THE ENTROPY OF A MARKOV INFORMATION SOURCE

S. Kullback

George Washington University

Prepared for:
Office of Naval Research

19 April 1974

DISTRIBUTED BY:

NTIS
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road, Springfield, Va. 22151
THE ENTROPY OF A MARKOV INFORMATION SOURCE

For a first-order and m-th order Markov source with a finite alphabet there are derived various relations and limiting behavior of expressions for the entropy.
<table>
<thead>
<tr>
<th>KEY WORDS</th>
<th>LINK A</th>
<th>LINK B</th>
<th>LINK C</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARKOV SOURCE</td>
<td>ROLE</td>
<td>ROLE</td>
<td>ROLE</td>
</tr>
<tr>
<td>ENTROPY</td>
<td>WT</td>
<td>WT</td>
<td>WT</td>
</tr>
<tr>
<td>STATIONARY</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DISTRIBUTION</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
THE ENTROPY OF A MARKOV INFORMATION SOURCE

by

S. KULLBACK

TECHNICAL REPORT NO. 18
APRIL 19, 1974

PREPARED UNDER CONTRACT N00014-67-A-0214-0015
(NR-042-267)
OFFICE OF NAVAL RESEARCH

Herbert Solomon, Project Director

Reproduction in Whole or in Part is Permitted for any Purpose of the United States Government

DEPARTMENT OF STATISTICS
THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, D.C. 20006
THE ENTROPY OF A MARKOV INFORMATION SOURCE

Let $S$ be a first-order Markov source with alphabet \( \{s_1, s_2, \ldots, s_q\} \), time-homogeneous transition probabilities \( P(s_i | s_j) \) and stationary distribution \( p_i = \text{Prob}(s = s_i) \), \( i = 1, 2, \ldots, q \). These define a simple stationary Markov chain with the matrix of transition probabilities

$$
(1) \quad P = \begin{pmatrix}
P(s_1 | s_1) & P(s_2 | s_1) & \cdots & P(s_q | s_1) \\
P(s_1 | s_2) & P(s_2 | s_2) & \cdots & P(s_q | s_2) \\
\vdots & \vdots & \ddots & \vdots \\
P(s_1 | s_q) & P(s_2 | s_q) & \cdots & P(s_q | s_q)
\end{pmatrix}
$$

where

$$
(2) \quad P(s_i | s_j) + P(s_2 | s_j) + \ldots + P(s_q | s_j) = 1, \quad j = 1, 2, \ldots, q
$$

and

$$
(3) \quad p_j = p_1 P(s_1 | s_j) + p_2 P(s_2 | s_j) + \ldots + p_q P(s_q | s_j), \quad j = 1, 2, \ldots, q
$$

If we set \( \mathbf{p}' = (p_1, p_2, \ldots, p_q) \) then in matrix notation the relations are expressed by

$$
(4) \quad \mathbf{p}' P = \mathbf{p}'.
$$
If the source is in state \( s_i \), then its transitions to the different states \( s_j, \quad j = 1, 2, \ldots, q \) form a finite scheme

\[
\begin{align*}
  & P(s_1 | s_i) \quad P(s_2 | s_i) \quad \ldots \quad P(s_q | s_i), \\
  & P(s_1 | s_i) + P(s_2 | s_i) + \ldots + P(s_q | s_i) = 1.
\end{align*}
\]

The entropy of the finite scheme in (5) we write as

\[
(6) \quad H(S | s_i) = - \sum_{j=1}^{q} P(s_j | s_i) \log P(s_j | s_i)
\]

and may be regarded as a measure of the amount of information obtained when the source (Markov process) advances one step forward, starting from the state \( s_i \). The mean value of the quantity in (6) over all states \( s_i \), that is,

\[
(7) \quad H(S | S) = \frac{1}{q} \sum_{i=1}^{q} H(S | s_i) = \frac{1}{q} \sum_{i=1}^{q} P(s_i) H(S | s_i) \log P(s_i | s_i)
\]

may be regarded as a measure of the mean amount of information obtained when the source (Markov process) moves one step ahead. The quantity \( H(S | S) \) which we shall call the entropy of the source obviously characterizes the source,
and is uniquely determined by the absolute probabilities $p_i$ and the conditional probabilities $P(s_i|s_{i-1})$, $1 < i < q$, $1 < j < q$.

Note that (7) may be written as

\[(8) \quad H(S^2) = H(S) + H(SIS),\]

which is essentially a special case of the general result

\[(9) \quad H(S, U) = H(S) + H(U|S).\]

It had been shown that generally

\[(10) \quad H(U) > H(U|S),\]

which in this case of a Markov source becomes

\[(11) \quad \log q > H(S) \geq H(S|S).\]

If we now consider sequences of $(n + 1)$ successive signals which we may also consider as constituted of the pair consisting of a sequence of $n$ signals and a single signal then we have

\[(12) \quad H(S^{n+1}) = H(S^n, S) = H(S^n) + H(S|S^n).\]

But the Markov chain property that the conditional probability of a state depends only on the immediately preceding state
Implies that

(13) \( H(S^n S^n) = H(S|S) \)

or

(14) \( H(S^{n+1}) = H(S^n) + H(S|S) \)

and by successive application of (14) we get

(15) \( H(S^{n+1}) = H(S) + nH(S|S) \), that is

(16) \( H[(n + 1)\text{-}tuple] = H(\text{single}) + n H(\text{transition}) \).

We may also write

(17) \( H(S^{n+1}) = H(S,S^n) = H(S) + H(S^n|S) \)

so that comparing (15) and (17), we see that

(18) \( H(S^n|S) = n H(S|S) \)

and

(19) \( H(S^{n+1}|S) = (m + n)H(S|S) = m H(S|S) + nH(S|S) = 
     = H(S^n|S) + H(S^n|S) \)

that is, the entropy in \((m + n)\) transitions is the sum of the
entropy in \( m \) and \( n \) transitions.

From (15) we see that

\[
\frac{1}{n+1} H(S^{n+1}) = \frac{1}{n+1} H(S) + \frac{n}{n+1} H(S|S) \tag{20}
\]

so that

\[
\lim_{n \to \infty} \frac{1}{n+1} H(S^{n+1}) = H(S|S), \tag{21}
\]

that is, the mean entropy per signal in a long sequence of signals is simply the entropy of the Markov source.

For a source without memory (independence) we have that

\[
H(S^2) = 2H(S) \tag{22}
\]

and for a Markov (order 1) source we have that

\[
H(S^2) = H(S) + H(S|S). \tag{23}
\]

Since we had in (11) that \( H(S) > H(S|S) \), a measure of the correlation between successive signals may be taken as

\[
2H(S) - H(S) - H(S|S) = H(S) - H(S|S) = 2H(S) - H(S^2). \tag{24}
\]

Note that
(25) \[ I = \sum_{i,j} P(s_i, s_j) \log \frac{P(s_i, s_j)}{P_i, P_j} \geq (\sum_{i,j} P(s_i, s_j)) \log \frac{\sum_{i,j} P(s_i, s_j)}{\sum_{i,j} P_i, P_j} = 0 \]

and

(26) \[ I = \sum_{i,j} P(s_i, s_j) \log P(s_i, s_j) - \sum_{i,j} P(s_i, s_j) \log p_i - \sum_{i,j} P(s_i, s_j) \log p_j \]

\[ = \sum_{i,j} P(s_i, s_j) \log P(s_i, s_j) - \sum_i p_i \log p_i - \sum_j p_j \log p_j, \]

\[ = -H(S^2) + 2H(S) = H(S) - H(S|S), \]

where \( I = 0 \Leftrightarrow P(s_i, s_j) = p_i p_j \), that is, no memory.

The reader is reminded that all the preceding was relative to a first-order Markov source.

Let us now turn our attention to the case of an \( m \)-th order Markov source, that is, the conditional probability of a signal value or state depends only on the preceding \( m \) signal values or states. As before we have the relation in (12) (which we repeat here)

(12) \[ H(S^{n+1}) = H(S^n, S) = H(S^n) + H(S|S^n) \]

but now the specification of an \( m \)-th order source implies that

(27) \[ H(S|S^n) = H(S|S^m) \quad n \geq m \]
or

\[(28) \quad H(S^{n+1}) = H(S^n) + H(S \mid S^n)\]

and by successive application of (28) we get

\[(29) \quad H(S^{n+1}) = H(S^n) + (n - m + 1) H(S \mid S^n), \quad n > m.\]

As in (11) we may also write

\[(30) \quad H(S^{n+1}) = H(S^n, S^{n-
+1}) = H(S^n) + H(S^{n+1} \mid S^n)\]

so that comparing (29) and (30) we see that

\[(31) \quad H(S^{n+1} \mid S^n) = (n - m + 1) H(S \mid S^n), \quad n > m,\]

and

\[(32) \quad H(S^{n+1} \mid S^n) = (n_1 + n_2) H(S \mid S^n) = n_1 H(S \mid S^n) + n_2 H(S \mid S^n)\]

\[= H(S^{n+1} \mid S^n) + H(S^{n+1} \mid S^n), \quad n_1 > m, n_2 > m.\]

From (29) we see that

\[(33) \quad \frac{1}{n+1} H(S^{n+1}) = \frac{1}{n+1} H(S^n) + \frac{n-m+1}{n+1} H(S \mid S^n), \quad n > m\]

so that

\[(34) \quad \lim_{n \to \infty} \frac{1}{n+1} H(S^{n+1}) = H(S \mid S^n),\]
that is, the mean entropy per signal in a long sequence of signals is simply the entropy of the m-th order Markov source.

If we use the notation \( P(S^n) \) to represent the probability of a sequence of \( n \) successive signals, then by the convexity property

\[
I = \sum_{S} \sum_{S'} P(S^n, S) \log \frac{P(S^n, S)}{P(S^n) P(S)} = \sum_{S} \sum_{S'} P(S^n, S) \log \frac{\sum_{S'} P(S^n, S')}{\sum_{S'} P(S^n) P(S)}
\]

\[
= \log \frac{1}{1} = 0, \text{ or}
\]

\[
\sum_{S} \sum_{S} P(S^n, S) \log P(S^n, S) = \sum_{S} \sum_{S} P(S^n, S) \log P(S^n) - \sum_{S} \sum_{S} P(S^n, S) \log P(S)
\]

\[
= \sum_{S} P(S^{n+1}) \log P(S^{n+1}) - \sum_{S} P(S^n) \log P(S^n) - \sum_{S} P(S) \log P(S)
\]

\[
= -H(S^{n+1}) + H(S^n) + H(S) = -H(S^n) - H(S|S^n) + H(S^n) + H(S)
\]

so that

\[
I = H(S) - H(S|S^n) \geq 0
\]

with \( I = 0 \) a signal is independent of the preceding \( m \) signals, is a measure of the relation between a signal and the preceding \( m \) signals in an m-th order Markov source.

To assist the reader to relate the exposition and notation in these notes with that in the text by Abramson we indicate equivalent values and results in the following table
<table>
<thead>
<tr>
<th>Abramson</th>
<th>Kullback</th>
</tr>
</thead>
<tbody>
<tr>
<td>H(S)</td>
<td>H(S</td>
</tr>
<tr>
<td>H(S)</td>
<td>H(S</td>
</tr>
<tr>
<td>H(S)</td>
<td>H(S)</td>
</tr>
<tr>
<td>H(S^a)</td>
<td>H(S^a)</td>
</tr>
<tr>
<td>H(S^a)</td>
<td>H(S^a</td>
</tr>
<tr>
<td>H(S^a)</td>
<td>H(S^a</td>
</tr>
</tbody>
</table>

(2-29) p. 28 (11)
(2-37) p. 31 (18)
(2-41) p. 31, (2-40) p. 31 (15)
(2-42) p. 31 (29)
(2-45) p. 32 (34)
(2-44) p. 32 (10) with y=S^a, x=S^a, order m source

cf. On the entropy of Markov chains by G.A. Ambarcumjan
EXAMPLE - Entropy Markov Chain

<table>
<thead>
<tr>
<th></th>
<th>s_1</th>
<th>s_2</th>
<th>s_3</th>
<th>s_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>s_1</td>
<td>.2</td>
<td>.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s_2</td>
<td>0</td>
<td>0.1</td>
<td>.9</td>
<td></td>
</tr>
<tr>
<td>s_3</td>
<td>0</td>
<td>0.2</td>
<td>.8</td>
<td></td>
</tr>
<tr>
<td>s_4</td>
<td>.7</td>
<td>.3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Stationary Probabilities

\[ p_1 = .2p_1 + .7p_4 = p_1 \]
\[ p_2 = .8p_1 + .3p_4 = p_2 \]
\[ p_3 = .1p_2 + .2p_3 = p_3 \]
\[ p_4 = .9p_2 + .8p_3 = p_4 \]

Using (7)

\[ H(S|S) = \frac{7}{24}H(.2, .8) + \frac{1}{3}H(.9, .1) \]
\[ + \frac{1}{24}H(.2, .8) + \frac{1}{3}H(.7, .3) \]
\[ = \frac{7}{24} (.721928 + 1/3(.468996)) + \frac{1}{24}(.721928 + 1/3(.881291)) \text{ bits} \]
\[ = .690738 \text{ bits} \]

\[ H(S) = \frac{7}{24} \log 24/7 + \frac{1}{3} \log 3 + \frac{1}{24} \log 24 + \frac{1}{3} \log 3 \]
\[ = \frac{1}{3} \log 24 + 2/3 \log 3 - 7/24 \log 7 \]
\[ = \frac{1}{3}(4.584962) + 2/3(1.584962) - 7/24(2.807355) \text{ bits} \]
\[ = 1.76615 \text{ bits} \]
\[
\begin{array}{l}
\begin{array}{l}
 s_2^2 \\
 s_1 s_1 \\
 s_1 s_2 \\
 s_1 s_3 \\
 s_1 s_4 \\
 s_2 s_1 \\
 s_2 s_2 \\
 s_3 s_3 \\
 s_2 s_4 \\
 s_3 s_1 \\
 s_3 s_2 \\
 s_3 s_3 \\
 s_4 s_4 \\
 s_4 s_1 \\
 s_4 s_2 \\
 s_4 s_3 \\
 s_4 s_4 \\
 \end{array}
\end{array}
\]

\[
\begin{array}{l}
.2(7/24) = 14/240 \\
.8(7/24) = 56/240 \\
0 \\
0 \\
0 \\
0 \\
0 \\
.1(1/3) = 8/240 \\
.9(1/3) = 72/240 \\
0 \\
0 \\
.2(1/24) = 2/240 \\
.8(1/24) = 8/240 \\
.7(1/3) = 56/240 \\
.3(1/3) = 24/240 \\
0 \\
0 \\
\end{array}
\]

\[
\begin{array}{l}
\begin{array}{l}
 m & n & \log_2 n \\
14 & 53.30297 \\
56 & 325.2119 \\
8 & 24.0000 \\
72 & 444.2346 \\
2 & 2.0000 \\
24 & 110.0391 \\
\end{array}
\end{array}
\]

\[
\begin{align*}
2H(S) - H(S^2) &= \frac{14}{240} \log \frac{240}{14} + \frac{56}{240} \log \frac{240}{56} + \frac{8}{240} \log \frac{240}{8} + \frac{72}{240} \log \frac{240}{72} \\
&+ \frac{2}{240} \log \frac{240}{2} + \frac{8}{240} \log \frac{240}{8} + \frac{56}{240} \log \frac{240}{56} + \frac{24}{240} \log \frac{240}{24} \\
&= \log 240 - \frac{1308.0005}{240} \\
&= 5.321928 + 2.584962 - 5.45 \text{ bits} = 2.45689 \text{ bits} \\
H(S^2) - H(S) &= 2.45689 - 1.76615 = .69074 \text{ bits}
\end{align*}
\]