Coding (13.2, 13.3)

Setting: Want to transmit a message in a given "alphabet", X, a finite set, using symbols of a different alphabet A.

Example: $X = \{A, B, \ldots, Ø, Å, _\}$, $A = \{\cdot, -\}$

Use Morse code to define a function c from X to finite sequences of elements of A.

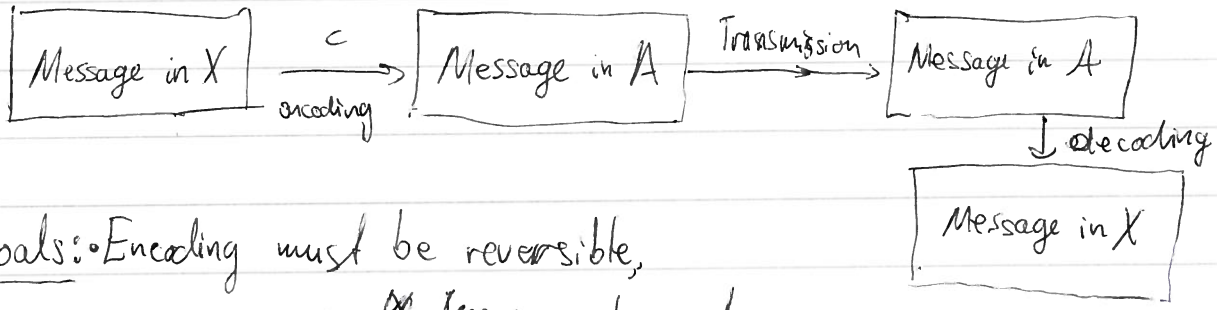
$A^* = \bigcup_{n \geq 0} A^n$, $c(S) = (\cdot, \cdot, \cdot)$, $c(O) = (-, -, -)$.

~~**Example:** $X = \{A, B, \ldots, Ø, Å, _\}$, $A = \{1, \ldots\}$~~

Encode SOS as $\cdots - - - \cdots$

Example: $X = \{A, B, \ldots, Å, _\}$ $A = \{0, 1\}$

Define c by $A = 1$, $B = 2$, $C = 3$, $\ldots$ $Å = 29$, $_ = 30$, then taking binary representation: $c(A) = 1$, $c(B) = 10$, $c(C) = 11$, $c(D) = 100$, $\ldots$



Goals:

- Encoding must be reversible, every message in A comes from at most one X-message
- Encoding should be robust to noise, transmission channel can make small changes to message, want decoding still possible
- Encoding should be efficient, messages in A as "short as possible".

Reversibility: Our letters-as-numbers-code fails this: $c(A) = 1$, $c(C) = 11$, then

$AAAA \rightsquigarrow 1\,1\,1\,1$

$CC \rightsquigarrow 1\,1\,1\,1$,

impossible to decode.

Morse code has $c(E) = \cdot$, $c(S) = \cdots$, distinguish EEE from S by pauses (so Morse code has $A = \{\cdot, -,$ pause$\}$).

Let $c : X \to A^*$ be a source encoding of X. Call $C = c(X)$ the code, elements of C the code words.

Def $C \subseteq A^*$ is a prefix code if no code word is the first part of another code word.

Ex If C is a prefix code & $0 \in C$, then no other code word can begin with a 0.

Ex $X = \{A, B, \ldots, Å, _\}$, $A = \{0, 1\}$

$c(A) = 1$, $c(B) = 01$, $c(C) = 001$, $c(D) = 0001$, ...

Prefix codes can be decoded: Read message from left, when a code word is found, decode that & continue:

101000010000000000001

A B E L .

Given frequency of letters $p : X \to [0, 1]$,

$p(x)$ = probability of given letter being x.

If $l : X \to \mathbb{N}$ is the length of codeword $c(x)$, want to minimise $\sum_{x \in X} l(x) p(x)$

Eg. In Morse code $c(E) = \cdot$, since E is most frequent letter in English.

Optimal solution to encoding problem above is Huffman code.

Noise & error-correction: Suppose we transmit a code word $w \in A^* = \{0,1\}^*$ over a channel. Physical disturbance, "noise", can change the word, turning a 0 into a 1 & vice versa. Want a code such that errors are ① detected & ② corrected.

Check symbols strategy: Given a message $w \in A^*$, decompose in blocks of length k, insert r check symbols between each block.

Example: Double repetition code, take $r = 2k$, repeat each block twice. E.g. $k=4$ $w = 10110110$

$\rightsquigarrow$ blocks 1011 0110

$\rightsquigarrow$ new message 10111011101101100110.

If at most one symbol in each repeated block changes, can still decode: 1001 1011 1011 0100 0110 $\rightsquigarrow$ 10110110

1 2 2 1 choose most common of 3.

If at most two symbols in each repeated block change, can spot errors

1001 1001 1011 must result from errors.

Small information rate: $\frac{\text{length of original message}}{\text{length of encoded message}} = \frac{k}{k+r} = \frac{k}{3k} = \frac{1}{3}$.

Example: Parity check, $r=1$, check symbol = sum of symbols in block mod 2.

w 1011 0110

$1011 \rightsquigarrow 1+0+1+1 \equiv 1 \pmod 2$

$0110 \rightsquigarrow 0+1+1+0 \equiv 0 \pmod 2$

new message 10111 01100.

Can detect 1 error - if number of 1s in message received is odd, error has occurred,

Can not correct error: 10011 could be 10010, 00011, 11011, 10111, 10001.

Information rate high: $\frac{k}{k+r} = \frac{k}{k+1} = 1 - \frac{1}{k+1}$.

Shannon's theorem: Gives a theoretical limit for how efficient a code can be while being decodable under a certain "noise level".

Roughly, probability of error in transmission determines "channel capacity" C, $0 \leq C \leq 1$.

For any $R < C$, & any $\varepsilon > 0$, there is a code with information rate R & "error rate" at most ε.

Example: Suppose channel flips 0 to 1 & 1 to 0 with probability 0.1. Then $C = 0.1 \log_2 0.1 + 0.9 \log_2 0.9 + 1 \approx 0.53$

Can achieve information rate of 0.5, e.g. encode a k-bit message in 2k bits, with as low an error rate as you want.

Remarks: Nonconstructive - doesn't say what code to take

Impractical,

Ex: In above example statement of Shannon's theorem is, if we want error rate ε, can find n & $c: \{0,1\}^n \to \{0,1\}^{2n}$ which gives this. But n may be huge & c is not prescribed.

Def Constructing good codes: A block code C of length n over alphabet A is a subset $C \subseteq A^n$, i.e. collection of code words of length n.

Often $A = \{0,1\}$

Ex Repetition code with blocks of length 2

$$C = \{00\,00\,00,\ 10\,10\,10,\ 11\,11\,11,\ 01\,01\,01\}$$

Ex Parity code with blocks of length 2

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

Remark: Block codes are prefix codes

Question: Receiver obtains a block not in code, what should they guess was sent?

"Closest codeword match" = codeword with fewest symbols differing.

E.g. 100 is closer to 000 (1 symbol differs) than to 111 (2 symbols differ).

Def Let $\vec{a} = (a_1, a_2, \ldots, a_n)$, $\vec{b} = (b_1, \ldots, b_n) \in A^n$. Then Hamming distance between $\vec{a}$ & $\vec{b}$ is

$$\Delta(\vec{a}, \vec{b}) = |\{i \mid a_i \neq b_i\}|$$

Ex $\Delta((1,0,0),(0,0,0)) = 1$, $\Delta((1,0,0),(1,1,1)) = 2$, $\Delta((0,0,0),(1,1,1)) = 3$.

This is a metric: $\Delta(\vec{a}, \vec{b}) \geq 0 \quad \forall\, \vec{a}, \vec{b}$

$\Delta(\vec{a}, \vec{b}) = 0 \Leftrightarrow \vec{a} = \vec{b}$

$\Delta(\vec{a}, \vec{b}) = \Delta(\vec{b}, \vec{a}) \quad \forall\, \vec{a}, \vec{b}$

$\Delta(\vec{a}, \vec{b}) + \Delta(\vec{b}, \vec{c}) \geq \Delta(\vec{a}, \vec{c}) \quad \forall\, \vec{a}, \vec{b}, \vec{c}$

For $\vec{a} \in A^n$, and $t \geq 0$, let

$$B_t(\vec{a}) = \{ \vec{b} \in A \mid \Delta(\vec{a}, \vec{b}) \leq t \},$$

the <u>ball of radius t around $\vec{a}$.</u>

Assume we have a code $C \subseteq A^n$, & know at most t errors can occur under transmission. &

Want: ① Receiver can tell that error in transmission has happened

② – " – _____ which message was sent.

① Sending $\vec{x} \in C$, at most t errors means that some $\vec{y} \in B_t(\vec{x})$ is received. If $y \notin C$, then receiver knows error has occurred.

So, if $B_t(\vec{x}) \cap C = \{\vec{x}\}$ then receiver detects errors.

Conversely if $B_t(\vec{x}) \cap C \neq \{\vec{x}\}$, there is some $\vec{y} \in B_t(\vec{x})$, and receiver would interpret message as $\vec{y}$.

Code is <u>t-error-detecting</u> if

$$\forall \vec{x} \in C,\ B_t(\vec{x}) \cap C = \{\vec{x}\} \iff \forall \vec{x}_1 \neq \vec{y} \in C,\ \Delta(\vec{x}, \vec{y}) \geq t+1$$

② Receiver sees $\vec{y} \in A^n$. Know that message sent $\vec{x} \in C$, has property $\Delta(\vec{y}, \vec{x}) \leq t$. Want this to be unique, that is

$$\forall \vec{y} \in A^n, \text{ at most one } \vec{x} \in C \text{ with } \Delta(\vec{y}, \vec{x}) \leq t$$

$$\Updownarrow$$

$$\forall \vec{x}_1 \neq \vec{x}_2 \in C,\ B_t(\vec{x}_1) \cap B_t(\vec{x}_2) = \phi$$

$$\Updownarrow$$

$$\forall \vec{x}_1 \neq \vec{x}_2 \in C,\ \Delta(\vec{x}_1, \vec{x}_2) \geq 2t+1.$$

Then say C is <u>t-error detecting</u>

Let $d(C) = \min_{\vec{a} \neq \vec{b} \in C} \Delta(\vec{a}, \vec{b})$, the distance of C

$d(C) \geq t+1 \overset{\text{def}}{\iff} C$ is t-error detecting

$d(C) \geq 2t+1 \iff C$ is t-error correcting

$C = \{000, 011, 101, 110\}$, $d(C) = 2$

$C = \{000000, 101010, 010101, 111111\}$ $d(C) = 3$

In code $C \subseteq A^n$, want

- large $d(C)$, i.e. good correction properties
- large $|C|$, i.e. good transmission rate.

Let $n, d, q \geq 0$, let $\mathcal{C}(n, d, A) = \{C \subseteq A^n \mid d(C) \geq d\}$

Thm (Hamming bound) Let $d = 2t+1$, $|A| = q$, $C \in \mathcal{C}(n, d, A)$.
Then

$$|C| \leq \frac{q^n}{\sum_{i=0}^{t} \binom{n}{i} (q-1)^i}$$

Pf Since $d(C) \geq d = 2t+1$, $\forall \vec{a}, \vec{b} \in C$, $\Delta(\vec{a}, \vec{b}) \geq 2t+1$.
This means $B_t(\vec{a}) \cap B_t(\vec{b}) = \emptyset$. Therefore

$$\left| \bigcup_{\vec{a} \in C} B_t(\vec{a}) \right| = \sum_{\vec{a} \in C} |B_t(\vec{a})|.$$

For $0 \leq i \leq t$, the set $\{\vec{b} \in A^n \mid \Delta(\vec{a}, \vec{b}) = i\}$ has
$\binom{n}{i} \times (q-1)^i$ elements — $\binom{n}{i}$ choices of coordinates to change, $(q-1)$ choices per coordinate.

Get

$$q^n = |A^n| \geq \left| \bigcup_{\vec{a} \in C} B_t(\vec{a}) \right| = \sum_{\vec{a} \in C} |B_t(\vec{a})| = |C| \sum_{i=0}^{t} \binom{n}{i} (q-1)^i.$$

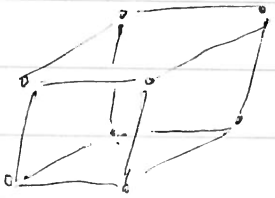
□

Def C is *t-perfect* if $d(C) \geq 2t+1$ and

$$|C| = \frac{q^n}{\sum_{i=0}^{t} \binom{n}{i}(q-1)^i}.$$

Equivalently, if A^n is the disjoint union of $B_t(\vec{a})$ for $\vec{a} \in C$.

Ex $A = \{0,1\}$, $C = \{\overbrace{000\ldots0}^{n=2t+1}, 11\ldots1\}$ is t-perfect. A word $a_1 a_2 \ldots a_n \in A^n$ with $\leq t$ 1s lies in $B_t(0\ldots0)$, a word with $\geq t+1$ 1s lies in $B_t(1\ldots1)$.



Ex $A = \{0,1\}$, $n = 7$.

Let $S = \{1,2,\ldots,7\}$

$P : A^7 \xleftrightarrow[\text{bijection}]{} 2^S$

$(a_1,\ldots,a_7) \longmapsto \{i \mid a_i = 1\}$, e.g. $(1,0,1,0,0,0,1) \mapsto \{1,3,7\}$.

$$\Delta(\vec{a},\vec{b}) = |\{i \mid i \in P(\vec{a}) \setminus P(\vec{b}) \text{ or } i \in P(\vec{b}) \setminus P(\vec{a})\}|.$$

Let $l_1, l_2, \ldots, l_7 \subseteq 2^S$ be lines of Fano plane, let $C = \{\emptyset, l_1, l_2, \ldots, l_7, S\setminus l_1, S\setminus l_2, \ldots, S\setminus l_7, S\}$.

Then $\Delta(\emptyset, l_i) = 3$, $\Delta(\emptyset, S\setminus l_i) = 4$, $\Delta(\emptyset, S) = 7$

$\Delta(l_i, l_j) = 4$, $\Delta(l_i, S\setminus l_i) = 7$, $\Delta(l_i, S\setminus l_j) = 3$

$\Delta(S\setminus l_i, S\setminus l_j) = 3$.

So $d(C) = 3$. Hamming bound for $t=1$ is

$$|C| \leq \frac{2^n}{1+\binom{n}{1}} = \frac{2^7}{1+7} = 16 = |C|,$$

C is 1-perfect.