**Entropy of Markov Information Sources**
**and Capacity of Discrete Input Constrained Channels**
(from Immink, Coding Techniques for Digital Recorders)

## 1. Entropy of Markov Chains

We have already introduced the notion of entropy in a conceptually simple situation: it was assumed that the symbols are independent and occur with fixed probabilities. That is, the occurrence of a specific symbol at a certain instant does not alter the probability of occurrences of symbols during any other symbol intervals. We need to extend the concept of entropy for more complicated structures where symbols are not chosen independently but their probabilities of occurring depend on preceding symbols. It is to be emphasized that nearly all practical sources emit sequences of symbols that are statistically dependent. Sequences formed by the English language are an excellent example. Occurrence of the letter Q implies that the letter to follow is probably a U. Regardless of the form of the statistical dependence, or structure, among the successive source outputs, the effect is that the amount of information coming out of such a source is smaller than from a source emitting the same set of characters in independent sequences. The development of a model for sources with memory is the focus of the ensuing discussion.

In probability theory the notation $P_r(A \mid B)$ means the probability of occurrence of event *A* given that ($\mid$) event *B* has occurred. Many of the structures that will be encountered in the subsequent chapters can usefully be modeled in terms of a Markov chain. A Markov chain is a special type of stochastic process distinguished by a certain Markov property A (discrete) Markov chain is defined as a discrete random process of the form

$$\{\ldots, Z_{-2}, Z_{-1}, Z_0, Z_1, \ldots\},$$

where the variables $Z_t$ are dependent discrete random variables taking values in the state alphabet $S= \{s_1,\ldots, s_N\}$, and the dependence satisfies the Markov condition

$$Pr(Z_t = \sigma_{i_t} \mid Z_{t-1} = \sigma_{i_{t-1}}, Z_{t-2} = \sigma_{i_{t-2}}, \ldots) = Pr(Z_t = \sigma_{i_t} \mid Z_{t-1} = \sigma_{i_{t-1}}).$$

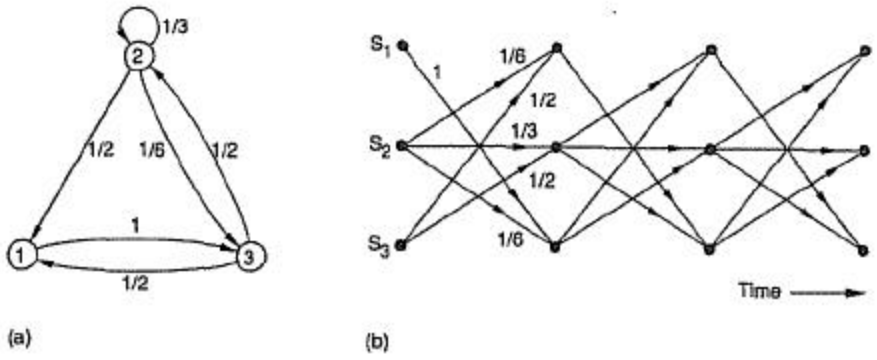
In words, the variable $Z_t$ is independent of past samples $Z_{t-2}, Z_{t-3}$ ... if the value of $Z_{t-1}$ is known. A (homogeneous) Markov chain can be described by a *transition probability matrix* Q with elements

$$q_{ij} = Pr(Z_t = \sigma_j \mid Z_{t-1} = \sigma_i), \quad 1 \le i,j \le N.$$

The transition probability matrix Q is a stochastic matrix, that is, its entries are non-negative, and the entries of each row sum to one. Any stochastic matrix constitutes a valid transition probability matrix.

Imagine the process starts at time $t = 1$ by choosing an initial state in accordance with a specified probability distribution. If we are in state $s_i$ at time $t = 1$, then the

process moves at $t = 2$ to a possibly new state, the quantity $q_{ij}$ is the probability that the process will move to state $s_j$ at time t = 2. If we are in state $s_j$ at instant t = 2, we move to $s_k$ at instant $t = 3$ with probability $q_{jk}$. This procedure is repeated ad infinitum.
The state-transition diagram of a Markov chain, portrayed in the following figure (a) represents a Markov chain as a directed graph where the states are embodied by the nodes or vertices of the graph; the transition between states is represented by a directed line, an edge, from the initial to the final state, The transition probabilities $q_{ij}$ corresponding to various transitions are shown marked along the lines of the graph. Another useful representation of a Markov chain is provided by a trellis (or lattice) diagram (see (b)).




Alternative representations of a three-state Markov chain. (a) state-transition diagram, (b) trellis diagram for the same chain. Permissible transitions from one state to the other are depicted by lines.

This is a state diagram augmented by a time axis so that it provides for easy visualization of how the states change with time.

In theory, there are many types of Markov chains; here we restrict our attention to chains that are *ergodic* and *regular*. Roughly speaking, ergodicity means that from any state the chain can eventually reach any other state; regularity means that the Markov chain is non- periodic. In the practical structures that will be encountered, all these conditions hold.

We shall now take a closer look at the dynamics of a Markov chain, To that end, let $w_j^{(t)} \equiv \Pr(Z_t = s_j)$ represent the probability of being in state $s_j$ at time $t$. Clearly,

$$\sum_{j=1}^{N} w_j^{(t)} = 1$$

The probability of being in state $s_j$ at time $t$ may be expressed in the state probabilities at instant t-1:

$$w_j^{(t)} = w_1^{(t-1)} q_{1j} + w_2^{(t-1)} q_{2j} + \cdots + w_N^{(t-1)} q_{Nj}.$$

The previous equation suggests the use of matrices. If we introduce the state distribution vector

$$\mathbf{w}^{(t)} = (w_1^{(t)}, \ldots, w_N^{(t)}),$$

then the previous equation can succinctly be expressed in an elegant matrix/vector notation, thus

$$\mathbf{w}^{(t)} = \mathbf{w}^{(t-1)}Q.$$

By iteration we obtain

$$\mathbf{w}^{(t)} = \mathbf{w}^{(1)}Q^{t-1}.$$

In other words, the state distribution vector at time t is the product of the state distribution vector at time $t = 1$, and the $(t-1)$th power of the transition matrix. It is easy to see that $Q^{t-1}$ is also a stochastic matrix. The previous formula (3.10) is equivalent to the assertion that the n-step transition matrix is the nth power of the single step transition matrix $Q$. We note also that $Q^0 = I$ is the ordinary identity matrix.

We shall concentrate now on the limiting behavior of the state distribution vector as $t \to \infty$. In many cases of practical interest there is only one such limiting distribution vector, denoted by $p = (p_1, \ldots, p_N)$. In the long run the state distribution vector converges to this *equilibrium distribution vector* from any valid initial state probability vector $w^{(1)}$ so

$$\pi = \lim_{t \to \infty} \mathbf{w}^{(1)}Q^{t-1}.$$

The number $p_1$, is called the steady, or stationary state probability of state $s_i$. The equilibrium distribution vector can be obtained by solving the system of linear equations in the N unknowns $p_1, \ldots, p_N$:

$$pQ = p$$

Only N-1 of these N equations are independent, so we solve the top N-1 along with the normalizing condition

$$\sum_{i=1}^{N} p_i = 1$$

The proof is elementary: we note that if $pQ = p$, then

$$\pi Q^t = \pi Q Q^{t-1} = \pi Q^{t-1} = \cdots = \pi.$$

Decomposition of the initial state vector $w^{(1)}$ in terms of the eigenvectors of $Q$ can be convenient to demonstrate the process of convergence. The matrix $Q$ has $N$ eigenvalues $\{l_1, \ldots, l_N\}$, that can be found by solving the characteristic equation

$$\det[Q - \lambda I] = 0,$$

where $I$ is the identity matrix, and $N$ (left) eigenvectors $\{u_1, \ldots, u_N\}$ each of which is a solution of the system

$$\mathbf{u}_i Q = \lambda_i \mathbf{u}_i, \quad i = 1, \ldots, N.$$

Provided that $l_i$, $i = 1,., N$, are distinct, there are $N$ independent eigenvectors, and the eigenvectors $l_i$, $i = 1, \ldots, N$, constitute a basis. The initial state vector may be written as

$$\mathbf{w}^{(1)} = \sum_{i=1}^{N} a_i \mathbf{u}_i.$$

We find the state distribution vector $w^{(t)}$ at instant $t$:

$$\mathbf{w}^{(t)} = \mathbf{w}^{(1)} Q^{(t-1)} = \sum_{i=1}^{N} a_i \lambda_i^{(t-1)} \mathbf{u}_i.$$

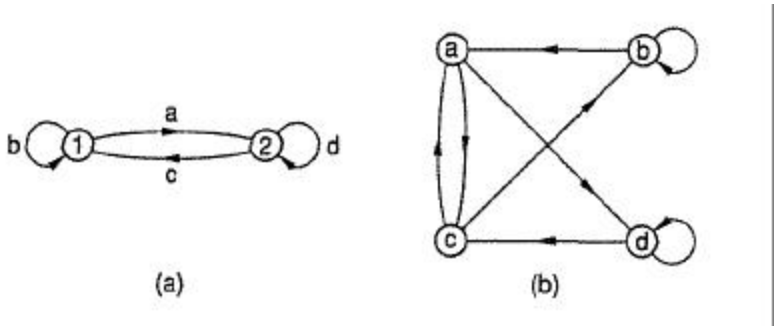
If it is assumed that the eigenvalues are distinct, the {?$_i$} can be ordered, such that $|l_1| > |l_2| > |l_3|$, etc. Combination of previous formulae reveals that p is an eigenvector with unity eigenvalue, thus $l_1 = 1$, We then have

$$\mathbf{w}^{(t)} = \pi + \sum_{i=2}^{N} a_i \lambda_i^{(t-1)} \mathbf{u}_i$$

and convergence to p is assured since $|l_1| < 1, i \neq 1$.

## 2. Entropy of Markov Information Sources

We are now in the position to describe a *Markov information source*. Given a finite Markov chain {$Z_t$} and a function ? whose domain is the set of states of the chain and whose range is a finite set G, the *source alphabet*, then the sequence {$X_t$} where $X_t = z(Z_t)$, is said to be the output of a Markov information source corresponding to the chain { $Z_t$ } and the function ? . In general, the number of states can be larger than the cardinality of the source alphabet, which means that one output symbol may correspond to more than one state. The essential feature of the Markov information source is that it provides for dependence between successive symbols, which introduces redundancy in the message sequence. Each symbol conveys less information than it is capable of conveying since it is to some extent predictable from the preceding symbol. In the foregoing description of an information source we assumed that the symbol emitted is solely a function of the state that is entered. This type of description is usually called a *Moore*-type Markov source, In a different description, called the *Mealy*-type Markov source, the symbols emitted are a function of the Markov chain $X_t = \hat{z}(Z_t, Z_{t+1})$. In other words, a Mealy-type Markov source is obtained by labelling the edges of the directed graph that represents the Markov chain. Mealy- and Moore-type descriptions are equivalent. Let a Mealy-type machine be given. By defining a Markov information source with state set composed of triples $\{s_i, s_j \hat{z}(s_i, s_j)\}$ and label $\hat{z}(s_i s_j)$ on the state $\{s_i, s_j \hat{z}(s_i, s_j)\}$, we obtain a Moore-type Markov source. The Moore-type model is referred to as the *edge graph* of the Mealy-type model. An example of a Mealy-type information source and its Moore-type equivalent are shown in the following figure.




(a) Example of a Mealy-type two-state Markov information source, and (b) its four-state Moore-type counterpart

The idea of a Markov source has enabled us to represent certain types of structure in streams of data. We next examine the information content, or entropy, of a sequence emitted by a Markov source, The entropy of a Markov information source is hard to compute in most cases. For a certain class of Markov information sources, termed *unifilar* Markov information source, the computation may be greatly simplified. The word unifilar refers to the following property.

Let a Markov information source with a set of states $\Sigma = \{s_1, \ldots s_N\}$, output alphabet G, and associated output function $z(Z_t)$ be given. For each state $s_k \in \Sigma$, let $s_{k1}, s_{k2}, \ldots, s_{kn_k}$ be the states that can be reached in one step from $s_k$, that is, the states $s_j$ such that $q_{kj} > 0$. We say $s_j$ is a *successor* of $s_k$, if $q_{kj} > 0$. The source is said to be unifilar if for each state $s_k$ the symbols $z(s_{k1}), \ldots, z(s_{kn_k})$ are distinct. In other words, each successor of k must be associated with a distinct symbol. Provided this condition is met and the initial state of the Markov information source is known, the sequence of emitted symbols determines the sequence of states followed by the chain, and a simple formula is available for the entropy of the emitted X-process. Given a unifilar Markov source, as above, let $s_{k1}, \ldots, s_{kn_k}$ be the successors of $s_k$, then it is quite natural to define the uncertainty of state $s_k$ as $H_k = H(p_{k1}, \ldots, p_{kn_k})$, with $H(p_{k1}, \ldots, p_{kn_k})$ defined as

$$H(p_1, \ldots, p_M) = -\sum_{i=1}^{M} p_i \log_2 p_i$$

Shannon defined the entropy of the unifilar Markov source as the average of these ilk weighed in accordance with the steady-state probability of being in a state in question, that is, by the expression

$$H\{X\} = \sum_{k=1}^{N} \pi_k H_k.$$

Note that we use the notation $H\{X\}$ to express the fact that we are considering the entropy of sequences $\{X\}$ and not the function $H(.)$. The next numerical example may serve to illustrate the theory.

*Example:* Consider the three-state unifilar Markov chain depicted in the figure above. From the diagram we may read the transition probability matrix

$$Q = \begin{bmatrix} 0 & 0 & 1 \\ 1/2 & 1/3 & 1/6 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

What is the average probability of being in one of the three states? We find the following system of linear equations that govern the steady-state probabilities:

$$\begin{aligned} \frac{1}{2}\pi_2 + \frac{1}{2}\pi_3 &= \pi_1 \\ \frac{1}{3}\pi_2 + \frac{1}{2}\pi_3 &= \pi_2 \\ \pi_1 + \frac{1}{6}\pi_2 &= \pi_3, \end{aligned}$$

from which we obtain $p_3 = \frac{4}{3}p_2$ and $p_3 = \frac{7}{6}p_2$. Since $p_1 + p_2 + p_3 = 1$ we have

$$\pi_1 = \frac{1}{3}, \quad \pi_2 = \frac{2}{7}, \quad \pi_3 = \frac{8}{21}.$$

The entropy of the information source is found to be equal

$$H\{X\} = \frac{1}{3}H(1) + \frac{2}{7}H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) + \frac{8}{21}H(\frac{1}{2}, \frac{1}{2}) \simeq 0.798.$$

In the next section we consider a problem which is central to the field of input-constrained channels. We focus on methods to compute the maximum amount of information that can be sent over an input-constrained channel per unit of time.

## 3. Capacity of Discrete Noiseless Channels

Shannon defined the capacity C of a discrete noiseless channel by

$$C = \lim_{T \to \infty} \frac{\log_2 N(T)}{T},$$

where *N(T)* is the number of admissible signals of duration *T*. The problem of calculation of the capacity for constrained channels is in essence a combinatorial problem, that is, finding the number of allowed sequences *N(T)*. This fundamental definition will be worked out in a moment for some specific channel models. We start, since virtually all channel constraints can be modeled as such, with the computation of the capacity of Markov information sources.

### 3.1. Capacity of Markov information sources

In the previous sections we developed a measure of information content of an information source that can be represented by a finite Markov model. As discussed, the measure of information content, entropy, can be expressed in terms of the limiting state-transition probabilities and the conditional entropy of the states. In this section we address a problem that provides the key to answer many questions that will emerge in the chapters to follow. Given a unifilar *N*-state Markov source with states $\{s_1, \ldots s_N\}$ and

transition probabilities $\hat{q}_{ij}$, we define the *connection matrix* $D = \{d_{ij}\}$ of the source as follows. Let

$$d_{ij} = 1 \text{ if } \hat{q}_{ij} > 0$$
$$d_{ij} = 0 \text{ if } \hat{q}_{ij} = 0, i, j = 1, \ldots, N.$$

The actual values of the transition probabilities are irrelevant; the connection (or adjacency) matrix contains binary-valued elements, and it is formed by replacing the positive elements of the transition matrix by l's. For an $N$-state source, the connection matrix D is defined by $d_{ij}$ = 1if a transition from state $i$ to state $j$ is allowable and $d_{ij} = 0$ otherwise. For a given connection matrix, we wish to choose the transition probabilities in such a way that the entropy

$$H\{X\} = \sum_{k=1}^{N} p_k H_k$$

is maximized. Such a source is called *maxentropic*, and sequences generated by a maxentropic unifilar source are called *maxentropic sequences*.

The maximum entropy of a unifilar Markov information source, given its connection matrix, is given by

$$C = \max H\{X\} = \log_2 l_{\max},$$

where ?$_{max}$ is the largest eigenvalue of the connection matrix D. The existence of a positive eigenvalue and corresponding eigenvector with positive elements is guaranteed by the Perron-Frobenius theorems. Essentially, there are two approaches to prove the preceding equation. One approach, provided by Shannon, is a straightforward routine, using Lagrange multipliers, of finding the extreme value of a function of several independent variables. The second proof of the above formula to be followed here, is established by enumerating the number of distinct sequences that a Markov source can generate.

The number of distinct sequences of length $m + 1$, $m > 0$, emanating from state $s_i$ denoted by $N_i(m + 1)$, equals the total of the numbers of sequences of unity length that emerge from $s_i$, and terminate in $s_j$ multiplied (since the source is unifilar) by the number of sequences of length $m$ that start in $s_j$. Thus we find the following recurrence equation

$$N_i(m+1) = \sum_{j=1}^{N} d_{ij} N_j(m), \quad i = 1, \ldots, N.$$

This is a system of $N$ linear homogeneous difference equations with constant coefficients, and therefore the solution is a linear combination of exponentials lambda$^m$. To find the particular {?$_i$}, we assume a solution of the form $N_i(m) = y_i * lambda^m$ to obtain

$$\lambda^m(\lambda y_i) = \lambda^m \sum_{j=1}^{N} d_{ij} y_j, \quad i = 1, \ldots, N,$$

or, letting $y^T = (y_1, \ldots, y_N)$, where the superscript $T$ stands for transposition, we have
lambda*y =Dy

Thus the allowable {lambda$_i$} are the eigenvalues of the matrix *D*. For large sequence length in we may approximate $N_i\ (m)$ by

$$N_i(m) \simeq a_i \lambda_{\max}^m,$$

where $a_i$ is a constant independent of *m* and lambda$_{max}$ is the largest real eigenvalue of the matrix D, or in other words, lambda$_{max}$ is the largest real root of the determinant equation

$$\det[D - zI] = 0.$$

Previoous equation states that for large enough m the number of distinct sequences grows exponentially with the sequence length *m*; the growth factor is lambda$_{max}$. This is not to say that $N_i\ (m)$ is accurately determined by the exponential term when *m* is small. We have

$$\frac{1}{m} \log_2 N_i(m) \simeq \frac{1}{m} (\log_2 a_i + m \log_2 \lambda_{\max}).$$

The maximum entropy of the noiseless channel may be evaluated by invoking the definition of the capacity, or

$$C = \lim_{m \to \infty} \frac{1}{m} \log_2 N_i(m) = \log_2 \lambda_{\max}.$$

The transition probabilities $q_{ij}$ associated with the maximum entropy of the source can be found with the following reasoning. Let $p = (p_1, \ldots P_N)^T$ denote the eigenvector associated with the eigenvalue lambda$_{max}$ or

$$D\mathbf{p} = \lambda_{\max}\mathbf{p}.$$

The state-transition probabilities that maximize the entropy are

$$q_{ij} = \lambda_{\max}^{-1} d_{ij} \frac{p_j}{p_i}.$$

To prove the above formula is a matter of substitution. According to the Perron-Frobenius theorems [5], the components of the eigenvector *p* are positive, and thus $q_{ij} = 0,\ 1 = i,j = N$. Since $p = (p_1, .., p_N)^T$ is an eigenvector for ?$_{max}$ we conclude

$$\sum_{j=1}^{N} q_{ij} = 1,$$

and hence the matrix Q is indeed stochastic. The entropy of a Markov information source is, according to definition

$$H\{X\} = \sum_{k=1}^{N} \pi_k H_k,$$

where $H_k$ is the uncertainty of state $s_k$ and ($p_1$, ... ,$p_N$) is the steady-state distribution. Thus,

$$H\{X\} = -\sum_{i,j=1}^{N} \pi_i q_{ij} \log_2 q_{ij} = \sum_{i,j=1}^{N} \pi_i q_{ij} (\log_2 \lambda_{\max} + \log_2 p_i - \log_2 p_j).$$

Since

$$\sum_{i,j} \pi_i q_{ij} \log_2 p_i = \sum_i \pi_i \log_2 p_i$$

and

$$\sum_{i,j} \pi_i q_{ij} \log_2 p_j = \sum_j \log_2 p_j \sum_i \pi_i q_{ij} = \sum_j \pi_j \log_2 p_j,$$

we obtain

$$H\{X\} = -\sum_{i,j=1}^{N} \pi_i q_{ij} \log_2 q_{ij} = \sum_{i,j} \pi_i q_{ij} \log_2 \lambda_{\max} = \log_2 \lambda_{\max}.$$

This demonstrates that the transition probabilities given by

$$q_{ij} = \lambda_{\max}^{-1} d_{ij} \frac{p_j}{p_i}.$$

are indeed maximizing the entropy.

*Example (Continued):* We revert to the three-state unifilar Markov chain with transition probability matrix

$$Q = \begin{bmatrix} 0 & 0 & 1 \\ 1/2 & 1/3 & 1/6 \\ 1/2 & 1/2 & 0 \end{bmatrix}.$$

The adjacency matrix D is

$$D = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

The characteristic equation is

$$\begin{aligned} \det[D - zI] &= -z(z^2 - z - 2) \\ &= -z(z - 2)(z + 1) = 0, \end{aligned}$$

from which we conclude that the largest root $?_{max} = 2$, and the capacity is $C = \log_2 ?_{max} = 1$. The eigenvector associated with the largest eigenvalue is $p = (1, 3, 2)^T$. The transition probabilities that maximize the entropy of the Markov information source are found with

$$Q = \begin{bmatrix} 0 & 0 & 1 \\ 1/6 & 1/2 & 1/3 \\ 1/4 & 3/4 & 0 \end{bmatrix}.$$