Date __/__/__ DNA Storage and Security

CANVAS

→ IT495 - Course Code

Natural computing - bridge between computer science and natural sciences.

→ Mid term test 1 - 30%.
Scribe notes - 20%.
Project - 50%.

→ You should try to do 1 course on biology because computer science and biology combined can do wonders.

→ Eric lender course - secrets of life (Beautiful course)

→ The next century is for biology because it can do wonders which we have never thought earlier.

→ "If you are doing some theory, you should know how to sell it".

→ Nature uses DNA for storage.

→ Genome: In molecular biology and genetics, genome is all genetic information of an organism. It consists of nucleotide sequences of DNA

→ RNA computing, DNA computing, Bacterial computing

V. Imp.

CANVAS

Date __/__/__

→ In DNA,
A, T → complement
G, C → complement
(short form)

→ ACGT is acronym for four types of bases found in DNA molecules.
A → adenine
C → cytosine
G → guanine
T → thymine

→ cloud - bunch of computers distributed across networks

→ storing the data is very much challenging task now, because data generated every day is very huge.

→ for 1 Yottabyte storage, the cost is $100 Trillion which is more than GDP of whole world.

→ So, what is next generation storage device? DNA

→ 1 g of DNA can store 455 EB of data. (EB - Exabytes)

→ structure of DNA - search on Youtube
(Double helix structure)

→ DNA synthesis - write process
DNA sequencing - read process

Date ___/___/___

CANVAS

→ Anything that is ever produced in this world can be stored in 1 KG of DNA.

→ Access time for data stored on DNA is more than current storage technologies.

→ Synthesizer machine- produces physical DNA

→ Error correction is very very important in both reading from DNA and writing on DNA.

→ Codec : A codec compresses or decompresses media files such as storage or videos.

⇒ Lecture 2 :-

→ Codec - Encoding & decoding ~~method~~

→ DNA storage system steps :-

Encoding → synthesis → storage → Retrieval → Sequencing → decoding

(whole course is divided into above 6 parts)

→ Sequencing : reading of DNA

→ Error correcting codes are required for efficient storage and retrieval.

→ Claude E Shannon - father of Information Technology
(A mathematical Theory of Communication)
If you want to understand information theory, read this paper.

→ small molecules follow clifford algebra.

→ Regenerating codes technique is nowadays used in cloud computing

→ integer % n is field if and only if n is prime number.
This is very important because the elements data packets are nothing but elements of finite fields.

→ Z_n is a field, if and only if n is a prime number.
n = 2, 3, 4, 5, 6, 7, 8, 9, 10
✓ ✓ ✗ ✓ ✗ ✓ ✗ ✗ ✗

$GF(4) = \{ a + \alpha b \mid a, b \in Z_2, \alpha^2 + \alpha + 1 = 0 \}$

$= \{ 0, 1, \alpha, \alpha^2 \}$ $\{0, 1, \alpha, 1+\alpha\}$

$= \{00, 01, 10, 11\}$
(bit representations)

3 different representations

We can construct field with any prime power.
In general we can construct,
$GF(p^m) = GF(q)$ where q is prime power
p is prime.

Date ___/___/___

CANVAS

→ here we are talking about data which is discrete and that's why we require this discrete structure which is "finite field".

→ Search on youtube for finite fields

→ $Z_2^n = \{(x_1, x_2, \ldots, x_n) \mid x_i \in Z_2\}$

$R^n = \{(x_1, x_2, \ldots, x_n) \mid x_i \in R\}$

$n = 3$, $Z_2^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$ ← each vector in this space is combination of {001, 010, 100} ↳ this is basis of the vector space

In R^n, there are infinite vectors.
but Z_2^n is finite.

hence Basis for $Z_2^3 = \{001, 010, 100\}$
dimension of $Z_2^3 = 3$ ← size of Basis vector

→ any non-zero vector forms a one dimensional space.

→ If basis $B = \langle \bar{x} \rangle$, $\bar{x} \neq 0$ and $\bar{x} \in Z_2^3$

$C = \langle 110 \rangle$ space generated by 110
$= \{\lambda(110)\}$ $\lambda \in Z_2$
$= \{000, 110\}$

CANVAS

Date ___/___/___

→ here out of 8 vectors we can for 1 one-dimensional service spaces, and 1 ~~dim~~ 0-dimensional service space.

→ how many 2 dimensional service spaces are there? you can choose any 2 which are linearly independent

→ 2 dimensional service space - 4 vectors
1 " " " - 2 vectors
0 " " " - 1 vector

→ Search : what is dimension - 0, 1, 2 service spaces?

→ F_q = GF(q) (field with q number of elements)

$F_q^n = \{(x_1, x_2, \ldots x_n) \mid x_i \in F_q\}$

linear code is subspace of F_q^n.

$C \leq F_q^n$

IMP ✓ linear code which is used for correcting errors is mathematically subspace of a vector space over finite field.

→ If vector space is of dimension n → the basis has n vectors

Date ___/___/___

CANVAS

⇒ lecture-3

→ code $C = \{000, 111\}$ → is 1
here dimension of C because there is only 1 vector 111 which will generate whole space.
→ $\lambda(111)$ where $\lambda = 0$ or 1
so it generates whole space

→ Z_2^n → vector space of dimension n
(linear code)
codespace $C \leq Z_2^n$ k dimensional
→ Linear code C is simply a subservice space of Z_2^n.
In general ex: $C \leq Z_2^{n=3}$ (1-dimensional)
$\{000, 111\}$

→ since it is k dimensional service space it will have a basis
$\beta = \{\bar{v}_1, \bar{v}_2, \ldots \bar{v}_k\}$
(1) all these k vectors are linearly independent

$$\sum_{i=1}^{n} \lambda_i \bar{v}_i = 0 \Rightarrow \lambda_i = 0 \text{ for } \forall i$$

(2) $\langle \bar{v}_1, \bar{v}_2, \ldots, \bar{v}_k \rangle = C$
(span of $\bar{v}_1, \bar{v}_2, \ldots, \bar{v}_k$ is whole space C.

$\therefore \{\lambda_1 \bar{v}_1 + \lambda_2 \bar{v}_2 + \ldots + \lambda_k \bar{v}_k \mid \lambda_i \in Z_2\} = C$
(λ_i are scalars in Z_2)

∴ our code C which is used for correcting errors is nothing but a k dimensional service space.

Date ___/___/___

so, total vectors will be 2^k, because $\forall \lambda_i$ has 2 choices : either 0 or 1.

→ for our example,

$\{000, 111\} = \langle 111 \rangle = \{\lambda(111) \mid \lambda \in \mathbb{Z}_2\}$

We can write it in the form of matrix.

$$G = \begin{bmatrix} \overline{v_1} \\ \overline{v_2} \\ \vdots \\ \overline{v_k} \end{bmatrix}_{k \times n}$$

here in above example G is 111

$\therefore G = [111]$

→ how encoding is done?

k-bits n-bits

$(x_1, x_2, \ldots x_k)\, G_{k \times n} = (y_1, y_2, \ldots y_n)$

if bit is 0 ⇒ $0[111] = [000]$

if bit is 1 ⇒ $1[111] = [111]$

so, $0 \rightarrow 000$

$1 \rightarrow 111$

$C = \{000, 111\}$

so, error correcting codes are subspace of vector space.

Date ___/___/___

CANVAS

if the hamming distance is d, you can correct $\frac{(d-1)}{2}$ errors.

→ encoding is easy, decoding is difficult.

→ $C \leq \mathbb{Z}_2^n$

This linear code is described as,

$C : [n, k, d]_2$ (2 means binary)

n → length ~~of C (subspace)~~

k → dimension

d → hamming distance

minimum

if d is large we can correct many errors.

$$d = \min \{ d_k(\bar{x}, \bar{y}) \mid \bar{x}, \bar{y} \in C, \bar{x} \neq \bar{y} \}$$

example

$$G = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}_{2 \times 3} \rightarrow n = 3, \quad k = 2$$

⇒ $C = \{000, 110, 101, 011\}$

so, $C : [3, 2, 2]_2$

so, we can say it is binary code of length 3, dimension 2 and distance 2.

this is not useful because $\left(\frac{d-1}{2}\right) = 0$ so, you can correct 0 errors. d should be at least 3, so that you can correct errors

$$(0\ 1)\begin{bmatrix}1 & 1 & 0\\ 1 & 0 & 1\end{bmatrix} = [1\ 0\ 1]$$

So, (0 1) is encoded as 101

So, K is important which is dimension of service space and K information bits are converted to n bits.

→ Dual space : All those vectors in a vector space such that their inner product is 0.

$(C^{\perp})$ ← symbol

$$C^{\perp} = \{\bar{y} \in Z_2^n \mid \langle \bar{x}, \bar{y} \rangle = 0,\ \forall\, \bar{x} \in C\}$$

Where $\langle \bar{x}, \bar{y} \rangle = x_1y_1 + x_2y_2 + \cdots + x_ny_n = 0$

if $n=3$ and $C=\{000, 111\}$, then,

$Z_2^3 =$ { 000
001
100
010
110
101
011
111

So, $C^{\perp}$ is set of all vectors in Z_2^3 such that its inner product with all vectors in C is 0.

$$C^{\perp} = \{000, 110, 101, 011\}$$

↓

because $\langle 110, 000 \rangle = (0+0+0) \% 2 = 0$
$\langle 110, 111 \rangle = (1+1+0) \% 2 = 0$

Date ___/___/___

CANVAS

What is dimension of $C^{\perp}$?

ans : 2

Remember that,

$$\dim C + \dim C^{\perp} = \dim Z_2^3$$

$$\therefore 1 + 2 = 3$$

(note: here we are ignoring 000 in calculating dimension)

→ (1) C is self orthogonal if $C \subseteq C^{\perp}$

(2) C is self dual if $C = C^{\perp}$

→ $G_n = \underbrace{[I_k | A]}_{I_{k \times k}}{}_{k \times n}$ (G is generator matrix)

$H_{n-k \times n}$ is parity check matrix (PCM) and its size is $(n-k) \times n$

$t \mapsto$ transpose

$$H = [-A^t | I_{n-k}]$$

note: H is matrix such that $GH^T = 0$

→ example

$$G = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}_{k \times n} = [I_2 | A]_{2 \times 4}$$

$k = 2 \quad n = 4$

$$\Rightarrow A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \Leftrightarrow H = [-A^t | I_{n-k}]$$

$$H = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$ → Parity check matrix

Date ___/___/___

So, any subspace can be described in two ways. (2 matrices)
(1) Generator matrix (G)
(2) Parity check Matrix (H)

$C = \langle G \rangle$
$= \langle \bar{v}_1, \bar{v}_2, \bar{v}_{2^n} \rangle$

$C = \{ \bar{x} \in Z_2^n \mid H\bar{x}^t = 0 \}$

So, many times you describe subspace by describing H. C is null space of H. i.e. all those vectors $\bar{x}$ such that $\boxed{H\bar{x}^t = 0}$

⇒ Lecture -4 Date: 25/08/2022

→ code is collection of binary strings of some length. code is basically subset of Z_2^n.

→ code is called linear if it is subspace. for example:
$C = \{000, 111\}$ is subspace of Z_2^3.

→ $Z_2^n = \{ (x_1, x_2, \ldots, x_n) \mid x_i \in Z_2 \}$

(code)
$C \subseteq Z_2^n$ (subset)

$C : (n, M, d)$
- n ← length
- M ← $|C|$ ← Size of C = # of codewords = ~~dimension~~
- d → Minimum Hamming distance

So, code is simply subset of whole space. You can do encoding, decoding for these kind of codes and these are called nonlinear codes. Nonlinear means if you take two vectors

Date ___/___/___

CANVAS

out of the set C, their sum may not belong to same set C.

and linear code is simply k-dimensional service space.

see the diff in symbol → $C \leq Z_2^n$ (k-dim subspace)

$C : [n, k, d]$ (n: length, k: dimension, d: minimum hamming distance)

(HIP) for ex: $C = \{000, 111\}$ is $[3, 1, 3]$ code (3: length, 1: dim (because there is only 1 vector in basis), 3: min hamming distance)

C can be described by 2 kind of matrices,

(1) Generator matrix (G)

(2) Parity check matrix (H)

Method-1

→ generator matrix : $G_{k \times n} = [I_k | A]$ (standard form of generator matrix)

When you write down all the basis vectors of C in matrix for you will get generator matrix $G_{k \times n}$.

→ as there are k-linearly independent vectors in basis of C, there will be in total 2^k vectors in C. The k-linearly independent vectors span wholespace.

→ C is linear combination of k-linearly independent vectors.

→ Cardinality of C is 2^k.

Date ___/___/___

Method-2 : Parity check Matrix

$H_{(n-k)\times n}$ P.C.M. binary matrix

$C = \{\bar{x} \in Z_2^n \mid H\bar{x}^t = 0\}$ → nullspace of matrix H

If, $G = [I_k \mid A]$ then, (1-bcoz)

$H = [-A^t \mid I_{n-k}]_{n-k \times n}$

→ Rank of G? (How many L.I rows or colums?)
ans : k

→ Rank of H? ans : n-k

→ Relationship between G and H is,
$GH^T = 0$

example :-

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

↓

$C = \{0000, 1011, 0101, 1110\}$ (by taking linear combination from G)

also, $C = \{\bar{x} \in Z_2^n \mid H\bar{x}^t = 0\}$ (we want to verify this).
(here n is 4)

$H\bar{x}^t = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = 0 \Rightarrow \begin{matrix} x_1 + x_3 = 0 \\ x_1 + x_2 + x_4 = 0 \end{matrix}$

Date ___/___/___

CANVAS

remember we want to encode k information bits to n bits

So, matrix H works like,
we have k info. bits, we extend it to n info. bits by adding (n-k) bits more such that $H\bar{x}^t = 0$ for some matrix H.
Which is $(n-k) \times n$
↳ here $k=2$ $n=4$

So, we are adding parity check.

→ $x_1 + x_3 = 0$
⇒ $x_3 = -x_1 = x_1$

→ $x_4 + x_2 + x_4 = 0$
⇒ $x_4 = -x_1 - x_2 = x_1 + x_2$

x_1, x_2 we can control and from that we can get x_3 and x_4.

x_1	x_2	x_3	x_4
0	0	0	0
0	1	0	1
1	0	1	1
1	1	1	0

You can see we are getting same vectors as in C.

If we are using generator matrix (G), multiply with G (vector of length k) which is $k \times n$ and we get n bits. (vector of length n)

If we have parity check matrix, we add extra $n-k$ bits such that $H\bar{x}^t = 0$.

3.1.21

CANVAS

Date ___/___/___

→ if minimum Hamming distance is d, you can correct $(\frac{d-1}{2})$ errors.

→ for d = 2, you may be able to correct some errors but not all errors.

Imp → $\boxed{\text{syn}(y) = Hy^t}$ (syndrome decoding)

syndrome ⇒ This is used to simplify decoding procedure. You can compute syndrome for each vector.

→ (See in slides the value of syn(y) for vectors)

→ y is received bits

→ for decoding using syndrome decoding,
If x is sent and y is received, just compute the syndrome,

$$\text{syn}(y) = Hy^t$$

and whichever bit is 1 in syn(y), flip that bit in received sequence.

Search : what is syndrome decoding?

$C = \{000, 111\} \subset Z_2^3$

find the sphere of radius t along vector $\bar{x}$,

$$S_t(\bar{x}) = \{\bar{y} \in Z_2^3 \mid d(\bar{y}, \bar{x}) \leq t\}$$

$$t = \left\lfloor \frac{d-1}{2} \right\rfloor$$

Date ___/___/___

CANVAS

$S_1(000) = \{\quad\}$

$S_2(111) = \{\quad\}$

decoding is NP-complete problem.

Lecture 5 :-

Date: 25/08/2022

Z_2^m consists 2^m vectors

for $m=2$, $Z_2^2 = \{00, 01, 10, 11\}$

$$H = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}_{2\times 3}$$ (all vectors other than 00)

for $m=3$, $Z_2^3 = \{000, 011, 101, 110, 111, 100, 010, 001\}$

$$H = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}_{3\times 7}$$ (just write binary expression of numbers other than 0 as columns)

(all vectors other than 000 are used as columns)

→ make sure that no 2 columns are multiple of each other.

→ this parity check matrix generates the code known as 'Hamming code'.

Parameters for this code is $[7, 4, 3]$ (n, k, d)

$\bar{x}$, here - indicates that it is vector.

Date ___/___/___

$G = \{ x \in Z_2^3 \mid H\bar{x}^t = 0 \}$

→ How many vectors will be there in C?

(HHP) ans $= 2^4$ because dimension is $k=4$. so total possible vectors are $2^4 = 16$.

$k=4$ indicates there are 4 vectors in generator matrix.

In general we can construct code of length $= 2^m - 1$, dimension $= 2^m - 1 - m$ and distance $= 3$.

$\therefore$ Hm : $[2^m - 1,\ 2^m - 1 - m,\ 3]_2$

Hm : $[2^m - 1,\ 2^m - 1 - m,\ 3]_2$

n k d

since distance is 3, it can correct $\left(\frac{d-1}{2}\right) = \frac{3-1}{2} = 1$ error.

→ Decoding

If $\bar{x} \in C$ is sent and $\bar{y}$ is received, then,

$\bar{y} = \bar{x} + \bar{e}$

($\bar{e}$ error vector) $\bar{e} = (e_1 e_2 \ldots e_n)$

→ $Syn(\bar{y}) = Syn(\bar{x} + \bar{e})$

$= H\bar{x}^t + H\bar{e}^t$

but we know that if $\bar{x} \notin C$ then $H\bar{x}^t = 0$

Date ___/___/___

CANVAS

$\therefore syn(\bar{y}) = H\bar{e}^t$

let $H = (H_1\ H_2 \dots H_n)$ and $e = (0\,0 \dots 0\,1\,0\,0\,1 \dots 0\,1\,0 \dots 0)$
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad a \quad b \quad c$

a, b, c are positions on which 1 is occuring in e vector

$$syn(\bar{y}) = \sum_{i=1}^{n} c_i H_i = H_a + H_b + H_c$$

so, syndrome is basically sum of columns in ~~which~~ parity check matrix where error occured.

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}_{3\times 7}$$

$H : [7\ 4\ 3]$
$\qquad\ \nearrow\ \uparrow\ \nwarrow$
$\qquad n \quad k \quad d$
(length) (dimension)

suppose received vector is $\bar{y} = 1\,1\,0\,1\,0\,1\,1$

$$syn(\bar{y}) = H\bar{y}^t = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \quad (6^{th} \text{ column})$$

this means that 6th bit is corrupted.

so, $\bar{x} = 1\,1\,0\,1\,0\,0\,1$ (we reversed 6th bit)

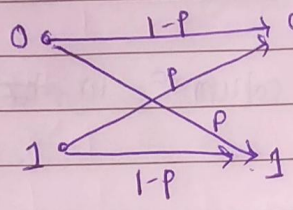
CANVAS

Date ___/___/___

→ weight of the code is $\frac{k}{n}$ → dimension / → length

→ see codetables.de website

→ There are only 3 examples of code which are perfect
(1) repetition code (2) Hamming code (3) Golay code



BSC (Binary symmetric channel)

see the example of probabilistic decoding in ppt.

→ coding theory and cryptography are connected, we can say coding theory is sister of cryptography

⇒ cyclic code

Properties :-
(1) Linear code (there should be basis vector)
(2) any cyclic shift of codeword is again a codeword
i.e $C_0C_1 \ldots C_{n-1} \in C \Rightarrow C_{n-1}C_0 \ldots C_{n-2} \in C$

ex:1 {000, 101, 011, 110}
this is cyclic code because it is linear code and also it follows 2nd property

⇒ {0000, 1001, 0110, 1111} ← ex:2
→ not cyclic because 2nd property is not followed

Date ___/___/___

CANVAS

⇒ {0000, 1010, 0101, 1111} ← ex:3
this is cyclic code because bothe the properties of cyclic codes are followed.

⇒ ex:4 {11111}
not cyclic code because it is not linear because 00000 is not there

⇒ $\{\bar{0}\} = \{000\ldots0\}$
this is ~~not~~ cyclic code

⇒ $\{\lambda I \mid \lambda \in \mathbb{F}_q\}$
It is cyclic code.

⇒ $\mathbb{F}_q^n$ (n-dimensional vector space over q elements)
It is also cyclic code

⇒ mapping to polynomial for representing cyclic codes

$c_0 c_1 \ldots c_{n-1} \longleftrightarrow c_0 + c_1 x + c_2 x^2 + \ldots + c_{n-1} x^{n-1}$
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \| \quad C(x)$

$xC(x) = c_0 x + c_1 x^2 + c_2 x^3 + \ldots + c_{n-1} x^n$

multiplied by x^1 means 1 cyclic shift

now x^n is going outside of space because degree of polynomial is going outwards
$\Rightarrow x^n = 1$

So, remember that whenever in polynomial you encounter x^n, it is 1.

Date ___/___/___

Lecture-6

Date: 26/08/2022

Z_2^3 $Z_2[x]$ $c_0 + c_1x + c_2x^2 + \dots + c_{n-1}x^{n-1}$

000 - 0
100 - 1
010 - x
001 - x^2
110 - $1+x$
011 - $x+x^2$
101 - $1+x^2$
111 - $1+x+x^2$

→ $f(x) = x^2+x+1$

→ $Z_2[x] / (x^2+x+1)$ ← (modulus) → GF(4)

$x^2 = x+1$ $(x^2+x+1=0)$

+	1	x	$x+1$	0
1	0	$x+1$	x	1
x	$x+1$	0	1	x
$x+1$	x	1	0	$x+1$
0	1	x	$x+1$	0

→ $Z_2[x] / (x^2+1)$

↑ (modulus)

*	1	x	1+x
1	1	x	1+x
x	x	1	1+x
1+x	1+x	1+x	0

here (1+x) does not have inverse. so it is not a field.
(there is no column as 1 for (1+x).)

→ $Z_2[x]/_{x^3-1}$ $= R_3$ (Ring) $\quad p(x) = x^3 - 1$

here $n = 3$ so, $x^3 = 1$ (should be 1)

x^3

$$\boxed{Z_2[x]/(x^3-1) = R_3 \text{ (Ring)}}$$

in general,

$Z_2[x]/_{x^n-1} = R_n$ (polynomial ring) → search *

→ lets see what happens for subspace

$$C = \begin{cases} 000 \\ 110 \\ 101 \\ 011 \end{cases} \quad \leq Z_2^3$$

for $C = \{0, 1+x, 1+x^2, x+x^2\} \underset{\text{ideal}}{\leq} R_3$

(it is called ideal code for R_3)

Ring $I \subseteq R$ is ideal if $(\neq \phi)$

(1) $a+b,\ a-b \in I$, $\forall\ a,b \in R$

(2) $r.a \in I$, $\forall a \in R$ & $G \in I$

→ For cyclic code just find Ideal subset.

$R_3 = Z_2[x]\ (x^3-1)$

$I = \{0, 1+x, 1+x^2, x+x^2\}$

$I = \langle 1+x \rangle$

$r(1+x)$, $r \in R_3$

$= I$

→ I is principle Ideal if $\exists\ a \in I$ s.t. $I = \langle g \rangle$

$= \{rg \mid r \in R_3\}$

→ C $(\neq 0)$ cyclic code in $R_n = Z_2[x]/x^n-1$

(1) $\exists\ a\ !$ monic polynomial $g(x)$ of smallest degree in C

(2) $C = \langle g(x) \rangle$

(3) $g(x)$ is factor of (x^n+1)

Date ___/___/___

CANVAS

→ If $g(x) = g_0 + g_1 x + \dots + g_r x^r$ then, C is generated by

$$G = \begin{bmatrix} g_0 & g_1 & g_2 & g_r & 0 \dots 0 \\ 0 & g_0 & g_1 & g_2, \dots, g_r \dots \\ 0 & 0 & \dots 0 & g_0 & g_1 \dots g_r \end{bmatrix} \leftarrow \text{1 NP}$$

How to construct G from gcds

→ How to find all binary cyclic codes of length 3?

$$x^3 - 1 = (x+1)(x^2+x+1)$$

Generator poly	code in R_3	code in Z_2^3
1	All of R_3	all of Z_2^3
$x+1$	$\{0, 1+x, 1+x^2, x+x^2\}$	$\{000, 110, 011, 101\}$
x^2+x+1	$\{0, 1+x+x^2\}$	$\{000, 111\}$
x^3-1	$\{0\}$	$\{000\}$

→ n factorization of $x^n - 1$, no of cyclic codes.

CANVAS

Date ___/___/___

(HW) → for $n=7$ which polynomial should we use for cyclic codes

Some interesting cyclic codes

Read Solomon code	$g(x) =$
Read muller code	$g(x) =$
Hamming code	$g(x) =$
BCH code	$g(x) =$

mother of all codes

⇒ Lecture 7

Binary string - 0,1
ternary " - 0,1,2
quatery " - 0,1,2,3

Most common error is deletion. chunk of letters are deleted.

→ If the encoded string has high minimum distance we can correct more and more errors.

→ Edit distance: Minimum number of insertion, deletion or substitution required to convert one string to another.

Goldman, church papers.

CD, DVD, QR code. High definition TV, deep space communication - uses Read Solomon code

$m_0 + m_1 x + m_2 x^2 + \dots + m_{k-1} x^{k-1}$ - degree k-1 polynomial from message bits

$\mathcal{C}_{DNA}(n, r, d) \subseteq \Sigma_{DNA}^n$

DNA code is subset of Σ_{DNA}^n with length n and has

Date ___/___/___

CANVAS

n codewords with minimum hamming distance d.

→ Main objective :- optimal DNA code with maximum minimum distance d (to correct more errors)

- Maximum storage capacity achieving codes.
- Designing the chunk architecture of DNA blocks.

→ Run length Encoding :

- lossless method
- text or Programs
- Simplest method of compression

Ex: I/P : AAA BBC DDD
Encoded : 3A2B1C3D

→ non Homopolymorphes mapping

previous nt written	next bit to encode		
	0	1	2
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

→ using this map, the corresponding DNA string has no repetition of A, C, G, T.

AA, CC, GG, TT never occur. (non Homopolymorphes)

Date ___/___/___

→ 3 perfect codes,

(1) repetition code
(2) Hamming code
(3) Golay code

→ Generator matrix for extended Golay code
$[24, 12, 8]_2$
total 2^{12} codewords.

→ How ternary Golay code can be constructed through quadratic residue?

→ $[11, 6, \overset{5}{3}]_3$ - Ternary Golay code

# of codewords $= 3^6 = 729$

⇒ Lecture - 8

coderate $= \frac{k}{n}$, k bits are encoded to n bits

challenge is to construct a code which has coderate close to 1.

→ Cd, dvd, optical storage devices, Blue ray disks, use constrained coding

flash memory

→ constrained coding: input is restricted.

Date ___/___/___

CANVAS

→ The # of consecutive 0's and 1's is called as runlength. for certain applications, the runlength value lies between some min value and maximum value.

→ suppose $x = (x_1, x_2 \ldots x_n) \in \{0,1\}^n$

A dk-sequence has 2 constraints.

d constraint : consecutive 1's are separated by at least d zeroes.

111...1 00...0 111...1

≥ d zeroes (d zeroes)

≤ K zeroes (at most K zeroes)

K constraint : consecutive 1's are separated by at most K zeroes.

→ RLL (Run length limited sequence)

$\bar{y} = (y_1 y_2 \ldots y_n)$

$y_i = y_{i-1} \oplus x_i$, $y_0 = 1$

here x is dk-sequence

→ so from dk-sequence we are constructing RLL sequence.

$\bar{x} = 010010001$

$\bar{y} = 100011110$

Run length of RLL sequence lies between d+1 and K+1

CANVAS

Date ___/___/___

→ capacity of constrained channel is,

$$C = \lim_{T \to \infty} \frac{1}{T} \log_2 N(T)$$

$N(T)$ = # of allowed signal sequences in the time interval T.

→ $N_d(n)$ = # of d sequence of length n
(k constraint is absent)

$N_d(n) \cong c\lambda^n \quad n \geq 1$

c is constant and λ is growth factor

$1 \leq \lambda \leq 2$, is the largest real root of characteristic equation.

$$z^{d+1} - z^d - 1 = 0$$

for such d constrained channel,
channel capacity $\boxed{C_d = \log_2 \lambda}$

→ code efficiency = $\frac{\text{code rate}}{\text{capacity of the constrained channel having same runlength constraint}}$ d-

⇒ lecture 9

net information density = $\frac{\text{number of input information bits}}{\text{number of bases in resulting DNA storing}}$

Date ___/___/___

CANVAS

→ GF(q) = {0, 1, α_2, ... α_{q-2}} $q = p^m$

T

Z_p

for $\beta (\neq 0) \in GF(q)$,

order of β: smallest +ve integer m such that $\beta^m = 1$

ex: $Z_5 = \{0, 1, 2, 3, 4\}$

$1^1 = 1 \Rightarrow$ order(1) = 1

$2^3 = 3$

$2^4 = 1 \Rightarrow$ order(2) = 4

Similarly, order(3) = 4 and order(4) = 2

→ here 3 and 2 are special because they are having order = 5 - 1 = 4. Basically they are generating whole Z_5 except 0.

So, 3 & 2 are known as "primitive elements".

→ If you take any element in the field, its order will divide the order of ~~multiplicative~~ group. (i.e. (q-1))

→ each finite field has one or more elements which are PRIMITIVE ELEMENTS.

→ The # of element in GF(q) of order t is $\phi(t)$

CANVAS

Date ___/___/___

$\phi(t)$ is euler totient function.

$\phi(20) = 4$
$\phi(6) = \phi(2*3) = \phi(2) * \phi(3) = 2$
$\phi(15) = 8$

→ GF(q): # of primitive elements is exactly $\phi(q-1)$

→ $GF(2^4) = \{0, 1, \alpha, \alpha^2, \ldots, \alpha^{14}\}$
where α is root of polynomial,
$\Pi(x) = 1 + x + x^4$.
$\Rightarrow \alpha^4 + \alpha + 1 = 0$

$\alpha, \alpha^2, \alpha^4$ are primitive element whereas α^3, α^5 are not.

→ A primitive polynomial (P.P) is a polynomial having primitive element as its root.

→ Finite fields - 2009 - 42 pg

→ If α is primitive element then α^{-1} is also primitive

$\alpha^{-1} = \alpha^{7-1} = \alpha^6$
$\alpha^{-2} = \alpha^{7-2} = \alpha^5$
$\alpha^{-4} = \alpha^{7-4} = \alpha^3$

Date ___/___/___

If $p(\alpha)$ is primitive, then $p^*(\alpha) = x^m p(x^{-1})$

→ Number of primitive polynomials of degree m is $\frac{\phi(q-1)}{m}$.

→ bchcodes- 1009, 4pg

→ set $Z_{p^m-1} \times P$
this divides Z_{p^m-1} into cosets.

cyclotonic coset containing s is,
$C_s = \{s, ps, p^2s, p^3s, \ldots, p^{m_s-1}s\}$
m_s is the smallest +ve integer such that

$$p^{m_s} \cdot s \equiv s \pmod{p^m - 1}$$

Example
mod 15, $p=2$ (Z_{15})

$C_0 = \{0\}$
$C_1 = \{1, 2, 4, 8\}$
$C_3 = \{3, 6, 12, 9\}$
$C_5 = \{5, 10\}$
$C_7 = \{7, 14, 13, 11\}$

These are all cyclotonic cosets.

⇒ Lecture 10

Reed solomon codes
Applications: Flash drive, CD, DVDs, QR code etc

CANVAS

Date ___/___/___

→ 8 bits corresponds to element of a field with 256 elements.

→ The integer representation is divided into blocks of size K.

→ GF(q)

message : $(m_0 m_1 m_2 \ldots m_{K-1}) \in (GF(q))^K$

bits $m_i \in Z_2$

polynomial $p(x) = m_0 + m_1 x + m_2 x^2 + \ldots + m_{K-1} x^{K-1}$

A Reed solomon codeword $\bar{c} \in RS$

$\{ \bar{c} = c_0 c_1 \ldots c_{q-1} = P(0)\, P(\alpha)\, P(\alpha^2) \ldots P(\alpha^{q-1}) \}$

$c_i = P(\alpha^i)$ and

$c_0 = P(0)$

→ $GF(q) = \{0, 1, \alpha, \alpha^2, \ldots, \alpha^{q-2}\}$

$|RS| = q^K$ (size of Reed solomon codes)

and RS codes are "linear codes"

$\dim(RS) = K$

$n = q$ (length of each codeword) ∴ [n K d]

→ $P(x) = m_0$

$P(\alpha) = m_0 + m_1 \alpha + \ldots + m_{K-1} \alpha^{K-1}$

$P(\alpha^2) = m_0 + m_1 \alpha^2 + \ldots + m_{K-1} \alpha^{2(K-1)}$

Date ___/___/___

CANVAS

$$p(\alpha^{q-1}) = m_0 + m_1 q\alpha^{q-1} + \dots + m_{k-1}\alpha^{(q-1)(k-1)}$$

Any K of these equations,

$$\textcircled{X} \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & \alpha & \alpha^2 & \dots & \alpha^{k-1} \\ \vdots & & & & \\ 1 & \alpha^{k-1} & \alpha^{2(k-1)} & \dots & \alpha^{(k-1)(k-1)} \end{bmatrix} \begin{bmatrix} m_0 \\ m_1 \\ \vdots \\ m_{k-1} \end{bmatrix} = \begin{bmatrix} p(0) \\ p(\alpha) \\ \vdots \\ p(\alpha^{k-1}) \end{bmatrix}$$

→ Vandermonde matrix is non-singular matrix as long as α_i are different.

→ here (X) is a nonsingular matrix. (out of q, t bits are corrupted)

→ Suppose that t of the codeword's coordinates are corrupted.

We might construct all possible distinct system of k equations. So there are total $_qC_k$ such systems. and $\binom{t+k-1}{k}$ of this will give incorrect solution.

→ Majority among all the solutions will give correct solution as long as,

$$\binom{t+k-1}{k} < \binom{q-t}{k} \quad \text{Why?}$$

$$\Rightarrow t+k-1 < q-t$$

$$\Rightarrow 2t < q-k+1 \quad \Rightarrow t = \left[\frac{q-k+1}{2}\right]$$

→ Read solomon code has property that distance d is $d = n-k+1$

$\Rightarrow$ $[n, k, n-k+1]$ ($n-k+1 = d$) → MDS (Maximum Distance separable codes)

condition for checking whether is RS code or not? you can find d if t is known.

→ If $m(x) = m_0 + m_1(x) + \dots + m_{k-1}x^{k-1}$

$c(x) = m(x)\, g(x) = \langle g(x) \rangle$

$g(x)$ is generator polynomial.

$$g(x) = \prod_{j=1}^{2t} (x - \alpha^j) \qquad n = q-1$$

→ A polynomial having primitive elements as root is "primitive polynomial". primitive elements has order $(q-1)$. (q is order of field)

→ Reed solomon codes → special case of BCH codes.

⇒ Lecture-11

PCR : polymerase chain Reaction (Makes copy of DNA)

DNA sequencer : machine for reading DNA

Hamming

Date ___/___/___

→ syndrome vector: $Syn(\bar{y}) = H\bar{y}^t = \begin{pmatrix} S_1 \\ S_2 \end{pmatrix}$

H is parity check matrix and $\bar{y}$ is received vector.

→ Hamming code is cyclic code and we can extend it to BCH code which can correct 2 errors.

→ If H is parity check matrix and d is minimum distance of the code then any (d-1) columns of H are linearly independent.

→ $m_i(x)$ is minimal polynomial of α^i.

→ Minimal polynomial is a polynomial of lowest degree having α as its root.
primitive polynomial is a polynomial which is having its root as primitive elements.
primitive polynomial can be minimal polynomial also. but every minimal polynomial may not be primitive polynomial.