Information Theory and Statistics
Lecture 4: Lempel-Ziv code

Łukasz Dębowski
ldebowsk@ipipan.waw.pl

Ph. D. Programme 2013/2014
Entroppy rate is the limiting compression rate

**Theorem**

For a stationary process \((X_i)_{i=-\infty}^{\infty}\), let \(L_n\) denote the minimal expected compression rate of a uniquely decodable code \(B_n : \mathbb{X}^n \rightarrow \{0, 1\}^*\) for the block of \(n\) variables. That is,

\[
L_n := \min_{B_n} \frac{1}{n} \mathbb{E} |B_n(X_1, \ldots, X_n)|.
\]

We claim that \(\lim_{n \to \infty} L_n = h\).

**Proof**

Assume that \(B_n\) is the Shannon-Fano code for the block \((X_1, \ldots, X_n)\). Then

\[
H(X_1^n) \leq nL_n \leq \mathbb{E} |B_n(X_1, \ldots, X_n)| \leq H(X_1^n) + 1.
\]

Hence the claim follows.
The problem of universal compression

- To compute the Shannon-Fano code we need to know the probability distribution of the block.
- Such a situation is unlikely in practical applications of data compression, where we have no prior information about the probability distribution of blocks.
- Fortunately, as an important corollary of the ergodic theorem, there exist universal codes whose compression rates tend to the entropy rate for any stationary process.
Universal codes

Definition (weakly universal code)
A uniquely decodable code $B : \mathbb{X}^* \rightarrow \{0, 1\}^*$ is called weakly universal if for any stationary process (not necessarily ergodic) we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E} |B(X_1^n)| = h.$$ 

Definition (strongly universal code)
A uniquely decodable code $B : \mathbb{X}^* \rightarrow \{0, 1\}^*$ is called strongly universal if for any stationary ergodic process inequality

$$\limsup_{n \rightarrow \infty} \frac{1}{n} |B(X_1^n)| \leq h.$$ 

holds with probability 1.
Strongly universal codes are better

\textbf{Theorem}

Let code $B$ be strongly universal. If there exists a constant $K$ such that

$$|B(x^n)| \leq Kn$$

for each string $x^n$ then code $B$ is weakly universal.
The problem of universal compression falls under the scope of statistics. Indeed, the interest of statisticians lies in identifying parameters of a stochastic process basing on the data typical for that process. Entropy rate of an ergodic process is an example of such a parameter. When we have a universal code, we may estimate the entropy rate as the compression rate.
Lempel-Ziv code

The code was derived by Abraham Lempel (1936–) and Jacob Ziv (1931–) in 1977 and is partly implemented in gzip and compress.

## Definition (LZ code)

For simplicity of the algorithm description we assume that the compressed data are binary sequences, that is $X = \{0, 1\}$. 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).
The length of the LZ code

- Let $C_n$ be the number of phrases in the compressed block $X_1^n$. If we know $C_n$, we need $\log C_n$ bits to identify the prefix index for each phrase and 1 bit to describe the following bit. Thus the LZ code uses $|B(X_1^n)| = C_n \left[ \log C_n + O(1) \right]$ bits in total.

- A splitting of a sequence into distinct phrases will be called a *distinct parsing* of the sequence.

**Theorem**

Let $(X_i)_{i=-\infty}^{\infty}$ be a stationary ergodic process and let $C_n$ be the number of phrases in a distinct parsing of block $(X_1, X_2, \ldots, X_n)$. With probability 1 we have

$$\limsup_{n \to \infty} \frac{C_n \left[ \log C_n + O(1) \right]}{n} \leq h.$$ 

**Remark:** Hence the LZ code is strongly universal. It can be also shown that the LZ code is weakly universal.
The first lemma

**Lemma**

The number of phrases $C_n$ in any distinct parsing of block $(X_1, X_2, \ldots, X_n)$ satisfies inequality

$$\lim_{{n \to \infty}} \frac{C_n \log n}{n} \leq 1.$$
Proof of the first lemma

Let \( n_k = \sum_{j=1}^{k} j2^j = (k - 1)2^{k+1} + 2 \) be the sum of lengths of distinct phrases that are not longer than \( k \). The number of phrases \( C_n \) in a distinct parsing will be maximal if the phrases are as short as possible. For \( n_k \leq n < n_{k+1} \) this happens if we take all phrases of length \( \leq k \) and \( \delta/(k + 1) \) phrases of length \( k + 1 \), where \( \delta = n - n_k \). Then

\[
C_n \leq \sum_{j=1}^{k} 2^j + \frac{\delta}{k + 1} \leq \frac{n_k}{k - 1} + \frac{\delta}{k + 1} \leq \frac{n}{k - 1}.
\]

In the following we will provide a bound for \( k \) given \( n \). We have \( n \geq n_k = (k - 1)2^{k+1} + 2 \geq 2^k \), so \( k \leq \log n \). Moreover \( n < n_{k+1} = k2^{k+2} + 2 \leq (\log n + 2)2^{k+2} \). Hence

\[
k + 2 > \log \frac{n}{\log n + 2}.
\]

Further transformations yield \( k - 1 > \log n - \log(\log n + 2) - 3 \). Hence we obtain the claim.
Ziv inequality

Let $P^k$ denote the measure of the $k$-th order Markov approximation of the process $(X_i)_{i=-\infty}^{\infty}$. That is

$$P^k(X_{n-k+1}^n|X_0^{0-k+1}) := \prod_{i=1}^{n} P(X_i|X_{i-k}^{i-1}).$$

Moreover, assume that sequence $(X_1, X_2, \ldots, X_n)$ is parsed into $C_n$ distinct phrases $(Y_1, Y_2, \ldots, Y_{C_n})$. Let $W_i$ denote the $k$ bits preceding $Y_i$. Next, let $C_{l,w}^n$ denote the number of phrases $Y_i$ that have length $l$ and context $W_i = w$.

Lemma (Ziv inequality)

We have inequality

$$- \log P^k(X_1, X_2, \ldots, X_n|W_1) \geq \sum_{l,w} C_{l,w}^n \log C_{l,w}^n.$$
Proof of Ziv inequality

Observe that

\[
- \log P^k(X_1, X_2, \ldots, X_n|W_1) = - \sum_{j=1}^{C_n} \log P(Y_j|W_j)
\]

\[
= - \sum_{l,w} C_{n}^{lw} \cdot \frac{1}{C_{n}^{lw}} \sum_{j:|Y_j|=l,W_j=w} \log P^k(Y_j|W_j)
\]

\[
\geq - \sum_{l,w} C_{n}^{lw} \log \left( \frac{1}{C_{n}^{lw}} \sum_{j:|Y_j|=l,W_j=w} P^k(Y_j|W_j) \right),
\]

where the inequality follows from the Jensen inequality because the logarithm function is concave. Because the phrases \( Y_j \) under the sum are distinct, we have \( \sum_{j:|Y_j|=l,W_j=w} P^k(Y_j|W_j) \leq 1 \). Hence the claim follows.
Third lemma

**Lemma**

Let \( \mathbf{L} \) be a nonnegative random variable taking values in integers and having expectation \( \mathbb{E} \mathbf{L} \). Then entropy \( H(\mathbf{L}) \) is bounded by inequality

\[
H(\mathbf{L}) \leq (\mathbb{E} \mathbf{L} + 1) \log (\mathbb{E} \mathbf{L} + 1) - \mathbb{E} \mathbf{L} \log \mathbb{E} \mathbf{L}.
\]

The proof of this lemma will be discussed after the lecture on maximum entropy modeling as an easy exercise.
Proof that LZ code is universal

Let $L$ and $W$ be random variables such that

$$P(L = l, W = w) = \frac{C_{lw}^n}{C_n^i}. $$

The expectation of $L$ is

$$E[L] = \sum_{l,w} l \frac{C_{lw}^n}{C_n^i} = \frac{n}{C_n^i}. $$

Hence by the third lemma, we obtain

$$H(L) \leq (E[L] + 1) \log (E[L] + 1) - E[L] \log E[L]$$

$$= \log \frac{n}{C_n^i} + \left( \frac{n}{C_n^i} + 1 \right) \log \left( \frac{C_n^i}{n} + 1 \right).$$
Proof that LZ code is universal (continued)

On the other hand, \( H(W) \leq k \), so

\[
H(L, W) \leq H(L) + H(W) \\
\leq \log \frac{n}{C_n} + \left( \frac{n}{C_n} + 1 \right) \log \left( \frac{C_n}{n} + 1 \right) + k.
\]

Then by the first lemma, we have

\[
\lim_{n \to \infty} \frac{C_n}{n} H(L, W) = 0.
\]
Proof that LZ code is universal (finished)

Now using the first lemma again, the Ziv inequality, and the ergodic theorem, we obtain

$$\limsup_{n \to \infty} \frac{C_n \left[ \log C_n + O(1) \right]}{n} = \limsup_{n \to \infty} \left( \frac{C_n \log C_n}{n} - \frac{C_n}{n} H(L, W) \right)$$

$$= \limsup_{n \to \infty} \frac{1}{n} \left( C_n \log C_n + C_n \sum_{l,w} \frac{C_{lw}}{C_n} \log \frac{C_{lw}}{C_n} \right)$$

$$= \limsup_{n \to \infty} \frac{1}{n} \sum_{l,w} C_{lw} \log C_{lw} \leq - \lim_{n \to \infty} \frac{1}{n} \log P^k(X_1^n|X_0^{n-k+1})$$

$$= - \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \log P(X_i|X_{i-k}^{i-1}) = H(X_i|X_{i-k}^{i-1}).$$

with probability 1. This inequality holds for any $k$. Considering $k \to \infty$, we obtain the claim.
Motivation for the R measure

- We can estimate the entropy rate of a stationary process as the length of the Lempel-Ziv code for a sequence of symbols drawn for the process divided by the sequence length.

- This method of estimation is far from satisfactory since the estimate of the entropy rate converges very slow as a function of the sequence length.

- We will present a method which seems to be better in case of empirical sources such as the natural language.
Let the frequency of substring $w_1^k$ in string $z_1^n \in \{0, 1, \ldots, D - 1\}^n$ be

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

**Definition (R measure)**

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

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

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

$$Q(x_1^n) = \sum_{k=-1}^{\infty} p_k B(x_1^n, k).$$
A probability distribution $Q$ is called *weakly universal* if for any stationary process (not necessarily ergodic) we have

$$\lim_{n \to \infty} \frac{1}{n} \mathbb{E} \left[ - \log Q(X_1^n) \right] = h.$$ 

A probability distribution $Q$ is called is called *strongly universal* if for any stationary ergodic process inequality

$$\limsup_{n \to \infty} \frac{1}{n} \left[ - \log Q(X_1^n) \right] \leq h.$$ 

holds with probability 1.

**Theorem**

*The R measure is both strongly and weakly universal.*
Proof of universality

Let \( P \) be a stationary ergodic distribution. Since the alphabet of \( X_i \) is finite, by the ergodic theorem differences \( B(X_n | X_1^{n-1}, k) - P(X_n | X_1^{n-1}) \) converge to 0 with \( P \)-probability 1. Hence

\[
\lim_{n \to \infty} \frac{1}{n} \left[ - \log B(X_1^n, k) \right] = \lim_{n \to \infty} \frac{1}{n} \left[ - \sum_{i=k+1}^{n} \log P(X_i | X_{i-k-1}^{i-1}) \right].
\]

Applying the ergodic theorem again, we obtain

\[
\lim_{n \to \infty} \frac{1}{n} \left[ - \sum_{i=k+1}^{n} \log P(X_i | X_{i-k-1}^{i-1}) \right] = E \left[ - \log P(X_{k+1} | X_1^k) \right].
\]

Hence

\[
\limsup_{n \to \infty} \frac{1}{n} \left[ - \log Q(X_1^n) \right] \leq \inf_{k \in \mathbb{N}} \lim_{n \to \infty} \frac{1}{n} \left[ - \log B(X_1^n, k) \right]
\]

\[
= \inf_{k \in \mathbb{N}} E \left[ - \log P(X_{k+1} | X_1^k) \right] = h.
\]

Hence the distribution \( Q \) is strongly universal. Since \(- \log Q(X_1^n) \leq - \log p - 1 + n \log D\), distribution \( Q \) is also weakly universal.
The R measure is effectively computable

Denote the maximal length of a substring that appears at least twice in $z_1^n$ as

$$L(z_1^n) := \max \left\{ k : \exists w_1^k : c(w_1^k|z_1^n) > 1 \right\}.$$  

For $k > L(x_1^n)$, 

$$B(x_1^n, k) = B(x_1^n, k - 1).$$

Hence the R measure is

$$Q(x_1^n) = \sum_{k=-1}^{L(z_1^n)} p_k B(x_1^n, k) + \left( 1 - \sum_{k=-1}^{L(x_1^n)} p_k \right) B(x_1^n, L(x_1^n)).$$