Lossy Compression & The Source Coding Theorem

Today we fix the number of bits but allow small error probability ("lossy"):

\[ X' \xrightarrow{C} \{0,1\}^2 \xrightarrow{D} X' \]

<table>
<thead>
<tr>
<th>Compressor, encoder</th>
<th>Decompressor, decoder</th>
</tr>
</thead>
</table>

\[ \text{WANT: } \Pr(X' \neq X) \leq \delta \]

How to achieve?

* Take set \( S \subseteq \mathcal{X} \) with \( \Pr(X \notin S) \leq \delta \).
* Then we can compress into \( l = \log |S| \) bits with error probability \( \leq \delta \). How?

Simply define \( C \) by sending all \( x \notin S \) to distinct bitstrings. (For \( x \in S \), pick arbitrary or fail.)

Define \( \delta \)-essential bit content by

\[ H_{\delta}(X') = H_{\delta}(C) = \min \left\{ \log |S| : \Pr(X \notin S) \leq \delta \right\} \]

\[ \Rightarrow H_{\delta}(X') \] is minimal \# bits required to compress \( X' \) with error \( \leq \delta \)

\[ H_{\delta}(X') \] is in general quite messy...

Amazingly, it simplifies dramatically if we compress blocks of symbols.

\[ \lim_{N \to \infty} \frac{H_{\delta}(X_1, X_2, \ldots, X_N)}{N} = H(P) \]

**Shannon’s Source Coding Theorem:** Let \( X_1, X_2, \ldots, X_N \) IID \( P \) and \( 0 < \delta < 1 \):

- Optimal compression rate for block size \( N \) and error prob \( \leq \delta \)
- Optimal asymptotic compression rate \( \delta \) — independent of \( \delta \)

(i.e. \( \forall \epsilon(\delta), \exists N_0 \forall N \geq N_0 : \left| \frac{H_{\delta}(X_1, X_2, \ldots, X_N)}{N} - H(P) \right| \leq \epsilon \))
If $R > H(P)$: $\exists N_0 \forall N \geq N_0$:

**CAN** compress at rate $R$ (= into $l = RN$ bits)

If $R < H(P)$: $\forall N \exists N_0$:

**CANNOT** compress at rate $R$

**Proof of the Source Coding Theorem**

**NOTATION:** $x_N = x_1 \cdots x_N = (x_1, \ldots, x_N)$ for strings of length $N$.

**Typical set:**

$T_{N,\varepsilon}(P) = \left\{ x_N \in \mathcal{X}^N : \left| \frac{1}{N} \log \frac{1}{P(x_N)} - H(P) \right| \leq \varepsilon \right\}$

$$\sum_{x_N \in T_{N,\varepsilon}} \frac{1}{P(x_N)} = \frac{1}{N} \sum_{k=1}^N \log \frac{1}{P(x_k)} - H(P) \leq \varepsilon$$

**Properties:**

1. $2^{-N(H(P) + \varepsilon)} \leq P(x_N) \leq 2^{-N(H(P) - \varepsilon)}$ (by definition)

2. $\# T_{N,\varepsilon} \leq 2^{-N(H(P) + \varepsilon)}$

   Let $P(x_N) \leq \sum_{x_N \in T_{N,\varepsilon}} \frac{1}{P(x_N)} \geq \# T_{N,\varepsilon} \cdot 2^{-N(H(P) + \varepsilon)}$. □

3. $\Pr(x_N \in T_{N,\varepsilon}) \leq \frac{\sigma^2}{N \varepsilon^2} \rightarrow 0$, where $\sigma^2 = \text{Var}(\log \frac{1}{P(x_N)})$.

   Let $L_k = \log \frac{1}{P(x_k)}$ and $\mu = \mathbb{E}[L_k] = H(X_k) = H(P)$. Then:

   $$\Pr\left(\frac{1}{N} \sum_{k=1}^N L_k - \mu > \varepsilon\right) \leq \frac{\text{Var}(L_k)}{N \varepsilon^2}.$$ □

   "Asymptotic Equipartition Property" (AEP)

   "For large $N$, typical probabilities are $2^{-N(H(P) + \varepsilon)}"$.

**Proof of the theorem:** Let $\delta(0,1)$ and $\varepsilon > 0$ be arbitrary.

1. $\Pr(x_N \in T_{N,\varepsilon}) \leq 1 - \frac{\sigma^2}{N \varepsilon^2} \geq 1 - \delta$ if $N$ large enough.

2. $\frac{H_S(x_N)}{N} \leq \log \# T_{N,\varepsilon} \leq H(P) + \varepsilon$ for large $N$. □
(5) Want to prove that \( \frac{H_S(X^N)}{N} \geq H(P) - \varepsilon \) for \( N \) large.

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

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

\[
\implies 1 - \delta \leq \Pr(X^N \in S_N) = \Pr(X^N \in S_{N \cap T_{N,\varepsilon/2}}) + \Pr(X^N \in S_{N \setminus T_{N,\varepsilon/2}})
\]

\[
\leq \Pr(X^N \in S_{N \cap T_{N,\varepsilon/2}}) + \Pr(X^N \notin T_{N,\varepsilon/2}) \to 0 \quad \text{by } 2
\]

\[
\leq \#S_N \cdot 2^{-N(H(P) - \varepsilon/2)} \to 0 \quad \text{by } 2
\]

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

**Remark:** \( T_{N,\varepsilon} \) is usually NOT the smallest set \( S_N \) w/ \( \Pr(X^N \in S_N) \geq 1 - \delta \).

... but small enough and easy to handle as \( N \to \infty \! \)!
Variations

A How to make it **LOSSLESS**?

When \( x^N \in T_{\text{in}} \), send uncompressed

using \( N \cdot \log \#Ax \) bits.

\[ \rightarrow \text{average rate } \bar{R} \leq \frac{1}{N} + \Pr(x^N \in T_{\text{in}}) \left( \log(N) + \frac{3 + \frac{1}{K}}{3} \right) \]

\[ + \Pr(x^N \notin T_{\text{in}}) \cdot \log \#Ax \]

\[ \approx H(C) + 2 \text{ for large } N \]

B How to also make it **UNIVERSAL**? (IID, but we do **NOT** know \( P \))

For simplicity: assume \( A \sim \text{LTI} \)'s i.e. data source of bits.

**FIX:** * block size \( N \)

* a way to order the sets

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

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

* Compute \( k := \# \text{ones in } x^N \)

* Determine index \( p \) of \( x^p \) in \( \mathcal{B}(N,k) \)

* Return \( k \) and \( p \) in binary.

\[ \approx \log(N) + 1 = \log \# \mathcal{B}(N,k) + 1 \text{ bits} \]

**DECOMPRESSOR:**

* Clear \( k \) and \( p \)

**Key idea:** \( \mathcal{B}(N,k) \) can be **much** smaller than \( \# \mathcal{B}(N,k) \)

(e.g. \( k \leq 1 \) for some \( k \))

* Just used in protocol only in the analysis!!

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

\( x^N \in T_{\text{in}} \iff \mathcal{B}(N,k) \in T_{\text{in}} \)

\[ \Rightarrow \# \mathcal{B}(N,k) \leq \# T_{\text{in}} \]

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

\[
R = \frac{\log(N)}{N} + \frac{\log \#B(n,x)}{N}
\]

dropping some \( \frac{1}{N} \) terms

so Can ignore

\[ \Rightarrow 0 \text{ as before} \]

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

\[
\leq \Pr(X^N \notin \text{Tue}) \cdot \frac{\log \#\text{Tue}}{N} + \Pr(X^N \notin \text{Tue}) \frac{\log 2N}{N}
\text{ \( \rightarrow 0 \) as before}
\]

\[ \Rightarrow H(\epsilon) + \epsilon \text{ for large } N! \]

\[ H(\epsilon) \]

Program this protocol & compress the donkey!

Discussion: Many disadvantages!

* Have to look at entire \( x^N \) to compress. Can we compress by looking at a few symbols at a time?

* Assume IID distribution...what if \( P \) changes? Or if we have local correlations!

\( \bar{\epsilon} \) Wednesday 😊