Lecture III: The Lempel-Ziv Algorithms

- A family of data compression algorithms presented in

  [LZ77] J. Ziv and A. Lempel, A universal algorithm for sequential data compression, IEEE Trans. Inform.Theory, vol. IT-23, pp. 337 – 343, May 1977

  [LZ78] J. Ziv and A. Lempel, Compression of individual sequences via variable rate coding, IEEE Trans. Inform. Theory, vol. IT-24, pp. 530 – 536, Sept. 1978.

- Many desirable features, the conjunction of which was unprecedented
  - simple and elegant
  - universal for individual sequences in the class of finite-state encoders
  - convergence to the entropy rate
  - string matching and dictionaries, no explicit probability model
  - very practical, with fast and effective implementations applicable to a wide range of data types

Two Main Variants

[LZ77] and [LZ78] present different algorithms with common elements

- The main mechanism in both schemes is pattern matching: find string patterns that have occurred in the past, and compress them by encoding a reference to the previous occurrence

- Both schemes are in wide practical use
  - many variations exist on each of the major schemes
  - we focus on LZ78, which admits a simpler analysis with a stronger result. Our proof follows [CT91]. It differs from the original proof in [LZ78]
  - we will also describe the [LZ77], and see a fundamental result of [Wyner&Ziv] providing insight into its workings
  - the scheme is based on the notion of incremental parsing

Incremental Parsing and the LZ78

- Parse the input sequence into phrases, each new phrase being the shortest substring that has not appeared so far in the parsing. E.g., for the string $x^n = 1011010100010$

$$1, 0, 11, 01, 010, 00, 10,$$

- Each new phrase is of the form $\mathbf{w}b$, where $\mathbf{w}$ is a previous phrase, $b \in \{0, 1\}$
  - a new phrase can be described as (i, b), where $i =$ index($\mathbf{w}$)
  - in the example: $(0, 1), (0, 0), (1, 1), (2, 1), (4, 0), (2, 0), (1, 0)$
  - let $c(n) =$ number of phrases in x^n
  - a phrase description takes $\leq 1 + \log c(n)$ bits
  - here describing 13 bits took us 28 but gets better as $n \to \infty$
  - another small overhead to indicate how many bits per description of phrase (in practice use increasing length codes)
  - So, all in all, bounding generously, the compression ratio attained is $\leq \frac{c(n)(\log c(n)+2)+\log n}{n}$

Performance Analysis

Lemma 1. *The number of phrases $c(n)$ in a distinct parsing of a binary sequence satisfies*

$$c(n) \leq \frac{n}{(1-\epsilon_n)\log n}, \tag{1}$$

where $\epsilon_n \to 0$ as $n \to \infty$.

Proof Idea: Letting n_k denote the sum of lengths of all distinct strings of length $\leq k$ and $k(n)$ denote the distinct value of k such that $n_k \leq n < n_{k+1}$, we show that for any distinct parsing

1. $c(n) \leq n/(k(n) - 1)$.

2. $k(n) = (1 \pm \epsilon_n)(\log n)$.

Ziv's Inequality

For fixed k let $P(\cdot|\cdot)$ be an arbitrary conditional distribution of X_0 given X_{-k}^{-1}. Define the probability distribution Q_k on X^n conditioned on $X_{-(k-1)}^0$ by

$$Q_k(x^n|x_{-(k-1)}^0) = \prod_{j=1}^{n} P(x_j|x_{j-k}^{j-1}). \tag{2}$$

Suppose now that x^n is parsed into c distinct phrases $y_1, y_2, \ldots, y_c$

Let ν_i be the index of the start of the i-th phrase, i.e., $y_i = x_{\nu_i}^{\nu_{i+1}-1}$

For each $i = 1, 2, \ldots, c$, define $s_i = x_{\nu_i-k}^{\nu_i-1}$

Thus s_i is the k bits of x preceding y_i

Let c_{ls} be the number of phrases y_i with length l and preceding state $s_i = s$ for $l = 1, 2, \ldots$ and $s \in \mathcal{X}^k$. So

$$\sum_{l,s} c_{ls} = c \quad \text{and} \quad \sum_{l,s} l c_{ls} = n. \tag{3}$$

Lemma 2. [Ziv's inequality] *For any distinct parsing of the string x^n*

$$\log Q_k(x^n|s_1) \leq -\sum_{l,s} c_{ls} \log c_{ls}.$$

Note right side does not depend on $P(\cdot|\cdot)$ through which Q_k was defined.
Proof:

$$\begin{aligned}
\log Q_k(x^n|x_{-(k-1)}^0) &= \sum_{i=1}^{c} \log Q_k(y_i|s_i) \\
&= \sum_{l,s} \sum_{i:|y_i|=l, s_i=s} \log Q_k(y_i|s_i) \\
&= \sum_{l,s} c_{ls} \sum_{i:|y_i|=l, s_i=s} \frac{1}{c_{ls}} \log Q_k(y_i|s_i) \\
&\leq \sum_{l,s} c_{ls} \log \left(\sum_{i:|y_i|=l, s_i=s} \frac{1}{c_{ls}} Q_k(y_i|s_i) \right) . \square
\end{aligned}$$

Maximum-Entropy Lemma

Lemma 3. *Let Z be a positive integer valued random variable with mean μ. Then*

$$H(Z) \leq (\mu + 1) \log(\mu + 1) - \mu \log \mu. \tag{4}$$

Proof: The maximum-entropy distribution over the positive integers under a constraint on the mean is the geometric one. The right hand side of (4) is readily checked to be the entropy of the geometric distribution with mean μ.
$\square$

Ziv's Inequality (another one)

Lemma 4. [Ziv's Inequality] *For all $\mathbf{x} \in \{0,1\}^\infty$*

$$\frac{c(n) \log c(n)}{n} \leq -\frac{1}{n} \log \max_{P \in \mathcal{P}_k} Q_k(x^n|x_{-(k-1)}^0) + \epsilon_k(n),$$

where $\epsilon_k(n) \to 0$ as $n \to \infty$ (uniformly in $\mathbf{x} \in \{0,1\}^\infty$).

Proof: Fix $P \in \mathcal{P}_k$ through which $Q_k(x^n|x_{-(k-1)}^0)$ is defined. By Ziv's inequality

$$\begin{aligned}
\log Q_k(x^n|x_{-(k-1)}^0) &\leq -\sum_{l,s} c_{ls} \log \frac{c_{ls} c}{c} && (5) \\
&= -c \log c - c \sum_{l,s} \frac{c_{ls}}{c} \log \frac{c_{ls}}{c}. && (6)
\end{aligned}$$

Denoting $\pi_{ls} = \frac{c_{ls}}{c}$, we have

$$\sum_{l,s} \pi_{ls} = 1, \quad \sum_{l,s} l\pi_{ls} = \frac{n}{c}. \tag{7}$$

Thus, defining the random variables U, V such that

$$\Pr(U = l, V = s) = \pi_{ls} \tag{8}$$

we have

$$EU = \frac{n}{c} \tag{9}$$

and, by (6),

$$-\frac{1}{n}\log Q_k(x^n|x^0_{-(k-1)}) \geq \frac{c}{n}\log c - \frac{c}{n}H(U,V). \tag{10}$$

Now

$$\begin{aligned} H(U) \quad &\leq \quad (EU+1)\log(EU+1) - EU\log EU & (11) \end{aligned}$$

$$\begin{aligned} &= \left(\frac{n}{c}+1\right)\log\left(\frac{n}{c}+1\right) - \frac{n}{c}\log\frac{n}{c} & (12)\\ &= \log\frac{n}{c} + \left(\frac{n}{c}+1\right)\log\left(\frac{c}{n}+1\right). & (13) \end{aligned}$$

Thus

$$\begin{aligned} & \frac{c}{n}H(U,V) & (14)\\ \leq\ & \frac{c}{n}(H(U)+H(V)) & (15)\\ \leq\ & \frac{c}{n}\log\frac{n}{c} + \left(\frac{c}{n}+1\right)\log\left(\frac{c}{n}+1\right) + \frac{c}{n}k & (16)\\ \leq\ & \epsilon_k(n), & (17) \end{aligned}$$

where (17) follows from Lemma 1 upon denoting

$$\epsilon_k(n) = -\frac{1}{(1-\epsilon_n)\log n}\log\frac{1}{(1-\epsilon_n)\log n} \tag{18}$$

$$+\left(\frac{1}{(1-\epsilon_n)\log n}+1\right)\log\left(\frac{1}{(1-\epsilon_n)\log n}+1\right) + \frac{k}{(1-\epsilon_n)\log n}. \tag{19}$$

Note, in particular, that

$$\epsilon_k(n) = O\left(\frac{\log\log n}{\log n}\right), \tag{20}$$

independently of x^n and $P \in \mathcal{P}_k$. The proof is completed by combining (10) with (17) and the arbitrariness of $P \in \mathcal{P}_k$. $\square$

The Key Result

Theorem 1. *Let $l(x^n)$ denote the Ziv-Lempel codeword length associated with x^n. Then, for all $\mathbf{x} \in \{0,1\}^\infty$,*

$$\limsup_{n\to\infty}\frac{1}{n}l(x^n) \leq \lim_{k\to\infty}\limsup_{n\to\infty}\left[-\frac{1}{n}\log\max_{P\in\mathcal{P}_k} Q_k(x^n|x^0_{-(k-1)})\right]. \tag{21}$$

Proof: The result is a direct consequence of the fact that $l(x^n) \leq c(n)(\log c(n) + 2) + \log n$, combined with Lemma 1 and Lemma 4. $\square$

Equipped with Theorem 1, the universality result in the stochastic setting is but a simple corollary:

Pointwise Universality of the LZ scheme

Corollary 1. *Let* $\mathbf{X} = \{X_i\}$ *be a stationary ergodic source. Then the Lempel-Ziv code satisfies*

$$\lim_{n\to\infty} \frac{1}{n} l(X^n) = \overline{H}(\mathbf{X}) \quad a.s. \tag{22}$$

Proof: For P denoting the true distribution of X_0 conditioned on X_{-k}^{-1} we have, with probability one,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} l(X^n) &\leq \limsup_{n\to\infty} \left[-\frac{1}{n} \log \max_{P\in\mathcal{P}_k} Q_k(X^n|X^0_{-(k-1)}) \right] & (23)\\
&\leq \limsup_{n\to\infty} \left[-\frac{1}{n} \sum_{i=1}^{n} \log P(X_i|X_{i-k}^{i-1}) \right] & (24)\\
&= H(X_0|X_{-k}^{-1}), & (25)
\end{aligned}$$

where the first inequality follows from Theorem 1, and the equality by ergodicity. The arbitrariness of k implies

$$\limsup_{n\to\infty} \frac{1}{n} l(X^n) \leq \overline{H}(\mathbf{X}) \quad a.s., \tag{26}$$

which, combined with exercise 2 of HW sheet 2, completes the proof. □

Universality for Individual Sequences

Another easy consequence of Theorem 1 is universality in the individual sequence setting. Define the *finite-memory compressibility*:

$$FM_k(x^n) = \inf_{P\in\mathcal{P}_k, s_1} \left[-\frac{1}{n} \log Q_k(x^n|s_1) \right]$$

$$FM_k(\mathbf{x}) = \limsup_{n\to\infty} FM_k(x^n)$$

$$FM(\mathbf{x}) = \lim_{k\to\infty} FM_k(\mathbf{x})$$

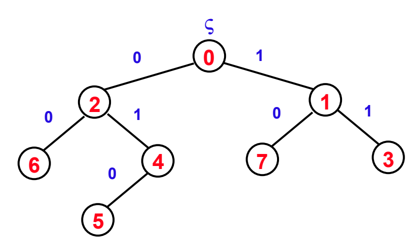
Corollary 2. *For all* $\mathbf{x} \in \{0,1\}^\infty$*, the LZ codeword lengths satisfy*

$$\limsup_{n\to\infty} \frac{1}{n} l(x^n) \leq FM(\mathbf{x}). \tag{27}$$

[LZ78] introduces a stronger notion of *finite-state compressibility* and shows that the LZ scheme attains that as well.

The Parsing Tree

x_1^n = **1,0,1 1,0 1,0 1 0,0 0,1 0 , ...**



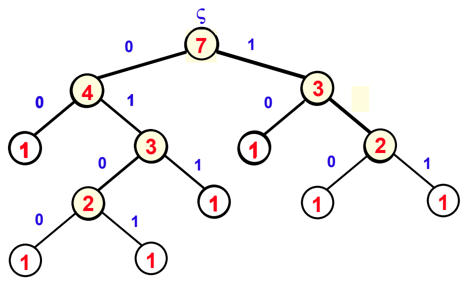
code	phrase
0	ς
1	0,1
2	0,0
3	1,1
4	2,1
5	4,0
6	2,0
7	1,0
*	*

dictionary

- **coding could be made more efficient by "recycling" codes of nodes that have a complete set of children (e.g., 1, 2 above)**
- **will not affect asymptotics**
- **many (many many) tricks and hacks exist in practical implementations**

The LZ Probability Assignment

x_1^n = 1,0,1 1,0 1,0 1 0, ...



- Slightly different tree evolution *anticipatory parsing*
- A *weight* is kept at every node
  - number of times the node was traversed through + 1
- A node act as a conditioning state, assigning to its children probabilities proportional to their weight
- Example: string s=101101010
  - $P(0|s)$ = 4/7
  - $P(1|s0)$ = 3/4
  - $P(1|s01)$ = 1/3
  - $P(011|s)$ = (4/7)*(3/4)*(1/3) = 1/7
  - *Notice `telescoping'*
- $P(s011)$ = 1/7!

- In general,

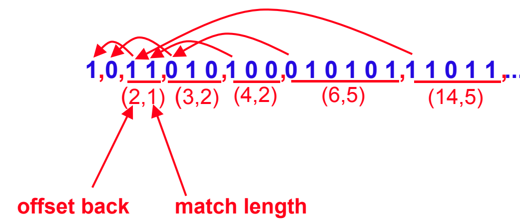
$$P(x_1^n) \approx \frac{1}{(c(n)+1)!}$$

$$-\log P \approx c(n)\log c(n) + o(c(n)\log c(n))$$ **LZ code length!**

every lossless compression algorithm defines a prob. assignment, even if it wasn't meant to!

LZ77

- ***Exhaustive parsing*** **as opposed to** ***incremental***
  - a new phrase is formed by the longest match *anywhere in a finite past window,* plus the new symbol
  - a pointer to the location of the match, its length, and the new symbol are sent
- **Has a weaker proof of universality, but actually works better in practice**



Analysis of LZ77

Think of X_{-n}^{-1} as a database. Then look into "positive time" at $X_0, X_1, \ldots$ and continue until the L-string X_0^{L-1} is *not* a substring of the extended database X_{-n}^{L-2}. Denote that L by $L_n(\mathbf{X})$.

In the [LZ77], if we set time to zero at the beginning of a new phrase after the algorithm has finished encoding the first n source symbols then the length of the new phrase will be $\stackrel{d}{\approx} L_n$, where the approximate (and not precise) relationship is due to the randomness in the time-shift.

Thus, the compression ratio on the new block is $\leq$

$$\frac{1}{L_n}\left(\log n + \log L_n + O(\log\log L_n) + \log(|\mathcal{A}|-1)\right).$$

Fundamental Result in Analysis of LZ77

Wyner and Ziv, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression", IEEE Trans. Info. Theory, vol. IT-35, pp. 1250 – 1258, November 1989.

Theorem 2. [WZ89] *For stationary ergodic* $\mathbf{X}$

$$\frac{\log n}{L_n} \to \overline{H}(\mathbf{X}) \quad \textit{in probability.} \tag{28}$$

Almost sure convergence in (28) was later established by:

Ornstein and Weiss, "Entropy and data compression schemes", IEEE Trans. Info. Theory, vol. IT-39, pp. 78 – 83, January 1993.

Theorem 2 can be restated in terms of waiting times as

Theorem 3. [WZ89] *Let $\mathbf{X}$ be stationary ergodic and define the random variable N_l as the smallest $N > 0$ such that*

$$X_0^{l-1} = X_{-N}^{-N+l-1}.$$

Then

$$\frac{1}{l}\log N_l \to \overline{H}(\mathbf{X}) \quad \text{in probability.} \tag{29}$$

Equivalence of theorems derives from the equivalence of events

$$\{N_l > n\} = \{L_n \le l\}.$$

Intuition can be gained via Kac's lemma. For stationary ergodic $\mathbf{Y}$, $Y_i \in \mathcal{B}$, $|\mathcal{B}| < \infty$ let

$$Q_k(b) = \Pr(Y_k = b; Y_j \ne b, 1 \le j \le k-1 | Y_0 = b)$$

and let

$$\mu(b) = \sum_{k=1}^{\infty} k Q_k(b)$$

denote the expected recurrence time for the symbol $b \in \mathcal{B}$.

Lemma 5. [Kac]

$$\mu(b) = 1/\Pr\{Y_0 = b\}.$$

Applied to our case Kac's lemma implies

$$E[N_l | X_0^{l-1} = x_0^{l-1}] \approx 2^{l(\overline{H}(\mathbf{X}) \pm \epsilon)}$$

or

$$\frac{\log E[N_l | X_0^{l-1} = x_0^{l-1}]}{l} \approx \overline{H}(\mathbf{X}) \pm \epsilon$$

for all typical x_0^{l-1}, which resembles (29).

See also

[A. Dembo and I. Kontoyiannis. "The asymptotics of waiting times between stationary processes, allowing distortion," Ann. Appl. Probab., 9, pp. 413-429, May 1999]

and references therein for analogues of Theorem 3 when distortion is allowed.