Lempel-Ziv 77 Algorithm

1 The Source Model

In the first main result, the authors establish that for infinite sequences emitted by sources of a particular kind, LZ77 attains optimal compression ratio.

The source in this particular case is not a probabilistic source, but one defined combinatorially. This is what makes the following result interesting, even though it is partial. (A full result will show that for every infinite sequence, LZ77 is optimal. This result was later established by Shields, and we will cover this later in the course. The proof of LZ77 is interesting in its own right, even if partial.)

Let A be the finite alphabet. Given a string $S \in A^*$, let $S\{m\}$ denote the set of all m-length substrings of S, and $S(m)$ denote the cardinality of this set.

If σ is a set of strings, let $\sigma\{m\}$ be the set of m-length strings in σ, and $\sigma(m)$ be the number of such strings.

Definition 1. A set $\sigma \subseteq A^*$ is called a *source* if the following hold.

1. $A \subset \sigma$ - *i.e.* all the digits of A are in σ.
2. If a string S is in σ, then so is SS.
3. If a string S is in σ, then $S\{m\} \subseteq \sigma\{m\}$.

Note that rules 1 and 2 force every source to be infinite. Examples of sources include the following

1. A^* - this is the largest possible source.
2. If $A = \{0, 1\}$, then $0^* \cup 1^*$ - *i.e.* the set of all single digit strings from A form a source. This is the smallest possible source.

2 Analysis: Lower Bound on the Compression Length

2.1 Block-to-Variable Coding

Consideer any information-lossless finite-state compressor which parses the input into fixed-length blocks of length L and produces output blocks. Since the output blocks are uniquely decipherable, we have

$$\{X_1, X_2, \dots, X_M\} = \sigma\{L\},$$

i.e. the L-length strings in σ correspond to the precise set of input blocks. Then, we have

$$M = \sigma(L).$$

Since

$$h(L) = \frac{1}{L} \log_\alpha \sigma(L)^1,$$

i.e.,

$$\sigma(L) = \alpha^{Lh(L)},$$

we have that

$$\max_{1 \le i \le M} \{\ell(Y_i)\} \ge \log_\alpha M = Lh(L).$$

Lemma 2.

$$\rho_{BV}(\sigma, M) \ge h(L).$$

Proof. The compression ratio ρ_i associated with the i^{th} word is given by

$$\rho_i = \frac{\ell(Y_i)}{L}.$$

The block-to-variable compression ration $\rho(\sigma, M)$ of the source σ is defined as

$$\rho(\sigma, M) = \min_{C(\sigma,M)} \max_{1 \le i \le M} \rho_i.$$

We know that

$$\min_{C(\sigma,M)} \max_{1 \le i \le M} \rho_i \ge \frac{\log M}{L} = \frac{Lh(L)}{L} = h(L).$$

Hence we have the result. □

2.2 Variable-to-Block Coding

Suppose now that $\ell(Y_i) = L$ for all $1 \le i \le M$. In this case, the compression ratio is given by

$$\rho_i = \frac{L}{\ell(X_i)}.$$

Then we have the following.

Lemma 3.

$$\rho_{VB}(\sigma, M) \ge h(L_M),$$

where

$$L_M = \max\{\ell \mid M \ge \sigma(\ell)\}.$$

Note that the above quantity L_M is not defined for all M. For example, if $\sigma(i) = 2$ for all i, then L_M is ∞ for all $M \ge 2$.

Proof. Assume without loss of generality that

$$\ell(X_i) \le \ell(X_{i+1}), \quad 1 \le i \le M - 1.$$

Hence, for each $C \in C_{VB}(\sigma, M)$, we have

$$\max \rho_i(C) = \frac{L(C)}{\ell(X_1)}.$$

[1] The base of the logarithm is α, the size of the alphabet A.

§a. (Upper bound on $\ell(X_1)$) If every string has the same length as X_1, then $\ell(X_1) = M$. Otherwise, if there is a longer string, then $\ell(X_1) < M$. Thus

$$M \geq \sigma(\ell(X_1)).$$

Fom the definition of L_M, it follows that $\ell(X_1) \leq L_M$. Thus

$$\max_{1 \leq i \leq M} \rho_i(C) = \frac{L(C)}{\ell(X_1)} \geq \frac{L(C)}{L_M}.$$

§b. (Lower Bound on $L(C)$) Since every code represents at most one X_i, we have $|C| \geq M$, or equivalently, $L(C) \geq \log_\alpha(M)$. Hence,

$$\max_{1 \leq i \leq M} \rho_i(C) \geq \frac{\log_\alpha(M)}{L_M}.$$

Since $M \geq \sigma(L_M) = \alpha^{L_M h(L_M)}$, we have

$$\max_{1 \leq i \leq M} \rho_i(C) \geq h(L_M).$$

□

3 Upper Bound on the Performance of LZ77

We will choose a buffer length carefully, to make the calculation simpler later.

Let

$$n = \sum_{m=1}^{\lambda} m\alpha^m + \sum_{m=\lambda+1}^{L_s-1} m\sigma(\ell) + L_s(N_{L_s} + 1), \tag{1}$$

where $\lambda = \lfloor (L_s - 1)h(L_s - 1) \rfloor = \lfloor \log \sigma(L_s - 1) \rfloor$, and

$$N_{L_s} = \sum_{m=1}^{\lambda} (L_s - m - 1)\alpha^m. \tag{2}$$

We will justify the need for this exact form of n later. For the moment, note that it is a constant that depends on L_s, σ and $h(L_s)$, but not on any individual string. (In particular, n depends on the source and its h parameters in addition to L_s.)

Let $Q \in \sigma\{n - L_s\}$ be a string whose parsing has the maximum number of phrases, and let the parsing be

$$Q = Q_1 Q_2 \ldots Q_N.$$

Unlike LZ78, we cannot claim that every phrase Q_j has a unique predecessor Q_i which is the longest proper prefix of Q_j. This is the claim that helped us derive the upper bound in LZ78. This property cannot hold here, hence the upper bound derivation is more complex. We proceed as follows.

First, we derive an upper bound for N, and then, we derive a lower bound for $n - L_s$.

3.1 Upper Bound for N

We can make the following observation about the phrases. If $\ell(Q_i) = \ell(Q_j)$, then $Q_i \neq Q_j$ - *i.e.*, the phrases are unique.

Let K_m be the number of phrases in the parsing of Q of length m. Then we have

$$N = 1 + \sum_{m=1}^{L_s} K_m,$$

since the longest a phrase can be is $L_s{}^2$.

We have $K_m \leq \sigma(m)$ for all $1 \leq m \leq L_s - 1$.

Since

$$n - L_s = \ell(Q) = \ell(Q_N) + \sum_{k=1}^{L_s} mK_m,$$

we can provide an upper bound for N by overestimating K_1, K_2, $\ldots$, and K_{L_s-1} at the expense of K_{L_s}. (*i.e.* Since the total length of the phrases is constant - the length of Q - we can maximize the number of phrases by assuming that there are more number of shorter phrases.)

Since $\sigma(m) \leq \sigma(m+1)$ for any source, and $\sigma(m) = \alpha^{mh(m)} \leq \alpha^m$, we obtain

$$N \leq K'_{L_s} + \sum_{m-1}^{L_s-1} K'_m = N'$$

where, for all $1 \leq m \leq L_s - 1$,

$$K'_m = \begin{cases} \alpha^m & \text{if } 1 \ \leq m \ \leq \lfloor (L_s-1)h(L_s-1) \rfloor \\ \sigma(L_s-1) & \text{if } \lfloor (L_s-1)h(L_s-1) \rfloor \ \leq m \ \leq L_s - 1. \end{cases}$$

and

$$K'_{L_s} = \left\lceil \frac{1}{L_s} \left(n - L_s - \sum_{m=1}^{L_s-1} mK'_m \right) \right\rceil.$$

This is obtained by trying to saturate the lower indices as much as possible, and assigning the remainder to K'_{L_s}. Note that $\lfloor (L_s-1)h(L_s-1) \rfloor$ is $\lfloor \log \sigma(L_s-1) \rfloor$, the upper threshold of those m which can be filled to the maximum.

Since we have appropriately chosen n, substituting the value of n from (1) into the expression above, e get $K_{L_s} = N_s$, and

$$N' = N_{\ell+1} + \sum_{m=1}^{\lambda} \alpha^m + \sum_{m=\lambda+1}^{L_s-1} \sigma(L_s-1).$$

Now, substituting it back in (1) and (**??**) gives

$$n - L_s - (L_s-1)N' = \sum_{m=1}^{\lambda} m\alpha^m + \sum_{m=\lambda+1}^{L_s-1} m\sigma(\ell) +$$
$$N_{L_s} - (L_s-1) \left[\sum_{m=1}^{\lambda} \alpha^m + \sum_{m=\lambda+1}^{L_s-1} \sigma(L_s-1) \right] = 0.$$

[2] Why the +1?

Hence,

$$N \leq N' = \frac{n - L_s}{L_s - 1}.$$

4 Lower Bound for $n - L_s$

We now prove the following theorem, during whose proof we derive a lower bound for $n - L_s$.

Theorem 4. *If the buffer length n for a source with known h-parameters is chosen according to (1), then we have*

$$\rho \leq h(L_s - 1) + \varepsilon(L_s),$$

where

$$\varepsilon(L_s) = \frac{1}{L_s - 1}\left(3 + 3\log(L_s - 1) + \log\frac{L_s}{2}\right).$$