Lempel-Ziv code

1 LZ code and switch distribution

Definition 1 (LZ code) *For simplicity of the algorithm description we assume that the compressed data are binary sequences, i.e., the input alphabet is* $\mathbb{X} = \{0, 1\}$. *The Lempel-Ziv compression algorithm is as follows.*

1. *The compressed sequence is parsed into a sequence of shortest phrases that have not appeared before (except for the last phrase). For example, the sequence* 001010010011100... *is split into phrases* $0, 01, 010, 0100, 1, 11, 00, ...$

2. *In the following, each phrase is described using a binary index of the longest prefix that appeared earlier and a single bit that follows that prefix. For the considered sequence, this representation is as follows:* $(0, 0)(1, 1)(10, 0)(11, 0)(0, 1)(101, 1)(1, 0)$.

Now let the input alphabet be $\mathbb{X} = \{0, 1, ..., D-1\}$. Let the frequency of substring $w_1^k \in \mathbb{X}^k$ in string $z_1^n \in \mathbb{X}^n$ be

$$c(w_1^k | z_1^n) = \sum_{i=0}^{n-k} \mathbf{1}\{w_1^k = z_{i+1}^{i+k}\}.$$

Definition 2 (R measure) *Define conditional probabilities* $B(x_{n+1}|x_1^n, -1) = D^{-1}$ *and*

$$B(x_{n+1}|x_1^n, k) = \frac{c(x_{n+1-k}^{n+1}|x_1^n) + B(x_{n+1}|x_1^n, k-1)}{c(x_{n+1-k}^{n}|x_1^{n-1}) + 1}.$$

We write $B(x_1^n, k) = \prod_{i=1}^n B(x_i|x_1^{i-1}, k)$. *Let* $p_k \in (0, 1)$ *satisfy* $\sum_{k=-1}^{\infty} p_k = 1$. *The R measure is*

$$Q(x_1^n) = \sum_{k=-1}^{\infty} p_k B(x_1^n, k).$$

2 Task

1. Download some texts, DNA sequences, or other discrete symbolic sequences (e.g., music in an appropriate format) in a sufficient amount (say, about 1MB) from the internet.

2. Write a program that computes the Lempel-Ziv code for a text and another program that computes probability $[-\log B(x_1^n, k)]$ for a text x_1^n.

3. Estimate the entropy rate of natural language as the length of the Lempel-Ziv code for the text divided by the text length. Compare this estimate with estimates of the entropy rate given by $\frac{1}{n}\left[-\log B(x_1^n, k)\right]$ for $k = -1, 0, 1, ..., 10$. Which quantity is the lowest?

4. Describe what you have obtained in a report, attach the used scripts, and send it to me (`ldebowsk@ipipan.waw.pl`).