In lossy compression, we fix the number of bits but allow small error probability:

\[
X' \xrightarrow{G} Y \xrightarrow{D} X''
\]

\[\text{Compressor, encoder} \quad \xrightarrow{0.15} \quad \text{decompressor, decoder} \quad \xrightarrow{X'}\]

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

How to achieve?

* Take set \( S \subseteq \Omega \) with \( \Pr(X \notin S) < \delta \).

* Then we can compress into \( e = \lceil \log \#S \rceil \) bits
  with error probability \( \leq \delta \). How?

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

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

\[
H_\delta(X') = H_\delta(P) = \min \left\{ \log \#S \mid \Pr(X \notin S) \leq \delta \right\}
\]

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

\( H_\delta(X') \) is in general quite messy... Amazingly, it simplifies dramatically if we compress blocks of symbols.

### Shannon's Source Coding Theorem

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

11D (memoryless) information source

Optimal compression rate for block size \( N \) and error prob \( \leq \delta \)

\[\text{optimal asymptotic compression rate } \delta \text{ independent of } \delta\]

(i.e. \( \forall \varepsilon(\delta), \exists \delta \in N_\varepsilon(0) \): \( \left| \frac{H_\delta(X_1, \ldots, X_N)}{N} - H(P) \right| < \varepsilon \))
If $R > H(CP)$, then $\exists N_0$ s.t. $H(CP) > 0$, strings of bit length $N > N_0$ can compress at rate $R$.

If $R < H(CP)$, then $\exists N_0$ s.t. $H(CP) > -\log_2 N$, strings of bit length $N > N_0$ cannot compress at rate $R$.

**Proof of the Source Coding Theorem**

**Notation:** $X_N = x_1, \ldots, x_N$ for strings of length $N$.

**Typical set:**

$$T_{N,\varepsilon}(CP) = \left\{ \frac{1}{N} \log \frac{1}{P(x_N)} - H(CP) \right\} \leq \varepsilon \right\}$$

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

**Properties:**

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

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

$$P_t \geq \Pr(X^N \in T_{N,\varepsilon}) = \sum_{X^N \in T_{N,\varepsilon}} P(X^N) \geq \#T_{N,\varepsilon} \cdot 2^{-N(H(CP)+\varepsilon)}$$

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

$$L_t = \log \frac{1}{P(x_N)}$$

$$\Pr \left( \left| \frac{1}{N} \sum_{k=1}^{N} L_t - \mu \right| > \varepsilon \right) \leq \frac{\text{Var}(L_t)}{N \varepsilon^2} \leq \frac{\sigma^2}{N \varepsilon^2} \rightarrow 0.$$  

"Asymptotic Equipartition Property" (AEP)

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

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

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

2. $\frac{H_S(X^N)}{N} \leq \log \#T_{N,\varepsilon} \leq H(CP) + \varepsilon$ for large $N$. [1]
Want to prove that \( \frac{H_S(X^N)}{N} \geq H(P) - \varepsilon \) for \( N \) large.

\[
\text{If not: } \exists \text{ sets } S_N \text{ for } N \to \infty \text{ s.t.} \\
\Pr(X^N \in S_N) \geq 1 - \varepsilon \text{ and } |S_N| < 2^{N(H(P) - \varepsilon)}.
\]
\[
\implies 1 - \varepsilon \leq \Pr(X^N \in S_N) = \Pr(X^N \in S_N \cap T_{N, \varepsilon/2}) + \Pr(X^N \in S_N \setminus T_{N, \varepsilon/2})
\]
\[
\leq \Pr(X^N \in S_N \cap T_{N, \varepsilon/2}) + \Pr(X^N \not\in T_{N, \varepsilon/2}) \to 0 \\
\leq |S_N| \cdot 2^{-N(H(P) - \varepsilon)} \to 0 \text{ by } 2
\]
\[
\leq 2^{-N^2} \to 0
\]

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 string 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 \( x_N \not\in T_{N, \varepsilon} \): FAIL
* Determine index \( p \) of \( x_N \) in \( T_{N, \varepsilon} \)
* Return \( p \) in binary.

This is a lossy compression protocol:

* Error probability: \( \Pr(X^N \not\in T_{N, \varepsilon}) \leq \frac{\varepsilon^2}{2N^2} \to 0 \) as \( N \to \infty \)

* **Rate** \( R = \frac{\# \text{bits required to represent } p}{N} \)

\[
\leq \frac{\log |T_{N, \varepsilon}| + 1}{N} \leq (H(P) + \varepsilon) + \frac{1}{N} \to 0
\]

**DECOMPRESSOR:**

Input: A binary string \( S \)

* Interpret \( S \) as integer \( p \)
* Return \( p \)-th element of \( T_{N, \varepsilon} \)

AEP
Variations

A. How to make it **LOSSLESS**?

When $X \in T_{\text{MIE}}$, send uncompressed
using $N \cdot \log \#A \times 7$ bits.

$$\bar{R} \leq \frac{1}{N} + \Pr(X \in T_{\text{MIE}}) \left( H(C) + \epsilon + \frac{1}{2} \right)$$

$$+ \Pr(X \notin T_{\text{MIE}}) \cdot \log \#A \times 7$$

$$\approx H(C) + \epsilon \text{ for large } N$$

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

For simplicity: assume $A = \{0, 1\}$, i.e., data source of bits.

**FIX**: * block size $N$

* a way to order the sets

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

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

* Compute $k := \#$ ones in $x^N$

* Determine index $p$ of $x^k$ in $B(N, k)$

* Return $k$ and $p$ in binary.

Average rate $\bar{R} \approx H(C) + \epsilon$ Assume that $X_1, \ldots, X_N \overset{\text{iid}}{\sim} P$. Then:

$x^N \in T_{\text{MIE}} \implies B(N, k) \in T_{\text{MIE}} \implies \#B(N, k) \leq \#T_{\text{MIE}}$.

**DECOMPRESSOR**: Clear! ?

* Key idea: $B(N, k)$ can be much smaller than $\{0, 1\}^N$.

Just used in protocol only in the analysis!!

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

\[ R = \frac{\log(N)}{N} + \frac{\log \#(\text{nic})}{N} \]

\( \leq \Pr(X \not\in \text{Tue}) \cdot \frac{\log \#(\text{Tue})}{N} + \Pr(X \in \text{Tue}) \cdot \frac{\log 2N}{N} \to 0 \) as before

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

**HW:** 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?

To Thursday ☝️