ITCT Lecture IV.3:
Markov Processes and Sources with Memory

4.1 Markov Processes

Thus far, we have been occupied with memoryless sources and channels. We must now turn our attention to sources with memory. By this, we mean information sources, where the successive symbols in a transmitted sequence are correlated with each other, i.e., the sources in a sense "remember" what symbols they have previously emitted, and the probability of their next symbol depends on this history.

Sources with memory arise in a number of ways. First, natural languages such as English have this property. For example, the letter “q” in English is almost always followed by the letter “u”. Similarly, the letter “t” is followed by the letter “h” approximately 37% of the time in English text.

Many real-world signals, such as a speech waveform, are also heavily time correlated. If a signal is bandlimited, that is, if most of the energy of the signal is limited to a specific range of frequencies in its Fourier spectrum, then this signal will be time-correlated. Any time-correlated signal is a source with memory. Finally, we sometimes wish to deliberately introduce correlation (redundancy) in a source for purposes of block coding, as discussed in the previous section.

Let $A = \{a\}$ be the alphabet of a discrete source having $|A|$ symbols, and suppose this source emits a time sequence of symbols $(s_0, s_1, \cdots, s_t, \cdots)$ with each $s_t \in A$. If the conditional probability $p(s_t \mid s_{t-1}, s_{t-2}, \cdots, s_0)$ depends only on j previous symbols so that

$$\Pr(s_t \mid s_{t-1}, \cdots, s_0) = p(s_t \mid s_{t-1}, s_{t-2}, \cdots, s_{t-j}) \qquad (4.1)$$

then A is called a j^{th}-order Markov process.

The string of j symbols $S_t = (s_{t-1}, \cdots, s_{t-j})$ is called the state of the Markov process at time t. A j^{th}-order Markov process therefore has $N = |A|^j$ possible states.

Let us number these possible states from 0 to $N-1$ and let $\pi_n(t)$ represent the probability of being in state *n* at time *t*. the probability distribution of the system at time *t* can then be represented by the vector

$$\Pi_t \equiv \begin{bmatrix} \pi_0(t) \\ \pi_1(t) \\ \vdots \\ \pi_{N-2}(t) \\ \pi_{N-1}(t) \end{bmatrix} \qquad (4.2)$$

For each state at time *t*, there are $|A|$ possible next states at time $t+1$, depending on which symbol *a* is emitted next by the source.

If we let $p_{i|k}$ be the conditional probability of going to state i given that the present state is state k, the state probability distribution at time $t+1$ is governed by the transition probability matrix

$$P_{A|\Pi} \equiv \begin{bmatrix} p_{0|0} & p_{0|1} & \cdots & p_{0|N-1} \\ p_{1|0} & p_{1|1} & \cdots & p_{1|N-1} \\ \vdots & & \ddots & \\ p_{N-1|0} & & \cdots & p_{N-1|N-1} \end{bmatrix} \qquad (4.3)$$

and is given by $\Pi_{t+1} = P_{A|\Pi}\Pi_t$ (4.4)

Example 4.1

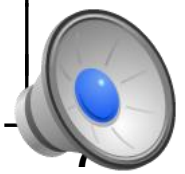
Let A be a binary 1^{st}-order Markov source with $A=\{0,1\}$. This source has 2 states labeled "0" and "1". Let the transition probabilities be

$$p_{0|0}=0.3,\quad p_{1|0}=0.7,\quad p_{0|1}=0.4,\quad p_{1|1}=0.6$$

What is the equation for the next probability state and find the state probabilities at time $t=2$, given that the probabilities at time $t=0$ are $\pi_0=1,\pi_1=0$.

Solution：$\Pi_{t+1}=\begin{bmatrix}0.3 & 0.4\\ 0.7 & 0.6\end{bmatrix}\Pi_t$, is the next-state equation for the state probabilities. The state probabilities at $t=2$ are found from

$$\Pi_1=\begin{bmatrix}0.3 & 0.4\\ 0.7 & 0.6\end{bmatrix}\begin{bmatrix}1\\ 0\end{bmatrix}=\begin{bmatrix}0.3\\ 0.7\end{bmatrix},\quad \Pi_2=\begin{bmatrix}0.3 & 0.4\\ 0.7 & 0.6\end{bmatrix}\begin{bmatrix}0.3\\ 0.7\end{bmatrix}=\begin{bmatrix}0.37\\ 0.63\end{bmatrix}$$



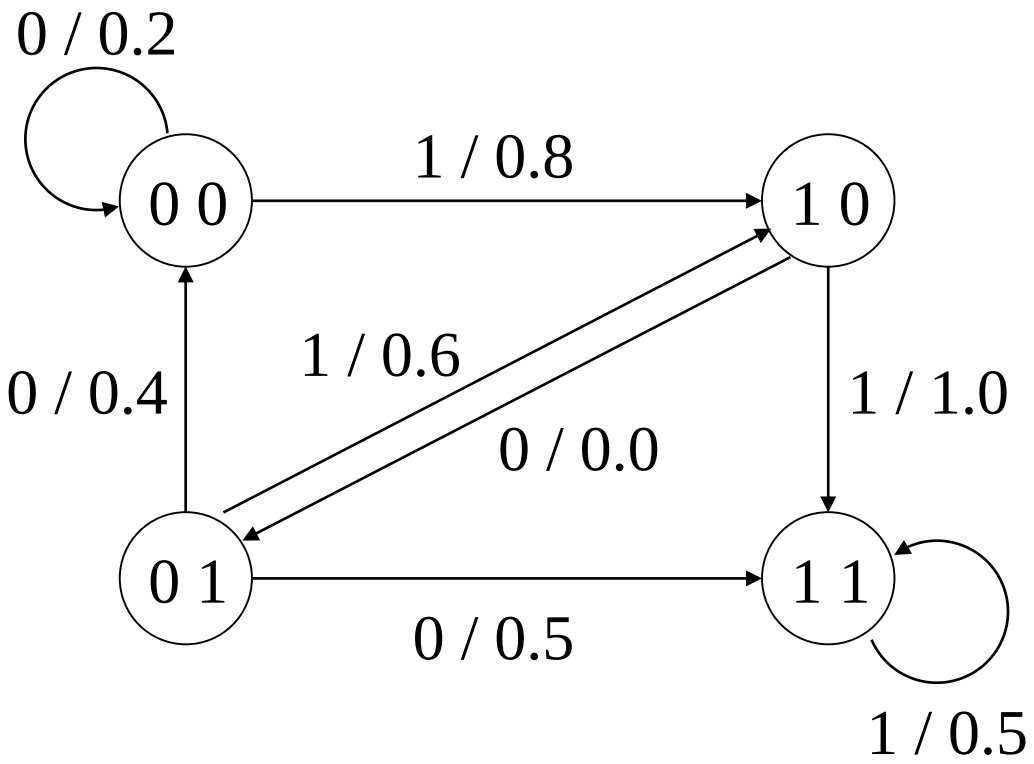
Example 4.2

Let A be a second-order binary Markov process with transition probabilities

$$\Pr(a=0\,|\,0,0)=0.2,\quad \Pr(a=1\,|\,0,0)=0.8,$$
$$\Pr(a=0\,|\,0,1)=0.4,\quad \Pr(a=1\,|\,0,1)=0.6,$$
$$\Pr(a=0\,|\,1,0)=0.0,\quad \Pr(a=1\,|\,1,0)=1.0,$$
$$\Pr(a=0\,|\,1,1)=0.5,\quad \Pr(a=1\,|\,1,1)=0.5,$$

and assume all states are equally probable at time $t=0$. What are the state probabilities at $t=1$?

Solution： Define the states as $S_0 = (0,0), S_1 = (0,1)$, etc. The possible state transitions and their associated transition probabilities can be represented using a state diagram. For this problem, the state diagram is



and the next-state probability equation is

$$\Pi_{t+1} = \begin{bmatrix} 0.2 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0.5 \\ 0.8 & 0.6 & 0 & 0 \\ 0 & 0 & 1.0 & 0.5 \end{bmatrix} \Pi_t$$

With all states equally probable at $t = 0$, we have

$\Pi_0^T = [0.25 \quad 0.25 \quad 0.25 \quad 0.25]$, so $\Pi_1^T = [.15 \quad .125 \quad .35 \quad .375]$.

Did you notice that every column of $P_{A|\Pi}$ adds to 1.0 in each of the previous two examples ? This is not accidental. Rather, it is a consequence of the requirement that, at every time *t*, the probabilities of all the states must sum to unity. Every properly constructed transition probability matrix has this property. Notice, however, that the rows of the matrix do not have to sum to unity. Can you explain why ?

4.2 Steady-State Probability and the Entropy Rate

The equation for the state probabilities is a homogeneous difference equation. It can be simply shown by induction that the state probabilities at time t are given by

$$\Pi_t = (P_{A|\Pi})^t \Pi_0$$

A Markov process is said to be ergodic if we can get from any initial state to any other state in some number of steps and if, for large t, Π_t approaches a steady-state value that is independent of the initial probability distribution Π_0. The steady-state value is reached when $\Pi_{t+1} = \Pi_t$. The Markov processes which model information sources are always ergodic.

Example 4.3 Find the steady-state probability distribution for the source in Example 4.2

Solution： In the steady state, the state probabilities become

$$\pi_0 = 0.2\pi_0 + 0.4\pi_1$$

$$\pi_1 = 0.5\pi_3$$

$$\pi_2 = 0.8\pi_0 + 0.6\pi_1$$

$$\pi_3 = \pi_2 + 0.5\pi_3$$

It appears from this that we have four equations and four unknowns, so solving for the four probabilities is no problem. However, if we look closely, we will see that only three of the equations above are linearly independent. To solve for the probabilities, we can use any three of the above equations and the constraint equation $\pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$. (This equation is a consequence of the fact that the total probability must sum to unity; it is certain the system is in some state!). Dropping the first equation above and using the constraint, we have

$$\pi_1 - 0.5\pi_3 = 0$$

$$0.8\pi_0 + 0.6\pi_1 - \pi_2 = 0$$

$$\pi_2 - 0.5\pi_3 = 0$$

$$\pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$$

which has solution $\pi_0 = 1/9, \pi_1 = \pi_2 = 2/9, \pi_3 = 4/9$

This solution is independent of the initial probability distribution

---

This situation illustrated in the previous example, where only $N-1$ of the equations resulting from the transition probability expression are linearly independent and we must use the "sum to unity" equation to obtain the solution, always occurs in the steady-state probability solution of an ergodic Markov process.

Now let's consider the entropy rate of an ergodic Markov process. From Equation (3.4), we have

$$R = \lim_{t \to \infty} \frac{1}{t} H(A_0, A_1, \cdots, A_{t-1})$$

However, as *t* grows very large, the state probabilities converge to a steady-state value, π_n, for each of the *N* possible states. Since *R* is the average information per symbol in this block of symbols, as *t* becomes large this average will be determined by the probability of occurrence of the symbols in *A* after the state probabilities have converges to their steady-state values.

Suppose we are in state S_n at time t. The conditional entropy of the next symbol a is given by

$$H(A \mid S_n) = \sum_{a \in A} \Pr(a \mid S_n) \log_2 (1/\Pr(a \mid S_n))$$

Since each possible symbol a leads to a unique next state, S_n can lead to $|A|$ possible next states. The remaining $N - |A|$ states cannot be reached from S_n, and for these states the transition probability $p_{i|n} = 0$ in Equation (4.3). Therefore, the conditional entropy expression above can be expressed in terms of the transition probabilities as

$$H(A \mid S_n) = \sum_{i=0}^{N-1} p_{i|n} \log_2 (1/p_{i|n})$$

For large t, the probability of being in state S_n is given by its steady-state probability π_n. Therefore, the entropy rate of the system becomes

$$R = \sum_{n=0}^{N-1} \pi_n H(A|S_n)$$

This expression, in turn, is equivalent to

$$R = \sum_{n=0}^{N-1} p_n \sum_{i=0}^{N-1} p_{i|n} \log_2(1/p_{i|n})$$

where the $p_{i|n}$ the are the entries in the transition probability matrix and the p_n are the steady-state probabilities. The entropy rate of an ergodic Markov process is a function only of its steady state probability distribution and the transition probabilities.

Example 4.4

Find the entropy rate for the source in Example 4.2. Calculate the steady state probability of the source emitting a “0” and the steady-state probability of the source emitting a “1”. Calculate the entropy of a memoryless source having these symbol probabilities and compare the result with the entropy rate of the Markov source.

Solution： We have the steady-state probabilities from Example 4.3. The entropy rate is

$$R = \frac{1}{9}(.2\log_2(5) + .8\log_2(1.25)) + \frac{2}{9}(.4\log_2(2.5) + .6\log_2(1.667))$$

$$+ \frac{2}{9}\log_2(1) + \frac{4}{9}(2 \cdot 0.5\log_2(2)) = .740$$

The steady-state symbol probabilities are

$$\Pr(0) = \sum_{n=0}^{3} \pi_n \Pr(0 \mid S_n) = \frac{.2}{9} + \frac{0.4(2)}{9} + \frac{0.5(4)}{9} = \frac{1}{3}$$

$$\Pr(1) = 1 - \Pr(0) = \frac{2}{3}$$

The entropy of a memoryless source having this symbol distribution is

$$H(A) = \sum_{a=0}^{1} \Pr(a) \log_2 (1/\Pr(a)) = 0.9183$$

From this we have $R < H(A)$. This inequality was predicted by Equation (3.4).

In Section III, we discussed how introducing redundancy into a block of symbols might be used to reduce the entropy rate to a level below the channel capacity and how this technique might be used for the correction of errors in the information bits to achieve an arbitrarily small information bit error rate. In this section, we have seen that a Markov process also introduces redundancy into the symbol block. Can this redundancy be introduced in such a way as to be useful for error correction ? The answer to this question is "yes". This is the principle underlying a class of error correcting codes known as convolutional codes.

V. Markov Chains and Data Processing

In the first two sections of this chapter, we examined the process of transmitting information X through the channel to produce channel output Y and discovered that a noisy channel entails some information loss if the entropy rate exceeds the channel capacity. It is only natural to wonder if there might exist some (possibly complicated) form of data processing which can be performed on Y to recover the lost information. Unfortunately, as we will now show, the answer to this question is "no". Once the information has been lost, it's gone.

To show this, suppose we operate on the received signal y by some function f to produce a result $z = f(y)$. We may use any conceivable function. Let the set of possible values of z which can be produced from the domain Y be denoted as Z. let the joint probability of the ordered triplet $\langle x, y, z \rangle$ be $p_{x,y,z}$. Then

$$p_{x,y,z} = p_{x,y} p_{z|x,y}$$

Since z is a function of y, the probability of z given x and y is just $p_{z|x,y} = p_{z|y}$, and therefore,

$$p_{x,y,z} = p_{x,y} p_{z|y} \qquad (5.1)$$

Three variables that satisfy Equation (5.1) are said to form a Markov chain. We now ask if the mutual information $I(X;(Y,Z)) \equiv I(X;Y,Z)$ is larger than $I(X;Y)$.

Let us define the conditional mutual information

$$I(X;Y \mid Z) \equiv H(X \mid Z) - H(X \mid Y,Z) \qquad (5.2)$$

The conditional mutual information is the reduction in our uncertainty of X due to our knowledge of Y when Z is given to us. The reason for defining this concept now will be clear in a moment.

From our earlier definition of mutual information

$$I(X;Y,Z) = H(X) - H(X \mid Y,Z)$$

and substituting for $H(X \mid Y,Z)$ in terms of the conditional mutual information in Equation (5.2) we have

$$I(X;Y,Z) = H(X) - H(X \mid Z) + I(X;Y \mid Z)$$

Since

$$H(X)-H(X\,|\,Z)=I(X;Z)$$

we have

$$I(X;Y,Z)=I(X;Z)+I(X;Y\,|\,Z) \qquad (5.3)$$

Now, since the joint probability $p_{y,z}=p_{z,y}$, we can also write

$$I(X;Y,Z)=I(X;Z,Y)=I(X;Z)+I(X;Z\,|\,Y) \qquad (5.4)$$

But, from Definition (5.2)

$$I(X;Z\,|\,Y)=H(X\,|\,Y)-H(X\,|\,Y,Z)$$

Since $z=f(y)$, if we are given y, z is completely determined and $H(X\,|\,Y,Z)=H(X\,|\,Y)$, which means $I(X;Z\,|\,Y)=0$. This is a consequence of the fact that *x*, *y*, and *z* form a Markov chain.

On the other hand, $I(X;Y|Z)$ will not be zero unless the function $f(y)$ is one-to-one and onto. Since we are considering any possible function, we must allow any conceivable $f(y)$ so, combining Equation (5.3) and (5.4) gives us

$$I(X;Z)+I(X;Y|Z)=I(X;Y) \qquad (5.5)$$

Since $I(X;Y|Z)\geq 0$,

$$I(X;Y)\geq I(X;Z) \qquad (5.6)$$

This is known as the data-processing inequality. It states that additional processing of the channel output *y* can at best result in no further loss of information and may even result in additional information loss.

The latter can result if, for example, $f(y)$ is a many-to-one mapping from Y to Z such as $z = y^2$ or $z = |y|$. A very common example of this kind of information loss occurs when a real-valued (that is, an analog) channel output is quantized by an analog to digital converter (ADC). If the ADC has too few quantization levels, the resulting information loss can be quite severe. (In fact, this information loss is often called "quantizing noise").

Another source of information loss can be roundoff or truncation error during digital signal processing in a computer or microprocessor. Since digital processing of signals is cost effective and very commonplace, designers of these systems need to have an awareness of the possible impact of such design decisions as the word length of the digital signal processor or the number of bits of quantization in analog to digital converters have on the information content.