①

Last time

* KL vs. Chernoff bounds.

* Data Processing Ineq. & Markov Chains.

* Fano's inequality. (P_{err} of estimate of X from Correlated Y).

* AEP (Asymptotic Equipartition Property):

$$X_1 \ldots X_n \text{ iid } \sim p(X) \Rightarrow$$

$$-\frac{\log p(X_1, \ldots, X_n)}{n} \to H(X) \text{ in probability.}$$

* Typical Set $A_\varepsilon^{(n)} := \{ (x_1, \ldots, x_n) \mid p(x_1, \ldots, x_n) \in 2^{-n(H(X) \pm \varepsilon)} \}$

$$\Rightarrow |A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}.$$

Universal Source Coding : LZ algorithm.

[motivation]

* The algorithm: Generate a dictionary $\mathcal{D}$, constantly update it.

* Given u_1^n where $u_i \in \mathcal{U}$,
  $= u_1 \cdots u_n$

1 * Initialize $\mathcal{D}$ ~~dictionary~~ with elements of $\mathcal{U}$.

2 * Look for the longest token w in the dictionary.

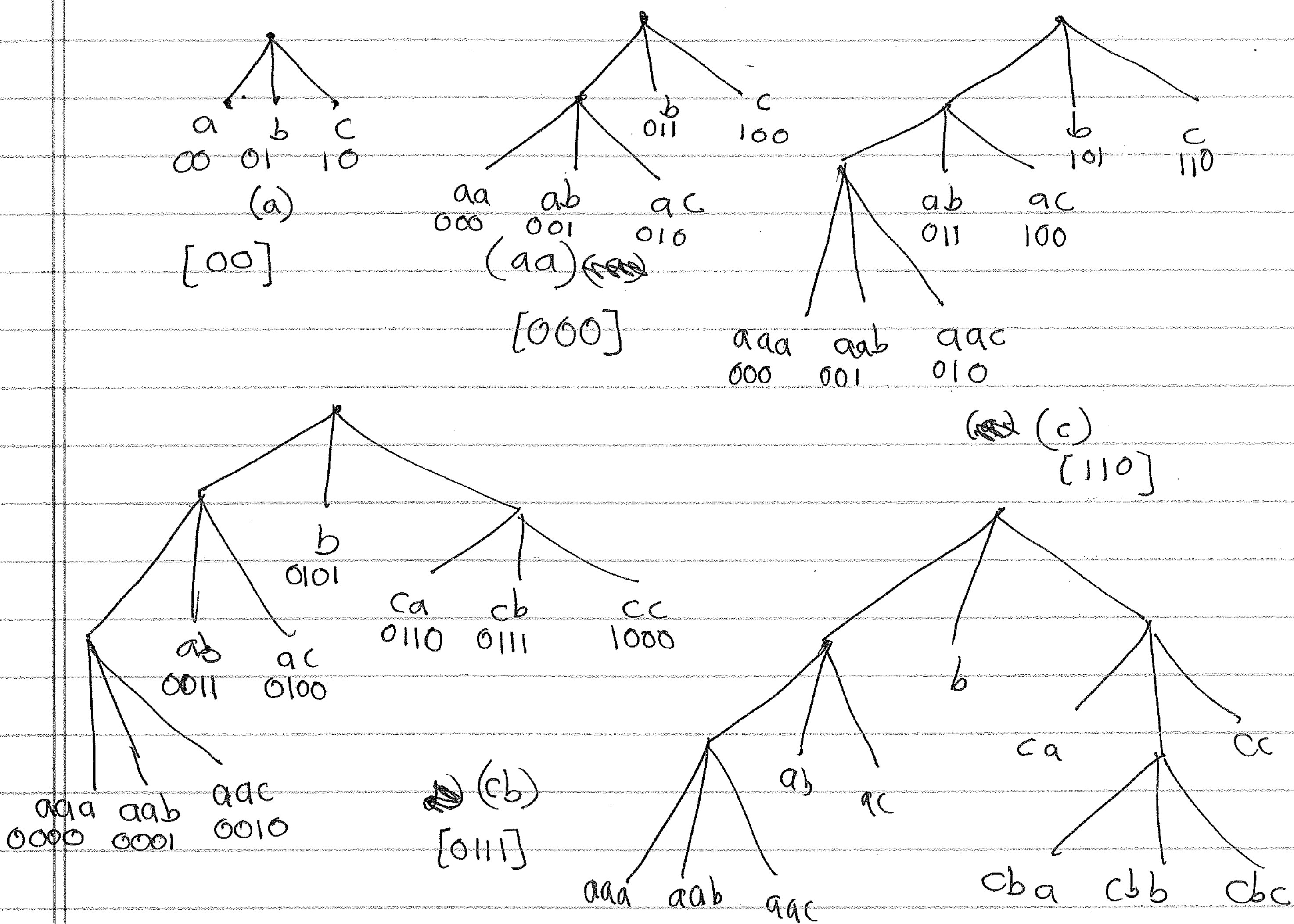
3 * Output the index of w in ~~the dictionary~~ $\mathcal{D}$ using $\lceil \log_2 \mathcal{D} \rceil$ bits.

4 * Remove w from $\mathcal{D}$, replace it with all its single-letter extensions.

5 * Repeat (goto 2) until u_n is reached.

$u_1^n =$

Ex: aaac~~b~~cb $\quad U = \{a, b, c\}$



$\Rightarrow$ output $= \underbrace{00}\,\underbrace{000}\,\underbrace{110}\,\underbrace{0111}$.

2

$$u_1^n \longrightarrow \boxed{LZ} \longrightarrow y_1^n \quad , \quad y_i \in \{0,1\}^*$$

def: $\quad |y_1^n| = \text{length}(y_1^n)$.

Analysis

Suppose u_1^n is parsed into c_{lz} words:

$$u_1^n = \underbrace{\lambda}_{\text{null}} w_1 w_2 \cdots w_{c_{lz}}.$$

$\lambda, w_1, \ldots, w_{c_{lz}-1}$ are distinct by construction.

Concatenate $(w_{c_{lz}-1}, w_{c_{lz}})$ together so we get a parsing into c_{lz} distinct words.

★ Def: $c^* :=$ max #distinct words (including λ) u_1^n can be parsed into $\quad \boxed{c^* \text{ depends on } u}$

$$\Rightarrow c_{lz} \le c^* \qquad (c^* \ge 1)$$

Let $q := |\mathcal{U}|$. Size of $\mathcal{D}$ in the end is

$$1 + \underbrace{(q-1)\, c_{lz}}_{[\text{the way } \mathcal{D} \text{ grows}]} \le 1 + (q-1)\, c^* \le q \cdot c^*.$$

$(q \ge 2, \; c^* \ge 1)$

$$\Rightarrow |y_1^n| \leq c_{l_2} \left\lceil \log_2 (q c^*) \right\rceil$$

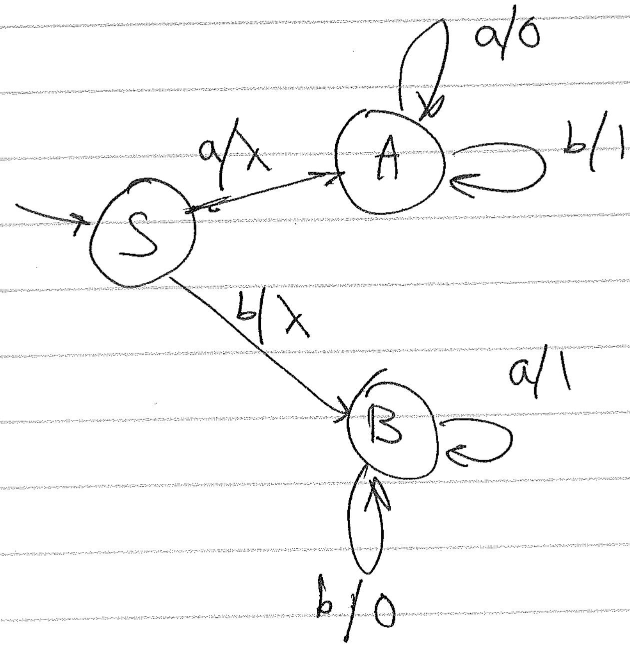
$$\leq c_{l_2} \log_2 (2 q c^*) \leq c^* \cdot \log_2 (2 q c^*).$$

$$\Rightarrow \limsup_{n \to \infty} \frac{|y_1^n|}{n} \leq \limsup_{n \to \infty} \frac{c^* (\log_2 c^* + \log_2 2q)}{n}$$

$$\boxed{= \limsup_{n \to \infty} \frac{c^* \cdot \log_2 c^*}{n}.} \quad * (1) *$$

How good is LZ?

Finite State machines:



Not uniquely decodable!

★ Huffman can be implemented as a F.S.M. and achieves the optimum <u>entropy rate</u> for iid sources (even "ergodic").

rate: $u_1^n \longrightarrow \boxed{\text{FSM}} \longrightarrow y_1^n$, $y_i \in \{0,1\}^*$

$$\rho_{FSM}(u_1^n) = \frac{|y_1^n|}{n} \longrightarrow H(X) \text{ for Huffman and optimal.}$$

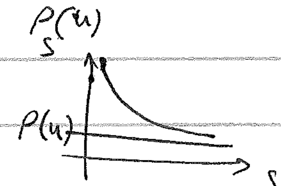
(u infinite)

★ Consider encoders with S or fewer states

<u>Def.</u>

$$\rho_s(u) = \limsup_{n\to\infty} \rho_s(u_1^n).$$

$$\rho(u) := \lim_{s\to\infty} \rho_s(u)$$



We prove (1) is a lowerbound for any F.S.M. $\Rightarrow$ LZ is optimal.

2)

Lower bound

Lemma: Let $Z = Z_1, \dots, Z_c$, $Z_i \in \mathcal{U}^*$, $|\mathcal{U}| = q$.

where Z_i are distinct. Then, $|Z| > c \log_q\left(\frac{c}{q^3}\right)$.

Proof: Write

(a) $c = \sum_{k=0}^{m-1} q^k + r \quad (m \geq 0,\ 0 \leq r < q^m)$

[Possible $\forall$ non-neg int].

(b) $|Z| \geq \sum_{k=0}^{m-1} k \underbrace{q^k}_{[\#\text{ str of len } k]} + mr$ (Since $|Z|$ is shortest if the z_i are as short as possible.)

* We use: $\sum_{k=0}^{m-1} q^k = \frac{q^m - 1}{q-1}$, and

$$\sum_{k=0}^{m-1} k q^k = m \frac{q^m}{q-1} - \frac{q}{q-1} \cdot \frac{q^m - 1}{q-1}$$

$$= \left(m - \frac{q}{q-1}\right) \sum_{k=0}^{m-1} q^k + \frac{m}{q-1}$$

$$\overset{(a)}{=} \left(m - \frac{q}{q-1}\right)(c - r) + \frac{m}{q-1}$$

$$(b) \Rightarrow |Z| \geq \left(m - \frac{q}{q-1}\right)(c-r) + \frac{m}{q-1} + mr$$

$$= \left(m - \frac{q}{q-1}\right)c \cancel{- rm} + \frac{qr}{q-1} + \frac{m}{q-1} \cancel{+ mr}$$

$$\geq (m-2) \cdot c. \quad (*)$$

* But: $c = \underbrace{\sum_{k=0}^{m-1} q^k}_{\frac{q^m-1}{q-1}} + \underbrace{r}_{< q^m} < q^{m+1}$

$$\Rightarrow m+1 > \log_q c$$

$$\overset{(*)}{\Rightarrow} n > (m-2)\, c > c \cdot \log_2 (c/q^3) . \qquad \square$$

Theorem: For any uniquely decodable FSM,

$$|y_1^n| \geq c^* \cdot \log c^* / (8 s^2) .$$

Proof: Parse u_1^n into c^* distinct words.

def: $c_{ij} :=$ # words which find the machine in state i and leave it in state j.

Unique decoding $\Rightarrow$ Output sequences corresponding to c_{ij} must be distinct.

def: $L_{ij} :=$ total length of encodings of words in c_{ij}.

by Lemma $\overset{(U=\{0,1\})}{\Rightarrow}$ $L_{ij} \geq c_{ij} \cdot \log_2 \left(c_{ij} / 8 \right)$.

$$\Rightarrow |y_1^n| = \sum_{1 \leq i,j \leq s} L_{ij} \geq \sum c_{ij} \log_2 \left(\frac{c_{ij}}{8} \right)$$

7

RHS is a symmetric convex function

& $\sum_{ij} c_{ij} = c^*$.

$\Rightarrow$ RHS is minimized when $\forall i,j,\ c_{ij} = c^*/s^2$.

$\Rightarrow |y_1^n| \geq c^* \cdot \log\left(c^*/8s^2\right)$. $\square$

by Theorem, $\rho_s(u) \geq \limsup_{n\to\infty} \frac{1}{n} c^* \log\left(c^*/8s^2\right)$

$= \limsup_{n\to\infty} \dfrac{c^* \log_2 c^* - c^* \log(8s^2)}{n}$

$= \limsup_{n\to\infty} \dfrac{c^* \log_2 c^*}{n} - O\left(\dfrac{c^*}{n}\right)$.

But by lemma: $n > c^* \log_q\left(c^*/q^3\right) \Rightarrow c^* = O\left(\dfrac{n}{\log n}\right)$

$\overset{\forall s,}{\Rightarrow} \boxed{\rho_s(u) \geq \limsup_{n\to\infty} \dfrac{c^* \log_2 c^*}{n}}$.

$\Rightarrow \rho(u) \geq \limsup_{n\to\infty} \dfrac{c^* \log_2 c^*}{n}$ ✓.