

Markov sources and search algorithms

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 processes

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

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

Definition (Stationary random process)

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

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

for all time shifts q .

Entropy rate

Definition

The **entropy rate** of a random process is defined as

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

Define the **alternative entropy rate** for a random process as

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

Theorem

The entropy rate and the alternative entropy rate are equivalent,

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

Special case – independent variables

We assume that X_i are I.I.D. and compute $H_\infty(X)$:

$$\begin{aligned} H_\infty(X) &= \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1 \dots X_n) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(X_i | X_1 \dots X_{i-1}) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n H(X_i) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} H(x) \sum_{i=1}^n 1 = H(X). \end{aligned}$$

Decreasing entropy chain

Considering a stationary (time-invariant) random process, it can be seen that:

$$H(X_n|X_1 \dots X_{n-1}) \leq H(X_n|X_2 \dots X_{n-1}) = H(X_{n-1}|X_1 \dots X_{n-2}),$$

where the last equality comes from stationary property (decreasing). Then

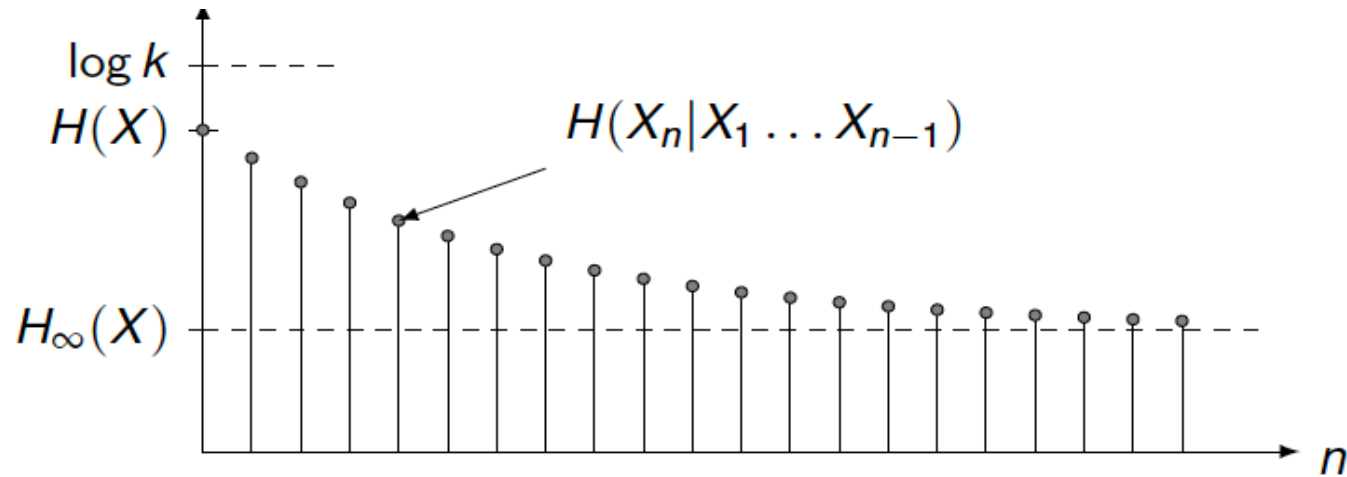
$$H(X_n|X_1 \dots X_{n-1}) \leq \dots \leq H(X_2|X_1) \leq H(X_1) = H(X) \leq \log_2 |\mathcal{X}|.$$

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 $\mathbf{X} = (X_1, \dots, X_N)$ 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 $N \rightarrow \infty$ the optimal codeword length per symbol becomes

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

Markov chain

Definition (Markov chain)

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

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

Definition (Stationary)

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

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

for all relevant n , l , x_a and x_b .

Markov chain

Theorem

For a Markov chain the joint probability function is

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

Markov chain characterization

Definition

A **Markov chain** is characterized by

- A **state transition matrix**

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

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

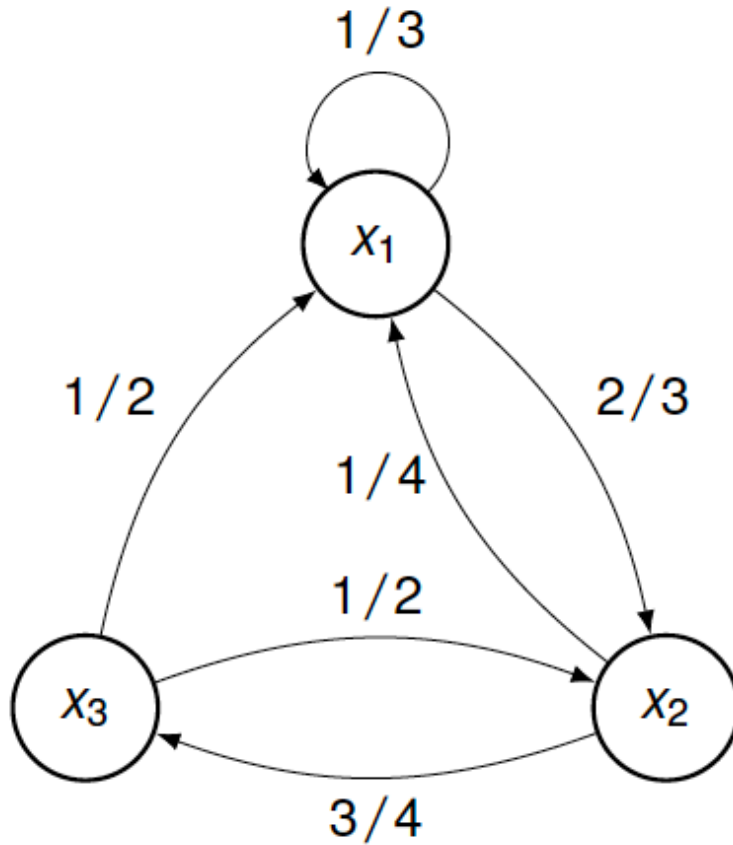
- A finite set of **states**

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

where the state determines everything about the past.

The **state transition graph** describes the behaviour of the process

Example of Markov chain



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

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

Computing the next state

- Assume at $t = 0$ we have
 $Pb(X_0 = x_1) = Pb(X_0 = x_2) = Pb(X_0 = x_3) = 1/3$. What about $t = 1$?
- We compute:

$$Pb(X_1 = x_1) = \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{4} + \frac{1}{3} \times \frac{1}{2} = \frac{13}{36}$$

$$Pb(X_1 = x_2) = \frac{1}{3} \times \frac{2}{3} + \frac{1}{3} \times 0 + \frac{1}{3} \times \frac{1}{2} = \frac{14}{36}$$

$$Pb(X_1 = x_3) = 0 + \frac{1}{3} \times \frac{3}{4} = \frac{9}{36}$$

Of course, $\sum_{i=1}^k Pb(X_t = x_i) = 1$, in our case $k = 3$.

Markov chain – state distribution

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.

Proof

Denote $\pi_j^{(n)} = P(X_n = x_j)$ - probability at time $t = n$ we are in state x_j .

Then,

$$\pi_j^{(n)} = \pi_1^{(n-1)} p_{1j} + \pi_2^{(n-1)} p_{2j} + \dots + \pi_k^{(n-1)} p_{kj},$$

for $j = 1, \dots, k$.

This implies

$$\begin{aligned} (\pi_1^{(n)}, \dots, \pi_k^{(n)}) &= (\pi_1^{(n-1)}, \dots, \pi_k^{(n-1)}) \times P \\ &= (\pi_1^{(n-2)}, \dots, \pi_k^{(n-2)}) \times P \times P \\ &= \dots \\ &= (\pi_1^{(0)}, \dots, \pi_k^{(0)}) \times P^n. \end{aligned}$$

Example- asymptotic distribution

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

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

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

Asymptotic distribution

- The entries in each column have approximately the same value !
- $p_{11}^{(8)} = p_{21}^{(8)} = p_{31}^{(8)}$ - coming to state x_1 is the same (does not depend on the initial state any more)
- In other words, for large n the probability p_{ij} does NOT depend on i .
- Alternatively, for large n , we get $P^{n+1} = P^n$!
- Think about $P \times P^n$ when P is of the form $P^n = \begin{pmatrix} A & B & C \\ A & B & C \\ A & B & C \end{pmatrix}$
- Make sense to define $\lim_{n \rightarrow \infty} p_{ij}^{(n)} = \pi_j$ for $j = 1, \dots, k$.

Markov chain – stationary distribution

Theorem

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

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

Proof

We show the second part (first one is obvious). The process being stationary we have $P \times P^n = P^{n+1} = P^n$. Then, multiplying with π we get

$$(\pi P) \times P^n = \pi \times P^n,$$

which implies that $\pi P = \pi$.

Which Markov chains are stationary ?

- Answer: If and only if there exists $N > 0$ such that P^N has a **positive column**, i.e. all elements in this column are > 0 !!

Computing stationary distribution

We need to compute the stationary distribution π using that $\sum_{i=1}^k \pi_i = 1$ and $\pi P = \pi$. The equation $\pi P = \pi$ gives:

$$1/3 \times \pi_1 + 1/4 \times \pi_2 + 1/2 \times \pi_3 = \pi_1$$

$$2/3 \times \pi_1 + 0 + 1/2 \times \pi_3 = \pi_2$$

$$0 + 3/4 \times \pi_2 + 0 = \pi_3$$

Solve and get : $\pi_1 = 0.3488$; $\pi_2 = 0.3721$; $\pi_3 = 0.2791$.

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 .

Proof

For Markov sources we have $H(X_n|X_1 \dots X_{n-1}) = H(X_n|X_{n-1})$ and we use the fact that

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

Then,

$$\begin{aligned} H_\infty(X) &= \lim_{n \rightarrow \infty} H(X_n|X_1 \dots X_{n-1}) = \lim_{n \rightarrow \infty} H(X_n|X_{n-1}) = H(X_2|X_1) \\ &= \sum_{i,j} P(X_1 = x_i, X_2 = x_j) \log_2 P(X_2 = x_j|X_1 = x_i) \\ &= \sum_i P(X_1 = x_i) \sum_j P(X_2 = x_j|X_1 = x_i) \log_2 P(X_2 = x_j|X_1 = x_i) \\ &= \sum_i H(X_2|X_1 = x_i) P(X_1 = x_i), \end{aligned}$$

where $H(X_2|X_1 = x_i) = \sum_j P(X_2 = x_j|X_1 = x_i) \log_2 P(X_2 = x_j|X_1 = x_i)$.

Entropy rate - example

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

Entropy per row:

$$H(X_2|X_1 = x_1) = h\left(\frac{1}{3}\right)$$

$$H(X_2|X_1 = x_2) = h\left(\frac{1}{4}\right)$$

$$H(X_2|X_1 = x_3) = h\left(\frac{1}{2}\right) = 1$$

Hence

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

Sorting algorithms

① SORTING ALGORITHM

INPUT: n different numbers A_1, A_2, \dots, A_n and simple test " $A < B$?" for any A and B .

GOAL: Efficient algorithm (thus MINIMAL NUMBER OF COMPARISONS) to SORT LIST IN INCREASING ORDER.

- List is completely unsorted, numbers can take any place in the list (their size is IRRELEVANT)
- FIRST PROBLEM: Determine a lower bound for average number of comparisons \bar{W} .
- Introduce INDEX ORDER I_1, I_2, \dots, I_n for A_1, A_2, \dots, A_n
- MEANING: $I_1 = 7 \Rightarrow A_1$ is the 7th least numb.
 $I_2 = 1 \Rightarrow A_2$ is the SMALLEST NUMBER in the list.

Sorting algorithms II

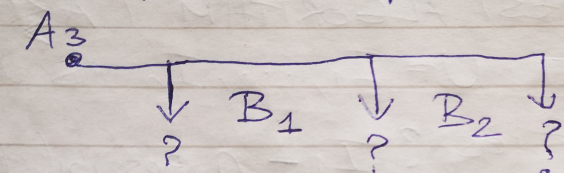
- This is a binary TEST ALGORITHM ($A \leq B?$) and UNSORTED LIST $U = I_1, \dots, I_n$ IS INPUT.
- Numb. of possibilities? $I_1 - n$ possibilities; given I_1 there are $n-1$ possib. for I_2 .
- So there are $n!$ possib. to choose $U = I_1, I_2, \dots, I_n$.
- All are equally likely, thus
$$H(U) = \log_2 n!$$
- Then, $\bar{W} \geq H(U) = \log_2 n!$
- For $n=64$, we get $\bar{W} \geq 296!$

Buble sort algorithm

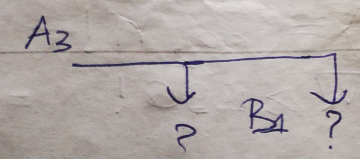
- One alternative is BUBBLE SORT.
 - Go through the list and SWAP neighboring entries if they are in wrong order
 - Repeat until no need for swapping
- In first cycle we get LARGEST NUMBER IN THE LAST POSITION OF THE LIST ...
$$W = (n-1) + (n-2) + \dots + 2 + 1 = \frac{1}{2}(n-1)n.$$
- For $n=64$, we get $W=2016$
- ANOTHER APPROACH - more efficient than BUBBLE SORTING.
- Ordering of partial list. Description as follows:
Because $P(A_1 < A_2) = \frac{1}{2} \Rightarrow H(X_1) = 1$ for our first test (X_1 -binary random variable for 1st test)

Description of the algorithm

- Let B_1, B_2 be THE CORRECT ORDER OF A_1, A_2
- Then A_3 fits at 3 places with same probability



- The test " $A_3 < B_2?$ " gives $H(x_2 | x_1 = x_1) = h(1/3) = 0.918$ (regardless of x_1). Now if " $A_3 < B_2?$ " is (X=1) NOT CORRECT $\Rightarrow A_3 > B_2$ and we have B_1, B_2, A_3 (are correctly ordered)
- But if $A_3 < B_2$ (X=0) we have



- Now test " $A_3 < B_1?$ " gives $H(x_3 | x_1, x_2 = x_1 0) = 1$

General conclusions

• In general, since B_1, B_2, \dots, B_{i-1} denotes the correct order of A_1, A_2, \dots, A_{i-1} , when we consider A_i have the following

$$\begin{array}{ccccccc}
 A_i & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
 & \downarrow & \downarrow & \downarrow & \dots & \downarrow & \downarrow \\
 & ? & B_1 & ? & B_2 & ? & \dots & ? & B_{i-1} & ?
 \end{array}$$

• A_i has the same prob. to fit anywhere.

• BUT if later perform " $A_i < B_{\lfloor \frac{i}{2} \rfloor}$ " (i EVEN) then for this test $h(1/2) = 1$.

• If i is odd then the entropy of the test " $A_i < B_{\lfloor \frac{i}{2} \rfloor}$ " is $h(\frac{i-1}{2i})$.

This value $h(\frac{i-1}{2i}) \geq h(1/3) = 0.918$ for $i \geq 3$.

• This test reduces the problem to (if " $A_i < B_{\lfloor \frac{i}{2} \rfloor}$ " is NOT correct

$$\begin{array}{ccccccc}
 A_i & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\
 & \downarrow & \downarrow & \downarrow & \dots & \downarrow & \downarrow \\
 & B_{\lfloor \frac{i}{2} \rfloor + 1} & B_{\lfloor \frac{i}{2} \rfloor + 2} & \dots & B_{i-1}
 \end{array}$$

Average number of tests - comparison

A_i
IS NOT correct

$\downarrow B_{\lfloor \frac{i}{2} \rfloor + 1}$
 $\downarrow B_{\lfloor \frac{i}{2} \rfloor + 2}$
 \dots
 $\downarrow B_{i-1}$

$?$
 $?$
 $?$
 $?$

If $A_i < B_{\lfloor \frac{i}{2} \rfloor}$ then

A_i

$\downarrow B_1$
 $\downarrow B_2$
 \dots
 $\downarrow B_{\lfloor \frac{i}{2} \rfloor - 1}$

$?$
 $?$
 $?$
 $?$

- The minimum entropy of ANY OF ABOVE TESTS IS $H_{min} = \log_2(1/3) = 0.918$. Then using

$$\bar{W} \leq \frac{H(U)}{H_{min}} = \frac{\log_2(n!)}{0.918} = 1.09 \log_2(n!)$$
- $n=64 \Rightarrow \bar{W} \leq 322$. Exact calculation gives $\bar{W} = 299$ for $n=64$
- Here $H_{min} = \min_{(x_1, \dots, x_i)} H(X_{i+1} | X_1, \dots, X_i = x_1, \dots, x_i)$