ENTROPY RATE of a sequence of random variables

$\{X_t\}_{t \in \mathbb{N}}$ joint distrib. $\underline{P}_N(x_1, \dots, x_N)$ for N variables

marginal distrib. over $A \subset \{1, \dots, N\}$; $\underline{x}_A = \{x_i; i \in A\}$

(obtained by summing over $\underline{x} \in \bar{A} = \{1, \dots, N\}/A$

$$\underline{P}(\underline{x}_A) = \sum_{\underline{x} \in \bar{A}} \underline{P}_N(x_1, \dots, x_N)$$

Simple examples:

$$\underline{P}_N(x_1, \dots, x_N) = \prod_1^N p(x_i) \quad (\text{i.i.d})$$

Markov chains $\underline{P}_N(x_1, \dots, x_N) = p_1(x_1) \prod_1^{N-1} W(x_t \to x_{t+1})$

initial state — transition probabilities

$$\sum_{y \in X} W(x \to y) = 1 \quad \forall x \in X$$

ENTROPY RATE for $\underline{X}_N = \{X_t\}_{t=\{1,\dots,N\}}$

$$h_X = \lim_{N \to \infty} H_{\underline{X}_N}/N$$

(entropy grows linearly with the number of variables)

- i.i.d with $p(x)$ — $h_\infty = H(p) = -\sum_x p(x) \log_2[p(x)]$

- Markov chain with $p^*(x)$ stationary distr.

$p^*(x) = \lim_{t \to \infty} p(x_t)$ → empirical of X_t

$$h_X = - \sum_{x,y \in \mathcal{X}} p^*(x)\, w(x \to y) \log_2 [w(x \to y)]$$

Correlated variables X, Y

(conditional entropy & mutual information

joint prob. distr. $p(x,y)$

conditional prob. $p(y/x)$

Bayes theorem → $p(y/x) = \dfrac{p(x,y)}{p(x)}$

$p(x) = \sum_y p(x,y)$

X, Y are independent → $p(y/x) = p(y)$

X, Y are correlated → measure of correlation?

Conditional entropy: Shannon entropy of the conditional prob (over y) averaged over x

$$H_{Y/X} = - \sum_x p(x) \sum_y p(y/x) \log_2 p(y/x)$$

joint entropy $H_{X,Y} = -\sum_{x}\sum_{y} p(x,y) \log_2[p(x,y)]$

CHAIN RULE → $\boxed{H_{X,Y} = H_X + H_{Y/X}}$

$(= H_Y + H_{X/Y})$

X, Y are independent → $H_{Y,X} = H_Y$

X, Y are completely correlated → $\underline{H_{X,Y} = H_X + H_Y}$

$Y = f(X) \rightarrow H_{Y/X} = 0 \quad (H_{X/Y} = 0)$

$\Rightarrow H_{X,Y} = H_X = H_Y$

MUTUAL INFORMATION

$$\boxed{I_{X,Y} = \sum_{x,y} p(x,y) \log_2\left[\frac{p(x,y)}{p(x)p(y)}\right]}$$

$$= D(p_{X,Y} \,||\, p_X p_Y) = -H_{X,Y} + H_X + H_Y$$

$$I_{X,Y} = H_Y - H_{Y/X} = H_X - H_{X/Y}$$

$I_{X,Y}$ is symmetric for $X \leftrightarrow Y$

$I_{X,Y} \geq 0$

- X, Y independent $\iff I_{X,Y} = 0$

- complete correlation ($X = f(Y)$) $\iff I_{X,Y} = H_Y = H_X$ (= $H_{X,Y}$)

Mutual information is degraded when data are transmitted

DATA-PROCESSING INEQUALITY:

Markov chain $X \to Y \to Z$

$p(x, y, z) = p_1(x)\, w_2(x \to y)\, w_3(y \to z)$

$$\boxed{I_{X,Z} \leq I_{X,Y}} \qquad \boxed{I_{X,f(Y)} \leq I_{X,Y}}$$

This holds also if $Z = f(Y)$!

DATA COMPRESSION

- source (input data) $\underline{X} = \{X_1, \dots, X_N\}$ X_i are random variables

how to store the info in $\underline{X}$ efficiently?

optimal compression $\longrightarrow$ entropy rate of $\underline{X}$

- Source code: mapping $W: \mathcal{X}^N \to \{0,1\}^*$

- codeword: any binary string — set of binary strings of arbitrary length

$X_1, \dots, X_N$ part of a Compression

Coding in 3 steps: (compression)

1) break stream into blocks of length N
2) each block encoded through mapping W
3) codewords are glued together to form the output stream — compressed data

$\underline{x}^{(1)}, \underline{x}^{(2)}, \dots, \underline{x}^{(t)}$

$W(\underline{x}^{(1)}), W(\underline{x}^{(2)}), \dots W(\underline{x}^{(t)})$

Careful: "parsing" the output stream into codewords in a unique way for ANY concatenation of codewords!

↳ uniquely decodable code

Stronger condition — instantaneous code

$\forall \underline{x}, \underline{x}'$ $W(\underline{x})$ is not a prefix of $W(\underline{x}')$

(decoding can be done on the spot)

Quality of compression: length of the codewords

$$L(w) = \sum_{\underline{x} \in X^N} p(\underline{x})\, l_w(\underline{x})$$

$l_w(\underline{x})$: length of the string $w(\underline{x})$

$\sum$: sum over all possible codewords

"Best code": code with shortest $L(w)$

THEOREM L_N^* shortest average length (instantaneous codes) for $\underline{X} = \{X_1, \dots X_N\}$

$H_{\underline{X}}$ = Shannon entropy of $\underline{X}$

(1) $\forall N \geq 1 \quad H_{\underline{X}} \leq L_N^* \leq H_{\underline{X}} + 1$

(2) if the source $\underline{X}$ has a finite entropy rate

$$h = \lim_{N \to \infty} H_{\underline{X}}/N \Rightarrow h = \lim_{N \to \infty} L_N^*/N$$

Shannon entropy of source sets the optimal compression that can be achieved

(1) is the difficult part in the proof:

instantaneous code $\Rightarrow$ codewords not too short

instantaneous codes $\left(= \sum_{\underline{x}} 2^{-\ell_w(\underline{x})} \leq 1\right.$ Kraft's inequality

Minimize $L(w)$

subject to the constraint $\sum_{\underline{x}} 2^{-\ell_w(\underline{x})} = 1$

$$\ell_w(\underline{x}) = -\log_2 p(\underline{x}) \quad \rightarrow \quad \boxed{\ell_w(\underline{x}) = \lceil -\log_2 p(\underline{x}) \rceil}$$

(integer part)

optimal solution for $N \gg 1$ ⟵ SHANNON's CODE

NOT optimal for finite N $\left(p(\underline{x}) \ll 1 \Rightarrow \ell_w(\underline{x}) \gg 1 \right)$

Shannon code is impractical (for $N \gg 1$) from a computational perspective

- build the code once and for all:
  → requires $\mathcal{O}(|\mathcal{X}|^N)$ memory
- reconstruct the code every time
  → it takes $\mathcal{O}(|\mathcal{X}|^N)$ CPU operations

DATA TRANSMISSION → may be degraded by noise

to counter transmission noise → add redundancy in source coding

(compression prior to transmission)

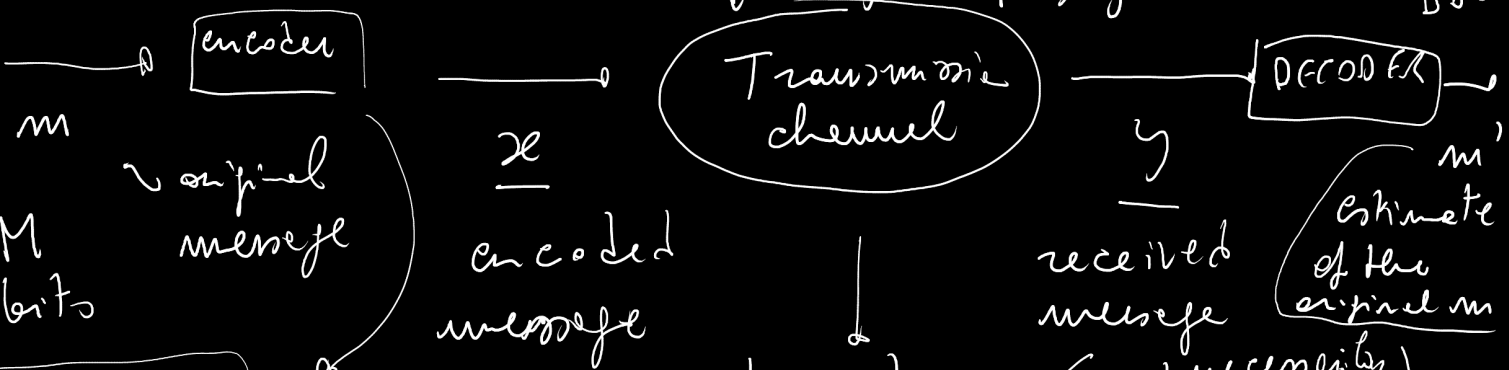
Key results: CHANNEL CODING THEOREM (Shannon 1948)

Level of redundancy needed = maximal level of noise that can be tolerated

entropy

COMMUNICATION SYSTEM

Alice → map $w: \{0,1\}^M \to \{0,1\}^N$ → Bob



$N > M$

redundancy

N = block length $\underline{y} = \{y_1, \dots, y_N\}$ from an alphabet $y_i \in \mathcal{Y}$

decoding map $d(y) = m'$

noise → $P(\underline{y}/\underline{x})$

Memoryless channels: noise on each bit

$$P(\underline{y}, \underline{x}) = \prod_1^N P(y_i / x_i)$$