Algorithms of Information Security: Error-correcting codes

Faculty of Information Technology
Czech Technical University in Prague

September 29, 2020

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 $|C|$, 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.

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

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

and defined

$$d(x, y) = \sum_{i=0}^{n} d(x_i, y_i).$$

Example. In the space ${Z_2}^5$ the Hamming distance satisfies $d(10111, 11001) = 3$ and in ${Z_3}^4$ we have $d(1122, 1220) = 2$.

Note. Hamming distance defines a metric on A^n.
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.
Proposition. The function d is a metric. 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) $d(x, z) \leq d(x, y) + d(y, z)$.

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 $d(C)$, 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 are 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 as $[n, k]$-code (or an $[n, k, d]$-code when the distance d is specified).

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 $\dim(C)$.

Theorem

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

1. $\mid C \mid = q^{dim(C)}$.
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 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 wt(C), is defined by

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

The following theorem only holds 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. In order 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 $k-$by$-k$ 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 $c = (c_0, \dots, c_{n-1})$ 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 $g(x)$.

Lemma

Let $c = (c_0, \dots, c_{n-1}) \in {F_q}^n$ *and* $c(x)$ *be its representing polynomial. Then,* $c \in C$ *if and only if* $c(\alpha^t) = 0$ *for every* $1 \leq t \leq n-k$.

Cyclic code

Definition

A code C is cyclic if every cyclic shift of a codeword in C is also a codeword. That is, $(c_0, c_1, \ldots, c_{n-1}) \in C$ implies that $(c_{n-1}, \ldots, c_{n-2}) \in C$.

In the notation of representing polynomials, a code C is cyclic if and only if $c(x) \in C$ implies

$$x \cdot c(x) \text{ mod } (x^n - 1) \in C.$$

If a code is linear, then equivalently we can say that $c(x) \in C$ implies

$$u(x) \cdot c(x) \text{ mod } (x^n - 1) \in C$$

for every $u(x) \in F_q[x]$. Hence, C is a linear cyclic code if and only if C is an ideal in the ring $F_q[x] \setminus (x^n - 1)$.

Cyclic code

Theorem

Let C be a cyclic code over F_q and g the monic polynomial in C of minimal positive degree (prove that it is unique!). Then g generates C, i.e., $c \in C$ iff $g \mid c$.

Theorem

A polynomial code is cyclic if and only if its generator polynomial divides $x^n - 1$.

Dual Codes of Cyclic Codes

Let C be an $[n, k]$ cyclic code with a generator $g(x) = \sum_{i=0}^{n-k} g_i x^i$.
We know that $g \bmod x^n - 1 = 0$ and therefore there exists $h(x) = \sum_{i=0}^{k} h_i x^i$ such that $gh = x^n - 1$.
Let $c \in C$. As g generates C we have $c = ga$ for some $a \in F_q[x]$.
Therefore

$$hc \bmod (x^n - 1) = hga \bmod (x^n - 1) = 0$$

This translates to the constraints:

$$c_0 h_i + c_1 h_{i-1} + \ldots + c_{n-k} h_{i-n+k} = 0,$$

for every $0 \leq i \leq n - 1$, where the indices are modulo n.

Dual Codes of Cyclic Codes

It follows that

$$H = \begin{pmatrix} h_k & h_{k-1} & \dots & h_0 & & & & \\ & h_k & h_{k-1} & \dots & h_0 & & & \\ \vdots & & \ddots & & & & & \\ & & & & h_k & h_{k-1} & \dots & h_0 \end{pmatrix}$$

is a $(n-k) \times n$ matrix of parity checks of C, and because it has the correct rank $n-k$ it is a parity check matrix of C.

Theorem

Let C be an $[n,k]$ cyclic code generated by $g(x)$ and let $h(x) = \frac{x^n-1}{g(x)}$. Then, the dualcode of C is a cyclic $[n, n-k]$code whose generator polynomial is $x^k h(x^{-1})$. The polynomial $h(x)$ is called the check polynomial of C.

The Binary Hamming Code

Definition

Let $r \geq 2$ and let C be a binary linear code with $n = 2^r - 1$ whose parity check matrix H is such that the columns are all of the non-zero vectors in F_2^r. This code C is called a binary Hamming code of length $2^r - 1$, denoted $\text{Ham}(r, 2)$.

Propositions.

1. All binary Hamming codes of a given length are equivalent.
2. For every $r \in N$, the dimension of $\text{Ham}(r, 2)$ is $k = 2^r - 1 - r$.
3. For every $r \in N$, the distance of $\text{Ham}(r, 2)$ is $d = 3$ and so the code can correct exactly one error.

Example. A generator matrix for Ham$(r, 2)$

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

We write the parity check matrix for Ham$(r, 2)$.

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

As can be seen, for Ham$(r, 2)$ we have $n = 7, k = 4$ and $H \in F_2^{7\times 3}$.

The Hamming code is cyclic

Any binary Hamming code is equivalent to a cyclic code.

Theorem

Fix a field F_{2^r} and let $n = 2^r - 1$. Then, there exists a $[n, k = n - r, 3]_2$ cyclic code. Since the only code with such length, dimension and distance is the Hamming code, the Hamming code is cyclic.

The Reed-Solomon code

Reed-Solomon codes were invented by Irving S. Reed and Gustave Solomon in 1960. In 1977 RS codes have been implemented in Voyager space program. The first commercial application of RS codes in mass-consumer products was in 1982.
Reed-Solomon codes were used in digital television, satellite communication, wireless communication, bar-codes, compact discs, DVD, ...
Fix a field F_q of size q with a generator α of F_q^*. The code $RS : F_q^k \to F_q^n$ corresponds to evaluating all degree $k-1$ polynomial (whose coefficients are given to us as the input message) on all nonzerofield elements. That is, $n = q-1$ and

$$RS(a_0, \ldots, a_{k-1}) = (p_a(\alpha^0), p_a(\alpha^1), \ldots, p_a(\alpha^{n-1})),$$

where $p_a(x) = \sum_{i=0}^{k-1} a_i x^i$.

Every nonzero polynomial of degree $k-1$ can have at most $k-1$ zeros in F_q, so the weight of every nonzero codeword is at least $d = n - (k-1) = n - k + 1$. Thus, the RS code is an $[n, k, n-k+1]_q$ code. By inspection, the $n \times k$ transpose of generating matrix is given by

$$G^T = \begin{pmatrix} (\alpha^0)^0 & (\alpha^0)^1 & \dots & (\alpha^0)^{k-1} \\ (\alpha^1)^0 & (\alpha^1)^1 & \dots & (\alpha^1)^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ (\alpha^{n-1})^0 & (\alpha^{n-1})^1 & \dots & (\alpha^{n-1})^{k-1} \end{pmatrix}$$

so for $0 \leq i \leq n-1$ and $0 \leq j \leq k-1$, $G^T[i, j] = \alpha^{ij}$.

Now, consider the $n \times (n-k)$ matrix

$$H^T = \begin{pmatrix} (\alpha^0)^1 & (\alpha^0)^2 & \dots & (\alpha^0)^{n-k} \\ (\alpha^1)^1 & (\alpha^1)^2 & \dots & (\alpha^1)^{n-k} \\ \vdots & \vdots & \ddots & \vdots \\ (\alpha^{n-1})^1 & (\alpha^{n-1})^2 & \dots & (\alpha^{n-1})^{n-k} \end{pmatrix}$$

so for $0 \leq i \leq n-1$ and $1 \leq j \leq n-k, H^T[i,j] = \alpha^{ij}$. H^T is a Vandermonde matrix, so in order to prove that H^T is indeed the parity-check matrix of the RS code, it is sufficient to prove that $H \cdot G^T = 0_{(n-k)\times k}$.

We have that

$$(H \cdot G^T)[a,b] = \sum_{k=0}^{n-1} H[a,k]G[b,k] = \sum_{k=0}^{n-1} (\alpha^{a+b})^k.$$

a ranges from 1 to $n-k$ and b ranges from 0 to $k-1$, so $1 \leq a+b \leq n-1$ and the above sums to zero.

Theorem

Reed-Solomon codes are linear codes.

Theorem

The Reed-Solomon code is a polynomial code.

Note. Reed-Solomon (RS) codes are non-binary cyclic codes.

An interesting property of Reed-Solomon codes

$$RS(k,q)^{\perp} = RS(q-k,q).$$

Singleton bound

Theorem

The minimum distance for an (n, k) linear code is bounded by

$$d \leq n - k + 1.$$

For an Reed-Solomon codes code $d \geq n - k + 1$, so $d = n - k + 1$ and all Reed-Solomon codes meet the Singleton bound – they are optimal $(n, k, n - k + 1)$ codes, $n = q - 1$.

Definition

Codes that meet the Singleton bound are called Maximum Distance Separable codes (MDS).

Note. The dual code of an Reed-Solomon code is also MDS.

BCH codes

BCH codes were discovered by independently by Bose and Ray-Chaudhuri and by Hocquenghem in the late 1950s.
BCH codes can be defined over any field, first we will focus on binary BCH codes:

Definition

For a length $n = 2^m - 1$, a distance d, and a primitive element $\alpha \in F^*_{2^m}$, we define the binary BCH code

$$BCH[n,d] = \Big\{(c_0, c_1, \dots, c_{n-1}) \in F_2^n \mid c(X) = c_0 + c_1 X + \dots + c_{n-1}X^{n-1}$$

$$\text{satisfies } c(\alpha) = c(\alpha^2) = \dots = c(\alpha^{d-1}) = 0\Big\}.$$

Lemma

The BCH codes form linear spaces.

Definition

For prime power q, integer m, and integer d, the BCH code $BCH_{q,m,d}$ is obtained as follows: Let $n = q^m$ and let F_{q^m} be an extension of F_q and let C' be the (extended) $[n, n-(d-1), d]_{q^m}$ Reed-Solomon code obtained by evaluating polynomials of degree at most $n-1$ over F_{q^m} at all the points of F_{q^m}. Then the code $BCH_{q,m,d}$ is the F_q-subfield subcode of C'. In other words, $BCH_{q,m,d} = C' \cap F_q^n$.

If we have $q = 2$ then $BCH_{2,m,d} = C \cap {F_2}^{2^n}$ where C is given as a Reed-Solomon code $C = RS[n = 2^m, n-(d-1), d]_{F_{2^m}}$.

The BCH code could be constructed in the following manner: Look at the Reed-Solomon code and only pick up the codewords that are in ${F_2}^{2^m}$.
Conjecture. Dimension of BCH code is at least $n - m(d-1)$.
The general idea of a BCH code is to identify its generating polynomial by the roots(instead of interms of the coeficients).

Theorem

For prime power q, integers m and d, the $BCH_{q,m,d}$ is an $[n, n-1-m\lceil\frac{(d-2)(q-1)}{q}\rceil, d]_q$ code, for $n = q^m$.

In the case of $q = 2$ (binary codes) we have the following corollary:

Corollary

For every integer m and t, the code $BCH_{2,m,2t}$ is an $[n, n-1-(t-1)\log n, 2t]$ code, for $n = 2^m$.

Note. Reed-Solomon codes are nonbinary BCH codes.

Example

Let $n = 8 - 1 = 7$. We want form F_8 from $x^3 + x + 1$.

$$\begin{array}{cc}
\alpha^0 & 1 \\
\alpha^1 & \alpha \\
\alpha^2 & \alpha^2 \\
\alpha^3 & \alpha + 1 \\
\alpha^4 & \alpha^2 + \alpha \\
\alpha^5 & \alpha^2 + \alpha + 1 \\
\alpha^6 & \alpha^2 + 1
\end{array}$$

Then

$$g(x) = (x - \alpha)(x - \alpha^2)(x - \alpha^3)(x - \alpha^4) =$$

$$= x^4 + \alpha^3 x^3 + x^2 + \alpha x + \alpha^3.$$

Then

$$H^T = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \alpha^4 & \alpha^5 & \alpha^6 \\ 1 & \alpha^2 & \alpha^4 & \alpha^6 & \alpha & \alpha^3 & \alpha^5 \\ 1 & \alpha^3 & \alpha^6 & \alpha^2 & \alpha^5 & \alpha & \alpha^4 \\ 1 & \alpha^4 & \alpha^8 & \alpha^5 & \alpha^2 & \alpha^6 & \alpha^3 \end{pmatrix}$$

Then

- RS: $n = q^m - 1 = 7$, where $q = 8$ and $m = 1$
- RS: $g(x) = (x - \alpha)(x - \alpha^2)(x - \alpha^3)(x - \alpha^4)$
- RS: we have $(7, 3, 5)$ code.
- RS: $q^k = 8^3 = 2^9 = 512$ codewords.