Lempel-Ziv Compression

Last time: Symbol codes (C: $a ightarrow \{0,1\}^*$), Kraft's inequality ($\sum 2^{-l_x} \leq 1$), $H(X) \leq H(X|C) < H(X) + 1$ for UD codes, achievable by $l_x = \lceil \log \frac{1}{P(x)} \rceil$

optimal code via Huffman algo

Today: Compression algo's that operate on "stream" of symbols, can emit < 1 bit/symbol, are asymptotically optimal for IID sources ($P \rightarrow H(X)$), but are also adaptive?  

Lempel-Ziv's coding algo (1978):

In short: stream ends with special symbol \(\perp\)

* phrases \(\leftarrow \emptyset\)

* while move to compress:
  - read symbols until we obtain "phrase" \(\perp\) \& phrases
  - \(P = \left[ \begin{array}{c} \perp \end{array} \right] \) where \(T\) \& phrases, \(x \in \mathbb{A}\)
  - append \(P\) to phrases
  - \(k\) \text{ } \text{ index of } T \text{ } \text{ in phrases}
  - write \((k,x)\) in bits

use \(\log_2(\ell)\) bits in j-th step \((j=1,2,..)\)  

(can skip if \(x=\perp\) (last step))

Example: Let's compress A|B|A|B|A|A|B|A|B|A|B|A|\(\perp\):

<table>
<thead>
<tr>
<th>Step</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>phrases</td>
<td>\emptyset</td>
<td>A</td>
<td>B</td>
<td>BA</td>
<td>BAA</td>
<td>BAAB</td>
<td>AB</td>
<td>A(\perp)</td>
</tr>
<tr>
<td>((k,x))</td>
<td>-</td>
<td>(0,A)</td>
<td>(0,B)</td>
<td>(2,A)</td>
<td>(3,A)</td>
<td>(4,B)</td>
<td>(1,B)</td>
<td>(1,(\perp))</td>
</tr>
<tr>
<td>compression</td>
<td>-</td>
<td>0</td>
<td>0,1</td>
<td>10,0</td>
<td>11,0</td>
<td>100,1</td>
<td>001,1</td>
<td>001,1</td>
</tr>
<tr>
<td>#bits for (k)</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

\(\Rightarrow\) 14 bits compressed into 20 bits... but the principle is sound.

Used in GIF; very similar to LZ78 which is used in ZIP, PNG,... (together with Huffman)

Splits input into minimal distinct "phrases"
**Q:** Intuition? Clear how to decompress?

**Analysis?** Let \( R = \frac{L}{n} \) the compression rate, where \( L = \text{bits of compression}. 

* Average case: For IID source, where \( X^N = X_1 \cdots X_N \overset{\text{iid}}{\sim} P: \)

\[
E[R] \leq H(P) + O\left(\frac{1}{\log N}\right) \rightarrow H(P)
\]

* Worst case: For any string \( X^N = x_1 \cdots x_N \),

\[
R \leq \log \#(\mathcal{A}) + O\left(\frac{1}{\log N}\right) \rightarrow \log \#(\mathcal{A})
\]

\( f = O(g) \) means \( \exists C > 0 : f(n) \leq Cg(n) \forall n \)

**Warmup:** Fix \( X^N \) and assume \( L \) decomposes it into \( C \) phrases:

\[
X^N = x_1 \cdots x_N = \Pi_1 \cdots \Pi_C
\]

\( \text{Disjoint phrases} \)

\[
\Rightarrow L = \sum_{j=1}^{C} \left( \log C_j + \log \#(\mathcal{A}_j) \right) \leq C \log(C) + C(1 + \log \#(\mathcal{A}) - \text{need to understand})
\]

To make progress, we need to upper-bound \( C \). For worst-case analysis, we would try to bound \( C \) in terms of \( N \) \( \rightarrow \text{EX CLASS}. \)

We focus on the average case, so want to relate \( C \) to \( P(X^N) \)

For simplicity: Assume all \( P(X) \leq \frac{1}{2} \) — but arbitrary \( \#(\mathcal{A}) \)

For our fixed string \( X^N \), consider:

\[
\Pi_k = \{ \Pi_i \mid 2^{-k} < P(\Pi_i) \leq 2^{-k+1/2} \}
\]

* for any phrase: \( P(\Pi) = P(\text{\(x^N\) has prefix } \Pi) \)

* any \( x^N \) has at most one prefix in \( \Pi_k \) (if both \( \Pi_i \) & \( \Pi_j \) are prefix, then \( \Pi_i = \Pi_j \cdots \Pi_k \) (or vice versa) \( \rightarrow P(\Pi_i) \leq P(\Pi_j) \frac{1}{2^{k-1}} \))

\[
\Rightarrow 1 \geq P(\text{\(x^N\) has prefix in } \Pi_k) \geq \sum_{\Pi \in \Pi_k} P(\Pi) \geq \#\Pi_k \cdot 2^{-k-1}
\]
Thus: \( \#\Pi_k < 2^{k+1} \)

How large can \( P(x^N) \) be if we know it has \( \circ \) phrases?

\[
P(x^N) = \prod_k \prod_{\Pi \in \Pi_k} P(\Pi) \leq (2^{-c})^{2^{k+1}} \cdots (2^{-c+1})^{2^L} (2^{-c})^{2^{L+1}} (2^{-c})^{2^{L+2}}
\]

where \( L \) is maximal with \( \sum_{k=0}^{L} 2^k = 2^{L+1} - 2 \leq c \). Note:

1. \( c \geq 2^{L+1} - 2 \Rightarrow L \leq \log(c+2) - 1 \leq \log(c) \) if \( c \geq 2 \), i.e. \( N > 1 \)
2. \( c < 2^{L+2} - 2 < 2^{L+2} \Rightarrow L \geq \log(c) - 2 \)

Thus:

\[
\left( \frac{1}{\log P(x^N)} \right) \geq \sum_{k=0}^{L} \frac{(k-1)2^k}{c} + L \left( c - 2^{L+1} + 2 \right)
\]

Check by inclusion:

\[
(L-2)2^{L+1} + 4 + L \left( c - 2^{L+1} + 2 \right)
\]

\[
= -42^L + 4 + cL + 2L
\]

1. \( \geq -4c + 4 + c(\log c - 2) + 2(\log c - 2) \)
2. \( = c \cdot \log c - 6c \)

Take expectation values and use \( H(x^N) = N \cdot H(P) \):

\[
N \cdot H(P) \geq E[c \cdot \log c] - 6E[c]
\]

\[
\Rightarrow E[|\log|] = \frac{1}{N} E[\log] \leq \frac{1}{N} E[c \cdot \log c] + \frac{1}{N} \left( \log(\#A) + 1 \right) E[c] \leq H(P) + \Theta \left( \frac{1}{N} E[c] \right)
\]

How to deal with \( E[c] \)?

\( \text{want that} \to 0 \)
E[C]\log E[C] \leq E[C \cdot \log C] \leq (HCP) + C \cdot N \quad \text{Since } \text{Jensen: } f(x) = x \cdot \log x \text{ is convex}

...so E[C] has to grow slower than linear? In fact:

\[ \implies E[C] = O\left(\frac{N}{\log N}\right) \]

and so we arrive at

\[ \implies E[R] \leq HCP + O\left(\frac{N}{\log N}\right) \]

**Why is this true?**

Assume that \( f(N) \cdot \log f(N) \leq g \cdot N \) for some \( N \).

We claim that \( f(N) < (g+1) \frac{N}{\log N} \). Indeed, otherwise we have

\[ f(N) \geq (g+1) \frac{N}{\log N} \quad \text{for a subsequence of } N \to \infty. \]

Then:

\[ f(N) \cdot \log f(N) \geq (g+1) \frac{N}{\log N} \cdot \log \left( (g+1) \frac{N}{\log N} \right) \]

\[ \geq (g+1)N \left(1 - \frac{\log \log N}{\log N}\right) \]

\[ \geq (g+1)N \]