Making the correct mistakes: Towards practical, universal lossy compression

Dharmendra S. Modha
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
Email: dmodha@us.ibm.com

Narayana P. Santhanam
Electrical & Computer Engineering
University of California, San Diego
La Jolla, CA 92037-0407
Email: prasad@talk.ucsd.edu

***Abstract*— The central problem of lossy source coding is to find an universal (for stationary, ergodic sources), sequential, adaptive, and polynomial-time (practical) algorithm. No such algorithm is yet known.**

We propose a new sequential, adaptive, quadratic-time algorithm for variable-rate lossy compression of memoryless sources at a fixed distortion that does not require any *a priori* information about the source statistics. The algorithm uses approximate pattern matching and is modeled after the Lempel-Ziv (LZ78) algorithm. Like the LZ78 algorithm, the algorithm *sequentially* parses the source sequence into non-overlapping phrases, mapping each phrase to a *codeword* in a *dictionary*, which in turn is updated sequentially. The algorithm ensures that each new codeword is a one-letter extension of a previously emitted codeword, and, hence, the output distorted sequence is naturally parsed using LZ78. Thus, the distorted sequence can be easily compressed without any further loss. The per-letter distortion between any source phrase and its associated codeword does not exceed the desired distortion. Typically, there are multiple ways to parse the incoming source string and map it into codewords. Multiple matches are a sign of underlying redundancy, and, hence, are a symptom of the fact that we are not operating near the rate-distortion curve. We focus on the following key question:

How to select between multiple parsings?

The choice of a codeword affects the future parsing, since the chosen codeword is used to update the dictionary. Moreover, the chosen codeword affects the per-letter code length for the phrase in question. We would like to carefully choose the codeword that best balances between the per-letter code rate in the current step and the quality of the resulting codebook for future steps.

As the key new idea, the algorithm uses *lower mutual information* to judiciously select "good" codewords. We empirically demonstrate that a greedy algorithm that randomly selects one of the longest matches in each step is unlikely to be asymptotically optimal. For Bernoulli sources with Hamming distortion, we empirically demonstrate that the new algorithm (a) discovers the optimal reproduction type, (b) leads to absence of multiple matches, and (c) seems to approach the rate-distortion coding rate. Based on empirical observations, we formulate two conjectures that could imply that the algorithm is asymptotically optimal for memoryless sources.

For details, please see [1].

REFERENCES

[1] D. S. Modha and N. P. Santhanam, "Making the correct mistakes," *Data Compression Conference*, March 28-30, 2006.