EE5585 Data Compression February 7, 2013

Lecture 6

Instructor: Arya Mazumdar *Scribe: Joshua Krist*

Main Topic: Proof that Lampel-Ziv is Optimal

Lampel-Ziv background

Lampel-Ziv is a universal code, and as such does not depend on foreknowledge of the occurrence probablities of the symbols being sent. The Lampel-Ziv code has two main types- a ”sliding window” and a ”tree structure” algorithm. The proof will focus on the ”tree structure” algorithm.

To encode a message using the Lampel-Ziv algorithm you must:

1) Start at the beginning and break the message into unique chunks (called phrases) truncating each chunk whenever a unique phrase is found.

For Example:
The Message: 0100011101001001110011010100... would be broken into the following phrases
Phrases: 0, 1, 00, 01, 11, 010, 0100, 111, 001, 10, 101, 0...

2) Each phrase is then encoded with a prefix followed by a final bit. The prefix is the index of a previous phrase that contains the first part of the current phrase, with the exception of the last bit, which is contained in the last bit part of the encoded phrase. A zero is used to indicate that there is no previous phrase.

Consider the above example: 0 1 00 01 11 010 0100 111 001 10 101 0...
The binary encoding would be: (0,0) (0,1) (1,0) (1,1) (2,1) (4,0) (6,0) (5,1) (3,1) (2,0) (10,1) ...

Some Notes:
Let $c(n)$ be the number of distinct phrases
The size of each encoded phrase would then be: $\log c(n) + 1$
To Decode: Just take the encoded message and evaluate in blocks of $\log c(n) + 1$, where the first $\log c(n)$ bits is the reference to a previous phrase and the last bit is the last bit of the decoded phrase.

Proof: Lampel-Ziv is Optimal

Notes on proof:
Length of phrase is $\log c(n) + 1$
Length of compressed file $c(n)(\log c(n) + 1)$
Rate of compression $\frac{c(n)(\log \ c(n)+1)}{n}$
For this proof we will assume that all the elements in the message are independently identically distributed (i.i.d.)
Note: there is a broader proof that deals with a stationary ergodic case, but that will not be dealt with here.

Let $x_1, x_2, ..., x_n$ be i.i.d.

Claim: $\frac{c(n)(\log c(n)+1}{n}$ converges to H(X) as n $\to \infty$

Lemma 1: $-\frac{1}{n}\log\ p(x_1, x_2, ..., x_n)$ converges to H(X) in probablity

This shows the Asymptotic Equipartition Property (AEP)

Let $x_1, x_2, ..., x_n$ be i.i.d.

Claim: $-\frac{1}{n}\log\ p(x_1, x_2, ..., x_n)$ converges to H(X) in probability

To show this convereance it needs to be shown that $|-\frac{1}{n}\log_2\ p(x_1, x_2, ..., x_n) - H(x)| < \epsilon, \forall\epsilon > 0$

Because $x_1, x_2, ..., x_n$ are i.i.d. it can be split thus $-\frac{1}{n}\log_2\ p(x_1, x_2, ..., x_n) = -\frac{1}{n}\sum_{i=1}^{n}\log_2\ p(x_i)$

Let $y_i = \log_2\ p(x_i)$ note this is also a random variable and i.i.d. This leads to

$-\frac{1}{n}\log_2\ p(x_1, x_2, ..., x_n) = -\frac{1}{n}\sum_{i=1}^{n}\log_2\ p(x_i) = -\frac{1}{n}\sum_{i=1}^{n} y_i$

Using the Weak Law of Large Numbers we can arrive at

$-\frac{1}{n}\log_2\ p(x_1, x_2, ..., x_n) = -\frac{1}{n}\sum_{i=1}^{n} y_i = -EY$ in probability

$= E\log_2 p(X)$

$= -\sum_{x\in X} p(X = x)\ \log_2$ p(X= x) = H(X)

Lemma 2: $c(n) \leq \frac{n}{(1-\epsilon_n)\log(n)}$ where $\epsilon_n \to 0$ as $n \to \infty$

Let n be the length of the binary sequence and c(n) be the number of distinct phrases

The sum of the length of distinct phrases that are less then or equal to k is $\sum_{j=1}^{k} j2^j$

This can be summed up like a geometric series where $\sum_{i=1}^{k} x^i = \frac{x^{k+1}-x}{x-1}$

$\sum_{j=1}^{k} jx^j = \frac{(k+1)x^k-1}{x-1} - x\frac{x^{k+1}-x}{(x-1)^2}$

placing x = 2 into the equation gives $\sum_{j=1}^{k} j2^j = 2[(k+1)2^k - 1 - 2^{n+1} + 2]$

$= 2[(k-1)2^k + 1]$

$= (k-1)2^{k+1} + 2$

$\equiv n_k$

now suppose n = n_k

then the maximum number of distinct phrases c(n) = $\sum_{j=1}^{k} 2^j = 2^{k+1} - 2 \leq 2^{k+1} \leq \frac{n_k}{k-1}$

for some k $n_k \leq n \leq n_{k+1}$

So $c(n) \leq c(n_k) + \frac{n-n_k}{k+1} \leq \frac{n_k}{k-1} + \frac{n-n_k}{k+1} \leq \frac{n}{k-1}$

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

$k \leq \log(n) - 1$

$n \leq n_{k+1} = k2^{k+2} + 2 \leq \log(n-1)2^{k+2} + 2$

$2^{k+2} \geq \frac{n-2}{\log(n)-1}$

$k - 1 \geq \log(\frac{n-2}{\log(n)-1}) - 3$

$c(n) \leq \frac{n}{\log \frac{n-2}{\log(n)-1} - 3}$

the denominator expanded out is: $\log(n-2) - \log(\log\ n - 1) - 3$

$= (\log\ n)[\frac{\log(n-2)}{\log(n)} - \frac{\log(\log(n-1)) - 3}{\log(n)}]$

$\geq (\log\ n)[\frac{\log(n)-1)}{\log(n)} - \frac{\log(\log(n-1)) - 3}{\log(n)}]$

$= (\log\ n)(1 - \frac{\log(\log(n-1) - 4}{\log(n)}) = (\log\ n)(1 - \epsilon_n)$

Note: $\epsilon \rightarrow 0$ as $n \rightarrow \infty$

Therefore $c(n) \leq \frac{n}{(1-\epsilon_n)\log(n)}$

Statement of Lemma 3

Lemma 3 is not proven here, and will be proved in the next set of notes. However, it is stated here for reference.

Given that z is a random varaiable that takes a non-negative integer value, with an expected value of $E(z) = \mu$

Then $H(z) \leq H(g)$ where $H(g)$ is a geometric random variable with an expected value of $E(g) = \mu$

Main Proof: $\frac{c(n)(\log c(n)+1}{n}$ converges to H(x) as n $\rightarrow \infty$

Starting with $-\frac{1}{n} \log p(x_1, x_2, ..., x_n)$ and $x_1, x_2, ..., x_n$ is i.i.d.

Let them be parsed as described by the Lampel-Ziv method above Let the distinct phrases be called $S_1, S_2, ..., S_{c(n)}$

So that $-\frac{1}{n} \log p(x_1, x_2, ..., x_n) = -\frac{1}{n} \log p(S_1, S_2, ..., S_{c(n)})$

$= -\frac{1}{n} \log \sum_{i=1}^{c(n)} p(S_i)$

$= -\frac{1}{n} \sum_{i=1}^{c(n)} \log\ p(S_i)$

Now by clumping phrases of the same length we can write

$= -\frac{1}{n}\sum_{l=1}^{l_{max}}\sum_{*}\log\ p(S)$ where $\sum_{*}$ is the summation of phrases that are of length l

$= -\frac{1}{n}\sum_{l=1}^{l_{max}} c_l \sum_{*}\frac{1}{c_l}\log\ p(S)$

where, c_l is the number of phrases of length l. By using Jensen's Inequality and the fact that the function is concave we can state that

$\geq -\frac{1}{n}\sum_{l=1}^{l_{max}} c_l \log\sum_{*}\frac{1}{c_l}p(S)$

The rest of the proof will be finished in the next set of notes.