CHAPTER 4

ENTROPY RATES OF A STOCHASTIC PROCESS

The asymptotic equipartition property in Chapter 3 establishes that $nH(X)$ bits suffice on the average to describe n independent and identically distributed random variables. But what if the random variables are dependent? In particular, what if the random variables form a stationary process? We will show, just as in the i.i.d. case, that the entropy $H(X_1, X_2, \ldots, X_n)$ grows (asymptotically) linearly with n at a rate $H(\mathcal{X})$, which we will call the *entropy rate* of the process. The interpretation of $H(\mathcal{X})$ as the best achievable data compression will await the analysis in Chapter 5.

4.1 MARKOV CHAINS

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 $\Pr\{(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)\} = p(x_1, x_2, \ldots, x_n)$, $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ 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,

$$\begin{aligned} &\Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} \\ &\quad = \Pr\{X_{1+l} = x_1, X_{2+l} = x_2, \ldots, X_{n+l} = x_n\} \end{aligned} \tag{4.1}$$

for every n and every shift l and for all $x_1, x_2, \ldots, x_n \in \mathcal{X}$.

Elements of Information Theory, Second Edition, By Thomas M. Cover and Joy A. Thomas

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.

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

$$\begin{aligned}\Pr(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) \\ = \Pr(X_{n+1} = x_{n+1} | X_n = x_n)\end{aligned} \tag{4.2}$$

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

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

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_n|x_{n-1}). \tag{4.3}$$

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 $n = 1, 2, \ldots,$

$$\Pr\{X_{n+1} = b | X_n = a\} = \Pr\{X_2 = b | X_1 = a\} \quad \text{for all } a, b \in \mathcal{X}. \tag{4.4}$$

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 $P_{ij} = \Pr\{X_{n+1} = j | X_n = i\}$.

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*.

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

$$p(x_{n+1}) = \sum_{x_n} p(x_n) P_{x_n x_{n+1}}. \tag{4.5}$$

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 4.1.1 Consider a two-state Markov chain with a probability transition matrix

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \tag{4.6}$$

as shown in Figure 4.1.

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. For the stationary distribution, the net probability flow across any cut set in the state transition graph is zero. Applying this to Figure 4.1, we obtain

$$\mu_1 \alpha = \mu_2 \beta. \tag{4.7}$$

Since $\mu_1 + \mu_2 = 1$, the stationary distribution is

$$\mu_1 = \frac{\beta}{\alpha + \beta}, \qquad \mu_2 = \frac{\alpha}{\alpha + \beta}. \tag{4.8}$$

If the Markov chain has an initial state drawn according to the stationary distribution, the resulting process will be stationary. The entropy of the

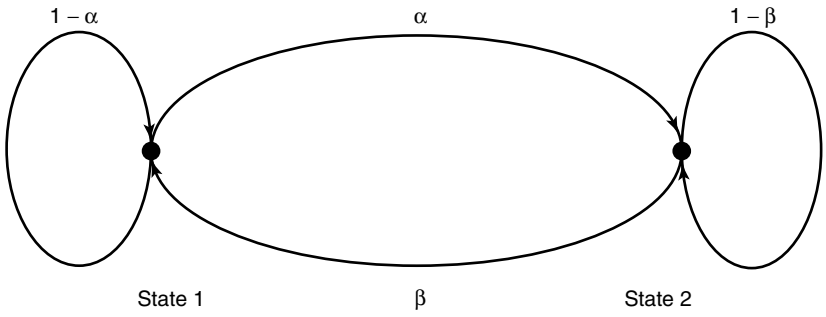

FIGURE 4.1. Two-state Markov chain.

state X_n at time n is

$$H(X_n) = H\left(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta}\right). \tag{4.9}$$

However, this is not the rate at which entropy grows for $H(X_1, X_2, \ldots, X_n)$. The dependence among the X_i's will take a steady toll.

4.2 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 $\{X_i\}$ is defined by

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{4.10}$$

when the limit exists.

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

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 $H(\mathcal{X}) = \log m$ bits per symbol.
2. $X_1, X_2, \ldots$ *are i.i.d. random variables*. Then

$$H(\mathcal{X}) = \lim \frac{H(X_1, X_2, \ldots, X_n)}{n} = \lim \frac{nH(X_1)}{n} = H(X_1), \tag{4.11}$$

   which is what one would expect for the entropy rate per symbol.
3. *Sequence of independent but not identically distributed random variables*. In this case,

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i), \tag{4.12}$$

   but the $H(X_i)$'s are all not equal. We can choose a sequence of distributions on $X_1, X_2, \ldots$ such that the limit of $\frac{1}{n}\sum H(X_i)$ does not exist. An example of such a sequence is a random binary sequence

where $p_i = P(X_i = 1)$ is not constant but a function of i, chosen carefully so that the limit in (4.10) does not exist. For example, let

$$p_i = \begin{cases} 0.5 & \text{if } 2k < \log\log i \le 2k+1, \\ 0 & \text{if } 2k+1 < \log\log i \le 2k+2 \end{cases} \tag{4.13}$$

for $k = 0, 1, 2, \ldots$.
Then there are arbitrarily long stretches where $H(X_i) = 1$, followed by exponentially longer segments where $H(X_i) = 0$. Hence, the running average of the $H(X_i)$ will oscillate between 0 and 1 and will not have a limit. Thus, $H(\mathcal{X})$ is not defined for this process.

We can also define a related quantity for entropy rate:

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) \tag{4.14}$$

when the limit exists.

The two quantities $H(\mathcal{X})$ and $H'(\mathcal{X})$ correspond to two different notions of entropy rate. The first is the per symbol entropy of the n random variables, and the second is the conditional entropy of the last random variable given the past. We now prove the important result that for stationary processes both limits exist and are equal.

Theorem 4.2.1 *For a stationary stochastic process, the limits in (4.10) and (4.14) exist and are equal:*

$$H(\mathcal{X}) = H'(\mathcal{X}). \tag{4.15}$$

We first prove that $\lim H(X_n|X_{n-1}, \ldots, X_1)$ exists.

Theorem 4.2.2 *For a stationary stochastic process, $H(X_n|X_{n-1}, \ldots, X_1)$ is nonincreasing in n and has a limit $H'(\mathcal{X})$.*

Proof

$$H(X_{n+1}|X_1, X_2, \ldots, X_n) \le H(X_{n+1}|X_n, \ldots, X_2) \tag{4.16}$$

$$= H(X_n|X_{n-1}, \ldots, X_1), \tag{4.17}$$

where the inequality follows from the fact that conditioning reduces entropy and the equality follows from the stationarity of the process. Since $H(X_n|X_{n-1}, \ldots, X_1)$ is a decreasing sequence of nonnegative numbers, it has a limit, $H'(\mathcal{X})$. □

We now use the following simple result from analysis.

Theorem 4.2.3 (*Cesáro mean*) *If* $a_n \to a$ *and* $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, *then* $b_n \to a$.

Proof: (*Informal outline*). 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.

Formal Proof: Let $\epsilon > 0$. Since $a_n \to a$, there exists a number $N(\epsilon)$ such that $|a_n - a| \leq \epsilon$ for all $n \geq N(\epsilon)$. Hence,

$$|b_n - a| = \left| \frac{1}{n}\sum_{i=1}^{n}(a_i - a) \right| \tag{4.18}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} |(a_i - a)| \tag{4.19}$$

$$\leq \frac{1}{n}\sum_{i=1}^{N(\epsilon)} |a_i - a| + \frac{n - N(\epsilon)}{n}\epsilon \tag{4.20}$$

$$\leq \frac{1}{n}\sum_{i=1}^{N(\epsilon)} |a_i - a| + \epsilon \tag{4.21}$$

for all $n \geq N(\epsilon)$. Since the first term goes to 0 as $n \to \infty$, we can make $|b_n - a| \leq 2\epsilon$ by taking n large enough. Hence, $b_n \to a$ as $n \to \infty$. □

Proof of Theorem 4.2.1: By the chain rule,

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1), \tag{4.22}$$

that is, the entropy rate is the time average of the conditional entropies. But we know that the conditional entropies tend to a limit H'. Hence, by Theorem 4.2.3, their running average has a limit, which is equal to the limit H' of the terms. Thus, by Theorem 4.2.2,

$$H(\mathcal{X}) = \lim \frac{H(X_1, X_2, \ldots, X_n)}{n} = \lim H(X_n | X_{n-1}, \ldots, X_1)$$
$$= H'(\mathcal{X}). \qquad \square \tag{4.23}$$

The significance of the entropy rate of a stochastic process arises from the AEP for a stationary ergodic process. We prove the general AEP in Section 16.8, where we show that for any stationary ergodic process,

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(\mathcal{X}) \tag{4.24}$$

with probability 1. Using this, the theorems of Chapter 3 can easily be extended to a general stationary ergodic process. We can define a typical set in the same way as we did for the i.i.d. case in Chapter 3. By the same arguments, we can show that the typical set has a probability close to 1 and that there are about $2^{nH(\mathcal{X})}$ typical sequences of length n, each with probability about $2^{-nH(\mathcal{X})}$. We can therefore represent the typical sequences of length n using approximately $nH(\mathcal{X})$ bits. This shows the significance of the entropy rate as the average description length for a stationary ergodic process.

The entropy rate is well defined for all stationary processes. The entropy rate is particularly easy to calculate for Markov chains.

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

$$\begin{aligned} H(\mathcal{X}) = H'(\mathcal{X}) &= \lim H(X_n|X_{n-1}, \ldots, X_1) = \lim H(X_n|X_{n-1}) \\ &= H(X_2|X_1), \end{aligned} \tag{4.25}$$

where the conditional entropy is calculated using the given stationary distribution. Recall that the stationary distribution μ is the solution of the equations

$$\mu_j = \sum_i \mu_i P_{ij} \quad \text{for all } j. \tag{4.26}$$

We express the conditional entropy explicitly in the following theorem.

Theorem 4.2.4 *Let $\{X_i\}$ be a stationary Markov chain with stationary distribution μ and transition matrix P. Let $X_1 \sim \mu$. Then the entropy rate is*

$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}. \tag{4.27}$$

Proof: $H(\mathcal{X}) = H(X_2|X_1) = \sum_i \mu_i \left(\sum_j -P_{ij} \log P_{ij} \right)$. □

Example 4.2.1 (*Two-state Markov chain*) The entropy rate of the two-state Markov chain in Figure 4.1 is

$$H(\mathcal{X}) = H(X_2|X_1) = \frac{\beta}{\alpha+\beta}H(\alpha) + \frac{\alpha}{\alpha+\beta}H(\beta). \tag{4.28}$$

Remark If the Markov chain is irreducible and aperiodic, it has a unique stationary distribution on the states, and any initial distribution tends to the stationary distribution as $n \to \infty$. In this case, even though the initial distribution is not the stationary distribution, the entropy rate, which is defined in terms of long-term behavior, is $H(\mathcal{X})$, as defined in (4.25) and (4.27).

4.3 EXAMPLE: ENTROPY RATE OF A RANDOM WALK ON A WEIGHTED GRAPH

As an example of a stochastic process, let us consider a random walk on a connected graph (Figure 4.2). Consider a graph with m nodes labeled $\{1, 2, \ldots, m\}$, with weight $W_{ij} \geq 0$ on the edge joining node i to node j. (The graph is assumed to be undirected, so that $W_{ij} = W_{ji}$. We set $W_{ij} = 0$ if there is no edge joining nodes i and j.)

A particle walks randomly from node to node in this graph. The random walk $\{X_n\}$, $X_n \in \{1, 2, \ldots, m\}$, is a sequence of vertices of the graph. Given $X_n = i$, the next vertex j is chosen from among the nodes connected to node i with a probability proportional to the weight of the edge connecting i to j. Thus, $P_{ij} = W_{ij} / \sum_k W_{ik}$.

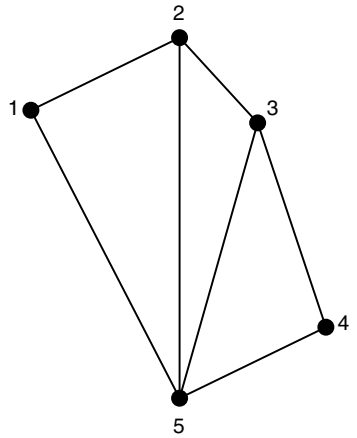

FIGURE 4.2. Random walk on a graph.