CS698D: Special topics in data compression

Instructor : Satyadev Nandakumar
Semester II, 2013-14

This course covers the introductory basics and some recent topics in the area of lossless data compression. We will cover the following topics.

First, we introduce the notion of a lossless compressor. We will cover the fundamental techniques like Shannon-Fano coding, Huffman coding and Arithmetic codes. We will cover Kolmogorov complexity as the "limit notion" of a lossless code, where the compressor can be any arbitrary program.

Special emphasis will be given to the class of "universal codes", including the Lempel-Ziv algorithm[1, 2] and the Burrows-Wheeler transform. We then study the optimality and rate of convergence results of the Lempel-Ziv algorithm, in the weak sense, and in the strong sense.

We then cover some current topics in the topic of universal codes. Specifically, the class of universal codes has been shown to be "non-robust" with respect to small perturbations. We will study the tools used to establish these results, and investigate some open problems in this area.[3]

Finally, we finish with some applications of the Lempel-Ziv algorithm to areas like Statistical inference, and algorithms.

Mathematical techniques used in these results include recurrence theorems from ergodic theory, and "Rokhlin-Kakutani tower constructions" from Ergodic theory (also known as "cutting and stacking".[4])

The text books we will use for the course include "Elements of Information Theory" by Cover and Thomas [5] , "Information Theory" by Csizar and Korner [6] , and "Text Compression" by Bell, Cleary and Witten[7]

---

[1] Ziv, Jacob and Lempel, Abraham, “A universal algorithm for sequential data compression,” *IEEE transactions in information theory* **23**(3), pp. 337-343 (May 1977).

[2] Ziv, Jacob and Lempel, Abraham, “Compression of individual sequences via variable rate coding,,” *IEEE transactions in information theory* **24**(5), pp. 530-536 (September 1978).

[3] Vyugin, Vladimir V., “On instability of ergodic limit theorems with respect to small violations of algorithmic randomness,” *Proceedings of IEEE International Symposium on Information Theory,* St. Petersburg, Russia (2011).

[4] Shields, Paul, “Cutting and stacking: a method for constructing stationary ergodic processes,,” *IEEE transactions on information theory* **37**(6), pp. 1605-1617 (November 1991).

[5] Cover, Thomas M. and Thomas, Joy A., “Elements of Infomation Theory,” *Wiley* (June 2006).

[6] Csiszar, Imre and K, “Information Theory,” *Cambridge.* (2010).

[7] Bell, Timothy C., Cleary, John G., and Witten, Ian H., “Text Compression,” *Prentice Hall Advanced Reference Series - Computer Science.* (1990).

The following gives a list of topics we intend to cover in the course.

1 Introduction to Lossless coding. Low probability of compression. Basic compression schemes - Run-length encoding, Differential compression,

2 Lossless Coding - Shannon's Noiseless Coding Theorem, Shannon-Fano Codes, Huffman Codes, Arithmetic Coding.

3 Mathematical Preliminaries - Entropy, Kullback-Leibler Divergence. Introduction to convergence - convergence in measure (weak convergence), convergence with probability 1 (strong convergence) and pointwise convergence. Chebyshev's inequality. Weak Law of Large Numbers, Strong Law of Large Numbers, Ergodicity, the Ergodic Theorem. Shannon-McMillan-Breiman Theorem.

4 Lempel Ziv compression - LZ77 and LZ78 variants. Optimality of LZ compressors - strong sense and pointwise sense [8] Lempel-Ziv Markov Chain variant. Non-robustness of universal compressors - Cutting and stacking [9] ,[10] Applications of Lempel-Ziv compression to statistical inference, portfolio selection and dynamic programming.

5 Burrows-Wheeler Transform - optimality.

[8] Shields, Paul, "Performance of LZ algorithms on individual sequences," *IEEE transactions on information theory* **45**(4), pp. 1283-1288 (May 1999).

[9] Shields, Paul, "Cutting and stacking: a method for constructing stationary ergodic processes,," *IEEE transactions on information theory* **37**(6), pp. 1605-1617 (November 1991).

[10] Vyugin, Vladimir V., "On instability of ergodic limit theorems with respect to small violations of algorithmic randomness," *Proceedings of IEEE International Symposium on Information Theory*, St. Petersburg, Russia (2011).