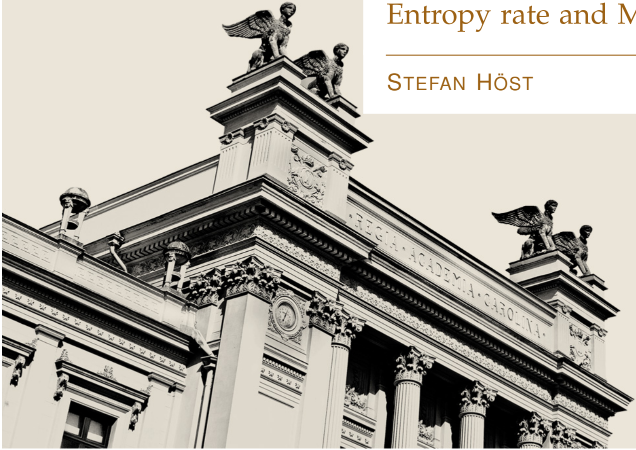
Information Theory

Lecture 5

Entropy rate and Markov sources

STEFAN HÖST



Universal Source Coding

Huffman coding is optimal, what is the problem?

In the previous coding schemes (Huffman and Shannon-Fano)it was assumed that

- The source statistics is known
- The source symbols are i.i.d.

Normally this is not the case.

How much can the source be compressed?
How can it be achieved?

Random process

Definition (Random process)

A random process $\{X_i\}_{i=1}^n$ is a sequence of random variables. There can be an arbitrary dependence among the variables and the process is characterized by the joint probability function

$$P\big(X_1, X_2, \dots, X_n = x_1, x_2, \dots, x_n\big) = p(x_1, x_2, \dots, x_n),\ n = 1, 2, \dots$$

Definition (Stationary random process)

A random process is stationary if it is invariant in time,

$$P\big(X_1, \dots, X_n = x_1, \dots, x_n\big) = P\big(X_{q+1}, \dots, X_{q+n} = x_1, \dots, x_n\big)$$

for all time shifts q.

Entropy rate

Definition

The entropy rate of a random process is defined as

$$H_\infty(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1 X_2 \dots X_n)$$

Define the alternative entropy rate for a random process as

$$H(X|X^\infty) = \lim_{n\to\infty} H(X_n|X_1 X_2 \dots X_{n-1})$$

Theorem

The entropy rate and the alternative enropy rate are equivalent,

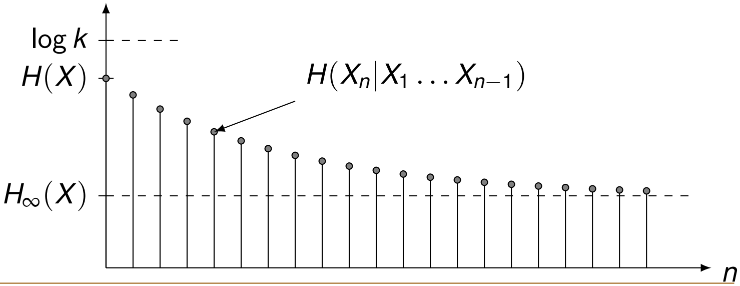
$$H_\infty(X) = H(X|X^\infty)$$

Entropy rate

Theorem

For a stationary stochastic process the entropy rate is bounded by

$$0 \leq H_\infty(X) \leq H(X) \leq \log k$$



Source coding for random processes

Optimal coding of process

Let $\boldsymbol{X} = (X_1, \dots, X_N)$ be a vector of N symbols from a random process. Use an optimal source code to encode the vector. Then

$$H(X_1 \dots X_N) \leq L^{(N)} \leq H(X_1 \dots X_N) + 1$$

which gives the average codeword length per symbol, $L = \frac{1}{N} L^{(N)}$,

$$\frac{1}{N} H(X_1 \dots X_N) \leq L \leq \frac{1}{N} H(X_1 \dots X_N) + \frac{1}{N}$$

In the limit as $N \to \infty$ the optimal codeword length per symbol becomes

$$\lim_{N \to \infty} L = H_\infty(X)$$

Markov chain

Definition (Markov chain)

A Markov chain, or Markov process, is a random process with unit memory,

$$P\big(x_n|x_1,\dots,x_{n-1}\big) = P\big(x_n|x_{n-1}\big), \qquad \text{for all } x_i$$

Definition (Stationary)

A Markov chain is stationary (time invariant) if the conditional probabilities are independent of the time,

$$P\big(X_n = x_a|X_{n-1} = x_b\big) = P\big(X_{n+\ell} = x_a|X_{n+\ell-1} = x_b\big)$$

for all relevant n, ℓ, x_a and x_b.

Markov chain

Theorem

For a Markov chain the joint probability function is

$$
\begin{aligned}
p(x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} p(x_i | x_1, x_2, \ldots, x_{i-1}) \\
&= \prod_{i=1}^{n} p(x_i | x_{i-1}) \\
&= p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1})
\end{aligned}
$$

Markov chain characterization

Definition

A Markov chain is characterized by

- A state transition matrix

$$P = \big[p(x_j|x_i)\big]_{i,j\in\{1,2,\dots,k\}} = \big[p_{ij}\big]_{i,j\in\{1,2,\dots,k\}}$$

  where $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$.

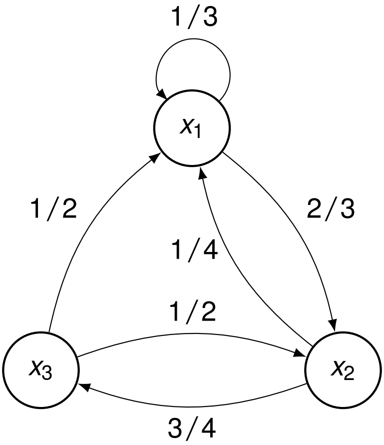
- A finite set of states

$$X \in \{x_1, x_2, \dots, x_k\}$$

  where the state determines everything about the past.

The state transition graph describes the behaviour of the process

Example



The state transition matrix

$$P = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

The state space is

$$X \in \{x_1, x_2, x_3\}$$

Markov chain

Theorem

Given a Markov chain with k states, let the distribution for the states at time n be

$$\boldsymbol{\pi}^{(n)} = (\pi_1^{(n)}\ \pi_2^{(n)} \dots \pi_k^{(n)})$$

Then

$$\boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)} P^n$$

where $\boldsymbol{\pi}^{(0)}$ is the initial distribution at time 0.

Example, asymptotic distribution

$$P^2 = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} = \begin{pmatrix} \frac{20}{72} & \frac{16}{72} & \frac{36}{72} \\ \frac{33}{72} & \frac{39}{72} & 0 \\ \frac{21}{72} & \frac{24}{72} & \frac{27}{72} \end{pmatrix}$$

$$P^4 = \begin{pmatrix} \frac{20}{72} & \frac{16}{72} & \frac{36}{72} \\ \frac{33}{72} & \frac{39}{72} & 0 \\ \frac{21}{72} & \frac{24}{72} & \frac{27}{72} \end{pmatrix} \begin{pmatrix} \frac{20}{72} & \frac{16}{72} & \frac{36}{72} \\ \frac{33}{72} & \frac{39}{72} & 0 \\ \frac{21}{72} & \frac{24}{72} & \frac{27}{72} \end{pmatrix} = \begin{pmatrix} \frac{1684}{5184} & \frac{1808}{5184} & \frac{1692}{5184} \\ \frac{1947}{5184} & \frac{2049}{5184} & \frac{1188}{5184} \\ \frac{1779}{5184} & \frac{1920}{5184} & \frac{1485}{5184} \end{pmatrix}$$

$$P^8 = \cdots\cdots\cdots\cdots\cdots = \begin{pmatrix} 0.3485 & 0.3720 & 0.2794 \\ 0.3491 & 0.3721 & 0.2788 \\ 0.3489 & 0.3722 & 0.2789 \end{pmatrix}$$

Markov chain

Theorem

Let $\boldsymbol{\pi} = (\pi_1 \dots \pi_k)$ be an asymptotic distribution of the state probabilities.Then

- $\sum_j \pi_j = 1$
- $\boldsymbol{\pi}$ *is a stationary distribution, i.e.* $\boldsymbol{\pi} P = \boldsymbol{\pi}$
- $\boldsymbol{\pi}$ *is a unique stationary disribution for the source.*

Entropy rate of Markov chain

Theorem

For a stationary Markov chain with stationary distribution π and transition matrix P, the entropy rate can be derived as

$$H_\infty(X) = \sum_i \pi_i H(X_2|X_1 = x_i)$$

where

$$H(X_2|X_1 = x_i) = -\sum_j p_{ij} \log p_{ij}$$

the entropy of row i in P.

Example, Entropy rate

$$P = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

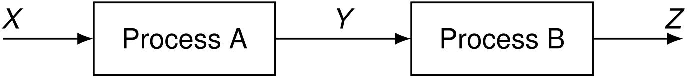
Entropy per row:

$$\begin{aligned} H(X_2|X_1 = x_1) &= h(\tfrac{1}{3}) \\ H(X_2|X_1 = x_2) &= h(\tfrac{1}{4}) \\ H(X_2|X_1 = x_3) &= h(\tfrac{1}{2}) = 1 \end{aligned}$$

Hence

$$H_\infty(X) = \tfrac{15}{43}h(\tfrac{1}{3}) + \tfrac{15}{43}h(\tfrac{1}{4}) + \tfrac{12}{43}h(\tfrac{1}{2}) \approx 0.9013 \text{ bit/source symbol}$$

Data processing lemma



Lemma (Data Processing Lemma)

If the random variables X, Y and Z form a Markov chain, $X \to Y \to Z$, we have

$$I(X;Z) \leq I(X;Y)$$
$$I(X;Z) \leq I(Y;Z)$$

Conclusion

The amount of information can not increase by data processing, neither pre nor post. It can only be transformed (or destroyed).