

LZ, MPM, GBC and GNU!

Decoding the title:

LZ ⇒ Lempel-Ziv
MPM ⇒ Multilevel Pattern Matching
GBC ⇒ Grammar Based Codes

So in a sense we do employ grammar based encoding every day, when we use acronyms.

GNU ⇒ **GNU** is Not Unix
 ⇒ **GNU** is Not Unix is Not Unix
 ⇒ **GNU** is Not Unix is Not Unix is Not Unix
 ⇒ . . .

Not all grammars are good for data compression, however!

Short Abstract

Why are Lempel-Ziv codes still an interesting research area, and what questions do the study of them rise?

- They solve a well defined mathematical problem.
- Possible to prove theorems about them both with and without probabilistic assumptions. Theorems frequently hard to prove.
- Optimal in more than one way. Links probabilistic and non-probabilistic setting.
- Not optimal in all aspects: Inspire search for new codes.
- Widely used. Performs well in practice.

Lempel-Ziv Algorithm

Idea: Parse the string to be encoded according to the following rule

The next phrase is the shortest new phrase

and then encode each phrase.

Example: The sequence

ababbabaaabaaabba

is parsed into

a | b | ab | ba | baa | aba | aa | bb | a

Observation: Each phrase is a previous phrase plus one new symbol. So we encode each phrase by a pair (i, s) where i is an index and $s \in A$:

<i>a</i>	<i>b</i>	<i>ab</i>	<i>ba</i>	<i>baa</i>	<i>aba</i>	<i>aa</i>	<i>bb</i>	<i>a</i>
<i>0, a</i>	<i>0, b</i>	<i>1, b</i>	<i>2, a</i>	<i>4, a</i>	<i>3, a</i>	<i>1, a</i>	<i>2, b</i>	<i>1, λ</i>
00000	00001	00011	00100	01000	00110	00010	00101	0001*

Codelength: We find

$$|\text{codeword}| \simeq c(x_1^n) (\log c(x_1^n) + \log |A|)$$

where $c(x_1^n)$ is the number of phrases of x_1^n .

Conclusion: We have $\varphi_{LZ} : A^* \rightarrow B^*$, a code.

Finite State Compressibility

Problem: Find a code which is universal with respect to all finite state encoders.

Lempel-Ziv codes is such a code. And the performance bound we derive is closely related to entropy!

Definition: Let $\varphi : A^* \rightarrow B^*$ be a code. Let

$$\rho_{\text{FS}(s)}(x_1^n) = \min_{\varphi \in \text{FS}(s)} \frac{|\varphi(x_1^n)|}{n}$$

be the minimal compression ratio possible when using finite state encoders of no more than s states. Finally, let $\mathbf{x} \in A^\infty$ and define

$$\rho_{\text{FS}(s)}(\mathbf{x}) = \limsup_{n \rightarrow \infty} \rho_{\text{FS}(s)}(x_1^n)$$

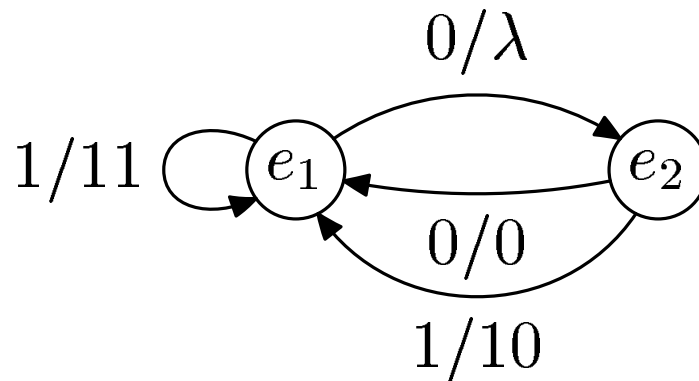
and

$$\rho(\mathbf{x}) = \lim_{s \rightarrow \infty} \rho_{\text{FS}(s)}(\mathbf{x})$$

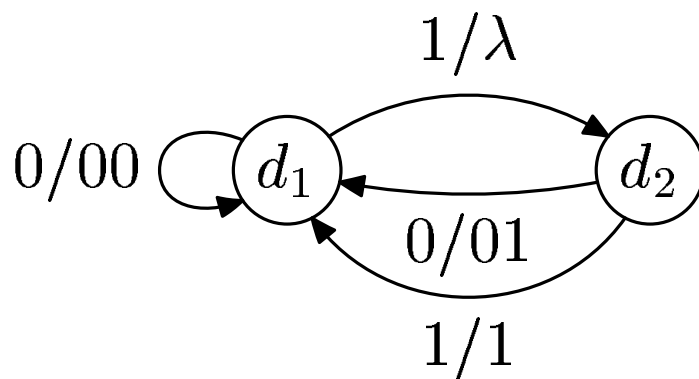
which is known as the *compressibility* of \mathbf{x} .

Example of an FS encoder

The encoder:



The decoder:



Example:

01	1	00	00	01	00	00		13 bits
10	11	0	0	10	0	0		10 bits

So we conclude, that

$$\rho_{\text{FS}(2)}(0110000010000) \leq 10/13.$$

Lempel-Ziv is universal FS encoder

Proposition: The Lempel-Ziv encoder may be implemented as a finite state encoder using $s(n)$ states in order to encode x_1^n .

Theorem 1 For every $n \in \mathbb{N}$ and $s \in \mathbb{N}$ and for every $x_1^n \in A^n$

$$\frac{|\varphi_{LZ}(x_1^n)|}{n} = \rho_{FS(s)}(x_1^n) + \delta_s(n)$$

with $\delta_s(n) \rightarrow 0$ as $n \rightarrow \infty$.

Furthermore, let $\mathbf{x} \in A^\infty$ and write

$$LZ_N(\mathbf{x}) = \limsup_{m \rightarrow \infty} \frac{1}{mN} \sum_{i=0}^{m-1} |\varphi_{LZ}(x_{iN+1}^{iN+N})|$$

then

$$\limsup_{N \rightarrow \infty} LZ_N(\mathbf{x}) = \rho(\mathbf{x}).$$

Proof: See [Ziv and Lempel, 1978, Theorem 2].

Probabilistic setup

Let μ be a stationary, ergodic measure on (A^∞, \mathcal{F}) .

A *probabilistic source* is measure; roughly speaking, $\mu(x_1^n)$ is the probability of observing $x_1^n \in A^n$.

Being *stationary* means that if $x_1^n = y_2^{n+1}$, then $\mu(x_1^n) = \mu(y_2^{n+1})$. That is, the probabilistic law does not change over 'time'.

Being *ergodic* means that the measure is concentrated (has support) on sequences with identical statistical properties.

Entropy: What is the entropy of μ (provided it is stationary)?

$$H(\mu) \stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} \frac{1}{n} \left(\sum_{a_1^n \in A^n} \mu(a_1^n) \log \frac{1}{\mu(a_1^n)} \right)$$

The AEP: Also known as the Entropy theorem:

$$h(\mu) = \lim_{n \rightarrow \infty} \frac{-\log \mu(x_1^n)}{n} \quad \text{almost surely}$$

Lempel-Ziv code is universal

Theorem 2 *Let μ be a stationary, ergodic measure on (A^∞, \mathcal{F}) . Then*

$$\mu\left(\{\mathbf{x} \in A^\infty \mid \rho(\mathbf{x}) = H(\mu)\}\right) = 1.$$

Proof: See e.g. [Ziv and Lempel, 1978, Theorem 4].

It follows, that the Lempel-Ziv algorithm compresses to entropy if block-size is big enough.

Lempel-Ziv variations

The two best known variations on Lempel-Ziv code is the LZ78 from [Ziv and Lempel, 1978] and the LZ77 from [Ziv and Lempel, 1977].

We have worked with LZ78 so far. The differences lies in the parsing.

LZ78 The next phrase is the shortest new block that has not been seen as a phrase previously.

LZ77 The next phrase is the shortest new block that does not start somewhere in the past.

Example (LZ78/SLZ):

a, b, ab, ba, baa, aba, aa, bb, a

Example (LZ77):

a, b, abb, abaa, abaaabb, a

I know how to spell 'banana', but I don't know when to stop.

– A little girl

Second order considerations

Second order analysis is study of subclasses of stationary, ergodic class.

Redundancy: Set

$$R_n = \frac{1}{n}(E(|\varphi_{LZ}(\cdot)|) - H_n(\mu))$$

where E denotes expectation and H_n is the n 'th order entropy.

In [Plotnik et al., 1992] it was shown that if μ is finite order Markov then

$$R_n = O\left(\frac{\log \log n}{\log n}\right)$$

and in [Louchard and Szpankowski, 1997] it was proven that if μ is a Bernoulli source, then

$$R_n = O\left(\frac{1}{\log n}\right)$$

These results are *not* flattering for the Lempel-Ziv code! In both cases a convergence at the speed of $O(\log n/n)$ is possible if one knows the source μ .

Tea-Time-Time-Line

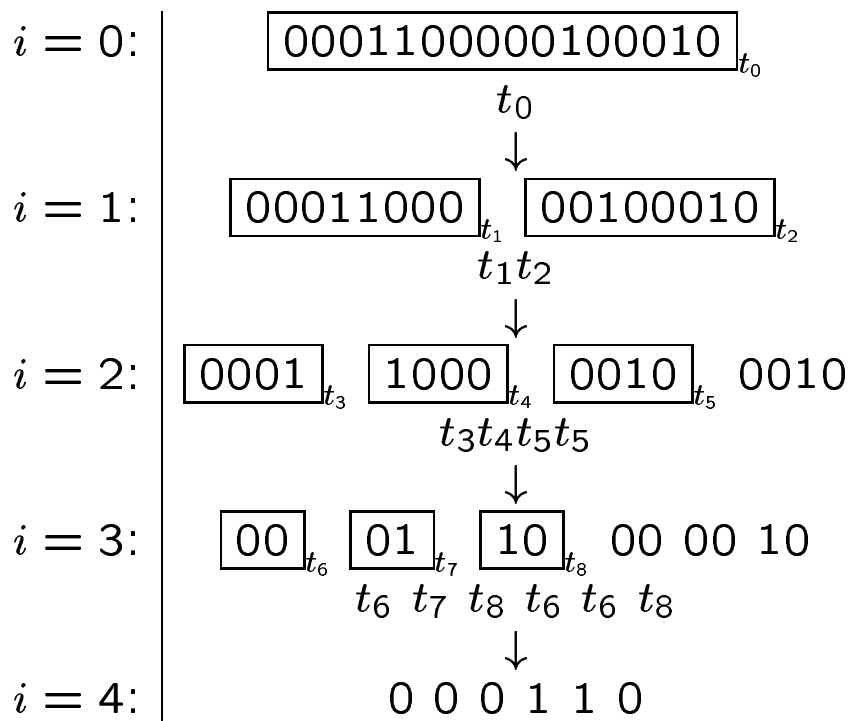
- 1976 The concept of complexity of individual sequences $x \in A^\infty$ is considered in [Ziv and Lempel, 1976].
- 1977 The LZ77 code emerge in [Ziv and Lempel, 1977].
- 1978 The LZ78 code is published in [Ziv and Lempel, 1978] and proven to be universal on the class of stationary ergodic sources as well as universal when competing with finite state encoders.
- 1983 Based on ideas in [Rissanen and Langdon, 1981] it is shown in [Langdon, 1983] that the Lempel-Ziv parsing (LZ78) can be considered as a sequential probability assignment.
- 1984 The well known LZW variation appears in [Welch, 1984].
- 1986 The *Context* algorithm is proposed in [Rissanen, 1986, 1983] as an attempt to fix some of the weaknesses in LZ78. It is somewhat twice-universal.
- 1991-4 New proofs of the universality on the stationary ergodic sources of LZ78/77 emerge in [Cover and Thomas, 1991] (LZ78), [Ornstein and Weiss, 1993] (LZ77 and LZ78), and [Wyner and Ziv, 1994] (LZ77).
- 1992 Second order results begin to appear. In [Plotnik et al., 1992] the subclass of finite state sources is considered.
- 1992 In the award winning paper [Feder et al., 1992] the ideas of using Lempel-Ziv parsing for probability assignment is employed as a method for universal prediction.
- 1997 Finally, [Louchard and Szpankowski, 1997] solves the problem of second order analysis of LZ78 applied to a Bernoulli source! Simultaneously the problem for Markov sources is solved by [Savari, 1997].
- 1998 Savari continues her work, this time on LZ77 and a number of other variants, in [Savari, 1998].
- 1999 A comparison of the different variations in the non-probabilistic case is carried out in [Shields, 1999].
- 2000 In a sequence of articles, [Kieffer and Yang, 2000; Kieffer et al., 2000], Kieffer and Yang explore the new class of GBC, and proposes a new code, known as MPM.
- Present Ryabko and Suzuki, (and Shields et al.) compares the performance of codes on infinite sequences and uses Hausdorff dimension to estimate the size of these sets. Still many open problems on second order analysis of MPM.

The Multilevel Pattern Matching (MPM) code

We introduce this by an example: We wish to encode

$$x_1^n = 0001100000100010$$

The algorithm works in $\log n + 1$ steps. Each step consist of a *splitting phase* and a *tokenization phase*.



The MPM code - II

From the codeword,

000110 , $t_6t_7t_8t_6t_6t_8$, $t_3t_4t_5t_5$, t_1t_2 , t_0

it is possible to construct the following grammar:

$$\begin{aligned}t_6 &\rightarrow 00 \\t_7 &\rightarrow 01 \\t_8 &\rightarrow 10 \\t_3 &\rightarrow t_6t_7 \\t_4 &\rightarrow t_8t_6 \\t_5 &\rightarrow t_6t_8 \\t_1 &\rightarrow t_3t_4 \\t_2 &\rightarrow t_5t_5 \\t_0 &\rightarrow t_1t_2\end{aligned}$$

And from this we derive the message:

$$t_0 \rightarrow t_1t_2 \rightarrow t_3t_4t_5t_5 \rightarrow$$
$$t_6t_7t_8t_6t_6t_8t_6t_8 \rightarrow 0001100000100010$$

Results on MPM code

Theorem 3 *For any stationary, ergodic measure μ*

$$\lim_{n \rightarrow \infty} \frac{1}{n} E_{\mu}(\varphi_{MPM}(\cdot)) = H(\mu)$$

Proof: See [Kieffer et al., 2000].

Second order analysis: For the class of finite state sources, the redundancy has faster decay than Lempel-Ziv, namely

$$R_n(MPM) = O\left(\frac{1}{\log n}\right)$$

where Lempel-Ziv has

$$R_n(LZ) = O\left(\frac{\log \log n}{\log n}\right)$$

Still many open problems, e.g. MPM performance on Bernoulli class.

Lempel-Ziv is also a GBC

Consider LZ78 encoding of

$a, b, ab, ba, baa, aba, aa, bb, a$

Recalling that each phrase is encoded using a pair (i, s) where i is an index pointing to a previous phrase, and s a symbol, we find the following grammar expresses the workings of Lempel-Ziv:

$$\begin{aligned} z_0 &\rightarrow z_1 z_2 z_3 z_4 z_5 z_6 z_7 z_8 z_9 \\ z_1 &\rightarrow a \\ z_2 &\rightarrow b \\ z_3 &\rightarrow z_1 b \\ z_4 &\rightarrow z_2 a \\ z_5 &\rightarrow z_4 a \\ z_6 &\rightarrow z_3 a \\ z_7 &\rightarrow z_1 a \\ z_8 &\rightarrow z_2 b \\ z_9 &\rightarrow z_1 \end{aligned}$$

A true Grammar Based Code result

If G is a *grammar transformation*

$$\lim_{n \rightarrow \infty} \max_{x_1^n \in A^n} \frac{|G(x_1^n)|}{n} = 0$$

we say that the grammar based code induced by G is *asymptotically compact*..

Lempel-Ziv codes and MPM codes are asymptotically compact.

Theorem 4 *Let μ be a stationary, ergodic measure on A^∞ . If a grammar based code φ is asymptotically compact, then*

$$\limsup_{n \rightarrow \infty} E_\mu(|\varphi(\cdot)|) = H(\mu).$$

Bibliography

- Cover, T. M. and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.
- Feder, M., N. Merhav and M. Gutman. Universal prediction of individual sequences. *IEEE Trans. Inform. Theory*, 38(4):1258–1270, 1992.
- Kieffer, J. and E.-H. Yang. Grammar-based codes: A new class of universal lossless source codes. *IEEE Trans. Inform. Theory*, 46:737–754, 2000.
- Kieffer, J., E.-H. Yang, Gregory Nelson and Pamela Cosman. Universal lossless compression via multilevel pattern matching. *IEEE Trans. Inform. Theory*, 46, 2000.
- Langdon, G.G. A note on the Ziv-Lempel model for compressing individual sequences. *IEEE Trans. Inform. Theory*, 29(2):284–287, 1983.
- Louchard, G. and W. Szpankowski. On the average redundancy rate of the Lempel-Ziv code. *IEEE Trans. Inform. Theory*, 43(1):1–8, 1997.
- Ornstein, Donald Samuel and Benjamin Weiss. Entropy and data compression schemes. *IEEE Trans. Inform. Theory*, 39(1):78–83, 1993.
- Plotnik, E., M.J. Weinberger and J. Ziv. Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel-Ziv algorithm. *IEEE Trans. Inform. Theory*, 38:66–72, 1992.
- Rissanen, J. A universal data compression system. *IEEE Trans. Inform. Theory*, 29:656–664, 1983.
- Rissanen, J. Complexity of strings in the class of Markov sources. *IEEE Trans. Inform. Theory*, 32(4):526–532, 1986.
- Rissanen, J. and G.G. Langdon. Universal modeling and coding. *IEEE Trans. Inform. Theory*, 27(1):12–23, 1981.
- Savari, S.A. Redundancy of the Lempel-Ziv incremental parsing rule. *IEEE Trans. Inform. Theory*, 43(1):9–21, 1997.
- Savari, S.A. Redundancy of the Lempel-Ziv string matching code. *IEEE Trans. Inform. Theory*, 44(2):787–792, 1998.
- Shields, P.C. Performance of LZ algorithms on individual sequences. *IEEE Trans. Inform. Theory*, 45(4):1283–1288, 1999.

- Welch, T.A. A technique for high-performance data compression. *IEEE Computer*, 17:8–19, 1984.
- Wyner, A.D. and J. Ziv. The sliding-window Lempel-Ziv is asymptotically optimal. *Proc. IEEE*, 82:872–877, 1994.
- Ziv, J. and A. Lempel. On the complexity of finite sequences. *IEEE Trans. Inform. Theory*, 22:75–81, 1976.
- Ziv, J. and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory*, 23(3):337–343, 1977.
- Ziv, J. and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory*, 24(5):530–536, 1978.