Coding thoery
오류 수정 부호 이론

Professor Sang-Mok Kim,
Department of Mathematics,
Kwangwoon University

Spring 2024

About this lecture...

1. TEXTBOOK :

① Code: A First course in Coding Theory (by R. Hill), Clarendon Press(1986)
② Crypto: PDF for this lecture / 이산수학(김상목 저, 드림미디어,2021)

2. 운영

① 100%대면강의 (♣ 참고: 휴일에 한하여 강의동영상 업로드)
② 영어강의 60% : 영문ppt + 국어강의 + 영문과제와 평가
③ Communication : 강의게시판예정/SNS(2주차까지 조교에 의해 카톡방 개설 후 이용가능 전망???)
④ 질의응답: 월/수 pm6:00~8:00 통화로 사전예약 바람-Webex를 통해 비대면 질의응답가능

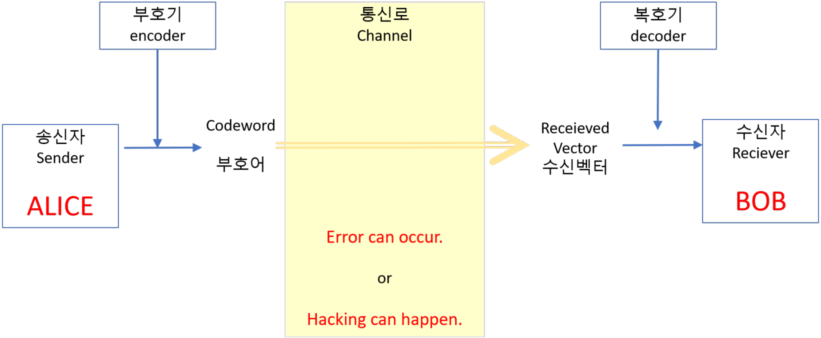
3. 평가

① 출석(10%)
② 과제(10%)
③ 중간시험(40%) + 과제/도전과제(extra α)
④ 기말시험(대면) (40%)

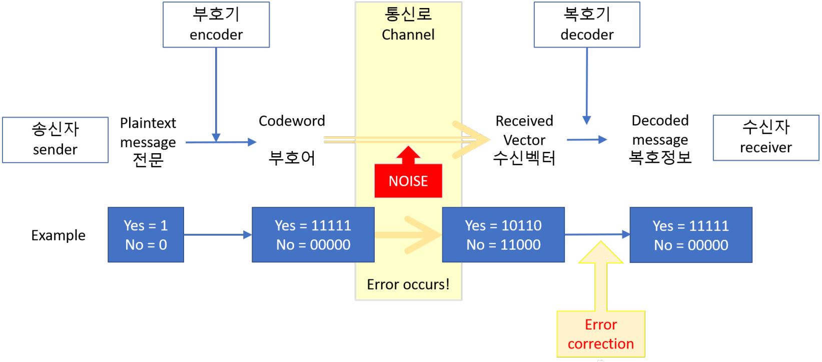
4. 담당교수 정보

① 김상목, 010-6797-1767, smkim@kw.ac.kr, 옥의관 604
② 조교 선생님 : 이원홍(010-8342-1304)

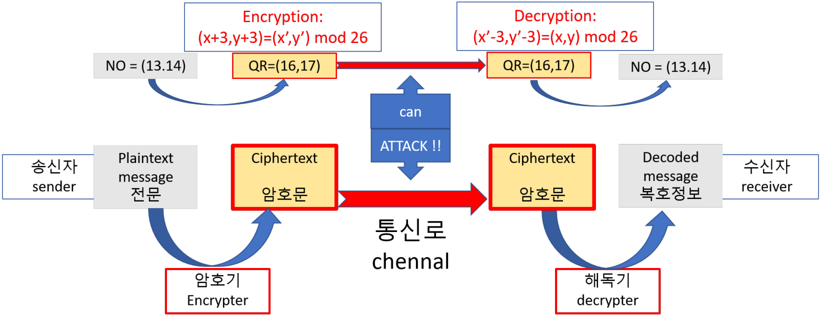
A to B : Infromation transferring



Infromation transfer on codes

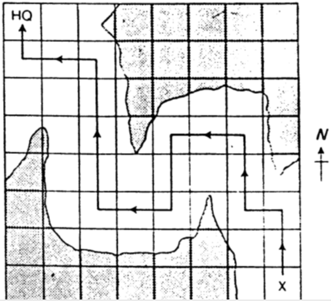


Infromation transfer on crypto



Infromation transfer

$$C_1 = \begin{cases} 0\ 0 = \mathrm{N} \\ 0\ 1 = \mathrm{W} \\ 1\ 0 = \mathrm{E} \\ 1\ 1 = \mathrm{S}. \end{cases}$$



Definition

1. C is a code $\Leftrightarrow C \subseteq Q^n = \underbrace{Q \times Q \times \cdots \times Q}_{n-\text{times}}$.
   Here, Q is a finite set of cardinal q.
   $*$ Mathmatically, C is defined as a subset of $\mathbb{F}_q^n$ where $\mathbb{F}_q^n$ is a finite field of order q.
2. $x \in C$ a code is written as $x = (x_1, \cdots, x_n)$ or $x = x_1, \cdots x_n$.

Note

1. Q can be any set of cardinal q.
2. When $Q^n = \mathbb{F}_2^n$,
   $q = 2$ or $Q = \{0, 1\} = \mathbb{Z}_2$: very typical case : $\{$on, off$\}$ system
   $q = p^k \quad \Leftrightarrow \quad Q = \mathbb{F}_q$: finite field exists.
   $\Leftrightarrow \quad Q = GF(q)$
3. $Q \neq \mathbb{F}_q$, $q = 4$
   $\Rightarrow \exists$ genetic code $Q = \{A, G, C, T\}$ Code(genetic)
   $Q^3 = \{(x, y, z) \mid x, y, z \in Q\}$ = the set of 3-mers
   $\Rightarrow A \subset Q^3$, A= the set of amino acides where $\exists$ 22 amino acides(decoded messages)

Genetic Code=Codon Table

Standard genetic code

1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*)[B]	TGA	Stop (*Opal*)[B]	A
	TTG[A]		TCG		TAG	Stop (*Amber*)[B]	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG[A]		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

Examples of code(mathematically simplified)

1. Messages: 동 서 남 북
   $C_1 = \{00, 01, 10, 11\} = \mathbb{Z}_2^2$
   $C_2 = \{000, 011, 101, 110\}$
   $C_3 = \{00000, 01101, 10110, 11011\}$
2. Repeatition

Yes	No	Error detected	Error corrected
0	1	0	0
0 0	1 1	1	0
0 0 0	1 1 1	2	1
0 0 0 0	1 1 1 1	3	1
0 0 0 0 0	1 1 1 1 1	4	2

Note that

1. We suppose that (the probability of error) $\leq \frac{1}{2}$.
   (infact, real prob(e) $<< \frac{1}{2}$)
2. This error correcting method is called **the nearest neighborhood decoding.**

Definition

(metric or distance on X) A distance on X is a function

$$d \,:\, X \times X \rightarrow \mathbb{R}^{+} \cup \{0\}$$

such that for $x, y \in X$,

(i) $d(x,y) = 0 \Leftrightarrow x = y$

(ii) $d(x,y) = d(y,x)$

(iii) $d(x,y) + d(y,z) \geq d(x,y), \quad \forall x, y \in X$

† Ex) 강의실에서 수강생 사이에 있을 수 있는 distances

- 이름 철자의 다른 정도 = hamming distance
- 좌석의 좌표 간의 Eulid distance

Definition

(Hamming distance) Let Q be a q-set. For $x, y \in Q^n$, define $d : Q^n \longrightarrow \mathbb{N} \cup \{0\}$ called "the Hamming distance" between x and y to be denoted by $d_H(x, y)$, as the number of places in which $\pi_i(x) \neq \pi_i(y)$ for $i = 1, \cdots, n$.

Example

In $\mathbb{Z}_2^5$, let $x = (00111)$, and $y = (11100)$
Then $d_H(x, y) = 4$.

Example

1. (Shown in the next page)Check [Hamming distance] d is a distance on Q^n.

$$\begin{array}{llll} \text{(i)} & x = y & \Leftrightarrow & x_i = y_i \, \forall i \in [n] \\ & & \Leftrightarrow & \text{d(x,y)=0} \end{array}$$

(ii) clear

(iii) $d(x) \leq d(x,z) + d(z,y)$

2.
$$\begin{array}{ll} \text{For } x, y \in \mathbb{R}, & d(x,y) = |x - y| \\ \text{For } x, y \in \mathbb{R}^2, & d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} = ||x - y|| \\ \quad \vdots & \\ \text{For } x, y \in \mathbb{R}^n, & d(x,y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}} \end{array}$$

: [Euclidean distance]

Example

3. For $x, y \in \mathbb{R}^2$, $d(x, y) = |x_1 - y_1| + |x_2 - y_2|$
4. In $\mathbb{R}^2$, $d_\infty((x_1, x_2), (y_1, y_2)) = max\{|x_1 - y_1|, |x_2 - y_2|\}$
   $(\because)$ δ's rule
$$\begin{aligned} & d((x_1, x_2), (z_1, z_2)) + d((z_1, z_2), (y_1, y_2)) \\ = \ & max\{|x_1 - z_1|, |x_2 - z_2|\} + max\{|z_1 - y_1|, |z_2 - y_2|\} \\ \geq \ & max\{|x_1 - z_1| + |z_1 - y_1|, |x_2 - z_2| + |z_2 - y_2|\} \\ \geq \ & max\{|x_1 - y_1|, |y_1 - y_2|\} \\ = \ & d((x_1, x_2), (y_1, y_2)) \end{aligned}$$
5. discrete metric
   X : any set $x, y \in X$
$$d(x, y) = \begin{cases} 0 & x = y \\ 1 & x \neq y \end{cases}$$

Example

6. $C[0,1]$ = the set of all real valued continuous functions on $[0,1]$

$$
\begin{aligned}
d_1(f,g) &= \int_0^1 |f(x)-g(x)|\mathrm{d}x \\
d_2(f,g) &= \left(\int_0^1 |f(x)-g(x)|^2\mathrm{d}x\right)^{\frac{1}{2}} \\
&\vdots \\
d_\infty(f,g) &= max\{|f(x)-g(x)| : 0 < x < 1\}
\end{aligned}
$$

Note

(i) Definition(Hamming distance) Let $x, y \in Q$ and $|Q| = q$. Define

$$d_o = \begin{cases} 0 & x = y \\ 1 & x \neq y \end{cases}$$

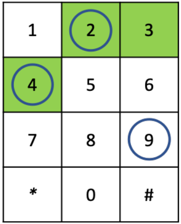
Define $d_H : Q^n \times Q^n \to \mathbb{Z}$ (Hamming distance on Q^n) as

$$d_H((x_1, \cdots x_n), (y_1 \cdots y_n)) = \sum_{k=1}^{n} d_o(x_k, y_k)$$

(ii) $d_H(x, y) = |\{i \in [n] \,|\, x_i \neq y_i\}|$

Remark

In this lecture, Hamming distance is considered as the only metric. (not in general : poset metric



$$d_H(423,422) = 1 = d_H(423,239).$$

Do you think it is reasonable?

Hamming distance was named by Richard Hamming.(Error correcting and error correcting codes, Bell system technical Jonrnal ; 1950)

Parameters of a code

1. A code $C \subset Q^n$ is represented as follows.
   $C = (n, m, d)$ q-ary code
   where

$$
\begin{aligned}
n &= \text{the length of a codeword i.e.,} \\
  &\quad x \in C \Rightarrow n = x_1 \cdots x_n \in Q^n \\
M &= |C| \text{ and} \\
d &= \text{the min distance of } C \\
  &= min\{d(x, y) \,|\, x, y \in C \quad x \neq y\}
\end{aligned}
$$

Symmetric channel

2. Decoding = the nearest neighbor decoding(NND)
   $y \in Q^n$: transmitted signal distorted by noise through the channel.
   (i) $y \in C \Rightarrow$ decode y to y.
   (ii) $y \notin C \Rightarrow$ decode y to $x \in C$.
   if $\exists!\ x \in C$ such that $d(x, y) = min\{d(x', y) \mid x' \in C\}$
   (Note $d(x, y) \geq d$) Otherwise, we can not decode y.

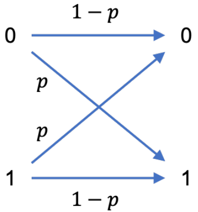
Note that the errors are supposed to the following assumptions made about the chennal.

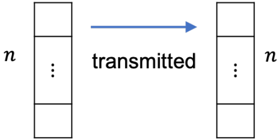
(i) For each symbol x, it has the same prob. $p < \frac{1}{2}$

(ii) If a symbol is received in error, then each of the q-1 possible errors is equally likely.
$\Rightarrow$ q-ary 'symmetric' channel.

3. Assumptins of errors of code for NND in $|Q| = q$.
   (i) Each symbol has the same prob($< \frac{1}{2}$) of being received in error.
   (ii) If a symbol is received in error, then each of the $q-1$ possible errors is equally likely
   On received symbol in error
   - each position in a received vector is same.
   - each alphabet has same error rate
     ex) in Q with $|Q| = q$, $q-1$ alphabet has same error rate from one genuine symbol.

4. Let $p =$ the symbol error probability of binary symmetric channel. Then





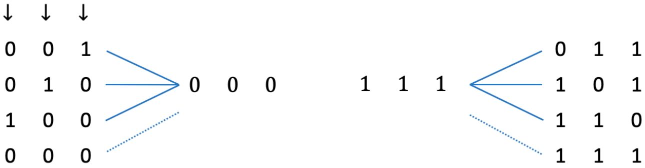
$$\begin{array}{lcl} \text{no error} & \Rightarrow & prob = (1-p)^n \\ \text{one error} & \Rightarrow & prob = p(1-p)^{n-1} \\ & \vdots & \end{array}$$

Example

$C_1 = \{000, 111\} \in \mathbb{Z}_2^3$.   By NND,
The probability that A received vector is decoded as 000 or 111 is

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

ex) $P = 0.01 \Rightarrow P_{error}(C) = 0.000298$.



??

Examples of codes for EWSN

Message	C_1	C_2	C_3
동	10	101	10110
서	01	011	01101
남	11	110	11011
북	00	00	00000
	$d=1$	$d=2$	$d=3$
	$2-(4,2,1)$	$2-(4,3,2)$	$2=(4,5,3)$

Theorem

Under NND,

(i) If $d = d(C) \geq s + 1$, then we detect up to s errors.

(ii) If $d = d(C) \geq 2t + 1$, then we correct up to t errors.

Proof.

(ii) Suppose $d = d(C) \geq 2t + 1$.
Let y be a received vector such that $d(x, y) \leq t$, for some $x \in C$.($\because$ $d(C) \geq 2t + 1$) Asuume that $\exists x' \in C$ with $x \neq x'$ such that $d(y, x') \leq t$. (둘다 $\geq t$ 임에도 불구하고 y가 수정될 수 없는 상황) Then

$$\begin{aligned} d(x, x') &\leq d(y, x) + d(x', y) \\ &\leq t + t \leq 2t \end{aligned}$$

which implies a contradiction since $d(C) \geq 2t + 1$

Corollary

Under NND, if $d(C) = d$ then C can be used either

(i) to detect up to $d - 1$ errors, or

(ii) to correct up to $\lfloor \frac{d-1}{2} \rfloor$ errors.

Examples of Codes q-ary (n, M, d)

(i) C_1 = 2-ary (2,4,1) code
C_2 = 2-ary (3,4,2) code
C_3 = 2-ary (5,4,3) code

(ii) (q-ary repeatition code: q-ary (n, q, n) code)

$$\begin{cases} 00\cdots 0 \\ 11\cdots 1 \\ \quad\vdots \\ q-1\cdots q-1 \end{cases}$$

(iii) 1^{st} order Reed-Muller code $\Rightarrow$ 2-ary (32,64,16)-code
: used by Mariner 9 for transmision of pictures from Mars in (1971, 30 May)

(History of Mariner 9)

5/30	1971	launhed
11/14	1971	reached to Mars
oct	1972	dead

- 1^{st} spacecratf to the ordit of another planet ;
  us beat USSR : Soviet Mars 2, 3 arrived with in a month to Mars.
- 7329 images had been sent for a year

(iv) Genetic Code : m-rna, t-ran $\cdots$
ISBN
MP3
Convolution code

Extra History

1963 Mariner 4 2-(6,64,1) code
- 100×100 화면
- 셀당 2^{60} 명암 000000 $\rightarrow$ 111111 까지
- no error correction
- 22장 전송 8.33 bits/sec : 장당 8시간 소요

1972 Mariner 6,7(8 disappeared) : $\exists$ some improvements

1972 Mariner 9
- Read Miiller 2-(32,64,16) code
- 7 error correction
- 장당 19분 (Mariner 4 대비 2000배 향상)

1976 - Viking 1
- 천연색사진 전송

Good code?

A good code (n, m, d) :

- Shorter n (short length of a codeword)
- Bigger M (more number of codewords)
- Bigger d (more error corrections)

Definition

$A_q(n, d)$ = the largest valuve of M such that $\exists$ a $q-(n, m, d)$ code (for given $q, n,$ and d)

Main coding theory problem

To optimize one of n, m, d Find the largest code of given n and d $= A(n, d)$

Theorem

(i) $A_q(n,1) = q^n$

(ii) $A_q(n,n) = q$

Proof.

(i) $\forall$ codewords are distinct $\Rightarrow \quad d \geq 1$
The largest $q-(n,M,1)$ is the whole of $\mathbb{F}_q^n$

(ii) Suppose C is a $q-(n,M,n)$-code Then $\forall x, y, \in C$,
$\forall i = 1, \cdots, n$, we have $x_i \neq y_i$. Thus $\forall q$ symbols appear at any position. $\therefore M \leq q \quad \therefore A_q(n,n) \leq q$.
Next, $\exists$ the repeatation code of length n and $|C| = q$
$\therefore A_q(n,n) = q$

Note that $A_q(n,d)$ is NP - complete.

Example

$A_2(5,3) = 4$

(i) C_3 is a 2-ary (5,4,3) code
$\therefore A_2(5,3) \leq 4$

(ii) To show $A_2(5,3) \geq 4$, we have to show that $\nexists$ 2-ary (5,5,3) code. It is not so simple to check every cases by hand.

- We have to check $\binom{2^5}{5} = 2^5(2^5-1)\cdots(2^5-4) = 201376$.
- For each case, we have to check $\binom{5}{2}$ distances.
- To calculate a hamming dstance, 5 comparisions of digits to check whether they are different or not.
  $\therefore$ Roughly calculating, there are about at least ten million (10,000,000) operation times.

Show that $A_2(5,3) \le 4$.

Proof.

(i) Suppose that C is a 2-ary (5,5,3) code and WLOG, let $(00000) = x_0 \in C$.

(ii) Then $\forall x (\neq x_0) \in C$, $w(x) \ge 3$.

(iii) $(11111) \notin C$: if so, then

$$\begin{aligned} x(\notin x_0) \in C \quad &\Rightarrow \quad w(x) \ge 3 \\ &\Rightarrow \quad d((11111,),x) \le 2 \quad (\rightarrow\leftarrow) \end{aligned}$$

(iv) $x \in C \Rightarrow x = x_0, \quad w(x) = 3$ or $w(x) = 4$.

(v) $\nexists$ two distinct 4-type codewords i n C (or more)
($\because$ If $y, z \in C$ are two distinct 4-type, then
$3 = d(C) \le d(y,z) \le 2$ $(\rightarrow\leftarrow)$)

(vi)-1 Case : $C = \{00000, y_4, x_1, x_2, x_3\}$ where y_4 is 4-type and x_i 3-type :

WLOG, let $x_1 = 11100 \in C$: then y_4 is one of the followings ;

$$\begin{cases} 11110 \\ 11101 \\ 11011 \\ 10111 \\ 01111 \end{cases}$$

- Let $y_4 = (11011)$. Up to now, we know that $00000, 11100, 11011 \in C$ compare $11100, 11011$ with $\binom{5}{2}$ vectors of weight 3.

$y_4 =$	11011		
$x_1 =$	11100	$d(x_1, x)$	$d(y_4, x)$
	11010	2	
	11001	2	
	10110	2	
	10101	2	
	10011	4	1
	01110	2	
	01101	2	
	01011	4	1
	00111	4	4

$\therefore$ 00111 is the only remaind 3-type codeword
$\therefore \rightarrow\leftarrow$ Since $C = \{00000, 11011, 11100, 00111\}$ which means $|C| = 4$. Similar arguments can be applied that to the otehr cases.

(vi)-2 Suppose that there are 4 3-type codewords in C, and that WLOG $x_1 = 11100 \in C$. Then, from the table, C could be as follows.

$$C = \{00000, 11100, 10011, 01011, 00111\}$$

However, $d(10011, 01011) = 2 < d(C) = 3$
Therefore there is no 2-ary(5,5,3) code.