6.896 Lecture 09

Note Title 3/9/2006

TODAY: LEMPEL ZIV ALGORITHM

+ Analysis

REFERENCES: - BOB GALLAGER'S NOTES 2/17/94

- COVER & THOMAS, Sec. 12.10

Review:

Last time: Introduced Universal Compression.

$C : \{0,1\}^* \to \{0,1\}^*$ is a universal compressor if $\forall$ ergodic Markov sources $\mathcal{X}$,

$$\lim_{n\to\infty} \frac{E\left[\,|C(x_1 \ldots x_t)|\,\right]}{t} \longrightarrow H(\mathcal{X}).$$

Recall:

Markov Sources: – State space : $\{1 \dots n\}$

– Trans. Prob. Mat $= \{P_{ij}\}_{i,j \in \{1..n\}}$

– Output sequence $f : \{1..n\} \times \{1..n\} \to \{0,1\}$.

– Start start $i_0 \in \{1 \dots n\}$.

Defines process

$X_1 \;..\; X_t \dots$ using the rule

$Y_0 = i_0$; $\Pr\{Y_t = j \mid Y_{t-1} = i, Y_{t-2} \dots Y_0\} = P_{ij}$

$X_t = f(Y_{t-1}, Y_t)$.

C-

Chain is (i) Periodic if every path of pos. prob. from $i \to \dots \to i$ is of length a multiple of c.

(2) Reducible if $\exists$ states i, j s.t. there is no path of pos. prob. from i to j.

M.C. is "nice" if <u>not</u> periodic

& <u>not</u> irreducible

"nice" M.C. $\iff$ Ergodic

general notion saying M.C. forgets where it started.

<u>Entropy Rate</u>:

$$H(\mathcal{X}) = \lim_{t \to \infty} \left\{ H(X_t \mid X_{t-1} \dots X_0) \right\}$$

AEP for Ergodic M.C.

(Typical Set Theorem)

$\forall\ t \quad \exists\ T_t \in \{0,1\}^t$

s.t. • $\Pr\left[(X_1 \ldots X_t) \in T_t\right] \to 1$

• $\dfrac{\log |T_t|}{t} \to H(\mathcal{X})$

• $\forall (x_1 \ldots x_t) \in T_t \quad \dfrac{-\log \Pr\left[(x_1 \ldots x_t)\right]}{\log |T_t|} \to 1$;

Algorithm from last time

$\text{SHABBY}_{L,k}(X_1 \ldots X_t)$

- Construct $(Y_1 \ldots Y_{t'}) \in \left\{\{0,1\}^L\right\}^{t'}$

where $t' = \dfrac{t}{L}$; $Y_i = (X_{i\cdot L+1} \ldots X_{(i+1)\cdot L})$

Count $\leftarrow 0$;

- Let $Z \in \{0,1\}^{2^L}$ be as follows

$$Z_w = 1 \text{ if } \#\{j \text{ s.t. } Y_j = w\} \geq k$$
$$= 0 \text{ o.w.}$$

if $Z_w = 1$ then $I(w) = \text{count}++;$

- For $j = 1$ to t' do

Let $w = Y_j$.

if $Z_w = 1$ $\quad U_j = (Z_w, I(w));$

else $\quad U_j = (Z_w, w);$

- Report $Z_1 \ldots Z_{2^L}$; $U_1 \ldots U_{t'}$

Shabby $(X_1 \ldots X_t)$

Pick (L,k) so as to minimize

$$|\text{Shabby}_{(L,k)}(X_1 \ldots X_t)|$$

& output $\text{Shabby}_{L,k}(X_1 \ldots X_t)$

Theorem:

& for $k = 2^{-(H(X))(1+\epsilon) \cdot L} \cdot \frac{t}{L}$

$$\lim_{\epsilon \to 0} \left\{ \lim_{L \to \infty} \left\{ \lim_{t \to \infty} \left\{ \frac{E\left[|\text{Shabby}(\bar{x})| \right]}{t} \right\} \right\} \right\} \to H(X)$$

Proof:

$$E\left[|\text{Shabby}(\bar{x})| \right] = 2^{L} +$$

$$t' +$$

$$(H(X))(1+\epsilon) \cdot L \cdot \frac{t}{L}$$

(AEP Error)

$$+ \delta \cdot L \cdot \frac{t}{L}$$

$$\frac{E\left[|\text{Shabby}(\bar{x})| \right]}{t} \to H(X).$$

Ziv Lempel '78

- Dynamic / Elegant / Natural Version of Above
- No fixing of (L, k) ...

Two pass algorithm

Pass 1: Breaks $X_1 \ldots X_t$ into
(Parse) substrings $Y_1 \ldots Y_{t'}$

Pass 2: Encodes substrings using
(Encode) "previous" substrings

Parse Phase

$$\underbrace{X_1 \cdots}_{Y_1}\ \underbrace{\cdots}_{Y_2}\ X \cdots \underbrace{\cdots X_i}_{Y_j}\ \Big|\ X_{i+1} \xrightarrow{\quad ? \quad} \underbrace{X_t}_{Y_{t'}}$$

$$Y_{j+1} = X_{i+1} \cdots X_{i+L}$$

such that

(1) $Y_{j+1} \neq Y_{j'}$ for $j' \leq j$

(2) $X_{i+1} \cdots X_{i+L-1} = Y_{j'}$ for some $j' \leq j$.

Report $Y_{j+1} = (Y_{j'}, b_{j+1})$, $b_{j+1} \in \{0,1\}$

Encode: - First output $t' = 0^{\lfloor \log t' \rfloor} \cdot \underbrace{t'}_{\text{in binary}}$

- Output t' sequences of the form

$$(j', b_j) \qquad j' \in \{0,1\}^{\lceil \log t' \rceil}$$
$$b_j \in \{0,1\}$$

["(" , ")" , ";" are for our help ; not needed by compressor/decompressor]

<u>Theorem</u> : ZL'78 is a universal compressing algorithm.

<u>Proof Steps</u> :

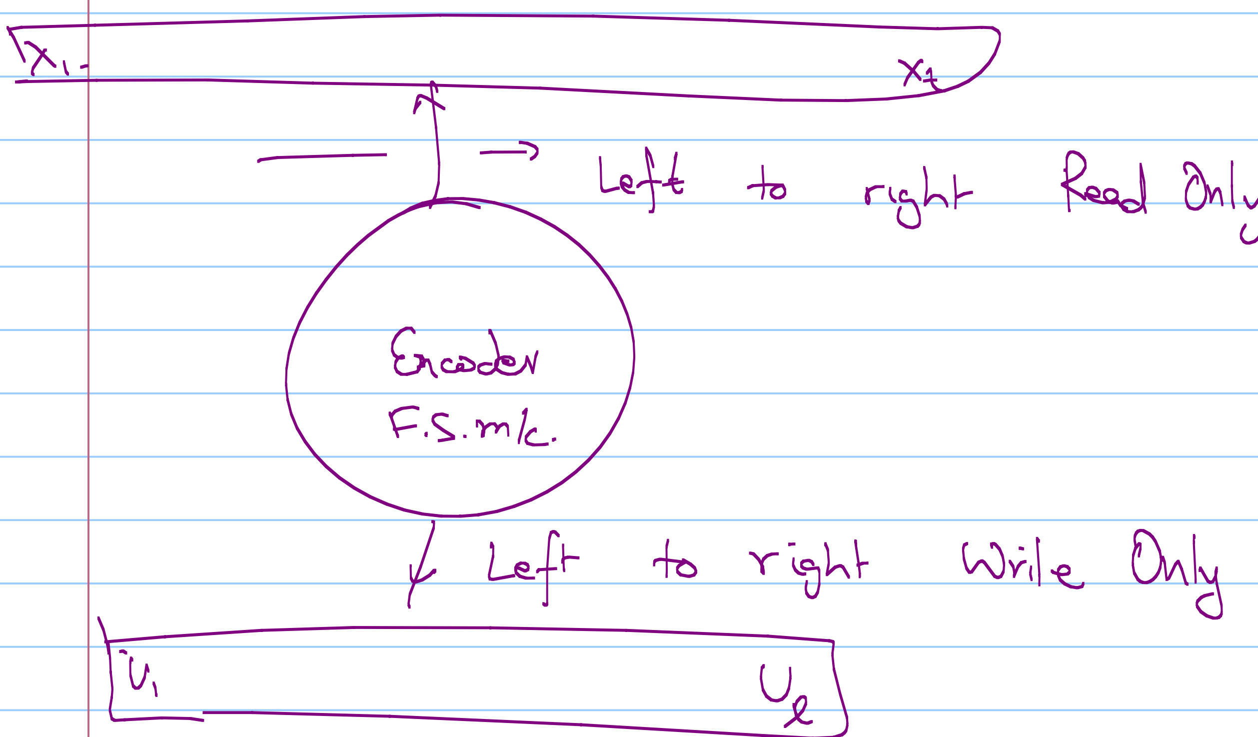
"<u>Lemma</u>" : ZL'78 performs at least as well as finite state encoders ....

"<u>Lemma</u>" : Finite State Encoders (e.g. Shabby) encode

Well.

X

What is a finite State Encoder?



Encoder in state $s \in S$

Reads $x_i \in \{0,1\}$

Writes $y_j \in \{0,1\}^*$ (poss. Empty)

$y_j = f(s, x_i)$

Moves to state $S' = g(s, x_i)$

Encoder defined by S, s_0, f, g.

---x---

Example finite state Encoders:

- Shubby$_{L,k}$ (after Z, I computed)

- ZL '78 (after Parsing).

$S = 2^L$

$S = ? \simeq t$.

Lemma 2: $\exists$ F.S. Encoders on S states

Can compress well

$$\lim_{s \to \infty} \left\{ \lim_{t \to \infty} \left\{ \frac{|C_s(\ )|}{t} \right\} \right\} \to H(\mathcal{X})$$

"Lemma 1": LZ does as well as FSE.

Definition:

$$c(\bar{x}) = \max \left\{ \ell \;\middle|\; \begin{array}{l} \exists\, Y_1 \dots Y_\ell \in \{0,1\}^* \text{ all distinct} \\ \text{s.t. } \bar{x} = Y_1 \dots Y_\ell \end{array} \right\}$$

✓ Claim 1: $\left| C_{LZ}(\bar{x}) \right| \le (1+o(1))\, c(\bar{x}) \cdot \log c(\bar{x})$

✓ Claim 2: $|\bar{x}| \geq c(\bar{x}) \cdot \log \left[\frac{c(\bar{x})}{4} \right]$

✓ Claim 3: $\left| C_s(\bar{x}) \right| \geq c(\bar{x}) \cdot \log \left[\frac{c(\bar{x})}{s^2} \right]$

Lemma 1: $\forall^{\mathcal{H},} S, \epsilon \;\exists\, t$ large enough

s.t. $\dfrac{E[\,|C_{LZ}(\bar{x})|\,]}{E[\,|C_s(\bar{x})|\,]} \le 1+\epsilon$

Proof of Claim 2 :

$Y_1 \ldots Y_\ell$ distinct

$$\Rightarrow \sum |Y_i| = \sum_k k \cdot \{\# i,\ |Y_i| = k\}$$

$$\geq \sum_{k=0}^{\lfloor \log_2 \ell \rfloor} k \cdot 2^k + \left(\ell - 2^{\lfloor \log_2 \ell \rfloor}\right)\left(\lfloor \log_2 \ell \rfloor + 1\right)$$

$$\geq \ell \log\left(\frac{\ell}{4}\right)$$ [OK ... so I didn't really work this out ...]

Proof of Claim 3

- Crucial Observation : if $\exists$ two paths from $i \to j$ in FSM then their outputs better be distinct ... or else no unique decodability

- Now given $Y_1 \ldots Y_\ell$

let $T_{ij} = \{ Y_{j'} \mid$ FSM started at i & ended at $j \}$;

$$L_{ij} = \sum_{j' \in T_{ij}} |Y_{j'}| ;$$

$$|L_{ij}| \geq |T_{ij}| \log \frac{|T_{ij}|}{4}$$

$$\sum |Y_j| \geq \sum_{ij} T_{ij} \log \frac{|T_{ij}|}{4}$$

$$\geq |S|^2 \cdot \frac{\ell}{|S|^2} \log \frac{\ell}{4|S|^2}$$

$$= \ell \log \frac{\ell}{4|S|^2} \qquad \boxtimes$$