

## 1. Entropy rates of a Stochastic Process

A stochastic process, also called a random process, is a set of random variables that model a non deterministic system. If the random variables are dependent or in particular, if the random variables form a stationary process, we will show, just as in the i.i.d. case, that the entropy  $H(X_1, X_2, \dots, X_n)$  grows (asymptotically) linearly with  $n$  at a rate  $H(X)$ , which we will call the entropy rate of the process.

An indexed sequence of random variables is called a stochastic process  $X_i$ . There can be arbitrary dependencies among random variables in general. The joint probability mass functions characterise the process.

$$Pr\{(X_1, X_2, \dots, X_n) = (x_1, x_2, \dots, x_n)\} = p(x_1, x_2, \dots, x_n), (x_1, x_2, \dots, x_n) \in X_n \text{ for } n = 1, 2, \dots$$

**Definition** A stochastic process is said to be stationary if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index; that is,

$$Pr\{X_1 = x_1, \dots, X_n = x_n\} = Pr\{X_1 + l = x_1, \dots, X_n + l = x_n\}$$

for every  $n$  and every shift  $l$  and  $\forall x_1, x_2, \dots, x_n \in X$ .

### 1.1. Markov chains

An example of a stochastic process with dependency is one in which each random variable is conditionally independent of all the other previous random variables. Markov is the name for such a process.

**Definition** A discrete stochastic process  $X_1, X_2, \dots$  is said to be a Markov chain if for  $n = 1, 2, \dots$ ,

$$Pr(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_1 = x_1) = Pr(X_{n+1} = x_{n+1} | X_n = x_n)$$

$$\forall x_1, x_2, \dots, x_n, x_{n+1} \in X.$$

In this case, the joint probability mass function of the random variables can be written as

$$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_n|x_{n-1}).$$

It is assumed that the Markov chains are time invariant unless otherwise stated.

**Definition** The Markov chain is said to be time invariant if the conditional probability  $p(x_{n+1}|x_n)$  does not depend on  $n$ ; that is, for  $n = 1, 2, \dots$ ,

$$Pr\{X_{n+1} = b | X_n = a\} = Pr\{X_2 = b | X_1 = a\}$$

$$\forall a, b \in X.$$

#### State and Transition matrix:

If  $X_i$  is a Markov chain,  $X_n$  is called the state at time  $n$ . A time-invariant Markov chain is characterized by its initial state and a probability transition matrix, given by  $P = [P_{ij}]$ , for  $i, j \in 1, 2, \dots, m$ , where  $P_{ij} = Pr\{X_{n+1} = j | X_n = i\}$ .

If it is possible to go with positive probability from any state of the Markov chain to any other state in a finite number of steps, the Markov chain is said to be **irreducible**. If the largest common factor of the lengths of different paths from a state to itself is 1, the Markov chain is said to be **aperiodic**.

**Example:** Consider a two-state Markov chain with a probability transition matrix given by:

$$\begin{bmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{bmatrix}$$

Let's say the stationary distribution is represented as a vector  $\mu$  with the stationary probabilities of states 1 and 2 as its components. Then, by solving the equation  $\mu P = \mu$  or, more simply, by balancing probabilities, the stationary probability can be obtained.

$$\mu P = \mu \implies \mu P = \mu I \implies \mu(P - I) = 0$$

The net probability flow across any cut set in the state transition graph is zero for the stationary distribution.

As  $\mu$  is a probability distribution it follows that  $\mu_1 + \mu_2 = 1$ . Solving for  $\mu$ , using the above and the following equation,

$\mu_1\alpha = \mu_2\beta$ . The stationary distribution is given by:

$$\mu_1 = \frac{\beta}{\alpha + \beta}, \mu_2 = \frac{\alpha}{\alpha + \beta}$$

The resulting process will be stationary if the Markov chain has an initial state drawn according to the stationary distribution. The entropy of the state  $X_n$  is:

$$H(X_n) = H\left(\frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta}\right).$$

## 1.2. Entropy Rate

Since a stochastic process defined by a Markov chain that is irreducible, aperiodic and positive recurrent has a stationary distribution, the entropy rate is independent of the initial distribution. is the asymptotic distribution of the chain.

**Definition:** The entropy of a stochastic process  $X_i$  is defined by

$$H(X) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, H_2, \dots H_3)$$

when the limit exists.

We can also define a related quantity for entropy rate:

$$H'(X) = \lim_{n \rightarrow \infty} H(X_n | X_{n-1}, X_{n-2}, \dots, X_1)$$

when the limit exists.

$H(X)$  and  $H'(X)$  are two different notions of entropy rate. The first is the entropy of the  $n$  random variables per symbol, and the second is the conditional entropy of the last random variable given the past.

We now show that both limits exist and are identical for stationary processes.

### Entropy of a stationary Markov Chain:

For a stationary Markov chain, the entropy rate is given by

$$\begin{aligned} H(X) &= H'(X) = \lim H(X_n | X_{n-1}, \dots, X_1) \\ &= \lim H(X_n | X_{n-1}) = H(X_2 | X_1) \end{aligned} \quad (1)$$

**Theorem:** Let  $\{X_i\}$  be a stationary Markov chain with stationary distribution  $\mu$  and transition matrix  $P$ . Let  $X_1 \sim \mu$  Then the entropy rate is

$$H(X) = - \sum_{ij} \mu_i P_{ij} \log P_{ij}$$

**Proof**

$$\begin{aligned} H(X) &= H(X_2 | X_1) \\ &= \sum_i \mu_i \left( \sum_j -P_{ij} \log P_{ij} \right) \end{aligned}$$

## 1.3. Functions of Markov chains

Let  $X_1, X_2, \dots, X_n, \dots$  be a stationary Markov chain, and let  $Y_i = \varphi(X_i)$  be a process each term of which is a function of the corresponding state in the Markov chain. To compute  $H(Y)$ , we might compute  $H(Y_n | Y_{n-1}, \dots, Y_1)$  for each  $n$  and find the limit. Upper and lower bounds converging to the limit from above and below might be advantageous computationally. When the difference between the upper and lower bounds is small, we can stop the computation and get a solid estimate of the limit. We know that  $H(Y_n | Y_{n-1}, \dots, Y_2, Y_1)$  converges to  $H(Y)$  i.e,  $H(Y_n | Y_{n-1}, \dots, Y_2, Y_1) \leq H(Y)$

The lemma shows that the interval between the upper and the lower bounds decreases in length.

**Lemma**

$$H(Y_n | Y_{n-1}, \dots, Y_1) - H(Y_n | Y_{n-1}, \dots, Y_1, X_1) \rightarrow 0.$$

**Proof**

The LHS can also be written as:

$$\begin{aligned} H(Y_n | Y_{n-1}, \dots, Y_1) - H(Y_n | Y_{n-1}, \dots, Y_1, X_1) \\ = I(X_1; Y_n | Y_{n-1}, \dots, Y_1) \end{aligned}$$

This mutual information is always less than or equal to  $H(X_1)$  i.e,

$$I(X_1; Y_1, Y_2, \dots, Y_n) \leq H(X_1)$$

As  $n$  tends to infinity,  $\lim I(X_1; Y_1, Y_2, \dots, Y_n)$  exists. From the chain rule,

$$\begin{aligned} H(X) &\geq \lim_{n \rightarrow \infty} I(X_1; Y_1, Y_2, \dots, Y_n) \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n I(X_1; Y_i | Y_{i-1}, \dots, Y_1) \\ &= \sum_{i=1}^{\infty} I(X_1; Y_i | Y_{i-1}, \dots, Y_1) \end{aligned}$$

As  $n$  tends to infinity, this sum of infinite terms is finite and the terms are non-negative, the terms must tend to 0; which proves the lemma, that is,

$$\lim I(X_1; Y_n | Y_{n-1}, \dots, Y_1) = 0$$

From this lemma, we have the following theorem:

**Theorem:**

If  $X_1, X_2, \dots, X_n$  form a stationary Markov chain, and  $Y_i = \varphi(X_i)$ , then

$$H(Y_n|Y_{n-1}, \dots, Y_1, X_1) \leq H(Y) \leq H(Y_n|Y_{n-1}, \dots, Y_1)$$

and

$$\lim H(Y_n|Y_{n-1}, \dots, Y_1, X_1) = H(Y)$$

similarly,

$$\lim H(Y_n|Y_{n-1}, \dots, Y_1) = H(Y)$$