ITCT Lecture 8.1: Lempel-Ziv Coding

Prof. Ja-Ling Wu

Department of Computer Science and Information Engineering
National Taiwan University
Algorithm:

- The source sequence is sequentially parsed into strings that have not appeared so far. For example, 1011010100010..., will be parsed as 1,0,11,0 1,010,00,10,...

- After every comma, we look along the input sequence until we come to the shortest string that has not been marked off before. Since this is the shortest such string, all its prefixes must have occurred easier. In particular, the string consisting of all but the last bit of this string must have occurred earlier.
We code this phrase by giving the location of the prefix and the value of the last bit.

Let $C(n)$ be the number of phrases in the parsing of the input $n$-sequence. We need $\log C(n)$ bits to describe the location of the prefix to the phrase and 1 bit to describe the last bit.

In the above example, the code sequence is:

$$(000,1), (000,0), (001,1), (010,1), (100,0), (010,0), (001,0)$$

where the first number of each pair gives the index of the prefix and the 2nd number gives the last bit of the phrase.
Decoding the coded sequence is straightforward and we can recover the source sequence without error!!

**Two passes-algorithm:**

**Pass 1:** parse the string and calculate $C(n)$

**Pass 2:** calculate the points and produce the code string

The two-passes algorithm allocates an equal number of bits to all the pointers. This is not necessary, since the range of the pointers is smaller at the initial portion of the string.

Definition:

A parsing $S$ of a binary string $x_1, x_2, \ldots, x_n$ is a division of the string into phrases, separated by commas. A distinct parsing is a parsing such that no two phrases are identical.

The L-Z algorithm described above gives a distinct parsing of the source sequence. Let $C(n)$ denote the number of phrases in the L-Z parsing of a sequence of length $n$.

The compressed sequence (after applying the L-Z algorithm) consists of a list of $c(n)$ pairs of numbers, each pair consisting of a pointer to the previous $\log c(n)$ occurrence of the prefix of the phrase and the last bit of the phrase.
the total length of the compressed sequence is:
\[ C(n)(\log c(n)+1) \]

It has been shown that
\[ \frac{C(n)(\log C(n)+1)}{n} \rightarrow H(\chi) \]

for a stationary ergodic sequence \( X_1, X_2, \ldots, X_n \).
Lemma 1 (Lempel and Ziv) :

The number of phrase $c(n)$ in a distinct parsing of a binary sequence $X_1, X_2, \ldots, X_n$ satisfies

$$C(n) \leq \frac{n}{(1 - \varepsilon_n) \log n}$$

where $\varepsilon_n \to 0$ as $n \to \infty$. 
Let $n_k = \sum_{j=1}^{k} j \cdot 2^j = (k-1)2^{k+1} + 2$

: the sum of the lengths of all distinct strings of length less than or equal to $k$.

$C(n)$: the no. of phrases in a distinct parsing of a sequence of length $n$

this number is maximized when all the phrases are as short as possible.

$n=n_k \rightarrow$ all the phrases are of length $\leq k$, thus

$C(n) \leq \sum_{j=1}^{k} 2^j = 2^{k+1} - 2 \leq 2^{k+1} \leq \frac{n_k - 2}{k-1} \leq \frac{n_k}{k-1}$

If $n_k \leq n < n_{k+1}$, we write $n=n_k + \Delta$, then

$\Delta = n - n_k < n_{k+1} - n_k = (k+1)2^{k+1}$
Then the parsing into shortest phrases has each of the phases of length $\leq k$ and $\frac{\Delta}{k+1}$ phrases of the length $k+1$. 

$$C(n) = C(n_k) + \frac{\Delta}{k+1} \leq \frac{n_k}{K-1} + \frac{\Delta}{k+1} \leq \frac{n_k + \Delta}{K-1} = \frac{n}{K-1}$$

$\Rightarrow C(n) \leq \frac{n}{K-1}$

We now bound the size of $k$ for a given $n$.

Let $n_k \leq n < n_{k+1}$. Then

$$n \geq n_k = (k-1)2^{k+1} + 2 \geq 2^k$$

$\Rightarrow k \leq \log n$

$$n \leq n_{k+1} = k \cdot 2^{k+2} + 2 \leq (k+2)2^{k+2}$$

$\leq [ (\log n+2) \cdot 2^{k+2} ]$

$\Rightarrow k + 2 \geq \log \frac{n}{\log n + 2}$
for all \( n \geq 4 \)

\[
k - 1 = (k + 2) - 3 \geq \log n - \log(\log n + 2) - 3
\]

\[
= \left(1 - \frac{\log(\log n + 2) + 3}{\log n}\right) \log n
\]

\[
\geq \left(1 - \frac{\log(2 \log n) + 3}{\log n}\right) \log n
\]

\[
= \left(1 - \frac{\log(\log n) + 4}{\log n}\right) \log n
\]

\[
= (1 - \varepsilon_n) \log n
\]

\[
\Rightarrow \quad C(n) \leq \frac{n}{K - 1} = \frac{n}{(1 - \varepsilon_n) \log n}
\]

where \( \varepsilon_n = \frac{\log(\log n) + 4}{\log n} \rightarrow 0 \) as \( n \rightarrow \infty \)
Lemma 2:

Let $Z$ be a positive integer valued r.v. with mean $\mu$. Then the entropy $H(z)$ is bound by

$$H(z) \leq (\mu+1) \log(\mu+1) - \mu \log \mu$$

pf : H W.
Let \( \{X_i\}_{i=-\infty}^\infty \) be a stationary ergodic process with pmf \( p(x_1,x_2,...,x_n) \). For fixed integer \( k \), defined the \( k \)th order Markov approximation to \( P \) as

\[
Q_k(x_{-(k-1)},...,x_0,x_1,...,x_n) \triangleq P(x_{-(k-1)}) \prod_{j=1}^n P(x_j|x_{j-k}^{j-1})
\]

where \( x_i^j \triangleq (x_i, x_{i+1},..., x_j), \ i \leq j \), and the initial state \( x_{-(k-1)}^0 \) will be part of the specification of \( Q_k \).

Since \( P(X_n|X_{n-k}^{n-1}) \) is itself an ergodic process, we have

\[
-\frac{1}{n} \log Q_k(X_1, X_2, ..., X_n | X_{-(k-1)}^0) = -\frac{1}{n} \sum_{j=1}^n \log P(X_j|X_{j-k}^{j-1})
\]

\[
\rightarrow -E \log P(X_j|X_{j-k}^{j-1}) \triangleq H(X_j|X_{j-k}^{j-1})
\]
We will bound the rate of the L-Z code by the entropy rate of the k-th order Markov approximation for all k. The entropy rate of the Markov approximation $H(X_j|X_{j-k}^{j-1})$ converges to the entropy rate of the process as $k \to \infty$ and this will prove the result.
S’pose \( X^n_{-(k-1)} = x^n_{-(k-1)} \), and s’pose that \( x_1^n \) is parsed into \( C \) distinct phrases, \( y_1, y_2, \ldots, y_c \). Let \( v_i \) be the index of the start of the \( i \)-th phrase, i.e., \( y_i = x_{v_i+1}^{v_i-1} \). For each \( i = 1, 2, \ldots, C \), defines \( S_i = x_{v_i-k}^{v_i-1} \). Thus, \( S_i \) is the \( k \) bits of \( x \) preceding \( y_i \), of course, \( S_1 = x^0_{-(k-1)} \).

Let \( C_{ls} \) be the number of phrases \( y_i \) with length \( l \) and preceding state \( S_i = S \) for \( l = 1, 2, \ldots \) and \( s \in X^k \), we then have:

\[
\sum_{l,s} C_{ls} = C
\]

and

\[
\sum_{l,s} lC_{ls} = n
\]
Lemma 3 : ( Ziv’s inequality )

For any distinct parsing (in particular, the L-Z parsing) of the string $x_1,x_2,\ldots,x_n$, we have

$$\log Q_k(x_1,x_2,\ldots,x_n|s_1) \leq -\sum_{l,s} C_{ls} \log C_{ls}$$

Note that the right hand side does not depend on $Q_k$.

proof : we write

$$Q_k(x_1,x_2,\ldots,x_n|s_1) = Q(y_1,y_2,\ldots,y_c|s_1)$$

$$= \prod_{i=1}^{C} P(y_i|s_i)$$
or \[ \log Q_k(x_1, x_2, \ldots, x_n | s_1) = \log P(y_i | s_i) \]

\[ = \log P(y_i | s_i) \]

\[ = C_{ls} \frac{1}{C_{ls}} \log P(y_i | s_i) \]

Jensen’ Inequality and the concavity of the logarithm. Now since the \( y_i \) are distinct, we have \( \sum_{i: y_i = l} P(y_i | s_i) \leq 1 \)

\[ \Rightarrow \log Q_k(x_1, x_2, \ldots, x_n | s_1) \leq \sum_{l,s} C_{ls} \log \frac{1}{C_{ls}} \]
Theorem:
Let \( \{X_n\} \) be a stationary ergodic process with entropy rate \( H(X) \), and let \( C(n) \) be the number of phrases in a distinct parsing of a sample of length \( n \) from this process. Then
\[
\limsup_{n \to \infty} \frac{C(n) \log C(n)}{n} \leq H(X)
\]
with probability 1.
pf: \log Q_k \left( x_1, x_2, \ldots, x_n \mid s_1 \right) \leq - \sum_{ls} C_{ls} \log \frac{C_{ls} \cdot C}{C}

= -C \log C - C \sum_{l,s} C_{ls} \log \frac{C_{ls}}{C}

writting \ \Pi_{ls} = \frac{C_{ls}}{C} , \ we \ have

\sum_{l,s} \Pi_{l,s} = 1, \ \sum_{l,s} l \Pi_{l,s} = \frac{n}{C}
We now define r.v.’s $U, V$, such that

$$\Pr(U=l, V=s) = \Pi_{l,s}$$

Thus $E(U) = n/c$, and

$$\log Q_k(x_1, x_2, \ldots, x_n | s_1) \leq CH(U,V) - C \log C$$

$$\frac{1}{n} \log Q_k(x_1, x_2, \ldots, x_n | s_1) \leq \frac{C}{n} \log C \leq H(U,V)$$

Since $H(U,V) \leq H(U) + H(V)$

and $H(V) \leq \log |X|^k = k$

From Lemma 2, we have

$$H(U) \leq (EU + 1) \log(EU + 1) - EU \log EU$$

$$= \left( \frac{n}{C} + 1 \right) \log \left( \frac{n}{C} + 1 \right) - \frac{n}{C} \log \frac{n}{C}$$

$$= \log \frac{n}{C} + \left( \frac{n}{C} + 1 \right) \log \left( \frac{C}{n} + 1 \right)$$

Thus, $\frac{C}{n} H(U,V) \leq \frac{C}{n} k + \frac{C}{n} \log \frac{n}{C} + O(1)$. 
For a given $n$, the maximum of $\frac{C \log \frac{n}{C}}{n}$ is attained for the maximum value of $C \left(\text{for } \frac{C}{n} \leq \frac{1}{e}\right)$. But from Lemma 1,

$$C \leq \frac{n}{\log n} \left(1 + O(1)\right).$$

Thus

$$\frac{C \log \frac{n}{C}}{n} \leq O\left(\frac{\log \log n}{\log n}\right)$$

and therefore $\frac{C}{n} H(U, V) \rightarrow 0$ as $n \rightarrow \infty$. 
Therefore,

\[
\frac{C(n) \log C(n)}{n} \leq -\frac{1}{n} \log Q_k \left(x_1, x_2, \ldots, x_n | s_1 \right) + \varepsilon_k(n)
\]

where \(\varepsilon_k(n) \to 0\) as \(n \to \infty\). Hence, with probability 1,

\[
\limsup_{n \to \infty} \frac{C(n) \log C(n)}{n} \leq \lim_{n \to \infty} -\frac{1}{n} \log Q_k \left(X_1, X_2, \ldots, X_n | X_{k-1}^0 \right)
\]

\[
= H \left(X_0 | X_{-1}, \ldots, X_{-k} \right) \to H(X) \quad \text{as} \quad k \to \infty.
\]
Theorem: Let \( \{X_i\}_{i=-\infty}^\infty \) be a stationary ergodic stochastic process. Let \( l(X_1,X_2,\ldots,X_n) \) be the L-Z codeword length associated with \( X_1,X_2,\ldots,X_n \). Then

\[
\lim_{n \to \infty} \frac{1}{n} l(X_1, X_2, \ldots, X_n) \leq H(X) \quad \text{with probability 1.}
\]

pf: \( l(X_1, X_2, \ldots, X_n) = C(n)(\log C(n) + 1) \)

by Lemma 1, \( \limsup \frac{C(n)}{n} = 0 \)

\[
\Rightarrow \limsup_{n \to \infty} \frac{l(X_1, X_2, \ldots, X_n)}{n} = \limsup_{n \to \infty} \left( \frac{C(n)\log C(n)}{n} + \frac{C(n)}{n} \right) \leq H(X) \quad \text{, with probability 1.}
\]

The L-Z code is a universal code, in which, the code does not depend on the distribution of the source!!
Optimal Variable Length-to-Fixed Length Code

Algorithm:
step 1. 永遠分最大的 node
step 2. 分到有 $2^l$ leaf nodes 則停

Example: Source data input: A B C C B A A A C
$P_A = 0.7$
$P_B = 0.2$
$P_C = 0.1$

code-tree

- AAAA
- AAAB
- AAAC
- AAA
- AAB
- AAC
- AA
- A
- AB
- AC
- root

A 0.7
B 0.2
C 0.1

AAAA 0.343
AAAB 0.098
AAAC 0.049
AAA 0.49
AAB 0.14
AAC 0.07
<table>
<thead>
<tr>
<th>Character</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1011</td>
</tr>
<tr>
<td>B</td>
<td>1010</td>
</tr>
<tr>
<td>C</td>
<td>1001</td>
</tr>
<tr>
<td>AA</td>
<td>1000</td>
</tr>
<tr>
<td>AB</td>
<td>0111</td>
</tr>
<tr>
<td>AC</td>
<td>0110</td>
</tr>
<tr>
<td>AAA</td>
<td>0101</td>
</tr>
<tr>
<td>AAB</td>
<td>0100</td>
</tr>
<tr>
<td>AAC</td>
<td>0011</td>
</tr>
<tr>
<td>AAAA</td>
<td>0010</td>
</tr>
<tr>
<td>AAAB</td>
<td>0001</td>
</tr>
<tr>
<td>AAAC</td>
<td>0000</td>
</tr>
</tbody>
</table>

A B, C, C, B, A A A C

0111 1001 1001 1010 0000

Tunstall ’77

GIT ph.D. Thesis (Tunstall Codes)