Entropy Rates of a Stochastic Process

Introduction

The AEP establishes that nH bits are sufficient on the average to describe n independent and identically distributed random variables. But, what if the random variables are dependent? In particular, what if they form a stationary process? Our objective is to show that the entropy grows (asymptotically) linearly with n at a rate $H(\chi)$, which we will call *entropy rate* of a process.

Stationary Process

A stochastic process $\{X_i\}$ is an indexed sequence of random variables. In general, there can be an arbitrary dependence among the random variables. The process is characterized by the joint probability mass functions:

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

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,

$\Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = \Pr\{X_{1+l} = x_1, X_{2+l} = x_2, \ldots, X_{n+l} = x_n\}$
for every n and every shift l and for all $x_1, x_2, \ldots, x_n \in \mathcal{X}$.

Markov Process

A simple example of a stochastic process with dependence is one in which each random variable depends only on the one preceding it and is *conditionally* independent of all the other preceding random variables. Such a process is said to be Markov.

Markov Chain

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

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

for all $x_1, x_2, \ldots, x_n, x_{n+1} \in X$.

In this case, the joint probability mass function of the random variables can be written as
$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2) \cdots p(x_n|x_{n-1})$.

Time Invariance

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 $n = 1, 2, \ldots,$
$\Pr\{X_{n+1} = b|X_n = a\} = \Pr\{X_2 = b|X_1 = a\}$ for all $a, b \in \mathcal{X}$.

We will assume that the Markov chain is time invariant unless otherwise stated.

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*

$P = [P_{ij}]$, $i, j \in \{1, 2, \ldots, m\}$, where $P_{ij} = \Pr\{X_{n+1} = j \,|X_n = i\}$.

Irreducible Markov Chain

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 *aperiodic*. This means that there are not paths having lengths that are multiple one of the other.

If the probability mass function of the random variable at time *n* is $p(x_n)$, the probability mass function at time *n* + 1 is:

$$p(x_{n+1}) = \sum_{x_n} p(x_n) P_{x_n x_{x+1}}$$

Where P is the probability transition matrix, and $p(x_n)$ is the probability that the random variable is in one of the states of the Markov chain, for example: $\Pr\{X_{n+1} = a\}$. This means that we can compute the probability of x_{n+1} by the knowledge of P and of $p(x_n)$.

Stationary Distribution

A distribution on the states such that the distribution at time *n* + 1 is the same as the distribution at time *n* is called a *stationary distribution*.

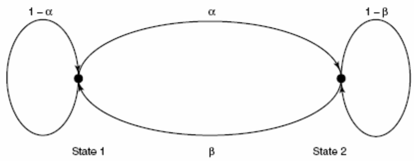
The stationary distribution is so called because if the initial state of a Markov chain is drawn according to a stationary distribution, the Markov chain forms a stationary process. If the finite-state Markov chain is irreducible and aperiodic, the stationary distribution is unique, and from any starting distribution, the distribution of X_n tends to the stationary distribution as $n \to \infty$.

Example

Consider a two state Markov chian with a probability transition matrix:

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

Let the stationary distribution be represented by a vector μ whose components are the stationary probabilities of states 1 and 2, respectively. Then the stationary probability can be found by solving the equation $\mu P = \mu$ or, more simply, by balancing probabilities. In fact, from the definition of stationary distribution, the distribution at time n is equal to the one at time n+1. For the stationary distribution, the net probability flow across any cut set in the state transition graph is zero.



Example

Referring to the Figure in the previous slide, we obtain: $\mu_1\alpha = \mu_2\beta$

Since $\mu_1 + \mu_2 = 1$, the stationary distribution is: $\mu_1 = \dfrac{\beta}{\alpha+\beta}, \mu_2 = \dfrac{\alpha}{\alpha+\beta}$

If this is true, then it should be true that: $p(x_{n+1}) = \sum_{x_n} p(x_n) P_{x_n x_{n+1}}$

That means:

$$\begin{aligned}
&\Pr\{X_{n+1} = state1\} = \\
&= \Pr\{X_n = state1\}\Pr\{X_{n+1} = state1 \mid X_n = state1\} + \\
&\Pr\{X_n = state2\}\Pr\{X_{n+1} = state1 \mid X_n = state2\} \\
&= p(state1)P_{state1,state1} + p(state2)P_{state2,state1} \\
&= \frac{\beta}{\alpha+\beta}(1-\alpha) + \frac{\alpha}{\alpha+\beta}\beta \\
&= \frac{\beta}{\alpha+\beta}
\end{aligned}$$

Example

If the Markov chain has an initial state drawn according to the stationary distribution, the resulting process will be stationary. The entropy of the state *Xn* at time *n* is

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

However, this is not the rate at which entropy grows for *H(X*1,*X*2*, . . . , Xn)*. The dependence among the *Xi*'s will take a steady toll.

Entropy Rate

If we have a sequence of n random variables, a natural question to ask is: How does the entropy of the sequence grow with *n*? We define the *entropy rate* as this rate of growth as follows.

Definition The *entropy* of a stochastic process {*Xi*} is defined by:

$$H(\chi) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots X_n)$$

when the limit exists.

We now consider some simple examples of stochastic processes and their corresponding entropy rates.

Example

1. **Typewriter.** Consider the case of a typewriter that has m equally likely output letters. The typewriter can produce m^n sequences of length n, all of them equally likely. Hence $H(X_1,X_2, \ldots, X_n) = \log m^n$ and the entropy rate is $H(X) = \log m$ bits per symbol.
2. $X_1,X_2, \ldots, X_n$ are i.i.d. random variables, then:

$$H(\chi) = \lim \frac{H(X_1, X_2, \ldots X_n)}{n} = \lim \frac{nH(X_1)}{n} = H(X_1)$$

3. Sequence of independent but not equally distributed random variables. In this case:

$$H(X_1, X_2, \ldots X_n) = \sum_{i=1}^{n} H(X_i)$$

but the $H(X_i)$ are all not equal. We can choose a sequence of distributions such that the limit does not exist.

Conditional Entropy Rate

We define the following quantity related to the entropy rate:

$$H'(\chi) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots X_1)$$

When the limit exists.

The two quantities entropy rate and the previous one correspond to two different notions of entropy rate. The first is the per symbol entropy rate of the n random variables, and the second is the conditional entropy rate of the last random variable given the past. We now prove that for stationary processes both limits exist and are equal:

Theorem: For a stationary stochastic process, the limits of $H(\chi)$ and $H'(\chi)$ exist and are equal.

Existence of the Limit of H'(χ)

Theorem: (*Existence of the limit*) For a stationary stochastic process, $H(X_n|X_{n-1},...X_1)$ is nonincreasing in n and has a limit H'(χ).

Proof:

$$H(X_{n+1}|X_1,X_2,....X_n) \le H(X_{n+1}|X_n,....X_2)$$
$$= H(X_n|X_{n-1},....X_1)$$

Where the inequality follows from the fact that conditioning reduces entropy (the first expression is more conditioned than the second one, because there is not X_1 anymore). The equality follows from the stationarity of the process. Since $H(X_n|X_{n-1},...X_1)$ is a decreasing sequence of nonnegative numbers, it has a limit, H'(χ).

Equality of H'(χ) and H(χ)

Let's first recall this result: if a_n->a and b_n= $\frac{1}{n}\sum_{i=1}^{n} a_i$ then b_n->a. This is because since most of the terms in the sequence a_k are eventually close to a, then b_n, which is the average of the first n terms, is also eventually close to a.

Theorem: (*Equality of the limit*) By the chain rule,

$$\frac{H(X_1,X_2,....X_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1},....X_1)$$

That is, the entropy rate is the average of the conditional entropies. But we know that the conditional entropies tend to a limit H'. Hence, by the previous property, their running average has a limit, which is equal to the limit H' of the terms. Thus, by the existence theorem:

$$H(\chi) = \lim \frac{H(X_1,X_2,....X_n)}{n} = \lim H(X_n|X_{n-1},....X_1) = H'(\chi)$$

Entropy Rate of a Markov Chain

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

$$H(\chi) = H(\chi)' = \lim H(X_n \mid X_{n-1}, \ldots, X_1)$$
$$= \lim H(X_n \mid X_{n-1}) = H(X_2 \mid X_1)$$

Where the conditional entropy is computed using the given stationary distribution. Recall that the stationary distribution μ is the solution of the equations:

$$\mu = \sum_i \mu_i P_{ij} \qquad \text{for all j.}$$

We explicitly express the conditional entropy in the following slide.

Conditional Entropy Rate for a SMC

Theorem (*Conditional Entropy rate of a MC*): Let $\{X_i\}$ be a SMC with stationary distribution μ and transition matrix P. Let $X_1 \sim \mu$. Then the entropy rate is:

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

Proof:

$$H(\chi) = H(X_2 \mid X_1) = \sum_i \mu_i \left(\sum_j -P_{ij} \log P_{ij}\right)$$

Example (*Two state MC*): The entropy rate of the two state Markov chain in the previous example is:

$$H(\chi) = H(X_2 \mid X_1) = \frac{\beta}{\alpha+\beta} H(\alpha) + \frac{\alpha}{\alpha+\beta} H(\beta)$$

If the Markov chain is irreducible and aperiodic, it has unique stationary distribution on the states, and any initial distribution tends to the stationary distribution as n grows.

Example: ER of Random Walk

As an example of stochastic process lets take the example of a random walk on a connected graph. Consider a graph with m nodes with weight $W_{ij} \geq 0$ on the edge joining node i with node j. A particle walk randomly from node to node in this graph.

The random walk is X_m is a sequence of vertices of the graph. Given $X_n=i$, the next vertex j is choosen from among the nodes connected to node i with a probability proportional to the weight of the edge connecting i to j.

Thus,

$$P_{ij} = \frac{W_{ij}}{\sum_k W_{ik}}$$

ER of a Random Walk

In this case the stationary distribution has a surprisingly simple form, which we will guess and verify. The stationary distribution for this MC assigns probability to node i proportional to the total weight of the edges emanating from node i. Let:

$$W_i = \sum_j W_{ij}$$

Be the total weight of edges emanating from node i and let

$$W = \sum_{i,j:j>i} W_{ij}$$

Be the sum of weights of all the edges. Then $\sum_i W_i = 2W$. We now guess that the stationary distribution is:

$$\mu_i = \frac{W_i}{2W}$$

ER of Random Walk

We check that μP=μ:

$$\sum_i \mu_i P_{ij} = \sum_i \frac{W_i}{2W}\frac{W_{ij}}{W_i} = \sum_i \frac{W_{ij}}{2W} = \frac{W_j}{2W} = \mu_j$$

Thus, the stationary probability of state i is proportional to the weight of edges emanating from node i. This stationary distribution has an interesting property of locality: It depends only on the total weight and the weight of edges connected to the node and therefore it does not change if the weights on some other parts of the graph are changed while keeping the total weight constant.

The entropy rate can be computed as follows:

$$H(\chi) = H(X_1 \mid X_2) = -\sum_{ij} \mu_i \sum_j P_{ij} \log P_{ij}$$

ER of Random Walk

$$= -\sum_i \frac{W_i}{2W} \sum_j \frac{W_{ij}}{W_i} \log \frac{W_{ij}}{W_i}$$

$$= -\sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_{ij}}{W_i}$$

$$= -\sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_{ij}}{2W} + \sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_i}{2W}$$

$$= H\left(\ldots, \frac{W_{ij}}{2W}, \ldots\right) - H\left(\ldots, \frac{W_i}{2W}, \ldots\right)$$

If all the edges have equal weight, , the stationary distribution puts weight $E_i/2E$ on node i, where E_i is the number of edges emanating from node i and E is the total number of edges in the graph. In this case the entropy rate of the random walk is:

$$H(\chi) = \log(2E) - H\left(\frac{E_1}{2E}, \frac{E_2}{2E}, \ldots, \frac{E_m}{2E}\right)$$

Apparently the entropy rate, which is the average transition entropy, depends only on the entropy of the stationary distribution and the total number of edges

Example

Random walk on a chessboard. Let's king move at random on a 8x8 chessboard. The king has eight moves in the interior, five moves at the edges and three moves at the corners. Using this and the preceding results, the stationary probabilities are, respectively, 8/420, 5/420 and 3/420, and the entropy rate is 0.92log8. The factor of 0.92 is due to edge effects; we would have an entropy rate of log8 on an infinite chessboard. Find the entropy of the other pieces for exercize!

It is easy to see that a stationary random walk on a graph is time reversible; that is, the probability of any sequence of states is the same forward or backward:

$$\Pr\{X_1 = x_1, X_2 = x_2, \ldots X_n = x_n\} = \Pr\{X_n = x_1, X_{n-1} = x_2, \ldots X_1 = x_n\}$$

The converse is also true, that is any time reversible Markov chain can be represented as a random walk on an undirected weighted graph.