IT & CS : CS 229r
LECTURE 5

2/12/2019

TODAY : - MARKOVIAN SOURCES
- LZ Compression Theorem

Source: My notes from Spring 2006 → $\boxed{\text{Gallager's Notes 1994}}$

RECALL MARKOVIAN SOURCES (= Hidden Markov Model):

~~$X_1 \ldots X_n \ldots$ is a hidden Markov Model~~
~~if $\exists Z \; \exists M$ ($k \times n$ state MC) &~~

- Hidden Markov Model given by matrix $M \in \mathbb{R}^{k \times k}$
  & $P_X^{(1)} \ldots P_X^{(k)}$ dist on Ω

~~s.t.~~

- Output of HMM = $X_1 \ldots X_n \ldots$ ~~are~~ generated as follows

~~Where X_i~~ ~~gene~~ $Z_1 \sim \pi(M)$
↑ stationary

$Z_{i-1} \xrightarrow{M} Z_i$

$Z_1, Z_2 \ldots Z_n \ldots$ generated according to ↑

$X_i \sim P_X^{(Z_i)}$.

[all MC's in lecture "irreducible" + "aperiodic"]

Entropy of Chain:

$$\lim_{n\to\infty} H(X_n \mid X_1 \dots X_{n-1}) \stackrel{\Delta}{=} H(M)$$

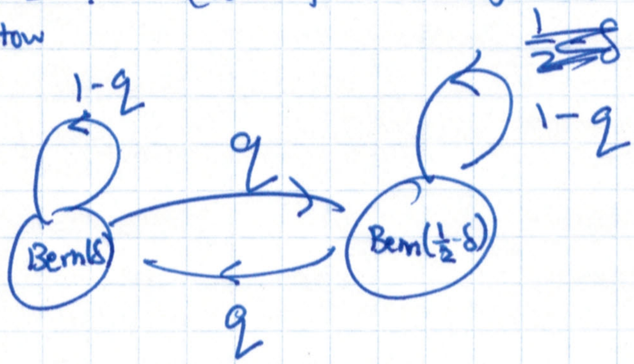
(*) Limit exists since

$$H(X_n \mid X_1 \dots X_{n-1}) \leq H(X_n \mid X_2 \dots X_{n-1})$$
$$= H(X_{n-1} \mid X_1 \dots X_{n-2})$$

} time invariance

Note limit exists but not known to be computable.

Eg. $H(M) = ?$ (as function of q, δ).
below



Theorem: $\forall$ markov chain M, $\forall \epsilon$, $\exists n_0$ $\forall n \geq n_0$

$$\Pr_{X_1 \dots X_n}\Big[\,|LZ(X_1 \dots X_n)| > (H(M)+\epsilon)\cdot n\,\Big] \leq \epsilon.$$

[Very "qualitative". Finite length analysis = "great project"]

Key Ingredients in Proof:

1. Universal-Huffman-Compressor achieves "optimal compression"
2. ⇒ Finite State Compressor compresses.
3. LZ no worse than finite state compressor.

"Universal-Huffman-Compressor" $C_{Huff,\ell}(X_1 \ldots X_n)$; $X_i \in \Omega$

1. ~~Comp for every~~
   divide $X = B_1 \ldots B_{n/\ell}$ $\quad B_i \in \Omega^\ell$
2. ~~Com~~ Compute frequencies ~~f...~~ $\{f_w\}_{w \in \Omega^\ell}$
   among $B_1 \ldots B_{n/\ell}$
3. ~~Huffman-Code ($B_1 \ldots B_{n/\ell}$;~~
   $Z_i = \text{Huffman}(B_i ; \{f_w\})$
4. Output $\{f_w\}, Z_1 \ldots Z_{n/\ell}$

Theorem': $\forall$ m.c. M, $\forall \epsilon$, $\exists n_0$ $\forall n \geq n_0$

$$\Pr_{X_1 \ldots X_n}\left[\,\left|\text{Univ.Huff}(X_1 \ldots X_n)\right| > (H(m)+\epsilon)\cdot n\,\right] < \epsilon$$

———— x ————

key insight to Proof of Theorem': AEP holds for Markov Sources

———— x ————

AEP Theorem: $\forall$ m.c. M, $\forall \epsilon$ $\exists\, l$ & set $S \subseteq \Sigma^l$ s.t. the following hold

① $|S| \leq 2^{(H(m)+\epsilon)\cdot n}$

$\cancel{② \ \frac{1}{2^{(H(m)+\epsilon)n}} \leq \Pr_{(X_1 \ldots X_l)\in S}\left[(X_1 \ldots X_l)\right]}$

② $\forall (\alpha_1 \ldots \alpha_l) \in S$

$$\frac{1}{2^{(H(m)+\epsilon)n}} \leq \Pr\left[(X_1 \ldots X_l) = (\alpha_1 \ldots \alpha_l)\right] \leq \frac{1}{2^{(H(m)-\epsilon)n}}$$

③ $\Pr_{(X_1 \ldots X_l)}\left[(X_1 \ldots X_l) \notin S\right] \leq \epsilon$.

Proof ~~to~~ of AEP Theorem: <u>Omitted</u>

<u>Ideas</u>: ~~(1) Consider States~~ $Z_\ell, Z_{2\ell}, Z_{3\ell}, \ldots$

(1) Prove for <u>Markov Chains</u> by decoupling walk into sequences that end at state (1) & do not contain (1) in between. Sequences now i.i.d.

$$\underbrace{Z_1 \ldots Z_a}_{}, Z_{a+1} \ldots Z_b, \ldots Z_c \ldots$$
$$Z_a = (1), \quad Z_b = (1), \quad Z_c = (1)$$

lengths i.i.d. $\longrightarrow$ both
Expected entropy iid $\longrightarrow$ converge $\Big\} \to$ ratio converges

(2) To get to HMM's, insert $Z_t, Z_{2t} \ldots$ at periodic intervals; results in slightly higher entropy but not much higher; now we have Markov chain!

AEP Theorem $\Rightarrow$ Theorem'

Essentially implies w.p. $1-\epsilon$ we get strings B_i of frequency

$\approx 2^{-H(m)\cdot \ell}$ $\Rightarrow$ Shannon coding will associate strings of length $\leq (H(m)+\epsilon)\cdot \ell$

w.p. ϵ use string of length $\approx \ell$

$\Rightarrow$ Expected compression length

$$= \frac{n}{\ell}\left(\epsilon + H(m)+\epsilon\right)\ell$$

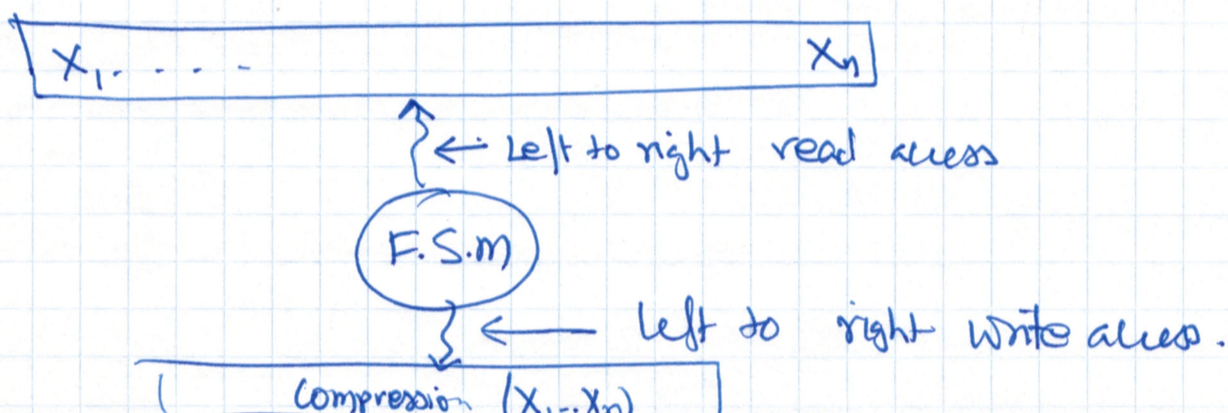
$$= (H(m)+2\epsilon)\, n. \qquad \boxtimes$$

[Full proof will involve union bound over S...]

—— x ——

Part II: Finite State Compressors:

C = finite state compressor if



Proposition: Univ.-Huffman$^{\ell}$ can be converted to ~~is~~ Ω^{ℓ}-state Finite State Compressor. [with preprocessing of M].

~~Proof: Figure out S before expected~~ (proof obvious).

~~Proof:~~

Theorem' $\Rightarrow$ M can be compressed by Ω^{ℓ}-state compressor.

Theorem'': if M compressed by S-state compressor then M compressed by Lempel-Ziv.

Key Ingredient in this proof

$$C(X_1 \ldots X_n) \triangleq \max \{ t \mid \exists Y_1 \ldots Y_t \text{ distinct in } \Omega^* \text{ s.t. } X_1 \ldots X_n = Y_1 \circ Y_2 \circ \cdots \circ Y_t \}$$

Intuition:

~~Rough Idea~~ (1) $C(X_1 \ldots X_n) \approx \underbrace{\frac{n}{\log n}}_{\text{high}} \Rightarrow X_1 \ldots X_n$ not compressible by any thing

(2) $C(X_1 \ldots X_n) =$ small $\Rightarrow$ LZ compresses well.

Formal Claims

Let $t = C(x_1 .. x_n)$ then

① ~~t~~ $n \geq t \log \frac{t}{4}$ (intermediate) [t is growing]

② $\forall$ s state compressors C^s

$$|C^s(x_1 .. x_n)| \geq t \cdot \log\left(\frac{t}{s^2}\right)$$

$$= t \log t - t \log s^2$$

$$\geq (1 - o(1))\, t \log t \quad \left[\text{if } s = O(1) \;\&\; t = \omega(1)\right]$$

③ $|C^{LZ}(x_1 ... x_n)| \leq (1 + o(1))\, t \log t$

$$\Rightarrow \quad |C^{LZ}(x_1 .. x_n)| \leq (1 + o(1)) \cdot |C^S(x_1 .. x_n)|$$

$$\leq (1 + o(1)) \cdot H(M) \cdot n \quad [\text{using Theorem}']$$

Proofs of Claims

1. Follow from # strings of length $\ell \leq \Omega^{\ell}$;

   $Y_1 \ldots Y_t$ distinct

$$\Rightarrow \sum |Y_i| = \sum_k k \cdot \{\# i \text{ s.t. } |Y_i| = k\}$$

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

$$\geq \ell \cdot \log_\Omega \ell \cdot (1 - o(1))$$

3.
$$\left|C^{LZ}(X_1 \ldots X_n)\right| \leq t\left(\lceil \log t \rceil + 1\right)$$

$$\leq t \cdot (1 + o(1)) \cdot \log t$$

2. Let $X_1 \ldots X_n = Y_1 \circ Y_2 \circ \cdots Y_t$, Y_i distinct

- Let $Y_1' \cdots Y_t'$ be outputs of FSM while traversing $Y_1 \ldots Y_t$.
- Note $Y_1' \cdots Y_t'$ need not be distinct; However (key) if Y_i & Y_j both start in state a & end in b, then $Y_i' \neq Y_j'$

- $T_{ab} \triangleq \{ j \mid$ FSM starts at a & ends at b on $Y_j \}$

- $L_{ab} \triangleq \sum_{j \in T_{ab}} |Y_j'|$

- $|L_{ab}| \geq |T_{ab}| \log \frac{|T_{ab}|}{4}$ (by ① applied to binary strings).

- $$|C^s(X_1 \ldots X_n)| = \sum_{j=1}^{n} |Y_j'|$$

$$= \sum_{a,b} \cancel{\sum} L_{ab}$$

$$\geq \sum_{a,b} |T_{ab}| \log \frac{|T_{ab}|}{4}$$

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

$$= t \cdot \log \frac{t}{S^2 \cdot 4} \qquad \boxtimes$$