

Lossless Data Compression with Error Correcting Codes

Giuseppe Caire
 Institut Eurécom
 06904 Sophia-Antipolis, France
 giuseppe.caire@eurecom.fr

Shlomo Shamai
 Technion
 Haifa 32000, Israel
 sshlomo@ee.technion.ac.il

Sergio Verdú
 Princeton University,
 Princeton, NJ 08544 USA
 verdu@ee.princeton.edu

Existing zero-error variable-length data compression algorithms suffer from sensitivity to transmission errors and error propagation across packets when used in packet-based data transmission through noisy channels. We propose a new approach to lossless data compression based on error correcting codes and the block-sorting transform.

It has long been recognized that linear source codes achieve the entropy rate of memoryless sources and that the parity check matrix of an error correcting code coupled with its associated syndrome decoder can be used as the linear compressor and decompressor, respectively. Such basic scheme is a constructive approach to Shannon's almost-noiseless fixed-length data compression, which despite its theoretical importance has had no impact in the practical world. Indeed, the reported embodiments of this approach with actual error correcting codes have not been able to compete favorably with existing data compression algorithms such as LZ. Not only the block error rates have been disappointingly high but, more importantly, the scope of this approach has been limited to memoryless sources (actually, biased coins for the most part), whose statistics are known to the decompressor.

In this paper, we propose a new approach that overcomes those shortcomings. The basic encoding scheme (Figure 1) takes a fixed-length block of data and passes it through the Burrows-Wheeler (Block sorting) transform. This is a one-to-one transformation that outputs the last column of the square matrix formed by lexicographic sorting of all the cyclic shifts of the original sequence. A key property of the BWT output [1] is that (as the blocklength grows) it is asymptotically piecewise i.i.d. (for stationary ergodic tree sources) with the length, location, and distribution of the i.i.d. segments depending on the statistics of the source. The segmentation block detects those segments whose entropy exceeds a certain threshold and those data segments are passed uncompressed to the decompressor. The remaining segments are multiplied by the parity-check matrix of a channel code designed for a nonstationary binary symmetric channel (block labelled as 'Syndrome'). In our preferred embodiment we use an irregular Low-Density Parity-Check (LDPC) matrix. The decompressor uses the iterative Belief Propagation (BP) algorithm modified to take into account that the parity-check equations now take the values of the compressed bits. The output is merged with the uncompressed segments and, finally, the inverse Burrows-Wheeler Transform is taken. Each of the building blocks can be implemented in linear time.

Key to the favorable performance of our scheme is the fact that the encoder can run an exact copy of the iterative BP algorithm. The block labelled "Iterative Doping" is present at the compressor and communicates the actual values of the bits that achieve the lowest reliability at certain iterations of the BP algorithm. Performance is further improved by having a library of parity-check matrices and tuning the choice of the matrix to the source realization.

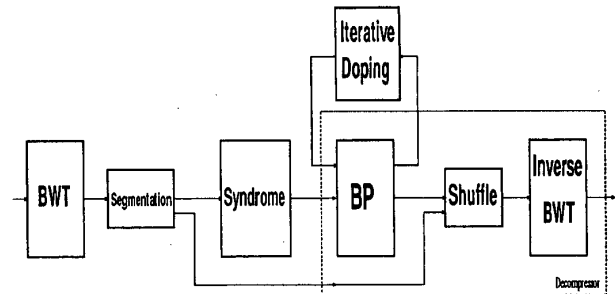
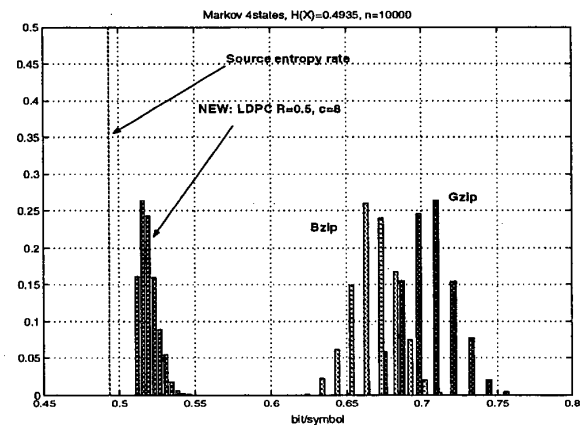


Figure 1: Basic Compression/Decompression Scheme



We have implemented a universal version of the scheme using an adaptive segmentation module that learns an approximation to the source tree in linear time via the Minimum-Description-Length principle. A coarse version of both the segmentation and the distributions within each segment is communicated to the decoder, which then refines iteratively the statistical model in conjunction with the Belief-Propagation algorithm. Figure 2 shows the histogram of achieved compression with the universal algorithm for a four-state Markov chain in comparison with LZ (gzip) and a conventional BWT-based compressor (bzip). Although the comparison is quite encouraging, the main motivation and advantages of our scheme over the conventional separation-based approach accrue in the joint source/channel setting, which will be reported elsewhere.

REFERENCES

- [1] M. Effros, K. Visweswariah, S. Kulkarni, and S. Verdú. Data compression based on the Burrows-Wheeler transform: Analysis and optimality. *IEEE Trans. on Information Theory*, 48:1061–1081, May 2002.