Coding Theory
UNIT-II
Error Control Coding
CSE XXX

Jayaprakash Kar

Table of contents

Basic Definitions

- A **Word** is a sequence of symbols
- A **Code** is a set of vectors called **codewords**
- The **Hamming weight** of a code is equal to the number of non-zero elements in the codeword. The Hamming weight of a codeword is denoted by **c** is denoted by $w(c)$.
- The **Hamming distance** between two codewords is the number of places by which the codeword differ. The Hamming distance between two codewords c_1 and c_2 is denoted by $d(c_1, c_2)$. It is easy to see that $d(c_1, c_2) = w(c_1 - c_2)$.

Example-I

Consider a code $c = \{0100, 1111\}$ which consists of two codewords *i*.e 0100 and 1111. The Hamming weight $w(0100) = 1$ and $w(1111) = 4$. The Hamming distance between two codewords is 3. They differ at 1^{st}, 3^{rd} and 4^{th} places. We can observe that

$w(0100 - 1111) = \ w(1011) = \ 3$

$= \ d(0100, 1111)$.

Example-II

Consider a code $c = \{01234, 43210\}$ which consists of two codewords *i*.e 01234 and 43210. The Hamming weight $w(01234) = 4$ and $w(43210) = 4$. The Hamming distance between two codewords is 4, because only the 3^{rd} component of the two codewords are identical while they differ at 4 places.

Block code

- A block code consists of a set of fixed length codewords. The fixed length of these codewords is called the **block length** and is typically denoted by n. Thus a code of block length n consists of a set of codewords having n components.
- A block code of size M defined over an alphabet with q symbols is set of M q-ary sequences, each of length n. If it is binary, then $q = 2$, the symbols are called bits and the code is said to be binary code. Usually, $M = q^k$ for some integer k. We called such a code an (n, k) code.
- Thus an (n, k) block code over an alphabet q is a set of q^k codewords of block length n.

Example-III

The code $c = \{00000, 10100, 11110, 11001\}$ is a block code of block length is 5. This code can be used to represent two-bit binary numbers which are as follows:

Uncoded bits	Codewords
00	00000
01	10100
10	11110
11	11001

Here $M = 4$, $k = 2$ and $n = 5$. Suppose we have to transmit a sequence of 1's and 0's using the given coding scheme. Let the sequence to be encoded is 1001010011 . . . .

Cont..

The procedure is

- The first step is to break the sequence in group of two bits. So we partitions the sequence as follows:
  10 01 01 00 11 ....
- Next replace each block by corresponding codeword.
  11110 10100 10100 00000 11001 ...
  Thus, 5 bits are sent for every two bits of uncoded message. We can observe that for every 2 bits of information, we are sending 3 extra bits(redundancy)

Linear code

A linear code has the following properties:

- The sum of two codewords belonging to the code is also a codeword belonging to the code.
- The all-zero codeword is always a codeword.
- The minimum hamming distance between two codewords of a linear code is equal to the minimum weight of any non-zero codeword *i*.e $d^* = w^*$.

Example

The code $c = \{0000, 1010, 0101, 1111\}$ is a linear block code of block length $n = 4$. This is a $(4, 2)$ code. We can observe that the following are sum of codewords.

There will be ten possible sum.

$0000 + 0000 = 0000$

$0000 + 1010 = 1010$

$0000 + 0101 = 0101$

$0000 + 1111 = 1111$

$1010 + 1010 = 0000$

$1010 + 0101 = 1111$

Example-Cont..

Proceeding in this way, we can observe that all are in c. The minimum distance of this code is $d^* = 2$. In oder to verify the minimum distance of this linear code , we can determine the distance between all pairs of codewords :

$d(0000, 1010) = 2, d(0000, 0101) = 2, d(0000, 1111) = 4$

$d(1010, 0101) = 4, d(1010, 1111) = 2, d(0101, 1111) = 2$

Generation of code using Galois field(GF)

- Let us define a vector space $GF(q^n)$, which is a set of n-tuples of elements from $GF(q)$. Linear block codes can be looked upon as a set of n-tuples of length n over $GF(q)$ such that the sum of two codewords is also a codeword and the product of any codeword by a field element is a codeword. Thus a linear block code is subspace of $GF(q^n)$.
- Let S be a set of vector of length n whose components are defined over $GF(q)$. The set of all linear combinations of vectors of S is called linear span and is noted by $< S >$. Thus a linear span is a subspace of $GF(q^n)$, generated by S.

Generation of code using Galois field(GF)-Cont..

It is possible to obtain a linear code $c = < S >$ generated by S consisting of precisely the following codewords:

1. All zero-word
2. All word in S
3. All linear combination of two or more words in S.

Example

Let $S = \{12, 21\}$ defined over $GF(3)$. The addition and multiplication tables of field $GF(3) = \{0, 1, 2\}$ are given by

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$*$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

Example-Cont..

All possible linear combination of 12 and 21 are as follows $12 + 21 = 00$, $12 + 2(21) = 21$, $2(12) + 21 = 12$. Therefore $c = < S >= \{00, 12, 21, 00, 21, 12\} = \{00, 12, 21\}$.

Example

Let $S = \{1100, 0100, 0011\}$. All possible linear combinations of S are
$1100 + 0100 = 1000$, $1100 + 0011 = 1111$, $0100 + 0011 = 0111$,
$1100 + 0100 + 0011 = 1011$
$c =< S >= \{0000, 1100, 0100, 0011, 1000, 1111, 0111, 1011\}$. The minimum distance of this code is $w(0100) = 1$

Matrix description of a Linear block

- Any code C is a subspace of $GF(q^n)$. Any set of basis vectors can be used to generate the code space. We can define a generator matrix G, the row of which form the basis vectors of the subspace. A linear combination of rows can be used to generate the codewords of C.
- The generator matrix converts(encodes) a vector of length k to a vector of length n. Let the input vector is i. The coded symbols will be given by
  $c = iG$. Where c is the codewords and i is the information word.
- The $n \times k$ matrix can generate q^k codewords.

Example

Consider a generator matrix

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

Information word is $\{00, 01, 10, 11\}$

$$c_1 = \begin{bmatrix} 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

$$c_2 = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$$

$$c_3 = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}$$

Cont..

Therefore this generator matrix generates the code $c = \{000, 010, 101, 111\}$. This is a $(3, 2)$ code from the fact that the dimension of the generator matrix $3x2$. The code rate is $r = 2/3$.

Decoding of a Linear Block Code

- The basic objective of channel coding is to detect and correct errors when the message are transmitted over a noisy channel.
- The noise, for example, changes just one of the symbols in the transmitted codeword, the erroneous codeword will be at a Hamming distance of one from the original codeword.
- If the noise transforms t symbols ($i.e$ t symbols in the codeword are in error), the Hamming distance of the received word will be at a Hamming distance of t from the originally transmitted codeword.

Detection and correction

Given a code, how many errors it can detect and how many it can correct?

- If the minimum distance between the codewords is d^*, the weight of the error pattern must be d^* or more to cause a transformation from one codeword to another.
- Therefore, an (n, k, d^*) code will detect at least all non-zero error patterns of weight less than or equal to $(d^* - 1)$.
- Moreover at least one error pattern of weight d^* which will not be detected.

Example-I

- For the code $C_1 = \{000, 111\}$ the minimum distance is 3. Therefore the error pattern of weight 2 or 1 can be detected. This means that the error pattern belonging to the set $\{011, 101, 110, 001, 010, 100\}$ will be detected by this code.
- Consider the code $C_2 = \{001, 110, 101\}$. Nothing can be said how many errors of this code can detect because $d^* - 1 = 0$.

Nearest Neighbour decoding

- Let we observe the codeword received and check with the originally transmitted codewords in term of Hamming distance.
- The valid codeword is the nearest to the received codeword. This technique is called **Nearest Neighbour decoding**.
- It may be possible that more than one codeword is at the same Hamming distance from the received word. In this case the receiver can do the following:

Nearest Neighbour decoding-Cont...

1. It can pick one of the equally distance neighbours randomly or
2. Request the transmitter to re-transmit.

To ensure that the received word that has at most t errors is closet to the original codeword and farther from the other codewords, we must have to follow the condition that $d^* \geq 2t + 1$.

Example-I

Let us consider the code $C = \{00000, 01010, 10101, 11111\}$. The minimum distance $d^* = 2$. Suppose the codeword 11111 was transmitted and received word is 11110, *i.e* $t = 1$. One error has occured in the fifth component. Now,
$d(11110, 00000) = 4$, $d(11110, 01010) = 2$
$d(11110, 10101) = 3$, $d(11110, 11111) = 1$
Using the nearest neighbour decoding, we can conclude that 11111 was transmitted.
In this case $d^* < 2t + 1 = 3$. Hence for certain scenarios, it is possible to correct error

Example-II

Suppose, 00000 was sent and 01000 was received.
$d(01000, 00000) = 1$, $d(01000, 01010) = 1$
$d(01000, 10101) = 4$, $d(01000, 11111) = 4$
In this case there cannot be clear-cut decision. We can randomly select one.

Example-III(**International Standard Book Number(ISBN))**

It is a linear block code of block length $n = 10$ over $GF(11)$. The symbol used are $0, 1, 2 \ldots 9, X$. Instead of using the symbol "10". The ISBN satisfies the following constraint:

$$\sum_{i=0}^{9}(10 - i)c_i \equiv 0 \bmod 11$$

e.g consider the ISBN $0 - 07 - 048297 - 7$
$10\times0+9\times0+8\times7+7\times0+6\times4+5\times8+4\times2+3\times9+2\times7+1\times7 = 176 \equiv 0 \bmod 11$

Cont..

Since ISBN is a linear block code, the all zero-codeword is a valid codeword. Also 1000000001 is a valid codeword.
Suppose one of the digit of the ISBN gets smudged and we have $0-07-048e97-7$. We can recover the erroneous digit e by solving
$10\times0+9\times0+8\times7+7\times0+6\times4+5\times8+4\times e+3\times9+2\times7+1\times7 = 168+4e \equiv 0 \bmod 11$. The solution of this equation with one unknown yields $e = 2$.

Equivalent Codes

Two linear q-ary codes are called **equivalent** if one can be obtained from the other by one or more operations listed below:

1. multiplication of components by a non-zero scalar.
2. permutation of position of the codeword.

Cont..

Two $k \times n$ matrices generate equivalent linear (n, k) codes over $GF(q)$ if one marix can be obtained from the other by sequence of the following operations:

1. permutation of rows
2. multiplication of rows by non-zero scalar
3. Addition of a scalar multiplication of one row to another
4. permutation of columns
5. multiplication of any column by non-zero scalar.

Systematic form

- A generator matrix can be reduced to **Systematic Form** also called **standard form** of the generator matrix.
- The standard form is represented as $G = [I|P]$, where I is a $k \times k$ identity matrix and P is a $k \times (n-k)$ matrix called **Parity Matrix**.

Example

Consider the generator matrix of $(4, 3)$ code over $GF(3)$

$$G = \begin{bmatrix} 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 2 & 2 & 1 \end{bmatrix}$$

Replacing r_3 by $r_3 - r_1 - r_2$

$$G = \begin{bmatrix} 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 2 & 2 & 0 \end{bmatrix}$$

Cont..

Next replace r_1 by $r_1 - r_3$ to obtain

$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}$$

Finally shifting $c_4 \rightarrow c_1$, $c_1 \rightarrow c_2$, $c_2 \rightarrow c_3$ and $c_3 \rightarrow c_4$

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix} = \left[\begin{array}{ccc:c} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 2 \end{array}\right] = [I|P]$$

Parity Check Matrix

- One objective of good code design is to have fast and efficient encoding and decoding techniques.
- Multiplying the generator matrix with the input vector (uncoded word), we obtain the codewords.
- Also we can detect a valid codeword using **parity check matrix** H for a given code c.
- For a parity check matrix $cH^T = 0$, where c is a valid codeword.

Parity Check Matrix(Cont..)

Since $c = iG$, therefore, $iGH^T = 0$
This is true for all valid codewords, we must have $GH^T = 0$

- A parity check matrix provides a simple method of detecting whether an error has occurred or not.
- If the multiplication of the received word (at the receiver) with the transpose of H yields a non-zero vector, it implies an error has occurred.
- However this methodology will not work if the number of errors in the transmitted codeword exceeds the number of error for which the coding scheme is designed.

Parity Check Matrix(Cont..)

Suppose the generetor matrix is reresented in its systemetic form $G = [I|P]$. P is the co-efficient matrix. The parity check matrix is defined as $H = [-P^T|I]$. This is because GH^T

$$GH^T = \left[\begin{array}{c|c} I & P \end{array}\right] \begin{bmatrix} -P \\ I \end{bmatrix} = 0$$

Since the choice of generator matrix is not unique for a code, the parity check matrix will not be unique. Given a generator matrix G, we can determine the corresponding parity check matrix.

Example

Let us consider a $(7, 4)$ linear block code. The generator matrix is given by

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

the parity matrix P is given by

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

Example-Cont..

$$P^T = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

Note the $-1 = 1$ for the case of binary. We can write the parity check matrix as

$H = [-P^T | I] =$

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

Coset

Let C be an (n, k) code over $GF(q)$ and a be any vector of length n. The the set

$$a + C = \{a + x | x \in C\}$$

is called a coset or translate of C. a and b are said to be in the same coset if $(a - b) \in C$. The vector having minimum weight in a coset is called **coset leader**.

Coset

Let C is an (n, k) code over $GF(q)$, then

1. every vector b of length n is in some coset of C.
2. each coset contains exactly q^k vectors.
3. two cosets are either disjoint or coincide (partially overlap is not possible).
4. if $a + C$ is a coset of C and $b \in a + C$, we have $b + C = a + C$.

Example

Let C be the binary $(3, 2)$ code with generator matrix given by

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

$i.e\ C = \{000, 010, 101, 111\}$. The cosets of C are
$000 + C = \ 000, 010, 101, 111$
$001 + C = \ 001, 011, 100, 110$
Note the all the eight vectors have been covered by these two cosets. If $a + C$ is a coset of C and $b \in a + C$, we have $b + C = \ a + C$.

Example-Cont..

Since two cosets are either disjoint or coincide, the set of all vectors $GF(q)^n$ can be written as

$$GF(q)^n = C \cup (a_1 + C) \cup (a_2 + C) \cup \cdots \cup (a_t + C),$$

where $t = q^{n-k} - 1$.

Standard array

A **Standard array** for an (n, k) code C is a $q^{n-k} \times q^k$ array of all vectors in $GF(q^n)$ in which the first row consists of the code C (with 0 on the extreme left) and the other rows are the cosets $a_i + C$, each arrange in corresponding order, with the coset leader on the left.

Steps for constructing a standard array

1. In the first row write down all the valid codewords, starting with the all-zero codeword.
2. Choose a vector a_1 which is not in the first row. Write down the coset $a_1 + C$ as the second row such that $a_1 + x$ is written under $x \in C$.
3. Next choose another vector a_2 (not present in the first two rows) of minimum weight and write down the coset $a_2 + C$ as the third row such that $a_2 + x$ is written under $x \in C$.
4. Continue the process until all the cosets are listed and every vector in $GF(q^n)$ appears exactly once.

Example

Consider the code $C = \{0000, 1011, 0101, 1110\}$. The corresponding standard array is

$$\begin{array}{cccc} 0000 & 1011 & 0101 & 1110 \\ 1000 & 0011 & 1101 & 0110 \\ 0100 & 1111 & 0001 & 1010 \\ 0010 & 1001 & 0111 & 1100 \end{array}$$

Coset leader are 0000, 1000, 0100 and 0010 respectively.

Decoding

- Since the standard array comprises all possible words belonging to $GF(q^n)$, the received word can always identified with one of the elements of the standard array.
- If the received codeword is a valid codeword, it is concluded that no errors have occurred.
- This conclusion may be wrong with a very low probability of error, when one valid codeword gets modified to another valid codeword due to noise!
- In this case, when received word v, does not belong to the set of the valid codewords, we deduce that, an error has occurred.
- The decoder then declares that the coset leader is the error vector e and decodes the codeword as $v - e$.
- This is the codeword at the top most column containing v

Example

- Let the codeword in the previous example $C = \{0000, 1011, 0101, 1110\}$ is used and the received word is $v = 1101$. Since it is not one of the valid codewords, we deduce that, and error has occurred.
- Next we try to estimate which one of the four possible codewords was actually transmitted.
- If we make use of the standard array of the earlier example, we find that 1101 lies in the 3^{rd} column.
- The top most entry of this column is 0101. Hence the estimated codeword is 0101.

Example(Cont..)

We can observe that

$$d(1101, 0000) = 3, d(1101, 1011) = 2$$
$$d(1101, 0101) = 1, d(1101, 1110) = 2$$

error vector $e = 1000$, the coset leader

Syndrome Decoding

Suppose H is a parity check matrix of an (n, k) code. Then for any vector $v \in GF(q)^n$, the vector

$$s = vH^T$$

is called the **syndrome** of v. The syndrome of v is explicitly written as $s(v)$. It is called syndrome, because it gives us the symptoms of the error. This helps to diagnose the error. The size of s is $1 \times (n - k)$.

Syndrome Decoding(Cont..)

Two vectors x and y are in the same coset of C if and only if they have the same syndrome.

The vectors x and y belongs to the same coset $\Longleftrightarrow$

$x + C = y + C$

$\Longleftrightarrow x - y \in C$

$\Longleftrightarrow (x - y)H^T = 0$

$\Longleftrightarrow xH^T = yH^T$

$\Longleftrightarrow s(x) = s(y)$

Thus there is one to one correspondence between corsets and syndromes

Probability of Error Correction

The **Probability of error** or the word error rate P_{err} for any decoding scheme is the probability that the decoder output is a wrong codeword. It is also called **Residual Error Rate**.

- Suppose there are M codewords of length n which are used with equal probability. Let the decoding be done by using a standard array.
- Let the number of coset leaders with weight i is denoted by α_i.
- Assume that, the channel is a binary symmetric channel(BSC) with symbol error probability p (error probability for a single bit).
- A decoding error occurs if the error vector e is not a coset leader.

Probability of Error Correction(Cont..)

Therefore, the probability of correct decoding will be

$$P_{cor} = \sum_{i=0}^{n} \alpha_i p^i (1-p)^{n-i}$$

Hence the probability of error will be

$$P_{err} = 1 - \sum_{i=0}^{n} \alpha_i p^i (1-p)^{n-i}$$

Example

Consider the standard array for code
$C = \{0000, 1011, 0101, 1110\}$.

$$\begin{array}{cccc} 0000 & 1011 & 0101 & 1110 \\ 1000 & 0011 & 1101 & 0110 \\ 0100 & 1111 & 0001 & 1010 \\ 0010 & 1001 & 0111 & 1100 \end{array}$$

Coset leader are 0000, 1000, 0100 and 0010 respectively.
$\alpha_0 = 1$, $\alpha_1 = 3$, $\alpha_2 = 1$ and all other $\alpha_i = 0$.

$$P_{err} = 1 - [(1-p)^4 + 3p(1-p)^3]$$

Perfect Codes

For any vector u in $GF(q^n)$ and an any integer $r \geq 0$, the sphere of radius r and center u, denoted by $S(u, r)$ is the set

$$\{v \in GF(q^n) | d(u, v) \leq r\}$$

A sphere of radius r $(0 \leq r \leq n)$ contains exactly

$$\binom{n}{0} + \binom{n}{1}(q-1) + \binom{n}{2}(q-1)^2 + \ldots \binom{n}{r}(q-1)^r \text{ vectors.}$$

Example

Consider a binary code $i.e\ q = 2$ and block length $n = 4$. The number of vectors at distance 2 or less from any codeword will be

$$\binom{4}{0} + \binom{4}{1}(1) + \binom{4}{2}(1)^2 = 1 + 4 + 6 = 11$$

Without loss of generality we can choose the fixed vector $u = 0000$. The vectors at a distance 2 or less are
vector at a distance $2 : 0011, 1001, 1010, 1100, 0110, 0101$
vector at a distance $1 : 0001, 0010, 0100, 1000$
vector at a distance $0 : 0000$
Total number of vectors $= 11$

Perfect Codes(Cont..)

A q-ary (n, k) code with M codewords and minimum distance $2t + 1$ satisfies

$$M\{\binom{n}{0} + \binom{n}{1}(q-1) + \binom{n}{2}(q-1)^2 + \ldots \binom{n}{t}(q-1)^t\} \leq q^n$$

This bound is called **Hamming Bound** or **Sphere Packing Bound**. For binary code, the Hamming bound will become

$$M\{\binom{n}{0} + \binom{n}{1}(q-1) + \binom{n}{2}(q-1)^2 + \ldots \binom{n}{t}(q-1)^t\} = 2^n$$

A **Perfect Code** is one which satisfies the Hamming bounds *i.e*

$$M\{\binom{n}{0} + \binom{n}{1}(q-1) + \binom{n}{2}(q-1)^2 + \ldots \binom{n}{t}(q-1)^t\} = q^n$$

Example

Consider the binary repeating code

$$C = \begin{cases} 00\ldots0 \\ 11\ldots1 \end{cases}$$

of block length n, where n is odd. In this case $M = 2$ and $t = (n-1)/2$

Substituting these values in the left hand side of inequality for Hamming bound

$$M\{\binom{n}{0} + \binom{n}{1}(q-1) + \binom{n}{2}(q-1)^2 + \ldots \binom{n}{t}(q-1)^t\} \leq q^n$$

We have $2\{\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \ldots \binom{n}{(n-1)/2}\} = 2 \cdot 2^{n-1} = 2^n$.

Hence the repeating code is a perfect code.

Hamming Codes

There are both binary and non-binary Hamming codes. Binary Hamming code have the property that

$$(n, k) = (2^m - 1, 2^m - 1 - m)$$

Where m is a positive integer.

Example

Let $m = 3$ we have $(7, 4)$ Hamming code. The parity check matrix H of (n, k) code has $n - k$ rows and n columns. For binary (n, k) Hamming code, the $n = 2^m - 1$ columns consisting of all possible binary vectors with $n - k = m$ elements, except all zero vector.