A suboptimal lossy data compression based on approximate pattern matching

Wojciech Szpankowski
Purdue University

December 11, 1995

[summary by Philippe Jacquet]

1. Introduction

A practical algorithm for lossy data compression is presented. It is derived from the lossless Lempel-Ziv data compression. The principle of the scheme consists in considering approximate pattern matching where no more than $D\%$ of mismatches are allowed.

An algorithm is considered to be lossless when $D = 0$. For example Hoffman's algorithm and the Lempel-Ziv algorithm are lossless. Such algorithms are extensively used for text or data transmission or storage every time it is required to have error-free recovery. In this case the compression is limited by information theory. With image or voice/sound compression, there is no need of exact recovery since the noise in the record and/or the limited sensitivity of our eyes or ears will hide the details of the data base. In this case the compression can be limitless, depending only on the degree of *fidelity* one wants to keep in the recovery. Examples of lossy algorithms are JPEG, GIF, and MPEG (for motion pictures), they are based on adaptation of Fourier or wavelet transform, or on self-similarity search as in fractal compression.

The new lossy algorithm can be adapted to numerous applications as image or voice compression. This universality of use simply comes from the fact that the new algorithm proceeds on the digital transcription of the data regardless of their origin. In particular it can be adapted to image compression provided some tuning. An adaptation for voice/sound is under study.

The scheme on image shows performance close to JPEG algorithms and outperforms fractal compression. More importantly, it benefits of a much simpler "on line" decompression algorithm. Another advantage is that the new algorithm is tractable to performance analysis when the database (the text or the image to compress) follows a stochastic model.

2. Measure of fidelity

Before describing the algorithm we will introduce the performance measurement called fidelity. Let x be a text of length n ($|x| = n$). On the transmitter side the compression algorithm encodes x into $c(x)$. The compression rate is the ratio $|c(x)|/n$. With lossless algorithms the average compression rate, $E|c(x)|/n$ cannot be better than the entropy h of the source from which the database is built. In general the lossless algorithms asymptotically attain this theoretical bound when $n \to \infty$. The better the algorithm is, the faster is the convergence:

$$\lim_{n\to\infty} E|c(x)|/n = h.$$

On the receiver side, the code c is decompressed into $\phi(c)$. With lossless compression $\phi(c(x)) = x$. With lossy compression $\phi(c(x)) = \hat{x}$ which is of the same length as x ($|\hat{x}| = |x| = n$) but in general differs from x. In the following, we use the Hamming distance: $d(x, \hat{x})$ is number of mismatches between x and $\hat{x}$, divided by n. We can also accommodate our results to more sophisticated distances where mismatches have different weight per pair of symbols.

3. Lossy Lempel-Ziv compression Algorithm

Let x be a text. We denote x_n the nth suffix of x (starting at position n) and x^n the nth prefix of x (ending at position n). We denote x_i^j the subword starting at position i and ending at position j.

The algorithm is a parsing algorithm. We suppose that at step k the text has been parsed up to position n, i.e. x^n has been compressed into $c(x^n)$. The step $k+1$ will consist in finding the largest prefix x_n^{n+j} of x_n which is a copy within distance D of a substring in x^n. Assume this copy is at position i in x^n. Therefore the new parsed position is $n+j$, and the encoded text is $c(x^n)$ plus the pair (i, j): $c(x^{n+j}) = c(x^n).$"(i, j)". The substring x_n^{n+j} is called the new parsed phrase and j is its length.

4. Results

4.1. Rate-distortion measure. Let A^n be the set of all sequences of length n and let S be a subset of A^n. We call $P(S)$ the probability weight of S in A^n.

The optimal compression ratio depends on the rate-distortion function $R(D)$, which is defined as follows. Let w be a text of length n, we define $B_D(w)$ as the D-ball of center w, i.e. $B_D(w) = \{x : d(x, w) \leq D\}$. We define $N(D, S)$ as the minimum number of D-ball needed to cover S. Then:

$$R_n(D, \varepsilon) = \min_{S \subset A^n, P(S) \geq 1-\varepsilon} \frac{\log N(D, S)}{n},$$

and the rate-distortion is defined as $R(D) = \lim_{\varepsilon \to 0, n \to \infty} R_n(D, \varepsilon)$.

4.2. Generalized entropy. The generalized b-order Rényi entropy $h_b(D)$ is defined as follows:

$$h_b = \lim_{n \to \infty} \frac{-\log E[P^b(B_D(x)) \mid |x| = n]}{bk} = \lim_{n \to \infty} \frac{-\log \sum_{x \in A^n} P^b(B_D(x)) P(\{x\})}{bk}.$$

For $b \to 0$ we understand $h_0(D) = \lim_{n \to \infty} E[-\log P_{(}B_D(x)) \mid |x| = n]/k$, provided the limit exists.

When $D = 0$ (lossless case) we naturally recover the known b-order entropies $h^{(b)}$ defined by $E[-P(\{x\}) \log P(\{x\}) \mid |x| = n]$.

4.3. Asymptotic results on lossy Lempel-Ziv. Under some probabilistic model (Bernoulli, Markov, Mixing conditions), about the already parsed part of the text x^n we can obtain the following result.

THEOREM 1. *The length of the new parsed phrase L_n satisfies:*

$$\lim_{n \to \infty} \frac{L_n}{\log n} = \frac{1}{h_0(D)}$$

The convergence is in probability and/or almost sure convergence.

For the Bernoulli model we prove that $r_0(D)$ is the compression rate of the lossy Lempel-Ziv scheme and that $\lim_{D \to 0} R(D) = \lim_{D \to 0} h_0(D) = h$. In the case of binary uniform database we have $h_0(D) = R(D)$

THEOREM 2. *In the Bernoulli model, the lossy Lempel-Ziv algorithm is asymptotically optimal when $D \to 0$ and is asymptotically optimal for all D in the binary uniform Bernoulli model.*