

Lossless Data Compression Via Error Correction

Sergio Verdú

Dept. Electrical Engineering, Princeton University,
Princeton, New Jersey 08540, USA
verdu@princeton.edu

Abstract. This plenary talk gives an overview of recent joint work with G. Caire and S. Shamai on the use of linear error correcting codes for lossless data compression, joint source/channel coding and interactive data exchange.

1 Summary

Over the last five decades, significant inventions have led to data compression and data transmission systems whose efficiency approaches Shannon's fundamental limits [1]. Error-correcting codes now exist (i.e. sparse-graph linear codes) that can achieve performance close to channel capacity with complexity and delay that are tolerable for many applications. Similarly, lossless data compression algorithms exist (most notably the Lempel-Ziv algorithm) that can provably achieve the entropy rate of a wide class of sources with very low complexity. Curiously, although Shannon's development of the theories of fundamental limits for data compression and transmission shared very strong commonalities, there has been essentially no intercourse between the respective constructive theories throughout their long histories.

While Shannon's separation principle establishes no loss in asymptotic performance when compression and transmission are performed separately, it has long been expected (but not fully realized) that, in the nonasymptotic regime, gains may accrue by joint design. Furthermore, in systems such as packet-oriented wireless high data rate systems, it is sometimes cumbersome to design systems based on the separation principle.

Lossless data compression algorithms find numerous applications in information technology, such as packing utilities (e.g. `gzip`), modem standards, fax standards, back-end of lossy compression algorithms (e.g. JPEG and MPEG), and compression of headers of TCP/IP packets in wireless networks.

Indeed, the field of lossless data compression has achieved a state of maturity, with algorithms that admit fast (linear-complexity) implementations and achieve asymptotically the fundamental information theoretic limits.

The availability of linear codes (such as the low-density parity check codes) that allow for very efficient encoding/decoding algorithms while operating near the Shannon limit makes their application in data compression competitive with state-of-the-art methods while not suffering from some of their shortcomings.

A series of recent papers [2, 3, 4, 5] presents a new approach to universal noiseless compression based on error correcting codes. The scheme is based on the concatenation of the Burrows-Wheeler block sorting transform (BWT) with the syndrome former of a Low-Density Parity-Check (LDPC) code. The proposed scheme has linear encoding and decoding times and uses a new closed-loop iterative doping (CLID) algorithm that works in conjunction with belief-propagation decoding.

Alternatively, fountain codes can replace the LDPC codes [6] to provide a streamlined design which is ideally suited for variable-length lossless compression.

One of the incentives to use error correcting codes for data compression is the natural extension of the schemes to joint source/channel encoding and decoding. Schemes for that purpose are explored in [7].

Building upon Slepian-Wolf coding [8], sparse-graph codes, belief propagation, and closed-loop iterative doping, new schemes for interactive data exchange between two agents who want to communicate losslessly their respective information via several rounds of communication are proposed in [9].

References

1. C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379–423, 623–656, Jul.-Oct. 1948.
2. G. Caire, S. Shamai, and S. Verdú, "A new data compression algorithm for sources with memory based on error correcting codes," *2003 IEEE Workshop on Information Theory*, pp. 291–295, Mar. 30- Apr. 4, 2003.
3. G. Caire, S. Shamai, and S. Verdú, "Lossless data compression with error correction codes," *2003 IEEE Int. Symp. on Information Theory*, p. 22, June 29- July 4, 2003.
4. G. Caire, S. Shamai, and S. Verdú, "Universal data compression with LDPC codes," *Third International Symposium On Turbo Codes and Related Topics*, pp. 55–58, Brest, France, September 1-5, 2003.
5. G. Caire, S. Shamai, and S. Verdú, "Noiseless data compression with low density parity check codes," in *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, P. Gupta and G. Kramer, Eds., pp. vol. 66, pp. 263–284. American Mathematical Society, 2004.
6. G. Caire, S. Shamai, A. Shokrollahi and S. Verdú, "Fountain Codes for Lossless Data Compression," in *DIMACS Series in Discrete Mathematics and Theoretical Computer Science: Algebraic Coding Theory and Information Theory*, A. Ashikhmin, A. Barg, I. Duursma, Eds., pp. vol. 68, pp. 1-20. American Mathematical Society, 2005.
7. G. Caire, S. Shamai, and S. Verdú, "Almost-noiseless joint source-channel coding-decoding of sources with memory," *Proc. Fifth International ITG Conference on Source and Channel Coding (SCC)*, pp. 295–304, Jan 14-16, 2004.
8. D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Information Theory*, IT-19:471–480.
9. G. Caire, S. Shamai, and S. Verdú, "Practical schemes for interactive data exchange," *Proc. 2004 Int. Symp. Information Theory and its Applications*, Parma, Italy, Oct. 2004.