*Elements of Information Theory*
Thomas M. Cover, Joy A. Thomas
Copyright © 1991 John Wiley & Sons, Inc.
Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1

# 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 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, i.e.,

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

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

A simple example of a stochastic process with dependence is one in which each random variable depends 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 $x_1, x_2, \ldots, x_n, x_{n+1} \in \mathscr{X}$.

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 $n$, i.e., for $n = 1, 2, \ldots$

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

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

If $\{X_i\}$ is a Markov chain, then $X_n$ 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, then the Markov chain is said to be *irreducible.*

If the probability mass function of the random variable at time $n$ is $p(x_n)$, then 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, then the Markov chain forms a stationary process.

If the finite state Markov chain is irreducible and aperiodic, then the stationary distribution is unique, and from any starting distribution, the distribution of $X_n$ tends to the stationary distribution as $n \rightarrow \infty$.

***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 $\mu$ whose components are the stationary probabilities of state 1 and state 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 0. Applying this to Figure 4.1, we obtain

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

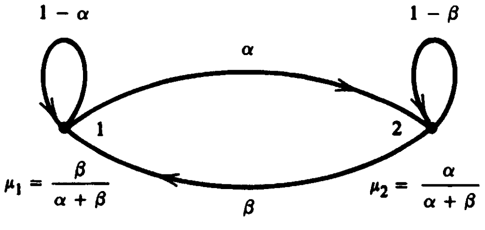
Since $\mu_1 + \mu_2 = 1$, the stationary distribution is

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

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



$1-\alpha$
$1-\beta$
$\alpha$
1
2
$\mu_1 = \frac{\beta}{\alpha+\beta}$
$\beta$
$\mu_2 = \frac{\alpha}{\alpha+\beta}$

Figure 4.1. Two-state Markov chain.

## 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 rate* of a stochastic process $\{X_i\}$ is defined by

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

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

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

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 $H(\mathscr{X})$ is not defined for this process.

We can also define a related quantity for entropy rate:

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

when the limit exists.

The two quantities $H(\mathscr{X})$ and $H'(\mathscr{X})$ 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 will now prove the important result that for stationary processes both the 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, i.e.,*

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

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

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

**Proof:**

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

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

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

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:** Since $a_n \to a$, there exists a number $N(\epsilon)$ such that $|a_n - a| \le \epsilon$ for all $n \ge N(\epsilon)$. Hence

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

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

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

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

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

**Proof of Theorem 4.2.1:** By the chain rule,

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

i.e., the entropy rate is the time average of the conditional entropies. But we know that the conditional entropies tend to a limit $H'(\mathscr{X})$. Hence, by Theorem 4.2.3, their running average has a limit, which is equal to the limit $H'(\mathscr{X})$ of the terms.

Thus, by Theorem 4.2.2.,

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

The significance of the entropy rate of a stochastic process arises from the AEP for a stationary ergodic process. We will prove the general AEP in Section 15.7, where we will show that for any stationary ergodic process,

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

with probability 1. Using this, the theorems of Chapter 3 can be easily 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 $2^{nH(\mathscr{X})}$ typical sequences of length $n$, each with probability about $2^{-nH(\mathscr{X})}$. We can therefore represent the typical sequences of length $n$ using approximately $nH(\mathscr{X})$ 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*

$$H(\mathscr{X}) = H'(\mathscr{X}) = \lim H(X_n|X_{n-1}, \ldots, X_1) = \lim H(X_n|X_{n-1}) = H(X_2|X_1)\,, \tag{4.25}$$

*where the conditional entropy is calculated using the given stationary distribution. We express this result explicitly in the following theorem:*

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

$$H(\mathscr{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij} \tag{4.26}$$

**Proof:** $H(\mathscr{X}) = H(X_2|X_1) = \Sigma_i\, \mu_i(\Sigma_j - P_{ij} \log P_{ij})$. □

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

$$H(\mathscr{X}) = H(X_2|X_1) = \frac{\beta}{\alpha+\beta} H(\alpha) + \frac{\alpha}{\alpha+\beta} H(\beta)\,. \tag{4.27}$$

**Remark:** If the Markov chain is irreducible and aperiodic, then 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 $H(\mathscr{X})$ as defined in (4.25) and (4.26).

## 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 $j$.

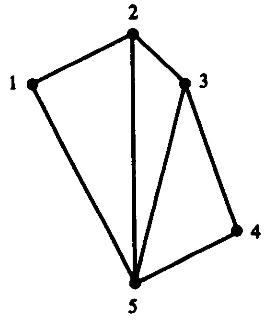


Figure 4.2. Random walk on a graph.

(The graph is assumed to be undirected, so that $W_{ij} = W_{ji}$. We set $W_{ij} = 0$ if the pair of nodes $i$ and $j$ are not connected.)

A particle randomly walks 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 $P_{ij} = W_{ij}/\Sigma_k W_{ik}$.

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

$$W_i = \sum_j W_{ij} \tag{4.28}$$

be the total weight of edges emanating from node $i$ and let

$$W = \sum_{i,j:\, j>i} W_{ij} \tag{4.29}$$

be the sum of the weights of all the edges. Then $\Sigma_i W_i = 2W$.

We now guess that the stationary distribution is

$$\mu_i = \frac{W_i}{2W}\,. \tag{4.30}$$

We verify that this is the stationary distribution by checking that $\mu P = \mu$. Here

$$\sum_i \mu_i P_{ij} = \sum_i \frac{W_i}{2W} \frac{W_{ij}}{W_i} \tag{4.31}$$

$$= \sum_i \frac{1}{2W} W_{ij} \tag{4.32}$$

$$= \frac{W_j}{2W} \tag{4.33}$$

$$= \mu_j\,. \tag{4.34}$$

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 hence does not change if the weights in some other part of the graph are changed while keeping the total weight constant.

We can now calculate the entropy rate as

$$H(\mathscr{X}) = H(X_2|X_1) \tag{4.35}$$

$$= -\sum_i \mu_i \sum_j P_{ij} \log P_{ij} \tag{4.36}$$

$$= -\sum_i \frac{W_i}{2W} \sum_j \frac{W_{ij}}{W_i} \log \frac{W_{ij}}{W_i} \tag{4.37}$$

$$= -\sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_{ij}}{W_i} \tag{4.38}$$

$$= -\sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_{ij}}{2W} + \sum_i \sum_j \frac{W_{ij}}{2W} \log \frac{W_i}{2W} \tag{4.39}$$

$$= H\left(\ldots, \frac{W_{ij}}{2W}, \ldots\right) - H\left(\ldots, \frac{W_i}{2W}, \ldots\right). \tag{4.40}$$

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

$$H(\mathscr{X}) = \log(2E) - H\left(\frac{E_1}{2E}, \frac{E_2}{2E}, \ldots, \frac{E_m}{2E}\right) \tag{4.41}$$

This answer for the entropy $X_1$e is so simple that it is almost misleading. 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 4.3.1*** *(Random walk on a chessboard)*: Let a king move at random on an $8 \times 8$ chessboard. The king has 8 moves in the interior, 5 moves at the edges and 3 moves at the corners. Using this and the preceding results, the stationary probabilities are respectively $\frac{8}{420}$, $\frac{5}{420}$ and $\frac{3}{420}$, and the entropy rate is $0.92 \log 8$. The factor of 0.92 is due to edge effects; we would have an entropy rate of $\log 8$ on an infinite chessboard.

Similarly, we can find the entropy rate of rooks ($\log 14$ bits, since the rook always has 14 possible moves), bishops and queens. The queen combines the moves of a rook and a bishop. Does the queen have more or less freedom than the pair?

**Remark:** 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:

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \Pr(X_n = x_1, X_{n-1} = x_2, \ldots, X_1 = x_n). \tag{4.42}$$

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

## 4.4 HIDDEN MARKOV MODELS

Here is an example that can be very difficult if done the wrong way. It illustrates the power of the techniques developed so far. Let $X_1, X_2, \dots, X_n, \dots$ be a stationary Markov chain, and let $Y_i = \phi(X_i)$ be a process, each term of which is a function of the corresponding state in the Markov chain. Such functions of Markov chains occur often in practice. In many situations, one has only partial information about the state of the system. It would simplify matters greatly if $Y_1, Y_2, \dots, Y_n$ also formed a Markov chain, but in many cases this is not true. However, since the Markov chain is stationary, so is $Y_1, Y_2, \dots, Y_n$, and the entropy rate is well defined. However, if we wish to compute $H(\mathcal{Y})$, we might compute $H(Y_n|Y_{n-1}, \dots, Y_1)$ for each $n$ and find the limit. Since the convergence can be arbitrarily slow, we will never know how close we are to the limit; we will not know when to stop. (We can't look at the change between the values at $n$ and $n+1$, since this difference may be small even when we are far away from the limit—consider, for example, $\Sigma \frac{1}{n}$.)

It would be useful computationally to have upper and lower bounds converging to the limit from above and below. We can halt the computation when the difference between the upper bound and the lower bound is small, and we will then have a good estimate of the limit.

We already know that $H(Y_n|Y_{n-1}, \dots, Y_1)$ converges monotonically to $H(\mathcal{Y})$ from above. For a lower bound, we will use $H(Y_n|Y_{n-1}, \dots, Y_2, X_1)$. This is a neat trick based on the idea that $X_1$ contains as much information about $Y_n$ as $Y_1, Y_0, Y_{-1}, \dots$.

**Lemma 4.4.1:**

$$H(Y_n|Y_{n-1}, \dots, Y_2, X_1) \le H(\mathcal{Y}) \tag{4.43}$$

**Proof:** We have, for $k = 1, 2, \dots$,

$$H(Y_n|Y_{n-1}, \dots, Y_2, X_1)$$

$$\overset{(a)}{=} H(Y_n|Y_{n-1}, \dots, Y_2, Y_1, X_1) \tag{4.44}$$

$$\overset{(b)}{=} H(Y_n|Y_{n-1}, \dots, Y_1, X_1, X_0, X_{-1}, \dots, X_{-k}) \tag{4.45}$$

$$\overset{(c)}{=} H(Y_n|Y_{n-1}, \dots, Y_1, X_1, X_0, X_{-1}, \dots, X_{-k}, Y_0, \dots, Y_{-k}) \tag{4.46}$$

$$\overset{(d)}{\leq} H(Y_n|Y_{n-1},\ldots,Y_1,Y_0,\ldots,Y_{-k}) \tag{4.47}$$

$$\overset{(e)}{=} H(Y_{n+k+1}|Y_{n+k},\ldots,Y_1), \tag{4.48}$$

where $(a)$ follows from that fact that $Y_1$ is a function of $X_1$, and $(b)$ follows from the Markovity of $X$, $(c)$ from the fact that $Y_i$ is a function of $X_i$, $(d)$ from the fact that conditioning reduces entropy, and $(e)$ by stationarity. Since the inequality is true for all $k$, it is true in the limit. Thus

$$H(Y_n|Y_{n-1},\ldots,Y_1,X_1) \leq \lim_k H(Y_{n+k+1}|Y_{n+k},\ldots,Y_1) \tag{4.49}$$

$$= H(\mathcal{Y}). \quad \square \tag{4.50}$$

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

**Lemma 4.4.2:**

$$H(Y_n|Y_{n-1},\ldots,Y_1) - H(Y_n|Y_{n-1},\ldots,Y_1,X_1) \to 0. \tag{4.51}$$

**Proof:** The interval length can be rewritten as

$$H(Y_n|Y_{n-1},\ldots,Y_1) - H(Y_n|Y_{n-1},\ldots,Y_1,X_1) = I(X_1;Y_n|Y_{n-1},\ldots,Y_1). \tag{4.52}$$

By the properties of mutual information,

$$I(X_1;Y_1,Y_2,\ldots,Y_n) \leq H(X_1), \tag{4.53}$$

and hence

$$\lim_{n\to\infty} I(X_1;Y_1,Y_2,\ldots,Y_n) \leq H(X_1). \tag{4.54}$$

By the chain rule,

$$\lim_{n\to\infty} I(X_1;Y_1,Y_2,\ldots,Y_n) = \lim_{n\to\infty} \sum_{i=1}^{n} I(X_1;Y_i|Y_{i-1},\ldots,Y_1) \tag{4.55}$$

$$= \sum_{i=1}^{\infty} I(X_1;Y_i|Y_{i-1},\ldots,Y_1). \tag{4.56}$$

Since this infinite sum is finite and the terms are non-negative, the terms must tend to 0, i.e.,

$$\lim I(X_1; Y_n | Y_{n-1}, \ldots, Y_1) = 0\,, \tag{4.57}$$

which proves the lemma. □

Combining the previous two lemmas, we have the following theorem:

**Theorem 4.4.1:** *If $X_1, X_2, \ldots, X_n$ form a stationary Markov chain, and $Y_i = \phi(X_i)$, then*

$$H(Y_n | Y_{n-1}, \ldots, Y_1, X_1) \le H(\mathcal{Y}) \le H(Y_n | Y_{n-1}, \ldots, Y_1) \tag{4.58}$$

*and*

$$\lim H(Y_n | Y_{n-1}, \ldots, Y_1, X_1) = H(\mathcal{Y}) = \lim H(Y_n | Y_{n-1}, \ldots, Y_1)\,. \tag{4.59}$$

## SUMMARY OF CHAPTER 4

**Entropy rate:** Two definitions of entropy rate for a stochastic process are

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)\,, \tag{4.60}$$

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1)\,. \tag{4.61}$$

For a stationary stochastic process,

$$H(\mathcal{X}) = H'(\mathcal{X})\,. \tag{4.62}$$

**Entropy rate of a stationary Markov chain:**

$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}\,. \tag{4.63}$$

**Functions of a Markov chain:** If $X_1, X_2, \ldots, X_n$ form a Markov chain and $Y_i = \phi(X_i)$, then

$$H(Y_n | Y_{n-1}, \ldots, Y_1, X_1) \le H(\mathcal{Y}) \le H(Y_n | Y_{n-1}, \ldots, Y_1) \tag{4.64}$$

and

$$\lim_{n\to\infty} H(Y_n | Y_{n-1}, \ldots, Y_1, X_1) = H(\mathcal{Y}) = \lim_{n\to\infty} H(Y_n | Y_{n-1}, \ldots, Y_1)\,. \tag{4.65}$$

## PROBLEMS FOR CHAPTER 4

1. *Doubly stochastic matrices.* An $n \times n$ matrix $P = [P_{ij}]$ is said to be *doubly stochastic* if $P_{ij} \geq 0$ and $\Sigma_j P_{ij} = 1$ for all $i$ and $\Sigma_i P_{ij} = 1$ for all $j$. An $n \times n$ matrix $P$ is said to be a *permutation* matrix if it is doubly stochastic and there is precisely one $P_{ij} = 1$ in each row and each column.

   It can be shown that every doubly stochastic matrix can be written as the convex combination of permutation matrices.

   (a) Let $\mathbf{a}^t = (a_1, a_2, \ldots, a_n)$, $a_i \geq 0$, $\Sigma\, a_i = 1$, be a probability vector. Let $\mathbf{b} = \mathbf{a}P$, where $P$ is doubly stochastic. Show that $\mathbf{b}$ is a probability vector and that $H(b_1, b_2, \ldots, b_n) \geq H(a_1, a_2, \ldots, a_n)$. Thus stochastic mixing increases entropy.

   (b) Show that a stationary distribution $\mu$ for a doubly stochastic matrix $P$ is the uniform distribution.

   (c) Conversely, prove that if the uniform distribution is a stationary distribution for a Markov transition matrix $P$, then $P$ is doubly stochastic.

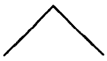
2. *Time's arrow.* Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary stochastic process. Prove that

$$H(X_0|X_{-1}, X_{-2}, \ldots, X_{-n}) = H(X_0|X_1, X_2, \ldots, X_n) .$$

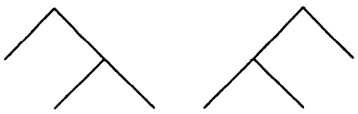
   In other words, the present has a conditional entropy given the past equal to the conditional entropy given the future.

   This is true even though it is quite easy to concoct stationary random processes for which the flow into the future looks quite different from the flow into the past. That is to say, one can determine the direction of time by looking at a sample function of the process. Nonetheless, given the present state, the conditional uncertainty of the next symbol in the future is equal to the conditional uncertainty of the previous symbol in the past.

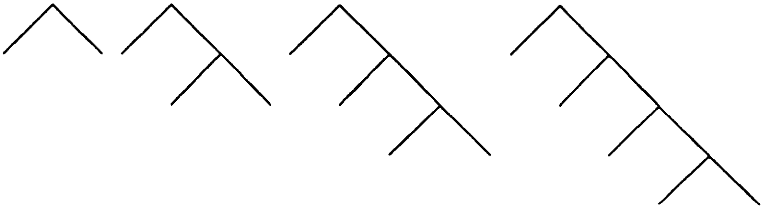
3. *Entropy of a random tree.* Consider the following method of generating a random tree with $n$ nodes. First expand the root node:



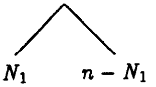
   Then expand one of the two terminal nodes at random:



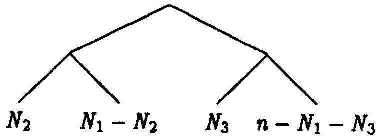
   At time $k$, choose one of the $k - 1$ terminal nodes according to a uniform distribution and expand it. Continue until $n$ terminal nodes have been generated. Thus a sequence leading to a five node tree might look like this:



Surprisingly, the following method of generating random trees yields the same probability distribution on trees with $n$ terminal nodes. First choose an integer $N_1$ uniformly distributed on $\{1, 2, \ldots, n-1\}$. We then have the picture.



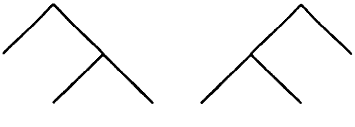
Then choose an integer $N_2$ uniformly distributed over $\{1, 2, \ldots, N_1 - 1\}$, and independently choose another integer $N_3$ uniformly over $\{1, 2, \ldots, (n - N_1) - 1\}$. The picture is now:



Continue the process until no further subdivision can be made. (The equivalence of these two tree generation schemes follows, for example, from Polya's urn model.)

Now let $T_n$ denote a random $n$-node tree generated as described. The probability distribution on such trees seems difficult to describe, but we can find the entropy of this distribution in recursive form.

First some examples. For $n = 2$, we have only one tree. Thus $H(T_2) = 0$. For $n = 3$, we have two equally probable trees:



Thus $H(T_3) = \log 2$. For $n = 4$, we have five possible trees, with probabilities 1/3, 1/6, 1/6, 1/6, 1/6.

Now for the recurrence relation. Let $N_1(T_n)$ denote the number of terminal nodes of $T_n$ in the right half of the tree. Justify each of the steps in the following:

$$H(T_n) \overset{(a)}{=} H(N_1, T_n) \tag{4.66}$$

$$\overset{(b)}{=} H(N_1) + H(T_n|N_1) \tag{4.67}$$

$$\overset{(c)}{=} \log(n-1) + H(T_n|N_1) \tag{4.68}$$

$$\overset{(d)}{=} \log(n-1) + \frac{1}{n-1}\sum_{k=1}^{n-1}[H(T_k) + H(T_{n-k})] \tag{4.69}$$

$$\overset{(e)}{=} \log(n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H(T_k)\,. \tag{4.70}$$

$$= \log(n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H_k\,. \tag{4.71}$$

(f) Use this to show that

$$(n-1)H_n = nH_{n-1} + (n-1)\log(n-1) - (n-2)\log(n-2)\,, \tag{4.72}$$

or

$$\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + c_n\,, \tag{4.73}$$

for appropriately defined $c_n$. Since $\Sigma c_n = c < \infty$, you have proved that $\frac{1}{n}H(T_n)$ converges to a constant. Thus the expected number of bits necessary to describe the random tree $T_n$ grows linearly with $n$.

4. *Monotonicity of entropy per element.* For a stationary stochastic process $X_1, X_2, \ldots, X_n$, show that

(a)
$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \le \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}\,. \tag{4.74}$$

(b)
$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \ge H(X_n|X_{n-1}, \ldots, X_1)\,. \tag{4.75}$$

5. *Entropy rates of Markov chains.*

(a) Find the entropy rate of the two-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1-p_{01} & p_{01} \\ p_{10} & 1-p_{10} \end{bmatrix}.$$

(b) What values of $p_{01}, p_{10}$ maximize the rate of part (a)?

(c) Find the entropy rate of the two-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix}.$$

(d) Find the maximum value of the entropy rate of the Markov chain of part (c). We expect that the maximizing value of $p$ should be less than 1/2, since the 0 state permits more information to be generated than the 1 state.

(e) Let $N(t)$ be the number of allowable state sequences of length $t$ for the Markov chain of part (c). Find $N(t)$ and calculate

$$H_0 = \lim_{t\to\infty} \frac{1}{t} \log N(t) .$$

*Hint*: Find a linear recurrence that expresses $N(t)$ in terms of $N(t-1)$ and $N(t-2)$. Why is $H_0$ an upper bound on the entropy rate of the Markov chain? Compare $H_0$ with the maximum entropy found in part (d).

**6.** *Maximum entropy process.* A discrete memoryless source has alphabet $\{1, 2\}$ where the symbol 1 has duration 1 and the symbol 2 has duration 2. The probabilities of 1 and 2 are $p_1$ and $p_2$, respectively. Find the value of $p_1$ that maximizes the source entropy per unit time $H(X)/El_X$. What is the maximum value $H$?

**7.** *Initial conditions.* Show, for a Markov chain, that

$$H(X_0|X_n) \geq H(X_0|X_{n-1}) .$$

Thus initial conditions $X_0$ become more difficult to recover as the future $X_n$ unfolds.

**8.** *Pairwise independence.* Let $X_1, X_2, \ldots, X_{n-1}$ be i.i.d. random variables taking values in $\{0, 1\}$, with $\Pr\{X_i = 1\} = \frac{1}{2}$. Let $X_n = 1$ if $\Sigma_{i=1}^{n-1} X_i$ is odd and $X_n = 0$ otherwise. Let $n \geq 3$.

(a) Show that $X_i$ and $X_j$ are independent, for $i \neq j$, $i, j \in \{1, 2, \ldots, n\}$.

(b) Find $H(X_i, X_j)$, for $i \neq j$.

(c) Find $H(X_1, X_2, \ldots, X_n)$. Is this equal to $nH(X_1)$?

**9.** *Stationary processes.* Let $\ldots, X_{-1}, X_0, X_1, \ldots$ be a stationary (not necessarily Markov) stochastic process. Which of the following statements are true? State true or false. Then either prove or provide a counterexample. Warning: At least one answer is false.

(a) $H(X_n|X_0) = H(X_{-n}|X_0)$.

(b) $H(X_n|X_0) \geq H(X_{n-1}|X_0)$.

(c) $H(X_n|X_1^{n-1}, X_{n+1})$ is nonincreasing in $n$.

**10.** *The entropy rate of a dog looking for a bone.* A dog walks on the integers, possibly reversing direction at each step with probability $p = .1$. Let $X_0 = 0$. The first step is equally likely to be positive or negative. A typical walk might look like this:

$$(X_0, X_1, \ldots) = (0, -1, -2, -3, -4, -3, -2, -1, 0, 1, \ldots) .$$

(a) Find $H(X_1, X_2, \ldots, X_n)$.

(b) Find the entropy rate of this browsing dog.

(c) What is the expected number of steps the dog takes before reversing direction?

**11.** *Random walk on chessboard.* Find the entropy rate of the Markov chain associated with a random walk of a king on the 3 × 3 chessboard

| | | |
|---|---|---|
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | 9 |

What about the entropy rate of rooks, bishops and queens? There are two types of bishops.

12. *Entropy rate.* Let $\{X_i\}$ be a discrete stationary stochastic process with entropy rate $H(\mathscr{X})$. Show

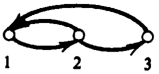
$$\frac{1}{n}H(X_n, \ldots, X_1|X_0, X_{-1}, \ldots, X_{-k}) \to H(\mathscr{X}), \qquad (4.76)$$

for $k = 1, 2, \ldots$.

13. *Entropy rate of constrained sequences.* In magnetic recording, the mechanism of recording and reading the bits imposes constraints on the sequences of bits that can be recorded. For example, to ensure proper synchronization, it is often necessary to limit the length of runs of 0's between two 1's. Also to reduce intersymbol interference, it may be necessary to require at least one 0 between any two 1's. We will consider a simple example of such a constraint.

Suppose that we are required to have at least one 0 and at most two 0's between any pair of 1's in a sequences. Thus, sequences like 101001 and 0101001 are valid sequences, but 0110010 and 0000101 are not. We wish to calculate the number of valid sequences of length $n$.

(a) Show that the set of constrained sequences is the same as the set of allowed paths on the following state diagram:



(b) Let $X_i(n)$ be the number of valid paths of length $n$ ending at state $i$. Argue that $\mathbf{X}(n) = [X_1(n)\ X_2(n)\ X_3(n)]^T$ satisfies the following recursion:

$$\begin{bmatrix} X_1(n) \\ X_2(n) \\ X_3(n) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1(n-1) \\ X_2(n-1) \\ X_3(n-1) \end{bmatrix} = A\mathbf{X}(n-1) \qquad (4.77)$$

with initial conditions $\mathbf{X}(1) = [1\ 1\ 0]^T$.

(c) Then we have by induction

$$\mathbf{X}(n) = A\mathbf{X}(n-1) = A^2\mathbf{X}(n-2) = \cdots = A^{n-1}\mathbf{X}(1). \qquad (4.78)$$

Using the eigenvalue decomposition of $A$ for the case of distinct eigenvalues, we can write $A = U^{-1}\ \Lambda U$, where $\Lambda$ is the diagonal matrix of eigenvalues. Then $A^{n-1} = U^{-1}\ \Lambda^{n-1}\ U$. Show that we can write

$$\mathbf{X}(n) = \lambda_1^{n-1}\mathbf{Y}_1 + \lambda_2^{n-1}\mathbf{Y}_2 + \lambda_3^{n-1}\mathbf{Y}_3 , \tag{4.79}$$

where $\mathbf{Y}_1$, $\mathbf{Y}_2$, $\mathbf{Y}_3$ do not depend on $n$. For large $n$, this sum is dominated by the largest term. Therefore argue that for $i = 1, 2, 3$, we have

$$\frac{1}{n} \log X_i(n) \rightarrow \log \lambda , \tag{4.80}$$

where $\lambda$ is the largest (positive) eigenvalue. Thus the number of sequences of length $n$ grows as $\lambda^n$ for large $n$. Calculate $\lambda$ for the matrix $A$ above. (The case when the eigenvalues are not distinct can be handled in a similar manner.)

(d) We will now take a different approach. Consider a Markov chain whose state diagram is the one given in part (a), but with arbitrary transition probabilities. Therefore the probability transition matrix of this Markov chain is

$$P = \begin{bmatrix} 0 & \alpha & 1 \\ 1 & 0 & 0 \\ 0 & 1-\alpha & 0 \end{bmatrix} . \tag{4.81}$$

Show that the stationary distribution of this Markov chain is

$$\mu = \left[ \frac{1}{3-\alpha}, \frac{1}{3-\alpha}, \frac{1-\alpha}{3-\alpha} \right]^T . \tag{4.82}$$

(e) Maximize the entropy rate of the Markov chain over choices of $\alpha$. What is the maximum entropy rate of the chain?

(f) Compare the maximum entropy rate in part (e) with $\log \lambda$ in part (c). Why are the two answers the same?

**14.** *Waiting times are insensitive to distributions.* Let $X_0, X_1, X_2, \ldots$ be drawn i.i.d. $\sim p(x)$, $x \in \mathcal{X} = \{1, 2, \ldots, m\}$ and let $N$ be the waiting time to the next occurrence of $X_0$, where $N = \min_n \{X_n = X_0\}$.

(a) Show that $EN = m$.

(b) Show that $E \log N \le H(X)$.

(c) (Optional) Prove part (a) for $\{X_i\}$ stationary and ergodic.

## HISTORICAL NOTES

The entropy rate of a stochastic process was introduced by Shannon [238], who also explored some of the connections between the entropy rate of the process and the number of possible sequences generated by the process. Since Shannon, there have been a number of results extending the basic theorems of information theory to general stochastic processes.