Information Theory & Coding

Oct 12th 2020

- Universal Compression
  - Lempel-Ziv method.
  - Finite-State-Machine, I.L. compressors.

- FSM: with m states.

g(starting state s, input letters $u_1 \dots u_n$) = t ← final state

f(" , ") = $y_1 \dots y_n$ ← $\in \{0,1\}^*$

$$\rho(M, u_1 u_2 \dots) = \limsup_{n\to\infty} \frac{1}{n} \text{ \# of bits in the } M \text{ output } (u_1 \dots u_n)$$

- IL: for any $s \in \mathcal{S}$, $u_1 \dots u_n \neq u_1' \dots u_\ell'$

  either $g(s, u_1 \dots u_n) \neq g(s, u_1' \dots u_\ell')$ ←

  or $f(s, u_1 \dots u_n) \neq f(s, u_1' \dots u_\ell')$ ←

Claim: LZ compresses better than any IL, FSM.

Def: given $u_1 \dots u_n$ we say that $w_1, \dots, w_c$ is a distinct parsing of $u_1 \dots u_n$ if $w_i \in \mathcal{U}^*$, $u_1 \dots u_n = w_1 \dots w_c$ & $w_i \neq w_j$ for $i \neq j$

Example: $\mathcal{U} = \{a, b, c\}$

$$\mathcal{U}^* = \{\text{null}, a, b, c, aa, ab, ac, ba, bb, bc, \ldots\}$$

$$\boxed{u_1 \ldots u_n = abaababc}$$

$w_1 = a$, $w_2 = b$, $w_3 = aa$, $w_4 = ba$, $w_5 = bc$

Note: LZ generates a distinct parsing

Recall (example) LZ would do in

$$\underbrace{a}_{w_1} | \underbrace{b}_{w_2} | \underbrace{a\,a}_{w_3} | b\,a | b\,c | - -$$

$\mathcal{D} = \{\cancel{a}, \cancel{b}, c\}$

~~aa~~ ~~ba~~ baa
ab bb bab
ac ~~bc~~ bac ...
aaa
aab
aac

Lemma: Suppose $u_1 u_2 \ldots\ldots$ is an infinite sequence, and let $q(n)$ be the number of words in a distinct (parsing) of $(u_1 \ldots u_n)$. Then

$$\lim_{n \to \infty} \frac{q(n)}{n} = 0.$$

(i.e. $q(n)$ grows slower than n.)

Pf: let $u_1 \ldots u_n = \underbrace{w_1 \ldots w_q}$ $q = q(n)$

how many of these w_i's can have length $\leq k-1$?

There are $1 + |U| + |U|^2 + \dots + |U|^{k-1} = F(k)$

So $q - F(k)$ of the w_i's have length $\geq k$

So $n \geq (q - F(k))\, k$

$\Rightarrow \quad q \leq \frac{n}{k} + F(k) \Rightarrow \frac{q(n)}{n} \leq \frac{1}{k} + \frac{F(k)}{n}$

$\overline{\lim_{n\to\infty}} \frac{q(n)}{n} \leq \frac{1}{k} + 0$ as k can be chosen arbitrarily large

we see $\lim_{n\to\infty} \frac{q(n)}{n} \leq 0$, ≥ 0 //.

(Also observe (as an aside) that we can make $q(n) \geq \sqrt{n}$

$\underbrace{u_1}_{w_1} \underbrace{u_2 u_3}_{w_2} \underbrace{u_4 u_5 u_6}_{w_3} \underbrace{u_7 \dots u_{10}}_{w_4} \dots$)

Suppose now $u_1 \dots u_n = w_1 \dots w_q$ (w's distinct)

is fed to a FSILM with $\leq m$ states.

$s_i \in \mathcal{L}, \ |\mathcal{L}| \leq m$

states : $s_0 \quad s_2 \quad s_3 \quad s_4 \quad \ldots \quad s_{q+1}$

input $\quad w_1 \quad w_2 \quad w_3 \quad \ldots \quad w_q$

output $\quad y_1 \quad y_2 \quad y_3 \quad \ldots \quad y_q$ - binary strings,

Claim : in the collection $y_1, y_2, \ldots, y_q$

any binary string can occur at most m^2 times.

(i.e. no binary string can occur $> m^2$ times).

Why? : Suppose a binary string t occurrence $> m^2$ times

among $y_1, \ldots, y_q$, let $I = \{ i : y_i = t \}$

$|I| > m^2$. so :

$f(s_i, w_i) = t \qquad \forall \ i \in I$

$g(s_i, w_i) = s_{i+1}$

$s_i \xrightarrow[t]{w_i} s_{i+1}$ $\qquad (s_i, s_{i+1})$ can take on at most m^2 values

$\exists \ i \neq j$ s.t $\qquad (s_i, s_{i+1}) = (s_j, s_{j+1}) = (\alpha, \beta)$

$f(\alpha, w_i) = t \qquad f(\alpha, w_j) = t$

$g(\alpha, w_i) = \beta \qquad g(\alpha, w_j) = \beta$

Contradicts IL property

Summary: $w_1 \dots w_q$ is a distinct parsing of $u_1 \dots u_n$

$\Rightarrow$ the output $y_1 \dots y_q$ of the FSM has the property that no binary string occurs $> m^2$ times in $y_1 \dots y_q$.

Lemma: if $y_1 \dots y_q$ is a collection of binary string s.t no string t can occur $> k$ times then:

- write $q = k[1 + 2 + \dots + 2^{j-1}] + r$
  with $0 \le r < k2^j$,

then $\sum_{i=1}^{q} \text{length}(y_i) \geq k[0 + 2.1 + 4.2 + \dots 2^{j-1}(j-1)] + rj$.

Example: Suppose 14 binary string s.t no string occurs > 3 times then

$14 = 3 + 3.2 + r$

$= 3(1+2) + 5$

total length of the 14 strings is $\geq 3[0.1 + 1.2) + 2.5$

$= 16$

Pf: set of binary strings: $\{0,1\}^* = \{\text{null}, 0, 1, 00, 01, 10, 11, 000, \ldots\}$

there

1 string of length 0 (null string)

2 " of " " 1 (0, 1)

4 " " " 2 (00, 01, 10, 11).

$$q = k + k \cdot 2 + k4 + \ldots k2^{j-1} + r$$

$$0 \leq r < k2^j.$$

total length of the "optimal" collection:

$$0 \cdot k + 1 \cdot k \cdot 2 + 2 \cdot k4$$
$$+ \ldots (j-1) k 2^{j-1} + jr.$$

Corollary: if $y_1 \ldots y_q$ has the properties in the previous lemma, then,

$$\sum_{i=1}^{q} \text{length}(y_i) \geq q \log_2 \frac{q}{8k}.$$

Pf: From the previous lemma

$$\boxed{q = k[2^j - 1] + r} \qquad 0 \le r < k2^j$$

$$\left(\sum_{i=0}^{j-1} x^i = \frac{x^j - 1}{x - 1} \right)$$

$$\text{total length} \ge k\left[\underbrace{\sum_{i=0}^{j-1} i\,2^i}_{2 + (j-2)2^j} \right] + rj$$

$$= (j-2)q + kj + 2r$$

$$\ge (j-2)q$$

$$q < k[2^{j+1} - 1) \le k2^{j+1}$$

$$j+1 \ge \log_2 \frac{q}{k} \Rightarrow j-2 \ge \log_2 \frac{q}{8k}$$

$$\Rightarrow \text{total length} \ge q \log_2 \frac{q}{8k} \quad //$$

Thm: Let $q^*(u_1 \ldots u_n)$ be the largest number of words in any distinct parsing of $u_1 \ldots u_n$. Then if M is a $\leq m$ state IL. FS. M,

$$\text{length output}(M, u_1 \ldots u_n) \geq q^*(u_1 \ldots u_n) \log \frac{q^*(u_1 \ldots u_n)}{8m^2}.$$

Pf: combine the Lemma above with $(**)$ above. //

Corollary: For any IL. FSM, M,

$$\rho(M, u_1 u_2 \ldots) \geq \limsup_{n \to \infty} \frac{1}{n} q^*(u_1 \ldots u_n) \log_2 q^*(u_1 \ldots u_n).$$

Pf: $\rho(M, u_1 u_2 \ldots)$

$$= \limsup_{n \to \infty} \frac{\text{length output}(M, u_1 \ldots u_n)}{n}$$

$$\geq \limsup_{n\to\infty} \frac{1}{n}\, q^*(u) \log \frac{q^*(u)}{8m^2} \qquad \text{(lemma above)}$$

$u_1 \ldots u_n$

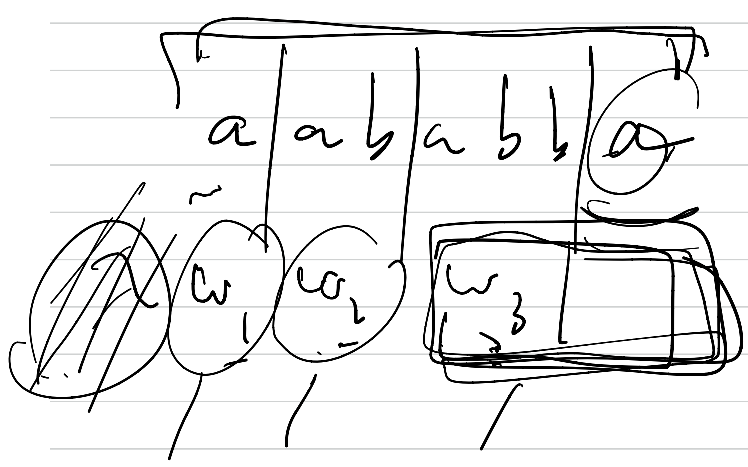
$$= \limsup_{n\to\infty} \frac{1}{n}\, q^*(u)\log q^*(u) - \underbrace{\frac{1}{n}\, q^*(u)\log(8m^2)}_{\to 0}$$

$$= \lim_{n\to\infty} \frac{1}{n}\, q^*(u)\log_2 q^*(u). \quad //$$

Corollary: $\rho_{FSM}(u_1u_2\ldots) \geq$. //

To show: $\rho_{LZ}(u_1u_2\ldots) \leq \limsup_{n\to\infty} \frac{1}{n}\, q^*(u_1\ldots u_n) \times \log_2 q^*(u_1\ldots u_n)$

To prove this, remember how LZ operates:



# of words LZ has created so far $\leq q^*(u_1\ldots u_n)$

How many bits has LZ produced?

w_1 is described by $\underbrace{\lceil \log_2 |\mathcal{U}| \rceil}$ bits.

w_2 is described by $\lceil \log_2 (|\mathcal{U}| - 1 + |\mathcal{U}|) \rceil$ bits

⋮

w_q is " " $\underbrace{\lceil \log_2 (1 + q(|\mathcal{U}| - 1)) \rceil}$

So the total output has length

$$\leq q \lceil \log_2 (1 + q(|\mathcal{U}| - 1)) \rceil$$

$$\leq q \log_2 2(1 + q(|\mathcal{U}| - 1))$$

$$\leq q \log_2 (2q|\mathcal{U}|)$$

$$\leq q^*(u_1 \ldots u_n)(\log_2 (2|\mathcal{U}|) + \log_2 q^*(u_1 \ldots u_n))$$

So:

$$\rho_{LZ}(u_1 u_2 \ldots) \leq \limsup_{n \to \infty} \frac{1}{n} q^*(u_1 \ldots u_n) \log_2 q^*(u_1 \ldots u_n) + \frac{1}{n} q^*(u_1 \ldots u_n) \log_2 (2|\mathcal{U}|)$$

$$= \boxed{\limsup_{n \to \infty} \frac{1}{n} q^*(u_1 \ldots u_n) \log_2 q^*(u_1 \ldots u_n)}$$

So, Thm: for any $u_1 u_2 u_3 \ldots$

$$\rho_{LZ}(u_1 u_2 \ldots) \leq \rho_{FSM}(u_1 u_2 \ldots)$$

Corollary: If $U_1 U_2 U_3 \ldots$ is an ergodic process, then

$$\rho_{LZ}(U_1 U_2 \ldots) \leq \text{entropy rate } \mathcal{H} \text{ of the process}$$

with prob. 1.