Basics of Coding Theory

1.) NOISLESS CODING THEORY

→ Shannon Entropies

→ Huffman Coding

2.) Noisy coding theory

→ Error correcting codes

Noiseless coding theory

Random variable X $\{x_0, \dots, x_{n-1}\}$
$p_0, \dots, p_{n-1}$

$\Sigma \rightarrow$ alphabet $\quad \Sigma = \{0,1\}$

codewords

$0 \rightarrow 00$
$1 \rightarrow 01$
$2 \rightarrow 10$
$3 \rightarrow 11$

is this the most efficient?

→ code C

$\log_2 n$ bits to find codewords for n different signals.

$$AVG(C) = \sum_i p_i \cdot |c_i|$$

→ length of codeword for ith signal.

1) given a probability distribution (random variable) what is the best achievable average $AVG(C)$?

2.) How to construct the best code?

1.) for a random variable X w.p. $(p_0, \dots, p_{n-1})$

$$\boxed{S(X) = -\sum_i p_i \log_2 p_i}$$

Shannon entropy

I. Average of code C for r.v. X $\geq S(X)$

II. Encoding multiple messages together helps

III. As the number of messages encoded together approaches infinity $S(X)$ is achievable

2n) Huffman coding

Alg.

INPUT: Probability distribution

OUTPUT: Optimal code

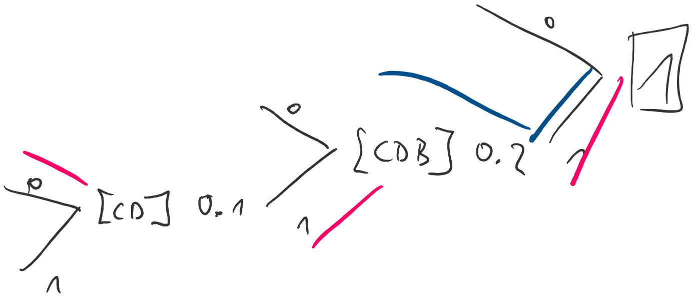
$\boxed{\text{Ex } 1.2}$

A 0.8

B 0.1

C 0.05

D 0.05



A → 0

B → 10

C → 110

D → 111

$$AVG(C) = (0.8)\cdot 1 + (0.1)\cdot 2 + (0.05)\cdot 3 + (0.05)\cdot 3$$

$$= 1.3$$

$$\vee$$

$$S(X) = -(0.8)\cdot \log_2(0.8) - (0.1)\cdot \log_2(0.1) - 2\cdot(0.05)\cdot \log_2(0.05)$$

AA $(0.8)^2$

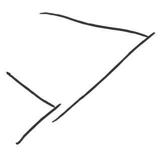
AB $(0.8)\cdot(0.1)$

⋮

A 1/3

B 1/3

C 13



A → 0

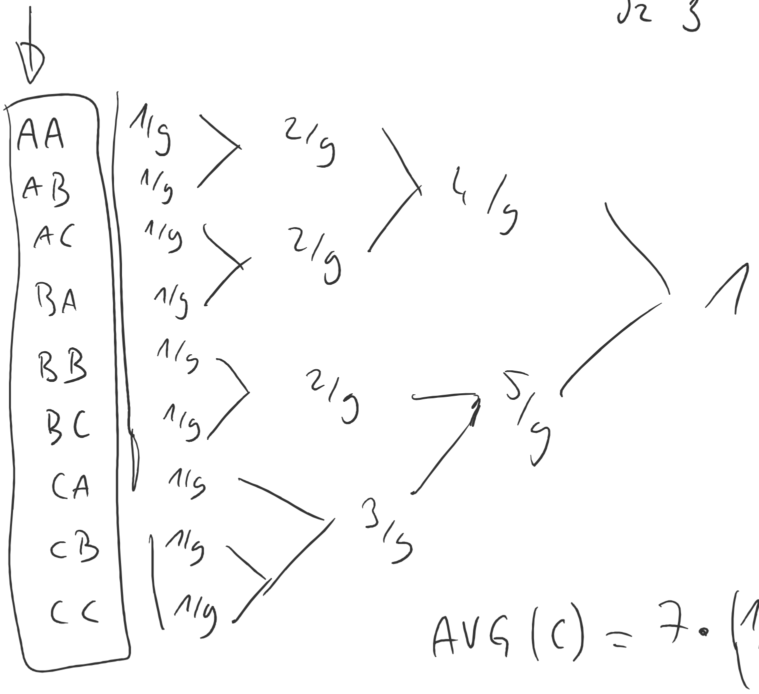
B → 10

C → 11

$$AVG(C) = \tfrac{1}{3}\cdot 1 + 2\cdot\left(\tfrac{1}{3}\cdot 2\right) = \frac{5}{3} \approx 1.66\overline{6}$$

$$S(X) = \left(-\tfrac{1}{3}\log_2 \tfrac{1}{3}\right)\cdot 3$$

$$= -\log_2 \tfrac{1}{3} = 1.5\ldots$$



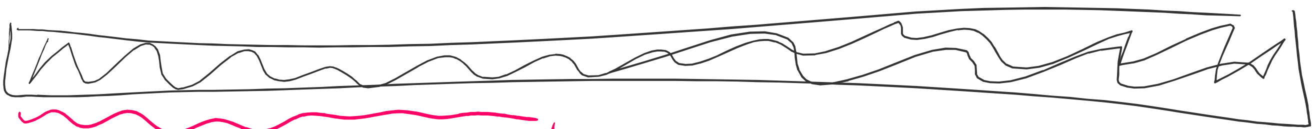
$$AVG(C) = 7\cdot\left(\tfrac{1}{9}\cdot 3\right) + 2\left(\tfrac{1}{9}\cdot 4\right)$$

$$= 3{,}22222$$

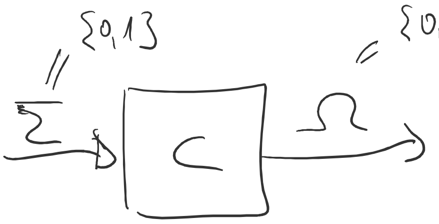
$$\frac{AVG(C)}{2} = 1{,}61\overline{11}$$

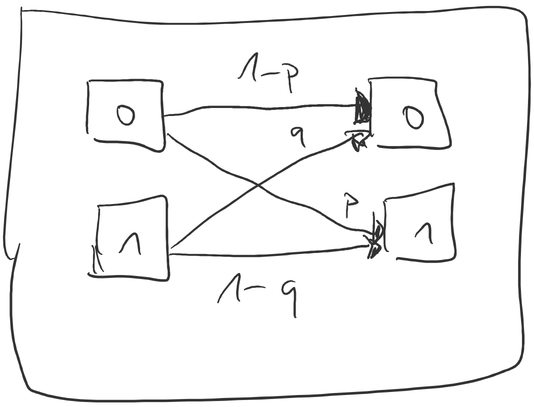
if you encode k symbols together

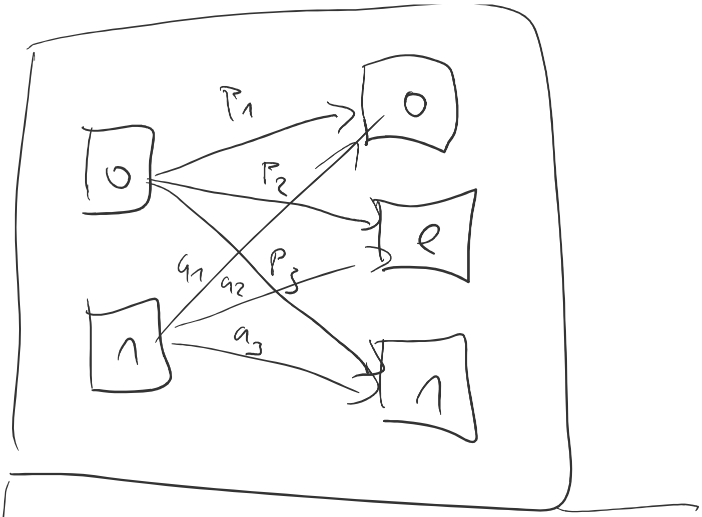
Then $$\lim_{k\to\infty} \frac{AVG(C_{Huff})}{k} = S(X)$$



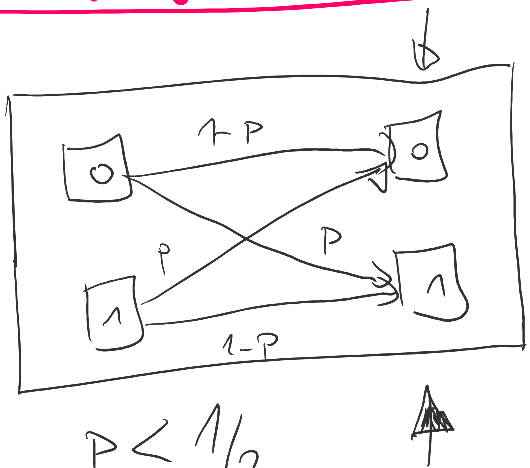
Noisy coding theory (ERROR CORRECTING CODES)



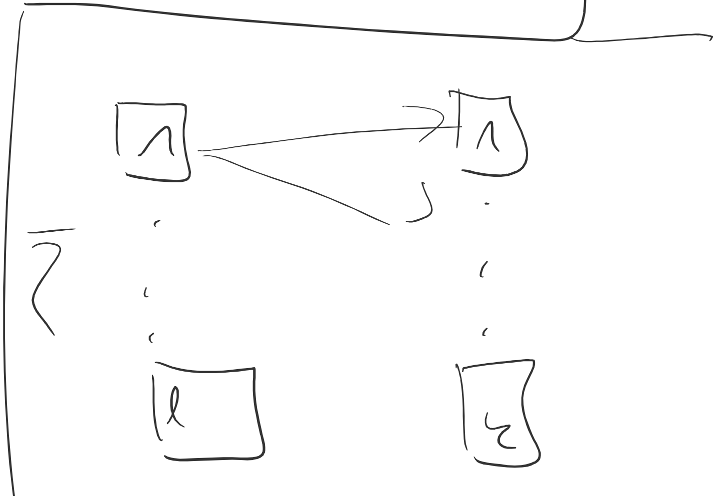




Binary symmetric channel



$p < 1/2$



$p(k|l)$ probability to recieve k if l was sent

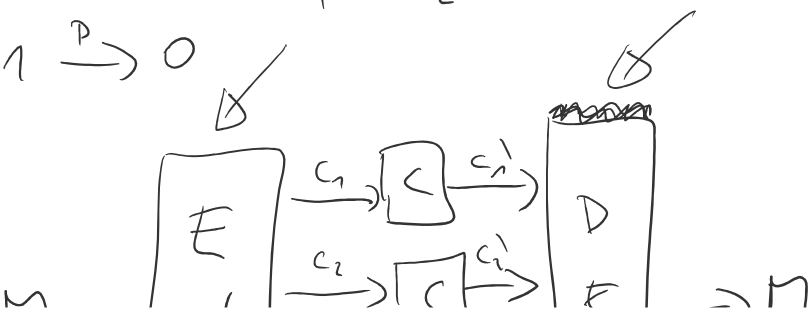
PRINCIPLE OF MAXIMUM LIKELIHOOD

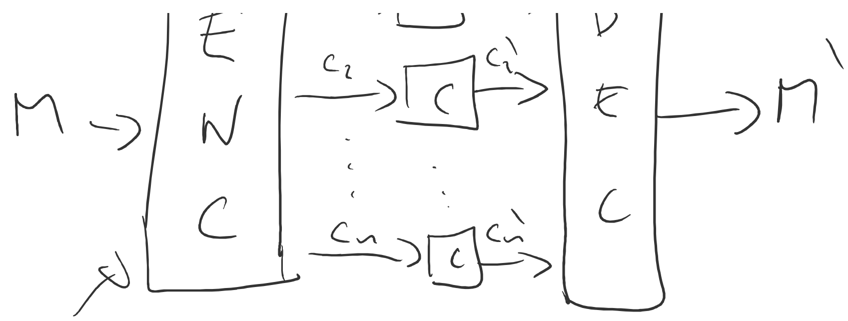
You receive 0 from a binary symmetric channel!

How do you interpret (decode) it?

$0 \xrightarrow{1-p} 0 \qquad p < 1/2 \Rightarrow \quad p < 1-p$

$1 \xrightarrow{p} 0$





$0 \to 000$ Decoding rule $\#1 \geq 2$ decode as 1

$1 \to 111$ $\#1 < 2$ decode as 0

$$\Pr(\underset{\text{input}}{000} \mid \underset{\text{output}}{001}) = (1-p)(1-p)\,p \Leftarrow p < 1/2$$

$$\vee$$

$$\Pr(111 \mid 001) = p \cdot p \cdot (1-p)$$

$\boxed{01}$

Repetition code can achieve arbitrary low probability of wrong decoding

$0 \to \overbrace{00 \ldots 0}^{2k+1}$ Decoding rule $\#1 \geq k+1 \Rightarrow 1$

$1 \to 11 \ldots 1$ $\boxed{\# < k+1 \Rightarrow 0}$

$$\Pr(\text{correct decoding}) = \sum_{i=0}^{k} \binom{2k+1}{i} p^i (1-p)^{2k+1-i}$$

000 100 010 001

111 110 101 011

$$\lim_{k \to \infty} = 1$$

$$\frac{\#\text{Messages}}{\text{length of codewords}} = \frac{2}{k} \xrightarrow[k \to \infty]{} 0 \quad \Big| \text{ code rate}$$

Hamming Distance

c_i – codewords

$C \subseteq \{0,1\}^n$ codes

$c_i \in C$

$Ham(c_i, c_j)$ the number of positions in which c_i and c_j differ.

$\boxed{\text{Ex } 1.6}$ $\boxed{11111}$

$\{10001, 00110, 11010, 01101\}$

$Ham(10001, 00110) = 4$ $Ham(00110, 11010) = 4$

$Ham(10001, 11010) = 3$ $Ham(00110, 01101) = 3$

$Ham(10001, 01101) = 3$ $Ham(11010, 01101) = 4$

$n = 5$ (length of codewords)

$M = 4$ (number of codewords)

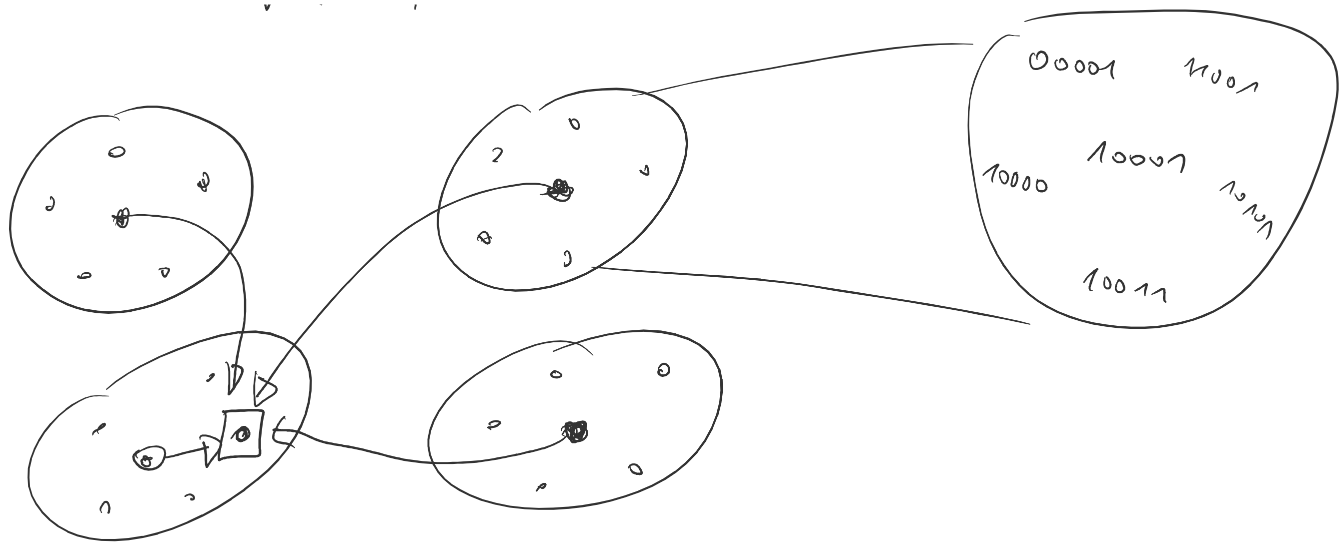
$d = 3$ (minimal distance)

Error detection → Output of a channel is not a codeword ($d-1$ errors)

Error correction → $d = 2t+1$ code can correct up to t errors

Pr of e (specific errors) $p^e(1-p)^{n-e}$

Pr (e specific errors) > Pr (e+1 specific errors)



$\boxed{10001}$ 1 error $\to$ 00001

$\boxed{11111}$

$n \downarrow$

$M \uparrow$

$d \uparrow$

$A_q(n, M)$ = the largest d of a code with M codewords of length n (over q-ary alphabet)

q: code alphabet size (usually $q=2$)

Sphere packing bound

$d = 2t+1$

$$\boxed{M} \cdot \left[\binom{n}{0} + \binom{n}{1}\cdot(q-1) + \binom{n}{2}\cdot(q-1)^2 + \dots + \binom{n}{t}(q-1)^t \right] \leq q^n$$

M: choose c_1

$\binom{n}{0}$: the number of codewords of distance 0

$\binom{n}{1}(q-1)$: the number of codewords of distance 1

2

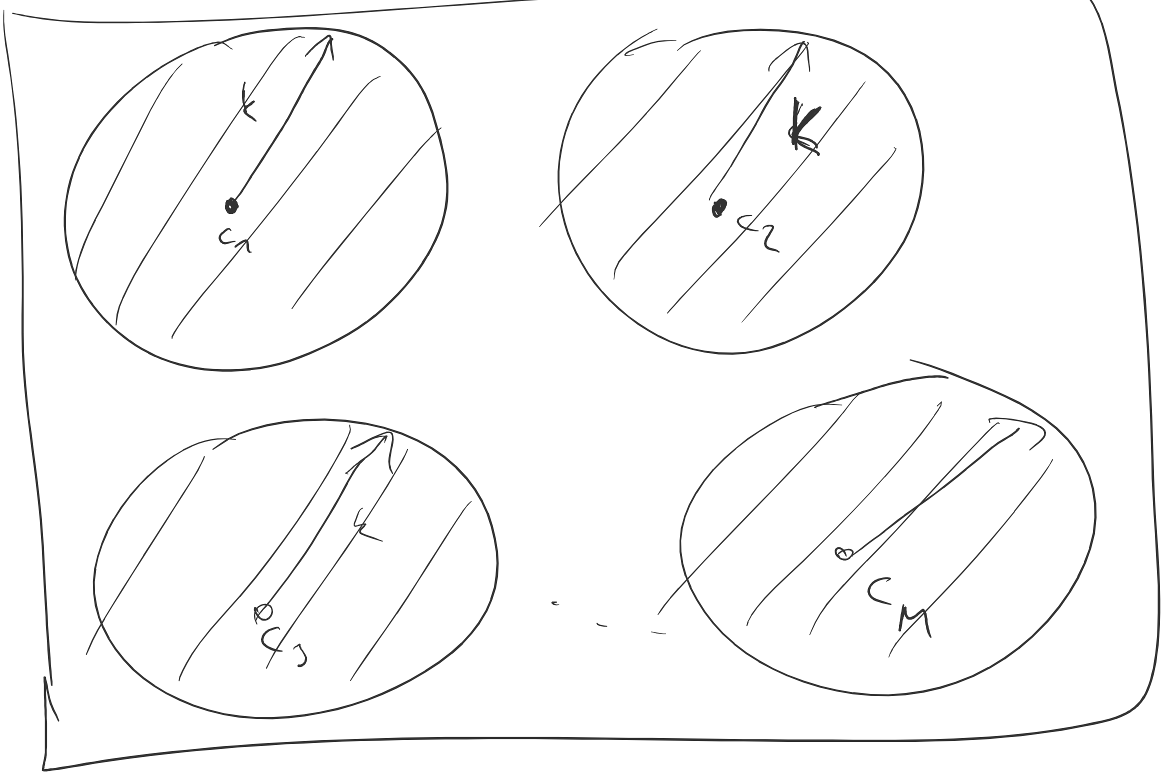
t

from c_1 from c_1

$\{0,\dots,q-1\}^n$

$\Updownarrow$

$\boxed{q^n}$ – total number of strings



$= 2\cdot 1+1 \Rightarrow \varepsilon = 1$ $q=2$

$5, 4, 3$ - code

n M d

$$4\left[\binom{5}{0}+\binom{5}{1}\right] \leq 2^5$$

$$4\left[1+5\right]$$

$$24 \leq 32$$

$0 \rightarrow 000$

$1 \rightarrow 111$