09. Shannon Information Theory

"Classical" Information Theory

=

Shannon Information Theory

Claude Shannon "A mathematical theory of communication" 1948

MIT → Princeton → Bell Labs → MIT

(a) Data compression: (Shannon) Coding Theorem

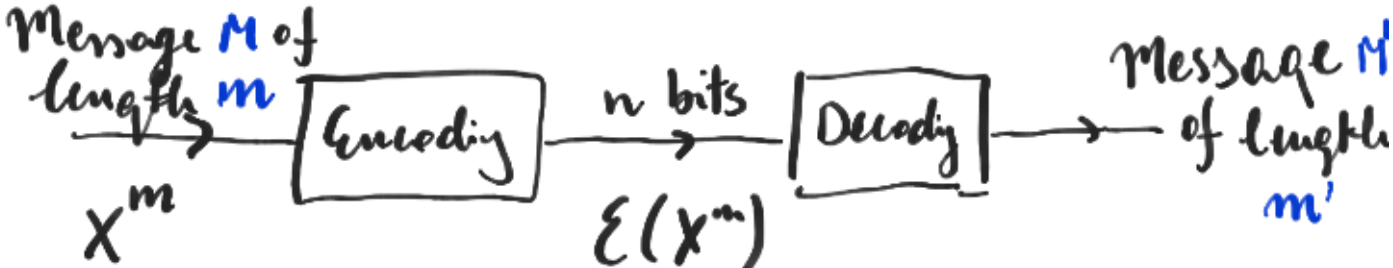
(b) Communication: (Shannon) Channel Capacity

(c) Cryptography: - Security of one-time pad (OTP)
- (Shannon) Secret Capacity

(#) Book: Cover Thomas: "Elements of Information Theory"

(a) Data Compression

Motivation: File compression {ZIP, ARJ, RAR, TTR, 7z, ...}



$X \in \{A, B, C, \ldots\}$ ← alphabet of d symbols

Q: How many bits (at least) per symbol?

for lossless compression $M = M'$

Assume: X are distributed according to

some PDF but i.i.d.!

[N.B. for "text" we can use also correlation e.g. "Ala ma kota" ...]
vowel after consonant

Example

alphabet = {'a','b','c','d'} — symbols ($d=4$)

	'a'	'b'	'c'	'd'
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

codewords

- "Simplest" encoding with 2 bits: 'a'=00, 'b'=01, 'c'=10, 'd'=11

twice as many: $\frac{\#\text{bits}}{\text{msg. length}} = \frac{n}{m} = \frac{m\cdot 2}{m} = 2$ ← average codeword length (in bits)

- "Smarter" encoding: 'a'=0, 'b'=10, 'c'=110, 'd'=111

"frequency" ≈ "probability"

unambiguous decoding!

1000 → b a a
11000 → c a a
111000 → d a a

$$\frac{n(m)}{m} \underset{m\to\infty}{=} \frac{m}{m}\left(\frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3\right) = 1.75$$

(1 ← bits, 2 ← pairs, 3 ← bits)

Q Shannon: What is the optimal $\frac{n(m)}{m}$ as $m\to\infty$?

Answer (Shannon Coding Thm):

$$\lim_{m\to\infty}\left\{\frac{n(m)}{m}\right\} = H(x)$$

Shannon Entropy

$$H(X) = -\sum_{x=1}^{d} p(x)\log_2 p(x)$$

$\log \equiv \log_{10}$, $\lg \equiv \log_2$

$\{ \ln = \log_e$

→ measure of randomness of the random variable X

d-outcomes (symbols in alphabet)

$X \sim p(x_i)$ $\quad p(x_i) = \delta_{ik} \Rightarrow 0 \leq H(X) \leq \lg d \Leftarrow p(x_i) = \frac{1}{d}$

Example: $H(X) = -\frac{1}{2}\lg\frac{1}{2} - \frac{1}{4}\lg\frac{1}{4} - \frac{1}{8}\lg\frac{1}{8} - \frac{1}{8}\lg\frac{1}{8} =$

$= \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = 1\frac{3}{4} = 1.75$

Intuition of the proof

X^m : as sequence length $m \to \infty$

the symbol $X = x$ expected to appear $mp(x)$ times

⇒ "typical" sequences

[consequence of "Law of Large Numbers"

i.e. $X_m = \sum_{i=1}^{m} \frac{X_i}{m} \xrightarrow[m\to\infty]{} \mathbb{E}[X]$

- Probability of obtaining a typical sequence:

$$P_{typ}^{m} = p(x_0 x_1 \dots x_m \in T) = \prod_{x=1}^{d} p(x)^{mp(x)}$$

(T: is typical)

$$= 2^{\lg \prod_x p(x)^{mp(x)}} = 2^{\sum_x \lg p(x)^{mp(x)}} = 2^{m\sum_x p(x)\lg p(x)}$$

- # all sequences (in bits) $\quad d^m = 2^{\lg d^m} = 2^{m \lg d} \to n$

- $\#_{typ} = \frac{1}{p_{typ}^{m}} = 2^{-m\sum_x p(x)\lg p(x)} = 2^{mH(x)} \to n$

"seq. Ptyp n

$\Rightarrow$ rather than n lgd we need $nH(X)$ bits to label all typical sequences

$$\lim_{n\to\infty}\left(\frac{n}{n}\right) = \frac{nH(X)}{n} = H(X)$$

$H(X)$ is the optimal compression rate $\Longleftrightarrow$ Shannon Coding Theorem

{How to construct the code? $\Rightarrow$ Huffman Coding}

Properties of entropies

Shannon Entropy: $X \sim p(x)$ $\quad H(X) = -\sum_x p(x) \lg p(x)$

Joint Entropy: $X, Y \sim p(X,Y)$ $\quad H(X,Y) = -\sum_{x,y} p(x,y) \lg p(x,y)$

$\Downarrow$

Conditional Entropy: $X \sim p(x|y)$ given y

$$H(X|Y=y) = -\sum_x p(x|y) \lg p(x|y)$$

$$H(X|Y) = \sum_y p(y)\, H(X|Y=y) =$$

$$= -\sum_{xy} p(x,y) \lg p(x|y) = -\sum_{x,y} p(x,y) \lg \frac{p(x,y)}{p(y)} =$$

$$= \sum_{x,y} p(x,y) \lg p(y) + \sum_{x,y} p(x,y) \lg p(x) =$$

$$= -\sum_{x,y} p(x,y) \lg p(x,y) + \sum_{x,y} p(x,y) \lg p(y) =$$

$$= -\sum_{x,y} p(x,y) \lg p(x,y) - \left(-\sum_{y} p(y) \lg p(y)\right) =$$

$$= H(X,Y) - H(Y)$$

自: (i) $0 \le H(X) \le \lg d$ ← uniform PDF

(ii) $H(X,Y) \le H(X) + H(Y)$ ← indep^{nt} variables

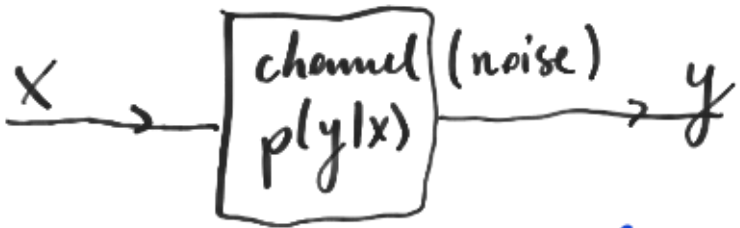
(iii) chain rule:

$$H(X,Y) = H(X|Y) + H(Y)$$

(iv) $H(X|Y) \le H(X)$

(v) $H(X,Y|Z) = H(X|YZ) + H(Y|Z)$

(b) Communication: Shannon Channel (Theorem) Capacity

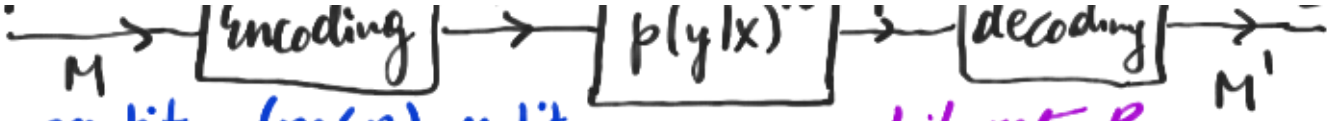


How to send a message reliably over a channel?

Encode m-bit message into n-bit codeword

$n > m$ (use redundant data)

A X^n Y^n B



m-bit message (m<n) n-bit

bit-rate R [přenosnost]

Channel capacity : $C = \lim_{m \to \infty} \left(\frac{m}{n} \right) \left(\leq 1 \right)$

Examples

1) error-correction codes eg. CD disc:

$C \approx \frac{1}{3}$ ⇐ 192 → 588 (m logical bits → n physical bits) ≈ 1 : ≈ 3

size: 650 MB → 2.4 GB

2) Transmission via an optical link eg. optical fibre

bandwidth

[bits/s] $C = B \log(1 + SNR)$

(SNR) Signal to noise ratio : Power of signal / Power of noise

in dB: $10 \log_{10} SNR$ {1000 → 30 dB}

Answer:

Shannon Channel Theorem

$C = \max_{p(x)} I(X:Y)$

{encoding}

channel capacity

(Shannon Mutual Information)

$I(X:Y) = H(X) - H(X|Y)$

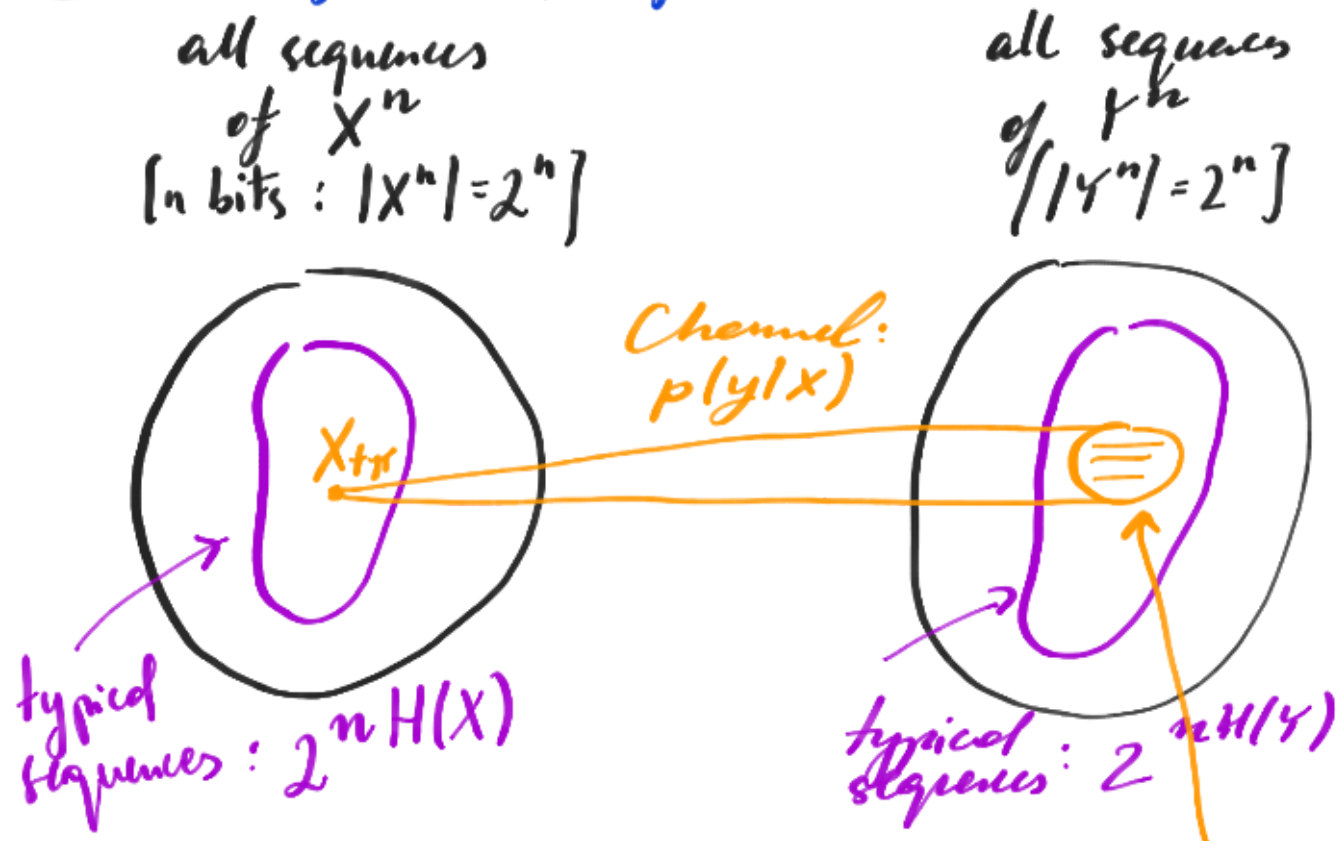
$X \to [p(y|x)] \to Y$ (channel)

How much about "X" we know from knowing "Y" ?

symetric:

$$I(X:Y) = H(X) + H(Y) - H(X,Y)$$

Intuition of the proof



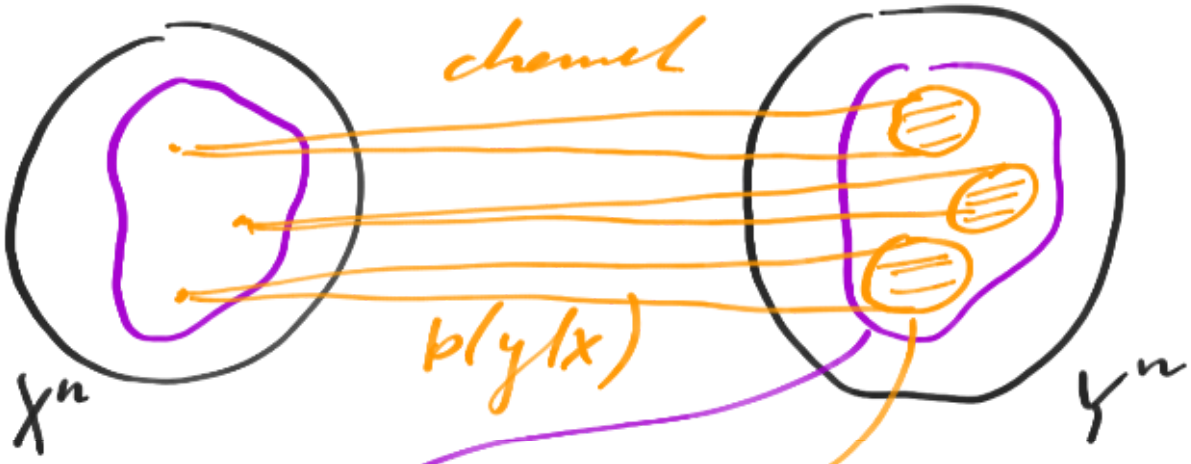
- given $X = x$ there will be $2^{H(Y|X=x)}$ ($n=1$) typical (fixed) Y it can lead to.
- ok. but there are $\approx np(x)$ (n) $X = x$ in a typical sequences

$X_{typ} \Rightarrow$ leads to

$$\prod_x \left(2^{H(Y|x)}\right)^{np(x)} = 2^{n\sum_x p(x) H(Y|x)} = 2^{nH(Y|X)}$$

How many different input (typical)

sequences can be distinguished from typical output sequences?



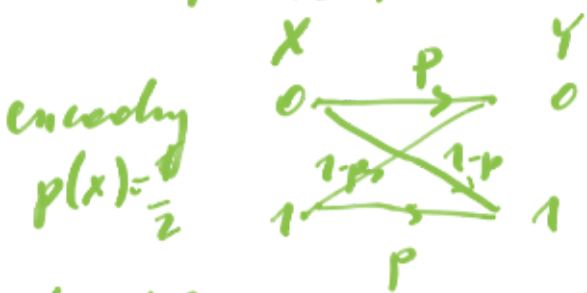
$$\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n\,I(X:Y)}$$

$I(X:Y)$: bit rate R

Still, freedom to optimise over encodings

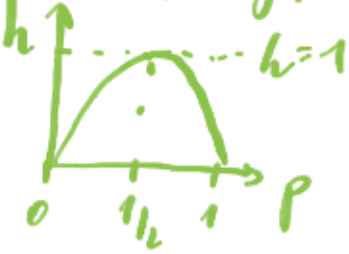
$$X \sim p(x) \Rightarrow C = \max_{p(x)} I(X:Y)$$

Example (4)



$$I(X:Y) = 1 - h(p)$$

$h(p) = -p\lg p - (1-p)\lg(1-p)$ binary entropy



$\Rightarrow$ $p = 1$ (100%) $I(X:Y) = 1$

$p = \frac{1}{2}$ (50%) $I(X:Y) = 0$

(no noise) perfect channel

completely

Last modified: 16:36