The Source Coding Theorem: Proof and Variations

Recall from Tuesday:

\[ H_s(Y) = \log \min \{ \# S : \Pr(\text{YES}) \geq 1 - \delta \} \]
\[ \delta = \text{essential bit content} \]
\[ = \text{minimum bits need to compress } Y \text{ w/ error probability } \leq \delta \]

Shannon’s Source Coding Theorem: Let \( X_1, X_2, X_3, \ldots \) IID \( \mathcal{P} \) and \( 0 < \delta < 1 \):

\[ \lim_{N \to \infty} \frac{H_s(X^N)}{N} = H(\mathcal{P}) \]

<table>
<thead>
<tr>
<th>IID (memoryless) information source</th>
</tr>
</thead>
</table>

**Typical set:**

\[ T_{N, \varepsilon}(\mathcal{P}) = \left\{ x^N \in \mathcal{A}_X : \left| \frac{1}{N} \log \frac{1}{\Pr(x^N)} - H(\mathcal{P}) \right| \leq \varepsilon \right\} \]
\[ = \left\{ x^N \in \mathcal{A}_X : \left| \frac{1}{N} \sum_{k=1}^{N} \log \frac{1}{\Pr(x_k)} - H(\mathcal{P}) \right| \leq \varepsilon \right\} \]

**Properties:**

1. \( 2^{-N(H(\mathcal{P}) + \varepsilon)} \leq \Pr(x^N) \leq 2^{-N(H(\mathcal{P}) - \varepsilon)} \) (by definition)
2. \( \# T_{N, \varepsilon} \leq 2^{N(H(\mathcal{P}) + \varepsilon)} \)
   \[ \Pr: 1 \geq \Pr(x^N \in T_{N, \varepsilon}) = \sum_{x^N \in T_{N, \varepsilon}} \Pr(x^N) \geq \# T_{N, \varepsilon} 2^{-N(H(\mathcal{P}) + \varepsilon)} \]
3. \( \Pr(x^N \notin T_{N, \varepsilon}) \leq \frac{\sigma^2}{N \varepsilon^2} \to 0 \), where \( \sigma^2 = \text{Var}(\log \frac{1}{\Pr(x_k)}) \).
   \[ \Pr: \text{let } L_k = \log \frac{1}{\Pr(x_k)} \text{ and } \mu := \mathbf{E}[L_k] = H(\mathcal{X}) = H(\mathcal{P}). \text{ Then:} \]
   \[ \text{LHS} = \Pr \left( \frac{1}{N} \sum_{k=1}^{N} L_k - \mu \right| > \varepsilon \right) \leq \frac{\text{Var}(L_k)}{N \varepsilon^2} \]

"Asymptotic Equipartition Property" (AEP)

For large \( N \ldots \) typical probabilities are \( 2^{-N(H(\mathcal{P}) + \delta)} \).
Proof of Shannon's theorem: Let $\mathcal{X} \in \{0,1\}$ and $\varepsilon > 0$ be arbitrary.

(≤): \[ \Pr(X^N \in T_{N\varepsilon}) \geq 1 - \frac{2^{-N\varepsilon^2}}{N^2} \geq 1 - \varepsilon \text{ if } N \text{ large enough} \]

\[ \frac{H_8(X^N)}{N} \leq \frac{\log \#T_{N\varepsilon}}{N} \leq H(P) + \varepsilon \]

(≥) Want to prove that $\frac{H_8(X^N)}{N} \geq H(P) - \varepsilon$ for $N$ large.

If not: \exists sets $S_N$ for $N \to \infty$ s.t.

\[ \Pr(X^N \in S_N) \geq 1 - \varepsilon \text{ and } \#S_N < 2^{N(H(P) - \varepsilon)} \]

\[ \frac{\log \#S_N}{N} \geq 2^{-N(H(P) - \varepsilon)} \]

\[ \leq 2^{-N\varepsilon^2} \to 0 \]

\[ \rightarrow \exists CLASS \]

Remark: $T_{N\varepsilon}$ is usually NOT the smallest set $S_N$ with $\Pr(X^N \in S_N) \geq 1 - \varepsilon$...

... but small enough and easy to handle as $N \to \infty$!

How to use this in practice?

Scenario: Want to compress IID (memoryless) data source $P$ (we know $P$, but NOT which samples will be emitted)

**FIX:**

* block size $N$
* parameter $\varepsilon > 0$
* a way to order the typical set $T_{N\varepsilon}$

**COMPRESSOR:** Input: A string $x^N = x_1 \ldots x_N$

* If $x^N \notin T_{N\varepsilon}$: FAIL
* Determine index $p$ of $x^N$ in $T_{N\varepsilon}$
* Return $p$ in binary.

**DECOMPRESSOR:**

Input: A binary string $s$

* Interpret $s$ as integer $p$
* Return $p$-th element of $T_{N\varepsilon}$.
This is a lossy compression protocol:

\[ R = \frac{\# \text{bits required to represent } p}{N} \rightarrow 0 \text{ as } N \rightarrow \infty \]

\[ \Pr (X^N \notin T_{H(E)}) \leq \frac{\sigma^2}{N^2} \]

Variations

How to make it \textbf{LOSSLESS}? Instead of failing, send \( x^N \) uncompressed!

\( \overline{R} \leq \Pr (X^N \notin T_{H(E)}) \cdot (H(P) + \varepsilon + \frac{1}{N}) + \Pr (X^N \in T_{H(E)}) H_0(P) \rightarrow 0 \)

\( \approx H(P) + \varepsilon \) for large \( N \)

Undesirable that we need to know \( P \) ... can we compress \&\& knowing \( P \)?

\textit{UNIVERSAL SCENARIO:} want to compress IID source, but do not know \( P \)

For simplicity: assume \( A=\{0,1\}^5 \) (i.e. data source of bits)

\( F(X) \): block size

* a way to order the sets

\[ B(N,k) := \{x^N \text{ with } k \text{ ones and } N-k \text{ zeros}\} \]

**COMPRESSOR:** Input: A bitstring \( x^N = x_1 \ldots x_N \)

* Compute \( k := \# \text{zeros in } x^N \)
* Determine index \( p \) of \( x^N \) in \( B(N,k) \)
* Return \( k \) and \( p \) in binary.

**DECOMPRESSOR:**

Clear!

\[ B(3,2) \]

<table>
<thead>
<tr>
<th>Index</th>
<th>String</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>011</td>
</tr>
<tr>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>2</td>
<td>110</td>
</tr>
</tbody>
</table>

Not used in protocol, only in the analysis!!!

Average rate \( \overline{R} \)? Assume that \( X_1, \ldots, X_N \sim P \). Then:

\[ x^N \in T_{H(E)} \quad \text{since} \quad B(N,k) \in T_{H(E)} \implies \#B(N,k) \leq \#T_{H(E)} \]

Typicality only depends on \#zeros and \#ones in \( x^N \)!
Thus we can argue as above:

\[
\bar{R} = \frac{\text{#bits required to represent } k}{N} + \frac{\text{#bits required to represent } p}{N}
\]

\[
\leq \frac{\log(N)}{N} + \frac{\log \#B(n,k)}{N}
\]

Use \(a\) to obtain the following bound:

\[
\leq \Pr(X^{\text{NETINE}}) \frac{\log \#T\text{INE}}{N} + \Pr(X^{\text{null}}) \frac{\log 2^N}{N} \rightarrow 0, \text{ as before}
\]

\(\approx H(p) + \varepsilon \) for large \( N \)!

\(\text{Homework:} \) Program this protocol & compress the chicken!

\(\text{Discussion:} \) Many disadvantages!

* Have to look at entire \( x^N \) to compress. Can we compress symbol by symbol?
* Assume IID distribution... what if \( P \text{ changes} \)? Or if we have local correlations?

\(\text{Q:} \) \( R \)

Next week \( \infty \)