3. Information-Theoretic Foundations

- **Founder:** Claude Shannon, 1940's
- **Gives bounds for:**
  - Ultimate data compression
  - Ultimate transmission rate of communication
- **Measure of symbol information:**
  - Degree of surprise / uncertainty
  - Number of yes/no questions (binary decisions) to find out the correct symbol.
  - Depends on the probability *p* of the symbol.

Choosing the information measure

- **Requirements for information function *I*(*p*):**
  - $I(p) \geq 0$
  - $I(p1p2) = I(p1) + I(p2)$
  - $I(p)$ is continuous with p.

- The solution is essentially unique:
  $I(p) = -\log p = \log (1/p)$.

- Base of log = 2 $\Rightarrow$ The unit of information is *bit*.

Examples

- Tossing a fair coin: P(heads) = P(tails) = ½
  - Information measures for one toss:
    Inf(heads) = Inf(tails) = $-\log_2 0.5$ bits = 1 bit
  - Information measure for a 3-sequence:
    Inf(<heads, tails, heads>) = $-\log_2(½ \cdot ½ \cdot ½)$ bits = 3 bits.
  - Optimal coding: heads → 0, tails → 1

- An unfair coin: P(heads) = 1/8 and P(tails) = 7/8.
  - Inf(heads) = $-\log_2(1/8)$ bits = 3 bits
  - Inf(tails) = $-\log_2(7/8)$ bits ≈ 0.193 bits
  - Inf(<tails, tails, tails>) = $-\log_2(7/8)^3$ bits ≈ 0.578 bits
  - Improving the coding requires grouping of tosses into blocks.

Entropy

- Measures the *average* information of a symbol from alphabet *S* having probability distribution *P*:

$$H(S) = \sum_{i=1}^{q} p_i I(p_i) = \sum_{i=1}^{q} p_i \log_2 \left(\frac{1}{p_i} \right)$$

- **Noiseless source encoding theorem (C. Shannon):** Entropy *H*(*S*) gives a lower bound on the average code length *L* for any instantaneously decodable system.

Example case: Binary source

- Two symbols, e.g. $S = \{0, 1\}$, probabilities p_0 and $p_1 = 1 - p_0$.
- Entropy = $p_0 \log_2 \frac{1}{p_0} + (1 - p_0) \log_2 \frac{1}{1 - p_0}$
- $p_0 = 0.5$, $p_1 = 0.5$ → $H(S) = 1$
- $p_0 = 0.1$, $p_1 = 0.9$ → $H(S) \approx 0.469$
- $p_0 = 0.01$, $p_1 = 0.99$ → $H(S) \approx 0.081$

- The *skewer* the distribution, the *smaller* the entropy.
- *Uniform* distribution results in *maximum* entropy

Example case: Predictive model

HELLO WOR ?

Already processed 'context'

Next char	Prob	Inf (bits)	Weighted information
L	0.95	$-\log_2 0.95 \approx 0.074$ bits	$0.95 \cdot 0.074 \approx 0.070$ bits
D	0.04	$-\log_2 0.04 \approx 4.644$ bits	$0.04 \cdot 4.644 \approx 0.186$ bits
M	0.01	$-\log_2 0.01 \approx 6.644$ bits	$0.01 \cdot 6.644 \approx 0.066$ bits

Weighted sum ≈ 0.322 bits

Code redundancy

- Average redundancy of a code (per symbol): $L - H(S)$.
- Redundancy can be made = 0, if symbol probabilities are negative powers of 2. (Note that $-\log_2(2^{-i}) = i$)
- Generally possible:

$$\log_2\left(\frac{1}{p_i}\right) \le l_i < \log_2\left(\frac{1}{p_i}\right) + 1$$

- **Universal code**: $L \le c1 \cdot H(S) + c2$
- **Asymptotically optimal code:** $c1 = 1$

Generalization: m-memory source

- *Conditional information:* $\log_2(1/P(s_i|s_{i_1},\cdots,s_{i_m}))$

- *Conditional entropy* for a given context:

$$H(S|s_{i_1},\cdots,s_{i_m}) = \sum_S P(s_i|s_{i_1},\cdots,s_{i_m}) \log_2\left(\frac{1}{P(s_i|s_{i_1},\cdots,s_{i_m})}\right)$$

- *Global entropy* over all contexts:

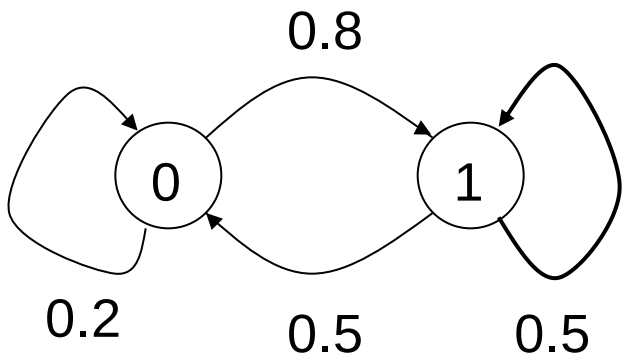
$$H(S) = \sum_{S^m}\sum_S P(s_{i_1},\cdots,s_{i_m}) P(s_i|s_{i_1},\cdots,s_{i_m}) \log_2\left(\frac{1}{P(s_i|s_{i_1},\cdots,s_{i_m})}\right)$$

$$= \sum_{S^{m+1}} P(s_{i_1},\cdots,s_{i_m},s_i) \log_2\left(\frac{1}{P(s_i|s_{i_1},\cdots,s_{i_m})}\right)$$

About conditional sources

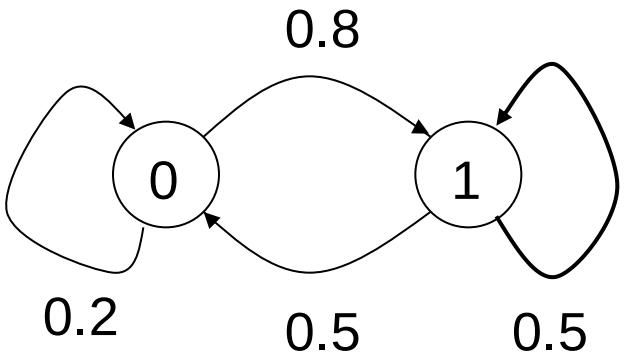
- **Generalized Markov process:**
  - Finite-state machine
  - For an *m*-memory source there are q^m states
  - Transitions correspond to symbols that follow the *m*-block
  - Transition probabilities are state-dependent

- **Ergodic source:**
  - System settles down to a limiting probability distribution.
  - *Equilibrium* state probabilities can be inferred from transition probabilities.



Solving the example entropy

$$\begin{cases} p_0 = 0.2\,p_0 + 0.5\,p_1 \\ p_1 = 0.8\,p_0 + 0.5\,p_1 \end{cases}$$



Solution: *eigenvector*

$p_0 = 0.385$, $p_1 = 0.615$

$$H(S) = \sum_{i=0}^{1}\sum_{j=0}^{1} p_i \Pr(j \mid i) \log \frac{1}{\Pr(j \mid i)} =$$

$$p_0\left(0.2 \log \frac{1}{0.2} + 0.8 \log \frac{1}{0.8}\right) + p_1\left(0.5 \log \frac{1}{0.5} + 0.5 \log \frac{1}{0.5}\right) \approx 0.893$$

Example application: compression of black-and – white images (black and white areas highly clustered)

Empirical observations

- Shannon's experimental value for the entropy of the English language ≈ ***1 bit per character***
- Current text compressor efficiencies:
  - gzip ≈ 2.5 – 3 bits per character
  - bzip2 ≈ 2.5 bits per character
  - The best predictive methods ≈ 2 bits per character
- Improvements are still possible!
- However, digital *images*, *audio* and *video* are more important data types from compression point of view.

Other extensions of entropy

- *Joint* entropy, e.g. for two random variables X, Y:

$$H(X,Y) = -\sum_{x,y} p_{x,y} \log_2 p_{x,y}$$

- *Relative* entropy: difference of using q_i instead of p_i:

$$D_{KL}(P \| Q) = \sum_i p_i \log_2 \frac{p_i}{q_i}$$

- *Differential* entropy for continuous probability distribution:

$$h(X) = -\int_X f(x) \log f(x) dx$$

Kolmogorov complexity

- Measure of message information = Length of the shortest binary program for generating the message.
- This is close to entropy $H(S)$ for a sequence of symbols drawn at random from a distribution that S has.
- Can be much smaller than entropy for artificially generated data: pseudo random numbers, fractals, ...
- Problem: Kolmogorov complexity is *not computable*! (Cf. Gödel's incompleteness theorem and Turing machine stopping problem).