Today: Markovian Sources
- L2 Compression Theorem

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

Recall Markovian Sources (\text{= Hidden Markov Model}): 

- Hidden Markov Model given by matrix \( M \in \mathbb{R}^{k \times k} \) and \( p(x_i | x_{i-1}) \), \( x \) is one of \( k \) states.

- Output of HMM = \( X_1 \ldots X_n \ldots \) where generated as follows:

Where \( Z_t \) is gene \( Z_t \sim \mathcal{U}(m) \) stationary \( Z_{i-1} \rightarrow Z_i \)

\( Z_1, Z_2, \ldots, Z_n, \ldots \) generated according to \( \uparrow \)

\( X_i \sim p(x_i | Z_i) \).

\( \text{All } \mathcal{M}_i \text{'s in lecture "irreducible" + "aperiodic"} \)
Entropy & Chain:

\[ \lim_{n \to \infty} H(X_n | X_1, \ldots, X_{n-1}) \leq H(M) \]

\( \text{dim} \quad \text{limit exists since} \)

\[ H(X_n | X_1, \ldots, X_{n-1}) = H(X_{n-1} | X_1, \ldots, X_{n-2}) \]

\( \text{time invariance} \quad \text{limit exists} \quad \text{but not known to be computable.} \)

\( \text{Eg. } H(M) = ? \) (as function of \( q, \delta \)).

\[ \begin{align*}
1-q & \quad 1-q \\
B_{\text{make}} & \quad B_{\text{flip} (\frac{1}{2}, \delta)} \\
q & \quad q
\end{align*} \]

Theorem: A markov chain \( M, \forall \epsilon, \exists n_0 \forall n \geq n_0 \)

\[ \Pr \left[ \left| \frac{1}{n} \sum_{i=1}^{n} (X_i, X_{i+1}, \ldots, X_{i+n}) \right| > (H(M) + \epsilon) \cdot n \right] \leq \epsilon. \]

[ 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" $C_{Huff, l}(x_1,...,x_n); x_i \in \Xi$

1. **Compute** for every
   divide $X = B_1,...,B_{n/l}$

2. **Compute** frequencies $f_i = \sum_{w \in \Xi^l} f(w)$
   among $B_1,...,B_{n/l}$

3. Huffman code $(B_1,...,B_{n/l})$
   $Z_i = \text{Huffman} (B_i; \sum_{w \in \Xi^l} f(w)^2)$

4. Output $Z_1,...,Z_{n/l}$
Theorem': \( \forall m \in M, \forall \epsilon > 0, \exists n_0 \forall n > n_0 \)

\[
\Pr \left[ \left| \text{Unif-Huff}(X_1, \ldots, X_n) \right| > (H(m) + \epsilon) \cdot n \right] < \epsilon
\]

---

Key insight to proof of Theorem': AEP holds for Markov sources

---

AEP Theorem: \( \forall m \in M, \forall \epsilon > 0 \) set \( S \subseteq \mathbb{Z}^n \) s.t. the following hold

1. \( |S| \leq (H(m) + \epsilon) \cdot n \)

2. \( \forall (x_1, \ldots, x_e) \in S \)

\[
\frac{1}{2^{(H(m) + \epsilon) \cdot n}} \leq \Pr \left[ (X_1, \ldots, X_e) = (x_1, \ldots, x_e) \right] \leq \frac{1}{2^{(H(m) - \epsilon) \cdot n}}
\]

3. \( \Pr \left[ (X_1, \ldots, X_e) \notin S \right] < \epsilon \).
Proof of AEPR Theorem: Omitted

Ideas:

1. Consider states $Z_1, Z_2, Z_3, \ldots$

   - Prove for Markov Chains by decoupling walk into sequences that end at state $1$ do not contain $0$ in between. Sequences now i.i.d.

   \[
   Z_1, Z_2, Z_3, \ldots, Z_n, \ldots, \quad 1, 0, 1, 
   \]

   - Length i.i.d. $\rightarrow$ both $\rightarrow$ ratio converges

   - Expected entropy iid $\rightarrow$ converges

2. To get to HTM's, insert $Z_b, Z_{2b}, \ldots$ 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. $1 - \varepsilon$ we get strings, of frequency

$\approx 2^{-H(m)} \cdot \varepsilon$ $\Rightarrow$ Shannon coding will associate string of length $\leq (H(m) + \varepsilon) \cdot \varepsilon$.

w.p. $\varepsilon$, use string of length $\approx \varepsilon \cdot l$

$\Rightarrow$ Expected compression length

$$= \frac{n}{l} (\varepsilon + H(m) + \varepsilon) \cdot l$$

$$= (H(m) + 2\varepsilon) \cdot n.$$

[Full proof will involve union bound over $\leq$...]

Part II: Finite State Compressors:

$C = \text{finite state compressor if}$

\[
\begin{array}{c}
X_1 \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots X_n \\
\end{array}
\]

\[\text{left to right read access} \]

F.S.M

\[\text{left to right write access.}\]

\[\text{Compress } (X_1, X_n)\]
Proposition: Univ.-Huffman \( L \) is an \( \Omega \)-state Finite State Compressor. [with preprocessing of \( M \)].

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

Theorem \( \Rightarrow \): \( M \) can be compressed by \( \Omega \)-state compressor.

Theorem \( \Rightarrow \): if \( M \) compressed by \( S \)-state compressor, then \( M \) compressed by Lempel-Ziv.

Key ingredient in this proof:

\[
C(X_1 \ldots X_n) \leq \max \left\{ \frac{1}{2} \left\lceil \log \frac{1}{\varepsilon} \right\rceil \mid \forall Y_1, Y_2 \text{ distinct in } \mathcal{S} \right\},
\]

\( \varepsilon > 0 \), \( X_1 \ldots X_n = Y_1 \circ Y_2 \ldots \circ Y_k \)

Intuition:

\( C(X_1 \ldots X_n) \approx \frac{n}{\log n} \) \( \Rightarrow \) \( X_1 \ldots X_n \) not compressible by any thing

\( C(X_1 \ldots X_n) \) small \( \Rightarrow \) \( LZ \) compresses well.
Formal Claims

Let $t = C_0(x_1 \ldots x_n)$ then

1. $n \geq t \log \frac{t}{4}$ 

(Intermediate) 

[It is growing]

2. $\forall s$ state compression $C^s$

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

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

$$\geq (1-o(1)) t \log t$$

[if $s = O(1)$

$\text{and } t = o(1)$]

3. $|C^{L_2}(x_1 \ldots x_n)| \leq (1+o(1)) t \log t$

$$\Rightarrow |C^{L_2}(x_1 \ldots x_n)| \leq (1+o(1)) |C^s(x_1 \ldots x_n)|$$

$$\leq (1+o(1)) \cdot H(M) \cdot n$$ [using Theorem']
Proofs of Claims

1. Follow from \# strings of length \( l \leq 5^l \);
   \( Y_1 \ldots Y_t \) distinct
   
   \[ \geq \sum_{i=1}^{t} |Y_i| = \sum_{k=0}^{\log_2 \lambda} k \sum_{i \in \mathbb{S}} 2^k \] 
   
   \[ \geq \sum_{k=0}^{\log_2 \lambda} k \cdot 2^k + (l - \frac{\log_2 \lambda}{2})(\log_2 + 1) \] 
   
   \[ \geq l \cdot \log_2 l \cdot (1-o(1)) \]

2. Let \( X_1 \ldots X_n = Y_0 Y_2 \ldots Y_t \), \( Y_i \) distinct
   
   - Let \( Y_1' \ldots Y_t' \) be outputs of FSM while traversing \( Y_1 \ldots Y_t \).
   
   - Note \( Y_1' \ldots Y_t' \) need not be distinct;

   However (key) if \( Y_j' , Y_j \) start in state a & end in b, then \( Y_j' = Y_j \).
- \( T_{ab} \triangleq \sum_{j} j \) for paths from \( a \) to \( b \) on \( Y_j \).

- \( L_{ab} \triangleq \sum_{j \in T_{ab}} \mid Y_j \mid \)

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

- \( |C^s(X_1 \ldots X_n)| = \sum_{i=1}^{n} \mid Y_i \mid \)

\[
\geq \sum_{a \leq b} \mid L_{ab} \mid \log \frac{|T_{ab}|}{4} \]

\[
\geq \frac{t}{181^2} \cdot \log \frac{t}{s^2 \cdot 4} \cdot 181^2 \]

\[
= t \cdot \log \frac{t}{s^2 \cdot 4} \quad \Box
\]