Derivation of LogFac Estimator of Bits per symbol Estimate

Glen G. Langdon Jr.
Department of Computer Engineering
University of California
Santa Cruz, CA  95064
langdon@cse.ucsc.edu

Example of Enumerative Probabilities

Laplace’s Law of Succession

Consider following 12-symbol sequence $S$ from alphabet $\mathcal{A} = \{a, b, c\}$:

\begin{verbatim}
a a a b a c a b b a b c
\end{verbatim}

An approach to probability estimation is to assign a count of 1 to each symbol, and increment each symbol’s count after encoding it. The algorithm is causal since the decoder initializes and increments the counts after decoding them. The technique has been called the “cumulative count” technique among others.

This method of estimating a distribution by counting instances as they occur was employed in the eighteenth century by the French mathematician Laplace, and is also called who wrote:

“Placing the most ancient epoch of history at five thousand years ago, or at 1826213 days, and the sun having risen constantly in the interval of each revolution of twenty-four hours, it is a bet that 1826214 to one that it will rise again tomorrow.”

By this adaptive method of dynamic counting, the decoder can initialize each symbol count to 1, and whenever a given symbol occurs, after encoding the symbol (encoder) or decoding the symbol (at the decoder), the respective symbol count is incremented by 1. There is no need to give

the decoder the probabilities in a header part of the encoded file. Moreover, a counting technique is used in an adaptive Huffman code that dynamically adjusts the code tree as the symbols are counted.

Example 12-symbol sequence $S$

\begin{verbatim}
a a a b a c a b b a b c
\end{verbatim}

<table>
<thead>
<tr>
<th>$i$</th>
<th>$s(i)$</th>
<th>$n(a)$</th>
<th>$n(b)$</th>
<th>$n(c)$</th>
<th>Total</th>
<th>$p(s(i))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1/3</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2/4</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>3/5</td>
</tr>
<tr>
<td>4</td>
<td>b</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>1/6</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>4/7</td>
</tr>
<tr>
<td>6</td>
<td>c</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>1/8</td>
</tr>
<tr>
<td>7</td>
<td>a</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>9</td>
<td>5/9</td>
</tr>
<tr>
<td>8</td>
<td>b</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>10</td>
<td>2/10</td>
</tr>
<tr>
<td>9</td>
<td>b</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>11</td>
<td>3/11</td>
</tr>
<tr>
<td>10</td>
<td>a</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>12</td>
<td>6/12</td>
</tr>
<tr>
<td>11</td>
<td>b</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>13</td>
<td>4/13</td>
</tr>
<tr>
<td>12</td>
<td>c</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>14</td>
<td>2/14</td>
</tr>
</tbody>
</table>

Table 1: Laplace’s Law of Succession (cumulative counts)

If we form the product of each probability we have:
\[
\frac{1}{3}(\frac{2}{4})\left(\frac{3}{5}\right)\left(\frac{4}{6}\right)\left(\frac{1}{7}\right)\left(\frac{5}{8}\right)\left(\frac{2}{9}\right)\left(\frac{3}{10}\right)\left(\frac{6}{11}\right)\left(\frac{4}{12}\right)\left(\frac{2}{13}\right)\left(\frac{1}{14}\right)
\]

The numerator is in the form of factorials: 6!4!2! The denominator is a sequence of factors: 3 \times 4 \times \cdots \times 14. The denominator can be represented by 14! if we put 2! in the numerator to cancel the extra factors 1 \times 2 when 14! replaces the actual denominator product.

\[
\frac{2!6!4!2!}{14!}
\]

Since the product of Eq 2 represents the value \( p(S) \). The enumerative codlength (EL) is obtained by calculating the self-information of symbol sequence \( S \).

**Self-information of \( S \)**

\[
EL(S) = - \log_2 \left( \frac{2!6!4!2!}{14!} \right)
\]

In the general case with alphabet \( A = \{a_1, \ldots, a_M\} \) and \( \text{Len}(S) = N \) (an N-symbol sequence), with respective counts \( \text{Ct}(a_m) \), we have:

\[
EL(S) = \log_2 [(N + M - 1)!] - \log_2 [(M - 1)!] - \left( \sum_{m=1}^{M} \log_2 \text{Ct}(a_m) \right)
\]

The first term of Eq 3 represents the denominator because the minus sign in front (the \(-\log_2 [(N + M - 1)!]\)) of the count ratio denominator is changed to a + sign when factor \( \log_2 (N + M - 1)! \) moved to the numerator. The numerator has two \(-\log_2 \) factors, shown as the second and third terms of Eq 3. Term \( \log_2 (M - 1)! \) corresponds to the numerator correction factor \( M - 1 \) based on alphabet size. Note that the succession of total counts in the denominator is actually:

\[
M \times M + 1 \times \cdots \times M + N - 1
\]

Therefore, when we apply the factorial, as in \( (M+N-1)! \) in the denominator, \( (M-1)! \) is needed for compensation in the numerator so the denominator becomes effectively Eq 4 as verified below:

\[
M \times M + 1 \times \cdots \times M + N - 1 = \frac{(N + M - 1)!}{(M - 1)!}
\]

Thus, taking \( \log_2 \) of each side of Eq 5:

\[
\begin{align*}
-\log_2 (M \times M + 1 \times \cdots \times M + N - 1) &= -\log_2 \left( \frac{(N + M - 1)!}{(M - 1)!} \right) \\
&= -\log_2 (N + M - 1)! + \log_2 (M - 1)!
\end{align*}
\]

When the terms of Eq 6 are moved to the numerator, we have Eq 3, repeated below:

\[
EL(S) = \log_2 [(N + M - 1)!] - \log_2 [(M - 1)!] - \left( \sum_{m=1}^{M} \log_2 \text{Ct}(a_m) \right)
\]