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?
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

\[
P(X_1, X_2, \ldots, X_n = x_1, x_2, \ldots, x_n) = p(x_1, x_2, \ldots, x_n), \quad n = 1, 2, \ldots
\]

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

\[
P(X_1, \ldots, X_n = x_1, \ldots, x_n) = P(X_{q+1}, \ldots, X_{q+n} = x_1, \ldots, x_n)
\]

for all time shifts \( q \).
Entropy rate

**Definition**

The entropy rate of a random process is defined as

\[ H_\infty(X) = \lim_{n \to \infty} \frac{1}{n} H(X_1 X_2 \ldots X_n) \]

Define the alternative entropy rate for a random process as

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

**Theorem**

*The entropy rate and the alternative entropy 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 $X = (X_1, \ldots, X_N)$ be a vector of $N$ symbols from a random process. Use an optimal source code to encode the vector. Then

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

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

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

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,

\[ P(x_n|x_1, \ldots, x_{n-1}) = P(x_n|x_{n-1}), \quad \text{for all } x_i \]

Definition (Stationary)

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

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

for all relevant \( n, \ell, x_a \) and \( x_b \).
Markov chain

Theorem

For a Markov chain the joint probability function is

\[ p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|x_1, x_2, \ldots, x_{i-1}) \]

\[ = \prod_{i=1}^{n} p(x_i|x_{i-1}) \]

\[ = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1}) \]
Markov chain characterization

Definition

A Markov chain is characterized by

- A state transition matrix

\[ P = \begin{bmatrix} p(x_j|x_i) \end{bmatrix}_{i,j \in \{1,2,\ldots,k\}} = \begin{bmatrix} p_{ij} \end{bmatrix}_{i,j \in \{1,2,\ldots,k\}} \]

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

- A finite set of states

\[ X \in \{x_1, x_2, \ldots, x_k\} \]

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

\[ \pi^{(n)} = (\pi_1^{(n)}, \pi_2^{(n)}, \ldots, \pi_k^{(n)}) \]

Then

\[ \pi^{(n)} = \pi^{(0)} P^n \]

where \( \pi^{(0)} \) is the initial distribution at time 0.
Example, asymptotic distribution

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

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

\[ P^8 = \begin{pmatrix} 0.3485 & 0.3720 & 0.2794 \\ 0.3491 & 0.3721 & 0.2788 \\ 0.3489 & 0.3722 & 0.2789 \end{pmatrix} \]
Markov chain

Theorem

Let $\pi = (\pi_1 \ldots \pi_k)$ be an asymptotic distribution of the state probabilities. Then

- $\sum_j \pi_j = 1$
- $\pi$ is a stationary distribution, i.e. $\pi P = \pi$
- $\pi$ is a unique stationary distribution for the source.
Entropy rate of Markov chain

Theorem

For a stationary Markov chain with stationary distribution $\pi$ 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:

\[ H(X_2 \mid X_1 = x_1) = h\left(\frac{1}{3}\right) \]
\[ H(X_2 \mid X_1 = x_2) = h\left(\frac{1}{4}\right) \]
\[ H(X_2 \mid X_1 = x_3) = h\left(\frac{1}{2}\right) = 1 \]

Hence

\[ H_\infty(X) = \frac{15}{43} h\left(\frac{1}{3}\right) + \frac{15}{43} h\left(\frac{1}{4}\right) + \frac{12}{43} h\left(\frac{1}{2}\right) \approx 0.9013 \text{ bit/source symbol} \]
Data processing lemma

Lemma (Data Processing Lemma)

If the random variables $X$, $Y$ and $Z$ form a Markov chain, $X \rightarrow Y \rightarrow Z$, 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).