Conditional Distributions

A conditional distribution is the ratio of a joint distribution and a marginal distribution. When the value of random variable X is conditioned on the value of random variable Y:

$$p_{X|Y}(x \mid y) = \frac{p_{XY}(x,y)}{p_Y(y)}.$$

This can be generalized so that the values of N random variables $X_1 \ldots X_N$ are conditioned on the values of M random variables $Y_1 \ldots Y_N$:

$$p_{X_1 \ldots X_N|Y_1 \ldots Y_M}(x_1 \ldots x_N \mid y_1 \ldots y_M) = \frac{p_{X_1 \ldots X_N, Y_1 \ldots Y_M}(x_1 \ldots x_N, y_1 \ldots y_M)}{p_{Y_1 \ldots Y_M}(y_1 \ldots y_M)}.$$

Higher Order Markov Processes

Let S be a set of states:

$$S = \{1, 2, 3 \ldots N\}$$

and let $i, j, k \ldots \in S$. A random process is an order one Markov process iff:

$$p_{t\,|\,t-1,t-2\ldots-\infty}(i\,|\,j,k\ldots) = p_{t\,|\,t-1}(i\,|\,j).$$

The probability that the Markov process is in state i at time t is given by the following update formula:

$$p_t(i) = \sum_{j=1}^{N} p_{t\,|\,t-1}(i\,|\,j)\,p_{t-1}(j).$$

Higher Order Markov Processes (contd.)

Let S be a set of states:

$$S = \{1, 2, 3 \ldots N\}$$

and let $i, j, k \ldots \in S$. A random process is an order two Markov process iff:

$$p_{t\,|\,t-1,t-2\ldots-\infty}(i\,|\,j,k\ldots) = p_{t\,|\,t-1,t-2}(i\,|\,j,k).$$

The probability that the Markov process is in state i at time t is given by the following update formula:

$$p_t(i) =$$

$$\sum_{j=1}^{N} \sum_{k=1}^{N} p_{t\,|\,t-1,t-2}(i\,|\,j,k)\, p_{t-1,t-2}(j,k).$$

Higher-Order Markov Processes (contd.)

A Markov process of order two can be thought of as a mapping between two joint distributions. Both of these joint distributions give the probability that the process visits two states in two successive times:

$$p_{t,t-1}(i,j) = \sum_{k=1}^{N} p_{t\,|\,t-1,t-2}(i\,|\,j,k)\,p_{t-1,t-2}(j,k).$$

The state i at time t is a marginal distribution (produced by summing over all possible states j at time $t-1$):

$$p_t(i) = \sum_{j=1}^{N} p_{t,t-1}(i,j).$$

Higher Order Markov Processes (contd.)

It follows that a Markov process of order two, with states, S:

$$S = \{1, 2, 3 \ldots N\}.$$

can be reduced to a Markov process of order one, with states, $S' = S \times S$:

$$S' = \{\langle 1,1 \rangle, \langle 1,2 \rangle \ldots \langle N,N \rangle\}$$

and transition probability matrix:

$$p'_{t\,|\,t-1}(\langle i,j \rangle \mid \langle j,k \rangle) = p_{t\,|\,t-1,t-2}(i \,|\, j,k)$$

so that:

$$p'_t(\langle i,j \rangle) = \sum_{k=1}^{N} p'_{t\,|\,t-1}(\langle i,j \rangle \mid \langle j,k \rangle) p'_{t-1}(\langle j,k \rangle)$$

and

$$p_t(i) = \sum_{j=1}^{N} p'_t(\langle i,j \rangle).$$

Information Source with Memory

An *information source with memory* generates messages using a source alphabet of length, M. If the source is modeled as a Markov process of order one, then the entropy of a message of length N is:

$$H_1 = H_0 + (N-1)H_{t|t-1}$$

where

$$H_0 = -\sum_{i=1}^{M} p_t(i) \log p_t(i)$$

is the entropy of the first symbol and

$$H_{t|t-1} = -\sum_{i=1}^{M}\sum_{j=1}^{M} p_{t,t-1}(i,j) \log p_{t\,|\,t-1}(i\,|\,j)$$

is the entropy of each of the remaining symbols.

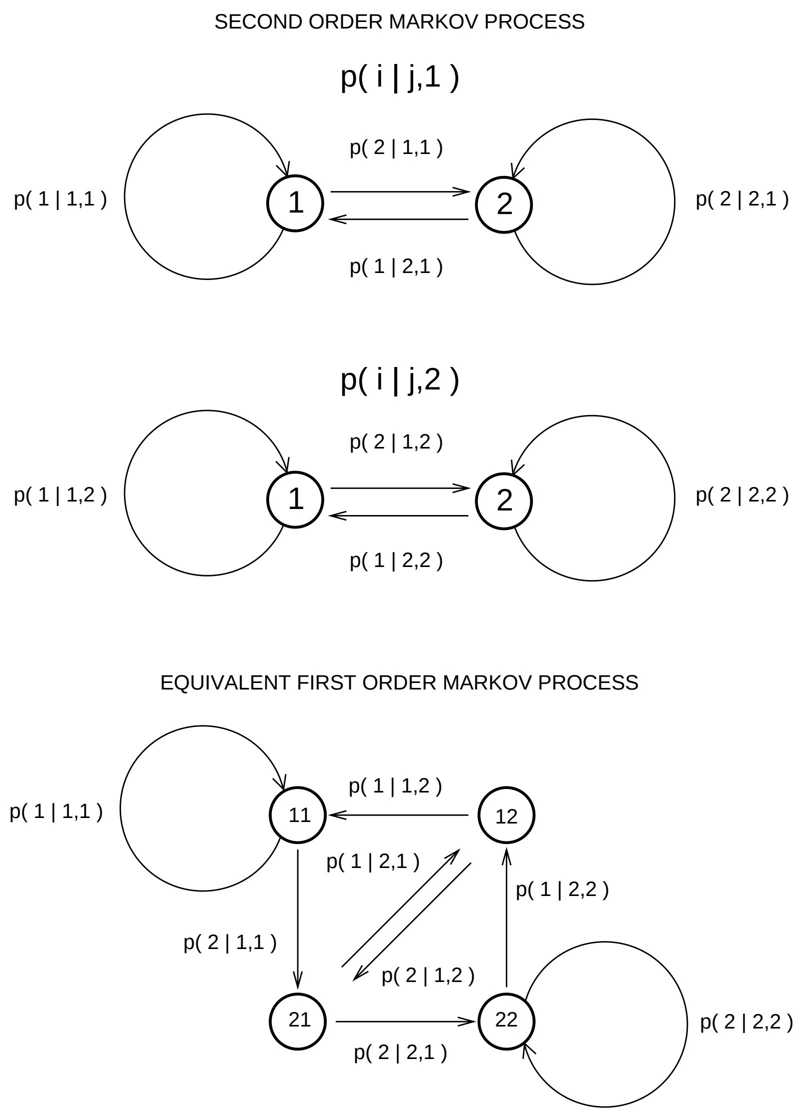

Figure 1: Second-order two-state Markov process and reduction to equivalent first-order Markov process.

Example One

On avg., how much information is provided by each character in a random string of zeros and ones? The distribution for the t-th character is:

$$p_t(0) = 0.5$$
$$p_t(1) = 0.5$$

$H_t = -0.5\log(0.5) - 0.5\log(0.5) = 1$ bit.

Each symbol delivers 1 bit of information on avg. in the memoryless case.

Example Two

Now let's consider a string where the first character is chosen at random, but the remaining characters follow a simple pattern:

$$0101...01 \;\text{ or }\; 1010...10$$

The distribution for the t-th character is:

$$\begin{aligned} p_t(0) &= 0.5 \\ p_t(1) &= 0.5 \end{aligned}$$

$H_t = -0.5\log(0.5) - 0.5\log(0.5) = 1$ bit.

The joint distribution is:

$$\begin{bmatrix} p_{t,t-1}(0,0) & p_{t,t-1}(0,1) \\ p_{t,t-1}(1,0) & p_{t,t-1}(1,1) \end{bmatrix} = \begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix}$$

Example Two (contd.)

The conditional distribution is:

$$p_{t\,|\,t-1}(i\,|\,j) = \frac{p_{t,t-1}(i,j)}{p_{t-1}(j)}$$

$$\begin{bmatrix} p_{t\,|\,t-1}(0|0) & p_{t\,|\,t-1}(0|1) \\ p_{t\,|\,t-1}(1|0) & p_{t\,|\,t-1}(1|1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

and the conditional entropy per character is:

$$\begin{aligned} H_{t\,|\,t-1} &= -\sum_{i=0}^{1}\sum_{j=0}^{1} p_{t,t-1}(i,j)\log p_{t\,|\,t-1}(i\,|\,j) \\ &= -0.5\log(1.0) + 0.5\log(1.0) \\ &= 0 \text{ bits.} \end{aligned}$$

This is less than in the memoryless case.

Information Source with Memory (contd.)

If the source is modeled as a Markov process of order two, then the entropy of a message of length N is:

$$H_2 = H_0 + H_{t|t-1} + (N-2)H_{t|t-1,t-2}$$

where H_0 and $H_{t|t-1}$ are the entropies of the first and second symbols and

$$H_{t|t-1,t-2} =$$

$$-\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} p_{t,t-1,t-2}(i,j,k)\log p_{t\,|\,t-1,t-2}(i\,|\,j,k)$$

is the entropy of each the remaining symbols.

Information Limit

Let H_0 be the entropy computed under the assumption that an information source is memoryless, and let H_1 be the entropy computed under the assumption that the source is a Markov process of order one, and H_2 be the entropy computed under the assumption that the source is a Markov process of order two, etc. Then

$$H_0 \geq H_1 \geq H_2 \geq \ldots \geq \lim_{k \to \infty} H_k.$$

Loss of Memory

Theorem The initial distribution and the limiting distribution of every irreducible, aperiodic Markov process have zero mutual information.

Proof Let I and L be discrete r.v.'s corresponding to the initial state and limiting state, then

$$p_{L|I}(i \mid j) = \lim_{n \to \infty} \left(\mathbf{P}^n\right)_{ij}$$

where $\mathbf{P}$ is the transition matrix. Because $\rho(\mathbf{P}) = 1$ for all stochastic matrices and the process is aperiodic and irreducible,

$$\lim_{n \to \infty} \mathbf{P}^n = \mathbf{x}_1 \mathbf{y}_1^{\mathrm{T}}$$

where $\mathbf{x}_1 = \mathbf{P}\mathbf{x}_1$ and $\mathbf{y}_1^{\mathrm{T}} = \mathbf{y}_1^{\mathrm{T}}\mathbf{P}$ by Perron's Theorem.

Loss of Memory (contd.)

Now, because $\mathbf{y}_1^{\mathrm{T}} = \left[\, 1 \;\; 1 \;\; 1 \;\; \ldots \;\; 1 \,\right]$ for all stochastic matrices:

$$\begin{aligned} p_{L|I}(i \mid j) &= \left(\mathbf{x}_1 \mathbf{y}_1^{\mathrm{T}}\right)_{ij} \\ &= \left(\left[\, \mathbf{x}_1 \,\middle|\, \mathbf{x}_1 \,\middle|\, \ldots \,\middle|\, \mathbf{x}_1 \,\right]\right)_{ij} \\ &= \left(\mathbf{x}_1\right)_i \\ &= p_L(i). \end{aligned}$$

It follows that L and I are statistically independent. Consequently I_{LI} equals zero.