ x_1 \rightarrow & (0 & 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1) \\
(\text{evaluation of})\ x_2 \rightarrow & (0 & 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1) \\
(\text{evaluation of})\ x_3 \rightarrow & (0 & 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1).
\end{array}$$

Therefore, the generator matrix of the code $R_2(1,3)$ is as follows:

$$\begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1
\end{pmatrix}$$