Lecture III: The Lempel-Ziv Algorithms

- A family of data compression algorithms presented in

- 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 $x^n = 10110101000101$, $1, 0, 11, 010, 00, 10$.

- Each new phrase is of the form $w b$, where $w$ is a previous phrase, $b \in \{0, 1\}$
  - a new phrase can be described as $(i, b)$, where $i = \text{index}(w)$
  - in the example: $(0, 1), (0, 0), (1, 1), (2, 1), (4, 0), (2, 0), (1, 0)$
  - let $c(n) = \text{number of phrases in } x^n$
  - 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
    \[ c(n) (\log c(n) + 2) + \log n \]

Performance Analysis

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

\[ c(n) \leq \frac{n}{(1 - \epsilon_n) \log n}, \]  

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 $X_{-(k-1)}$ by

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

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

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

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 $s \in X^k$. So

$$\sum_{l,s} c_{ls} = c \text{ and } \sum_{l,s} lc_{ls} = n. \hspace{1cm} (3)$$

---

**Maximum-Entropy Lemma**

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

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

**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 $\mu$. \hfill \Box

---

**Ziv’s Inequality (another one)**

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

$$\log Q_k(x^n|x_{-(k-1)}) \leq - \sum_{l,s} c_{ls} \log c_{ls}.$$ 

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

**Proof:**

$$\log Q_k(x^n|x_{-(k-1)}) = \sum_{i=1}^c \log Q_k(y_i|s_i)$$

$$= \sum_{i,s} \sum_{\nu_i:y_i=x_{\nu_i}} \log Q_k(y_i|s_i)$$

$$= \sum_{i,s} c_{ls} \sum_{\nu_i:y_i=x_{\nu_i}} \frac{1}{c_{ls}} \log Q_k(y_i|s_i)$$

$$\leq \sum_{i,s} c_{ls} \log \left( \sum_{\nu_i:y_i=x_{\nu_i}} \frac{1}{c_{ls}} Q_k(y_i|s_i) \right). \Box$$

---

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

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

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

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

$$\log Q_k(x^n|x_{-(k-1)}) \leq -\sum_{l,s} c_{ls} \log \frac{c_{ls} \epsilon^{c_{ls} \epsilon}}{c}$$

$$= -c \log c - c \sum_{l,s} \frac{c_{ls}}{c} \log \frac{c_{ls}}{c}. \hspace{1cm} (5)$$

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

$$\sum_{l,s} \pi_{ls} = 1, \sum_{l,s} l \pi_{ls} = \frac{n}{c}.$$  \hspace{1cm} (7)

Thus, defining the random variables $U, V$ such that

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

we have

$$EU = \frac{n}{c}.$$ \hspace{1cm} (9)

and, by (6),

$$-\frac{1}{n} \log Q_k(x^n|x^n_0|(k-1)) \geq \frac{c}{n} \log c - \frac{c}{n} H(U, V).$$ \hspace{1cm} (10)

Now

$$H(U) \leq (EU + 1) \log(EU + 1) - EU \log EU \leq EU + 1 \log EU + 1 - EU \log EU.$$ \hspace{1cm} (11)

Note, in particular, that

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

independently of $x^n$ and $P \in \mathcal{P}_k$.

Thus

$$\frac{c}{n} H(U, V) \leq \frac{c}{n} (H(U) + H(V)) \leq \frac{c}{n} \log \frac{n}{c} + \frac{c}{n} \log \left(\frac{c}{n} + 1\right) + \frac{c}{n} k.$$ \hspace{1cm} (14)

and

$$\frac{c}{n} H(U, V) \leq \frac{c}{n} \log \frac{n}{c} + \frac{c}{n} \log \left(\frac{c}{n} + 1\right) + \frac{c}{n} k.$$ \hspace{1cm} (15)

where (17) follows from Lemma 1 upon denoting

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

and

$$\epsilon_k(n) = -\frac{1}{n} \log \left(\frac{1}{(1 - \epsilon_n) \log n} + 1\right) \log \left(\frac{1}{(1 - \epsilon_n) \log n} + 1\right) + \frac{k}{(1 - \epsilon_n) \log n}.$$ \hspace{1cm} (19)

The Key Result

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

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

Equipped with Theorem 1, the universality result in the stochastic setting is but a simple corollary:
Corollary 1. Let $X = \{X_i\}$ be a stationary ergodic source. Then the Lempel-Ziv code satisfies
\[
\lim_{n \to \infty} \frac{1}{n} l(X^n) = \overline{H}(X) \quad a.s.
\] (22)

Proof: For $P$ denoting the true distribution of $X_0$ conditioned on $X_{-k}$ we have, with probability one,
\[
\limsup_{n \to \infty} \frac{1}{n} l(X^n) \leq \limsup_{n \to \infty} \left[ -\frac{1}{n} \log \max_{P \in \mathcal{P}_k} Q_k(X^n|X^{0_{(k-1)}}) \right]
\leq \limsup_{n \to \infty} \left[ -\frac{1}{n} \sum_{i=1}^{n} \log P(X_i|X_i^{i-1}_k) \right]
= H(X_0|X_{-k})
\] (25)

Universality for Individual Sequences

Another easy consequence of Theorem 1 is universality in the individual sequence setting. Define the finite-memory compressibility:
\[
FM_k(x^n) = \inf_{P \in \mathcal{P}_k,s_1} \left[ -\frac{1}{n} \log Q_k(x^n|s_1) \right]
\]

\[
FM_k(x) = \limsup_{n \to \infty} FM_k(x^n)
\]

\[
FM(x) = \lim_{k \to \infty} FM_k(x)
\]

Corollary 2. For all $x \in \{0, 1\}^\infty$, the LZ codeword lengths satisfy
\[
\limsup_{n \to \infty} \frac{1}{n} l(x^n) \leq FM(x).
\] (27)

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

The Parsing Tree

\[
x_i^n = 1,0,1,1,0,1,0,0,0,1,0,\ldots
\]

\[
\begin{array}{c}
\text{code} \quad \text{phrase} \\
0 & 1 \\
1 & 0,1 \\
2 & 0,0 \\
3 & 1,1 \\
4 & 2,1 \\
5 & 4,0 \\
6 & 2,0 \\
7 & 1,0 \\
\end{array}
\]

\[
\begin{array}{c}
\text{dictionary} \\
0 & 2 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & \ldots \end{array}
\]

where the first inequality follows from Theorem 1, and the equality by ergodicity. The arbitrariness of $k$ implies
\[
\limsup_{n \to \infty} \frac{1}{n} l(X^n) \leq \overline{H}(X) \quad a.s.,
\] (26)

which, combined with exercise 2 of HW sheet 2, completes the proof. □
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 $s = 101101010$

$P(0|s) = 4/7$  
$P(1|s0) = 3/4$  
$P(1|s01) = 1/3$  
$P(011|s011) = (4/7) \times (3/4) \times (1/3) = 1/7$

Notice 'telescoping'

In general,

$P(x^n) = \frac{1}{(c(n)+1)!}$

$LZ$ code length!

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

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

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.

Theorem 2. [WZ89] For stationary ergodic $X$

$$\frac{\log n}{L_n} \to H(X) \quad \text{in probability.}$$

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

Ornstein and Weiss, “Entropy and data compression schemes”, IEEE
Theorem 2 can be restated in terms of waiting times as

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

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

Then

$$\frac{1}{l} \log N_l \to \mathcal{H}(X) \text{ in probability.}$$

Equivalence of theorems derives from the equivalence of events

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

Intuition behind Theorem 3

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

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

and let

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

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

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

or

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

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

See also


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