9.3 SURPRISE, UNCERTAINTY, AND ENTROPY

Consider an event E that can occur when an experiment is performed. How surprised would we be to hear that E does, in fact, occur? It seems reasonable to suppose that the amount of surprise engendered by the information that E has occurred should depend on the probability of E. For instance, if the experiment consists of rolling a pair of dice, then we would not be too surprised to hear that E has occurred when E represents the event that the sum of the dice is even (and thus has probability $\frac{1}{2}$), whereas we would certainly be more surprised to hear that E has occurred when E is the event that the sum of the dice is 12 (and thus has probability $\frac{1}{36}$).

In this section, we attempt to quantify the concept of surprise. To begin, let us agree to suppose that the surprise one feels upon learning that an event E has occurred depends only on the probability of E, and let us denote by $S(p)$ the surprise evoked by the occurrence of an event having probability p. We determine the functional form of $S(p)$ by first agreeing on a set of reasonable conditions that $S(p)$ should satisfy and then proving that these axioms require that $S(p)$ have a specified form. We assume throughout that $S(p)$ is defined for all $0 < p \leq 1$, but is not defined for events having $p = 0$.

Our first condition is just a statement of the intuitive fact that there is no surprise in hearing that an event which is sure to occur has indeed occurred.

Axiom 1

$$S(1) = 0$$

Our second condition states that the more unlikely an event is to occur, the greater is the surprise evoked by its occurrence.

Axiom 2

$S(p)$ is a strictly decreasing function of p; that is, if $p < q$, then $S(p) > S(q)$.

The third condition is a mathematical statement of the fact that we would intuitively expect a small change in p to correspond to a small change in $S(p)$.

Axiom 3

$S(p)$ is a continuous function of p.

To motivate the final condition, consider two independent events E and F having respective probabilities $P(E) = p$ and $P(F) = q$. Since $P(EF) = pq$, the surprise evoked by the information that both E and F have occurred is $S(pq)$. Now, suppose that we are told first that E has occurred and then, afterward, that F has also occurred. Since $S(p)$ is the surprise evoked by the occurrence of E, it follows that $S(pq) - S(p)$ represents the additional surprise evoked when we are informed that F has also occurred. However, because F is independent of E, the knowledge that E occurred does not change the probability of F; hence, the additional surprise should just be $S(q)$. This reasoning suggests the final condition.

Axiom 4

$$S(pq) = S(p) + S(q) \qquad 0 < p \leq 1, \quad 0 < q \leq 1$$

We are now ready for Theorem 3.1, which yields the structure of $S(p)$.

Theorem 3.1. *If $S(\cdot)$ satisfies Axioms 1 through 4, then*

$$S(p) = -C \log_2 p$$

where C is an arbitrary positive integer.

Proof. It follows from Axiom 4 that

$$S(p^2) = S(p) + S(p) = 2S(p)$$

and by induction that

$$S(p^m) = mS(p) \tag{3.1}$$

Also, since, for any integral $n, S(p) = S(p^{1/n} \cdots p^{1/n}) = n\, S(p^{1/n})$, it follows that

$$S(p^{1/n}) = \frac{1}{n} S(p) \tag{3.2}$$

Thus, from Equations (3.1) and (3.2), we obtain

$$\begin{aligned} S(p^{m/n}) &= mS(p^{1/n}) \\ &= \frac{m}{n} S(p) \end{aligned}$$

which is equivalent to

$$S(p^x) = xS(p) \tag{3.3}$$

whenever x is a positive rational number. But by the continuity of S (Axiom 3), it follows that Equation (3.3) is valid for all nonnegative x. (Reason this out.)

Now, for any $p, 0 < p \leq 1$, let $x = -\log_2 p$. Then $p = \left(\frac{1}{2}\right)^x$, and from Equation (3.3),

$$S(p) = S\left(\left(\frac{1}{2}\right)^x\right) = xS\left(\frac{1}{2}\right) = -C \log_2 p$$

where $C = S\left(\frac{1}{2}\right) > S(1) = 0$ by Axioms 2 and 1. □

It is usual to let C equal 1, in which case the surprise is said to be expressed in units of *bits* (short for *binary digits*).

Next, consider a random variable X that must take on one of the values $x_1, \ldots, x_n$ with respective probabilities $p_1, \ldots, p_n$. Since $-\log p_i$ represents the surprise evoked if X takes on the value x_i,[†] it follows that the expected amount of surprise we shall receive upon learning the value of X is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

The quantity $H(X)$ is known in information theory as the *entropy* of the random variable X. (In case one of the $p_i = 0$, we take 0 log 0 to equal 0.) It can be shown (and we leave it as an exercise) that $H(X)$ is maximized when all of the p_i are equal. (Is this intuitive?)

Since $H(X)$ represents the average amount of surprise one receives upon learning the value of X, it can also be interpreted as representing the amount of *uncertainty* that exists as to the value of X. In fact, in information theory, $H(X)$ is interpreted as the average amount of *information* received when the value of X is observed. Thus, the average surprise evoked by X, the uncertainty of X, or the average amount of

[†]For the remainder of this chapter, we write $\log x$ for $\log_2 x$. Also, we use $\ln x$ for $\log_e x$.

information yielded by X all represent the same concept viewed from three slightly different points of view.

Now consider two random variables X and Y that take on the respective values $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ with joint mass function

$$p(x_i, y_j) = P\{X = x_i, Y = y_j\}$$

It follows that the uncertainty as to the value of the random vector (X, Y), denoted by $H(X, Y)$, is given by

$$H(X, Y) = -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j)$$

Suppose now that Y is observed to equal y_j. In this situation, the amount of uncertainty remaining in X is given by

$$H_{Y=y_j}(X) = -\sum_i p(x_i|y_j) \log p(x_i|y_j)$$

where

$$p(x_i|y_j) = P\{X = x_i|Y = y_j\}$$

Hence, the average amount of uncertainty that will remain in X after Y is observed is given by

$$H_Y(X) = \sum_j H_{Y=y_j}(X) p_Y(y_j)$$

where

$$p_Y(y_j) = P\{Y = y_j\}$$

Proposition 3.1 relates $H(X, Y)$ to $H(Y)$ and $H_Y(X)$. It states that the uncertainty as to the value of X and Y is equal to the uncertainty of Y plus the average uncertainty remaining in X when Y is to be observed.

Proposition 3.1.

$$H(X, Y) = H(Y) + H_Y(X)$$

Proof. Using the identity $p(x_i, y_j) = p_Y(y_j)p(x_i|y_j)$ yields

$$\begin{aligned}
H(X, Y) &= -\sum_i \sum_j p(x_i, y_j) \log p(x_i, y_j) \\
&= -\sum_i \sum_j p_Y(y_j) p(x_i|y_j) [\log p_Y(y_j) + \log p(x_i|y_j)] \\
&= -\sum_j p_Y(y_j) \log p_Y(y_j) \sum_i p(x_i|y_j) \\
&\quad - \sum_j p_Y(y_j) \sum_i p(x_i|y_j) \log p(x_i|y_j) \\
&= H(Y) + H_Y(X)
\end{aligned}$$

□

It is a fundamental result in information theory that the amount of uncertainty in a random variable X will, on the average, decrease when a second random variable Y is observed. Before proving this statement, we need the following lemma, whose proof is left as an exercise.

Lemma 3.1

$$\ln x \leq x - 1 \quad x > 0$$

with equality only at $x = 1$.

Theorem 3.2.

$$H_Y(X) \leq H(X)$$

with equality if and only if X and Y are independent.

Proof.

$$\begin{aligned}
H_Y(X) - H(X) &= -\sum_i \sum_j p(x_i|y_j) \log[p(x_i|y_j)]p(y_j) \\
&\quad + \sum_i \sum_j p(x_i, y_j) \log p(x_i) \\
&= \sum_i \sum_j p(x_i, y_j) \log \left[\frac{p(x_i)}{p(x_i|y_j)} \right] \\
&\leq \log e \sum_i \sum_j p(x_i, y_j) \left[\frac{p(x_i)}{p(x_i|y_j)} - 1 \right] \quad \text{by Lemma 3.1} \\
&= \log e \left[\sum_i \sum_j p(x_i)p(y_j) - \sum_i \sum_j p(x_i, y_j) \right] \\
&= \log e[1 - 1] \\
&= 0
\end{aligned}$$

□

9.4 CODING THEORY AND ENTROPY

Suppose that the value of a discrete random vector X is to be observed at location A and then transmitted to location B via a communication network that consists of two signals, 0 and 1. In order to do this, it is first necessary to encode each possible value of X in terms of a sequence of 0's and 1's. To avoid any ambiguity, it is usually required that no encoded sequence can be obtained from a shorter encoded sequence by adding more terms to the shorter.

For instance, if X can take on four possible values x_1, x_2, x_3, and x_4, then one possible coding would be

$$\begin{aligned}
x_1 &\leftrightarrow 00 \\
x_2 &\leftrightarrow 01 \\
x_3 &\leftrightarrow 10 \\
x_4 &\leftrightarrow 11
\end{aligned} \tag{4.1}$$

That is, if $X = x_1$, then the message 00 is sent to location B, whereas if $X = x_2$, then 01 is sent to B, and so on. A second possible coding is

$$\begin{aligned}
x_1 &\leftrightarrow 0 \\
x_2 &\leftrightarrow 10 \\
x_3 &\leftrightarrow 110 \\
x_4 &\leftrightarrow 111
\end{aligned} \tag{4.2}$$

However, a coding such as

$$x_1 \leftrightarrow 0$$
$$x_2 \leftrightarrow 1$$
$$x_3 \leftrightarrow 00$$
$$x_4 \leftrightarrow 01$$

is not allowed because the coded sequences for x_3 and x_4 are both extensions of the one for x_1.

One of the objectives in devising a code is to minimize the expected number of bits (that is, binary digits) that need to be sent from location A to location B. For example, if

$$P\{X = x_1\} = \frac{1}{2}$$
$$P\{X = x_2\} = \frac{1}{4}$$
$$P\{X = x_3\} = \frac{1}{8}$$
$$P\{X = x_4\} = \frac{1}{8}$$

then the code given by Equation (4.2) would expect to send $\frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{8}(3) + \frac{1}{8}(3) = 1.75$ bits, whereas the code given by Equation (4.1) would expect to send 2 bits. Hence, for the preceding set of probabilities, the encoding in Equation (4.2) is more efficient than that in Equation (4.1).

The preceding discussion raises the following question: For a given random vector X, what is the maximum efficiency achievable by an encoding scheme? The answer is that, for any coding, the average number of bits that will be sent is at least as large as the entropy of X. To prove this result, known in information theory as the *noiseless coding theorem*, we shall need Lemma 4.1.

Lemma 4.1

Let X take on the possible values $x_1, \ldots, x_N$. Then, in order to be able to encode the values of X in binary sequences (none of which is an extension of another) of respective lengths $n_1, \ldots, n_N$, it is necessary and sufficient that

$$\sum_{i=1}^{N} \left(\frac{1}{2}\right)^{n_i} \leq 1$$

Proof. For a fixed set of N positive integers $n_1, \ldots, n_N$, let w_j denote the number of the n_i that are equal to $j, j = 1, \ldots$. For there to be a coding that assigns n_i bits to the value $x_i, i = 1, \ldots, N$, it is clearly necessary that $w_1 \leq 2$. Furthermore, because no binary sequence is allowed to be an extension of any other, we must have $w_2 \leq 2^2 - 2w_1$. (This follows because 2^2 is the number of binary sequences of length 2, whereas $2w_1$ is the number of sequences that are extensions of the w_1 binary sequence of length 1.) In general, the same reasoning shows that we must have

$$w_n \leq 2^n - w_1 2^{n-1} - w_2 2^{n-2} - \cdots - w_{n-1} 2 \tag{4.3}$$

for $n = 1, \ldots$. In fact, a little thought should convince the reader that these conditions are not only necessary, but also sufficient for a code to exist that assigns n_i bits to $x_i, i = 1, \ldots, N$.

Rewriting inequality (4.3) as

$$w_n + w_{n-1}2 + w_{n-2}2^2 + \cdots + w_1 2^{n-1} \leq 2^n \quad n = 1, \ldots$$

and dividing by 2^n yields the necessary and sufficient conditions, namely,

$$\sum_{j=1}^{n} w_j \left(\frac{1}{2}\right)^j \leq 1 \quad \text{for all } n \tag{4.4}$$

However, because $\sum_{j=1}^{n} w_j \left(\frac{1}{2}\right)^j$ is increasing in n, it follows that Equation (4.4) will be true if and only if

$$\sum_{j=1}^{\infty} w_j \left(\frac{1}{2}\right)^j \leq 1$$

The result is now established, since, by the definition of w_j as the number of n_i that equal j, it follows that

$$\sum_{j=1}^{\infty} w_j \left(\frac{1}{2}\right)^j = \sum_{i=1}^{N} \left(\frac{1}{2}\right)^{n_i}$$

□

We are now ready to prove Theorem 4.1.

Theorem 4.1 The noiseless coding theorem
Let X take on the values $x_1, \ldots, x_N$ with respective probabilities $p(x_1), \ldots, p(x_N)$. Then, for any coding of X that assigns n_i bits to x_i,

$$\sum_{i=1}^{N} n_i p(x_i) \geq H(X) = -\sum_{i=1}^{N} p(x_i) \log p(x_i)$$

Proof. Let $P_i = p(x_i), q_i = 2^{-n_i} \Big/ \sum_{j=1}^{N} 2^{-n_j}, \ i = 1, \ldots, N$. Then

$$\begin{aligned}
-\sum_{i=1}^{N} P_i \log\left(\frac{P_i}{q_i}\right) &= -\log e \sum_{i=1}^{N} P_i \ln\left(\frac{P_i}{q_i}\right) \\
&= \log e \sum_{i=1}^{N} P_i \ln\left(\frac{q_i}{P_i}\right) \\
&\leq \log e \sum_{i=1}^{N} P_i \left(\frac{q_i}{P_i} - 1\right) \quad \text{by Lemma 3.1} \\
&= 0 \quad \text{since } \sum_{i=1}^{N} P_i = \sum_{i=1}^{N} q_i = 1
\end{aligned}$$

Hence,

$$\begin{aligned} -\sum_{i=1}^{N} P_i \log P_i &\le -\sum_{i=1}^{N} P_i \log q_i \\ &= \sum_{i=1}^{N} n_i P_i + \log\left(\sum_{j=1}^{N} 2^{-n_j}\right) \\ &\le \sum_{i=1}^{N} n_i P_i \quad \text{by Lemma 4.1} \end{aligned}$$

□

EXAMPLE 4a

Consider a random variable X with probability mass function

$$p(x_1) = \frac{1}{2} \quad p(x_2) = \frac{1}{4} \quad p(x_3) = p(x_4) = \frac{1}{8}$$

Since

$$\begin{aligned} H(X) &= -\left[\frac{1}{2}\log\frac{1}{2} + \frac{1}{4}\log\frac{1}{4} + \frac{1}{4}\log\frac{1}{8}\right] \\ &= \frac{1}{2} + \frac{2}{4} + \frac{3}{4} \\ &= 1.75 \end{aligned}$$

it follows from Theorem 4.1 that there is no more efficient coding scheme than

$$\begin{aligned} x_1 &\leftrightarrow 0 \\ x_2 &\leftrightarrow 10 \\ x_3 &\leftrightarrow 110 \\ x_4 &\leftrightarrow 111 \end{aligned}$$

■

For most random vectors, there does not exist a coding for which the average number of bits sent attains the lower bound $H(X)$. However, it is always possible to devise a code such that the average number of bits is within 1 of $H(X)$. To prove this, define n_i to be the integer satisfying

$$-\log p(x_i) \le n_i < -\log p(x_i) + 1$$

Now,

$$\sum_{i=1}^{N} 2^{-n_i} \le \sum_{i=1}^{N} 2^{\log p(x_i)} = \sum_{i=1}^{N} p(x_i) = 1$$

so, by Lemma 4.1, we can associate sequences of bits having lengths n_i with the $x_i, i = 1, \ldots, N$. The average length of such a sequence,

$$L = \sum_{i=1}^{N} n_i p(x_i)$$

satisfies

$$-\sum_{i=1}^{N} p(x_i)\log p(x_i) \le L < -\sum_{i=1}^{N} p(x_i)\log p(x_i) + 1$$

or

$$H(X) \le L < H(X) + 1$$

EXAMPLE 4b

Suppose that 10 independent tosses of a coin having probability p of coming up heads are made at location A and the result is to be transmitted to location B. The outcome of this experiment is a random vector $X = (X_1, \ldots, X_{10})$, where X_i is 1 or 0 according to whether or not the outcome of the ith toss is heads. By the results of this section, it follows that L, the average number of bits transmitted by any code, satisfies

$$H(X) \le L$$

with

$$L \le H(X) + 1$$

for at least one code. Now, since the X_i are independent, it follows from Proposition 3.1 and Theorem 3.2 that

$$\begin{aligned} H(X) = H(X_1, \ldots, X_n) &= \sum_{i=1}^{N} H(X_i) \\ &= -10[p\log p + (1-p)\log(1-p)] \end{aligned}$$

If $p = \frac{1}{2}$, then $H(X) = 10$, and it follows that we can do no better than just encoding X by its actual value. For example, if the first 5 tosses come up heads and the last 5 tails, then the message 1111100000 is transmitted to location B.

However, if $p \ne \frac{1}{2}$, we can often do better by using a different coding scheme. For instance, if $p = \frac{1}{4}$, then

$$H(X) = -10\left(\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}\right) = 8.11$$

Thus, there is an encoding for which the average length of the encoded message is no greater than 9.11.

One simple coding that is more efficient in this case than the identity code is to break up $(X_1, \ldots, X_{10})$ into 5 pairs of 2 random variables each and then, for $i = 1, 3, 5, 7, 9$, code each of the pairs as follows:

$$\begin{aligned} X_i = 0, X_{i+1} = 0 &\leftrightarrow 0 \\ X_i = 0, X_{i+1} = 1 &\leftrightarrow 10 \\ X_i = 1, X_{i+1} = 0 &\leftrightarrow 110 \\ X_i = 1, X_{i+1} = 1 &\leftrightarrow 111 \end{aligned}$$

The total message transmitted is the successive encodings of the preceding pairs.

For instance, if the outcome $TTTHHTTTTH$ is observed, then the message 01011 0010 is sent. The average number of bits needed to transmit the message with this code is

$$5\left[1\left(\frac{3}{4}\right)^2 + 2\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) + 3\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) + 3\left(\frac{1}{4}\right)^2\right] = \frac{135}{16}$$
$$\approx 8.44 \quad \blacksquare$$

Up to this point, we have assumed that the message sent at location A is received without error at location B. However, there are always certain errors that can occur because of random disturbances along the communications channel. Such random disturbances might lead, for example, to the message 00101101, sent at A, being received at B in the form 01101101.

Let us suppose that a bit transmitted at location A will be correctly received at location B with probability p, independently from bit to bit. Such a communications system is called a *binary symmetric channel*. Suppose further that $p = .8$ and we want to transmit a message consisting of a large number of bits from A to B. Thus, direct transmission of the message will result in an error probability of .20 for each bit, which is quite high. One way to reduce this probability of bit error would be to transmit each bit 3 times and then decode by majority rule. That is, we could use the following scheme:

Encode	Decode	Encode	Decode
$0 \rightarrow 000$	000, 001, 010, 100 $\rightarrow 0$	$1 \rightarrow 111$	111, 110, 101, 011 $\rightarrow 1$

Note that if no more than one error occurs in transmission, then the bit will be correctly decoded. Hence, the probability of bit error is reduced to

$$(.2)^3 + 3(.2)^2(.8) = .104$$

a considerable improvement. In fact, it is clear that we can make the probability of bit error as small as we want by repeating the bit many times and then decoding by majority rule. For instance, the scheme

Encode	Decode
$0 \rightarrow$ string of 17 0's	By majority rule
$1 \rightarrow$ string of 17 1's	

will reduce the probability of bit error to below .01.

The problem with this type of encoding scheme is that, although it decreases the probability of bit error, it does so at the cost of also decreasing the effective rate of bits sent per signal. (See Table 9.1.)

In fact, at this point it may appear inevitable to the reader that decreasing the probability of bit error to 0 *always* results in also decreasing the effective rate at which bits are transmitted per signal to 0. However, a remarkable result of information theory known as the *noisy coding theorem* and due to Claude Shannon demonstrates that this is not the case. We now state this result as Theorem 4.2.

TABLE 9.1: Repetition of Bits Encoding Scheme

Probability of error (per bit)	Rate (bits transmitted per signal)
.20	1
.10	.33 $\left(=\frac{1}{3}\right)$
.01	.06 $\left(=\frac{1}{17}\right)$

Theorem 4.2 The noisy coding theorem
There is a number C such that, for any value R which is less than C, and for any $\varepsilon > 0$, there exists a coding–decoding scheme that transmits at an average rate of R bits sent per signal and with an error (per bit) probability of less than ε. The largest such value of C—call it C^*[†] —is called the channel capacity, and for the binary symmetric channel,

$$C^* = 1 + p \log p + (1 - p) \log(1 - p)$$

SUMMARY

The *Poisson process* having rate λ is a collection of random variables $\{N(t), t \geq 0\}$ that relate to an underlying process of randomly occurring events. For instance, $N(t)$ represents the number of events that occur between times 0 and t. The defining features of the Poisson process are as follows:

(i) The number of events that occur in disjoint time intervals are independent.
(ii) The distribution of the number of events that occur in an interval depends only on the length of the interval.
(iii) Events occur one at a time.
(iv) Events occur at rate λ.

It can be shown that $N(t)$ is a Poisson random variable with mean λt. In addition, if $T_i, i \geq 1$, are the times between the successive events, then they are independent exponential random variables with rate λ.

A sequence of random variables $X_n, n \geq 0$, each of which takes on one of the values $0, \ldots, M$, is said to be a *Markov chain* with transition probabilities $P_{i,j}$ if, for all $n, i_0, \ldots, i_n, i, j$,

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} = P_{i,j}$$

If we interpret X_n as the state of some process at time n, then a Markov chain is a sequence of successive states of a process which has the property that whenever it enters state i, then, independently of all past states, the next state is j with probability $P_{i,j}$, for all states i and j. For many Markov chains, the probability of being in state j at time n converges to a limiting value that does not depend on the initial state. If we let $\pi_j, j = 0, \ldots, M$, denote these limiting probabilities, then they are the unique solution of the equations

$$\pi_j = \sum_{i=0}^{M} \pi_i P_{i,j} \quad j = 0, \ldots, M$$

$$\sum_{j=1}^{M} \pi_j = 1$$

[†]For an entropy interpretation of C^*, see Theoretical Exercise 9.18.

Moreover, π_j is equal to the long-run proportion of time that the chain is in state j.

Let X be a random variable that takes on one of n possible values according to the set of probabilities $\{p_1, \ldots, p_n\}$. The quantity

$$H(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

is called the *entropy* of X. It can be interpreted as representing either the average amount of uncertainty that exists regarding the value of X or the average information received when X is observed. Entropy has important implications for binary codings of X.

PROBLEMS AND THEORETICAL EXERCISES

9.1. Customers arrive at a bank at a Poisson rate λ. Suppose that two customers arrived during the first hour. What is the probability that
(a) both arrived during the first 20 minutes?
(b) at least one arrived during the first 20 minutes?

9.2. Cars cross a certain point in the highway in accordance with a Poisson process with rate $\lambda = 3$ per minute. If Al runs blindly across the highway, what is the probability that he will be uninjured if the amount of time that it takes him to cross the road is s seconds? (Assume that if he is on the highway when a car passes by, then he will be injured.) Do this exercise for $s = 2, 5, 10, 20$.

9.3. Suppose that in Problem 9.2 Al is agile enough to escape from a single car, but if he encounters two or more cars while attempting to cross the road, then he is injured. What is the probability that he will be unhurt if it takes him s seconds to cross? Do this exercise for $s = 5, 10, 20, 30$.

9.4. Suppose that 3 white and 3 black balls are distributed in two urns in such a way that each urn contains 3 balls. We say that the system is in state i if the first urn contains i white balls, $i = 0, 1, 2, 3$. At each stage, 1 ball is drawn from each urn and the ball drawn from the first urn is placed in the second, and conversely with the ball from the second urn. Let X_n denote the state of the system after the nth stage, and compute the transition probabilities of the Markov chain $\{X_n, n \geq 0\}$.

9.5. Consider Example 2a. If there is a 50–50 chance of rain today, compute the probability that it will rain 3 days from now if $\alpha = .7$ and $\beta = .3$.

9.6. Compute the limiting probabilities for the model of Problem 9.4.

9.7. A transition probability matrix is said to be doubly stochastic if

$$\sum_{i=0}^{M} P_{ij} = 1$$

for all states $j = 0, 1, \ldots, M$. Show that such a Markov chain is ergodic, then $\prod_j = 1/(M+1), j = 0, 1, \ldots, M$.

9.8. On any given day, Buffy is either cheerful (c), so-so (s), or gloomy (g). If she is cheerful today, then she will be c, s, or g tomorrow with respective probabilities .7, .2, and .1. If she is so-so today, then she will be c, s, or g tomorrow with respective probabilities .4, .3, and .3. If she is gloomy today, then Buffy will be c, s, or g tomorrow with probabilities .2, .4, and .4. What proportion of time is Buffy cheerful?

9.9. Suppose that whether it rains tomorrow depends on past weather conditions only through the last 2 days. Specifically, suppose that if it has rained yesterday and today, then it will rain tomorrow with probability .8; if it rained yesterday but not today, then it will rain tomorrow with probability .3; if it rained today but not yesterday, then it will rain tomorrow with probability .4; and if it has not rained either yesterday or today, then it will rain tomorrow with probability .2. What proportion of days does it rain?

9.10. A certain person goes for a run each morning. When he leaves his house for his run, he is equally likely to go out either the front or the back door, and similarly, when he returns, he is equally likely to go to either the front or the back door. The runner owns 5 pairs of running shoes, which he takes off after the run at whichever door he happens to be. If there are no shoes at the door from which he leaves to go running, he runs barefooted. We are interested in determining the proportion of time that he runs barefooted.
(a) Set this problem up as a Markov chain. Give the states and the transition probabilities.
(b) Determine the proportion of days that he runs barefooted.

9.11. This problem refers to Example 2f.
(a) Verify that the proposed value of $\prod_j$ satisfies the necessary equations.