Some Entropic Bounds for Lempel-Ziv Algorithms

S. Rao Kosaraju

Giovanni Manzini

We initiate a study of parsing-based compression algorithms such as LZ77 and LZ78 by considering the empirical entropy of the input string. For any string s, we define the k-th order entropy H_k(s) by looking at the number of occurrences of each symbol following each k-length substring inside s. The value H_k(s) is a lower bound to the compression ratio of a statistical modeling algorithm which predicts the probability of the next symbol by looking at the k most recently seen characters. Therefore, our analysis provides a means for comparing Lempel-Ziv methods with the more powerful, but slower, PPM algorithms.

We say that an algorithm is coarsely optimal, if for all k there exists a function f_k, with \( \lim_{n \to \infty} f_k(n) = 0 \), such that for all strings s the compression ratio \( \rho(s) \) is bounded by

\[ \rho(s) \leq H_k(s) + f_k(|s|) \]  

The coarse optimality of LZ78 has been proven by Plotnik et al. [IEEE Trans. Inf. Th., 38, 1992]. We give an alternative proof of the optimality of LZ78 and we extend our proof to show that LZ77 is coarsely optimal as well. Our proofs are based on a symbolwise model for LZ78 which can be used to prove the optimality of other Lempel-Ziv variants.

Having defined the optimality using (1), we can analyze an important phenomenon which went undetected using the previous approaches. If a string s is such that \( \lim_{|s| \to \infty} H_k(s) = 0 \), it is possible that (1) holds with a function f_k such that \( H_k(s) = o(f_k(|s|)) \). Hence, we can have a coarsely optimal algorithm with a compression ratio much bigger than the entropy. This fact suggests that further study is required to understand how Lempel-Ziv algorithms perform with low entropy strings, that is, with the strings such that \( H_k(s) \to 0 \) for some \( k \geq 0 \).

Our main contribution is a comparison of the compression ratio of Lempel-Ziv algorithms with the zeroth order entropy \( H_0 \). First we show that for low entropy strings LZ78 compression ratio can be much higher than \( H_0 \). Then, we present a modified algorithm which combines LZ78 with Run Length Encoding and is able to compress efficiently also low entropy strings. We prove that for any string s, the compression ratio of the modified algorithm is — ignoring lower order terms — bounded by \( 3H_0(s) \). Finally, we consider LZ77 and we show that this algorithm performs well also for low entropy strings; we prove that for any string s the compression ratio achieved by LZ77 is bounded asymptotically by \( 8H_0(s) \).