Outline

- Proof of the Channel Coding Theorem (Section 2.2).
And Now For Something Extraordinarily Important

The Noisy Channel Coding Theorem
The Three Regions

- Region I: Impossible (Done)
- Region II: The Noisy Channel Coding Theorem (In Progress)
- Region III: Lossy Data Compression (Coming Soon)
Overview of Shannon’s Proof

\[2^n R\] codewords

\[2^n H(X)\] typical X’s

\[2^n H(X|Y)\] reasonable causes for Y

causal codeword
noncausal codeword
non codeword

Received Y
The Probability of Failure, Simplified

- The probability that a randomly selected codeword will lie in $S(Y)$ is

\[
\frac{2^{nH(X|Y)}}{2^{nH(X)}} = 2^{-n(H(X)-H(X|Y))} = 2^{-nI(X;Y)} = 2^{-nC}
\]
• By the union bound, the probability that at least one of $2^{Rn}$ randomly selected codewords will lie in $S(Y)$ is upper bounded as follows:

$$\leq 2^{Rn} \cdot 2^{-nC} = 2^{-n(C-R)}.$$
And Now For Something Completely Rigorous
The Theorem

**Theorem.** Given a DMC with capacity $C$: For any $R < C$, and $\epsilon > 0$, for all sufficiently large values of $n$ there exists a channel code $\mathcal{C} = \{x_1, x_2, \ldots, x_M\}$ of length $n$ and rate $R$, and a suitable decoding rule, such that $P_E^{(i)} < \epsilon$ for all $i = 1, \ldots, M$. 
Theorem. Given a DMC with capacity $C$: For any $R < C$, and $\epsilon > 0$, for all sufficiently large values of $n$ there exists a channel code $C = \{x_1, x_2, \ldots, x_M\}$ of length $n$ and rate $R$, and a suitable decoding rule, such that $P_E^{(i)} < \epsilon$ for all $i = 1, \ldots, M$. 
The Theorem

**Theorem.** Given a DMC with capacity $C$: For any $R < C$, and $\epsilon > 0$, for all sufficiently large values of $n$ there exists a channel code $C = \{x_1, x_2, \ldots, x_M\}$ of length $n$ and rate $R$, and a suitable decoding rule, such that $P_E^{(i)} < \epsilon$ for all $i = 1, \ldots, M$.

**Definition.** The DMC has input alphabet $A$, output alphabet $B$, and transition probability matrix $Q$. The optimizing input distribution is $p$ and the corresponding output distribution is $q = pQ$. 
The Theorem

**Theorem.** Given a DMC with capacity $C$: For any $R < C$, and $\epsilon > 0$, for all sufficiently large values of $n$ there exists a channel code $C = \{x_1, x_2, \ldots, x_M\}$ of length $n$ and rate $R$, and a suitable decoding rule, such that $P_E^{(i)} < \epsilon$ for all $i = 1, \ldots, M$. 
The Theorem

**Theorem.** Given a DMC with capacity $C$: For any $R < C$, and $\epsilon > 0$, for all sufficiently large values of $n$ there exists a channel code $C = \{x_1, x_2, \ldots, x_M\}$ of length $n$ and rate $R$, and a suitable decoding rule, such that $P_E^{(i)} < \epsilon$ for all $i = 1, \ldots, M$.

**Definition.** A channel code of length $n$ and rate $R$ is a subset of size $M = 2^{Rn}$ of $A^n$. 
Theorem. Given a DMC with capacity $C$: For any $R < C$, and $\epsilon > 0$, for all sufficiently large values of $n$ there exists a channel code $\mathcal{C} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}$ of length $n$ and rate $R$, and a suitable decoding rule, such that $P_E^{(i)} < \epsilon$ for all $i = 1, \ldots, M$. 
Theorem. Given a DMC with capacity $C$: For any $R < C$, and $\epsilon > 0$, for all sufficiently large values of $n$ there exists a channel code $C = \{x_1, x_2, \ldots, x_M\}$ of length $n$ and rate $R$, and a suitable decoding rule, such that $P_E^{(i)} < \epsilon$ for all $i = 1, \ldots, M$.

Definition. A decoding rule is a mapping from $B^n$ to $C \cup \{?\}$. 
The Playing Field

\[ \Omega_n = A^n \times B^n, \quad \text{where } n \gg 1. \]
For each pair \((x, y) \in \Omega_n\), we define a kind of score:

\[
I(x; y) = \sum_{i=1}^{n} \log_2 \frac{Q(y_i|x_i)}{q(y_i)},
\]

where \(Q(y|x)\) is the channel transition probability matrix, and

\[
q(y) = \sum_{x \in A} p(x)Q(y|x)
\]

is the output density induced by the input density \(p(x)\).
The Playing Field

We make $\Omega_n$ into a probability space:

If $(x, y) = ((x_1, \ldots, x_n), (y_1, \ldots, y_n)) \in \Omega_n$, define

$$p_1(x, y) = p(x)Q(y|x),$$

where

$$p(x) = \prod_{i=1}^{n} p(x_i),$$

where $p(x)$ is a capacity-achieving input distribution, and

$$Q(y|x) = \prod_{i=1}^{n} Q(y_i|x_i),$$

where $Q(y|x)$ is the channel transition probability matrix.
The Playing Field

We make $\Omega_n$ into a probability space in another way:

If $(x, y) = ((x_1, \ldots, x_n), (y_1, \ldots, y_n)) \in \Omega_n$, define

$$p_2(x, y) = p(x)q(y),$$

where

$$p(x) = \prod_{i=1}^{n} p(x_i),$$

where $p(x)$ is a capacity-achieving input distribution, and

$$q(y) = \prod_{i=1}^{n} q(y_i).$$
The Playing Field

• $p_1(x, y)$: Choose $x \sim p(x)$ and then choose $y \sim Q(y|x)$.

\[ x \rightarrow \text{DMC} \rightarrow y. \]

• $p_2(x, y)$: Choose $x \sim p(x)$ and $y \sim q(y)$ independently.

\[ x \hspace{2cm} y. \]
“It is easy to see that”

\[ I(x; y) = \cdots = \log_2 \frac{p_1(x, y)}{p_2(x, y)}. \]
Typical Pairs

- Choose $R'$ so that

$$R < R' < C$$

- Define the set $T_n$ of typical pairs:

$$T_n = \{(x, y) \in \Omega_n : I(x; y) \geq nR'\}.$$
The Playing Field

Under $p_1(x, y)$, the $I$-score $I(x; y)$ is a sum of i.i.d. random variables, with common mean $C$:

$$\sum_{x,y} p_1(x, y) \log_2 \frac{Q(y|x)}{q(y)} = I_1(X; Y) = C.$$ 

Under $p_2(x, y)$, the $I$-score $I(x; y)$ is a sum of i.i.d. random variables, with common mean 0:

$$\sum_{x,y} p_2(x, y) \log_2 \frac{Q(y|x)}{q(y)} = I_2(X; Y) = 0.$$
Typical Pairs

• Thus for any $\delta > 0$,

$$\lim_{n \to \infty} \Pr_1 \left\{ \left| \frac{1}{n} I(x; y) - C \right| > \delta \right\} = 0.$$ 

• In particular, since $R' < C$,

$$\lim_{n \to \infty} \Pr_1 \left\{ (x, y) \notin T_n \right\} = \lim_{n \to \infty} \Pr_1 \left\{ \frac{1}{n} I(x; y) < R' \right\} = 0.$$
Typical Pairs

- Thus for any $\delta > 0$,
  \[
  \lim_{n \to \infty} \Pr_2 \left\{ \left| \frac{1}{n} I(x; y) \right| > \delta \right\} = 0.
  \]

- In particular, since $R' > 0$,
  \[
  \lim_{n \to \infty} \Pr_2 \{(x, y) \in T_n\} = \lim_{n \to \infty} \Pr_2 \left\{ \frac{1}{n} I(x; y) \geq R' \right\} = 0.
  \]

- Actually,
  \[
  \Pr_2 \{(x, y) \in T_n\} \leq 2^{-nR'}.
  \]
Here’s Why:

\((x, y) \in T_n \implies I(x; y) \geq nR'\)

\[\implies \log_2 \frac{p_1(x, y)}{p_2(x, y)} \geq nR'\]

\[\implies p_2(x, y) \leq p_1(x, y)2^{-nR'}\]

- And hence

\[
Pr_2 \{(x, y) \in T_n\} = \sum_{(x, y) \in T_n} p_2(x, y) \\
\leq 2^{-nR'} \sum_{x, y} p_1(x, y) = 2^{-nR'}
\]
Summary

\[ \Pr_1 \{(x, y) \in T_n \} \to 1 \]
\[ \Pr_2 \{(x, y) \in T_n \} \leq 2^{-nR'} \]
The Code

• Let

\[ C = \{ x_1, \ldots, x_M \} \subseteq A^n \]

be any code, good or bad, with \( M = 2^{Rn} \) codewords.
The Decoding Net

• Define, for $y \in B^n$,

$$S(y) = \{x \in A^n : I(x; y) \in T_n\}.$$ 

These are Shannon’s “reasonable causes” of the “effect” $y$.

• If the set $C \cap S(y)$ contains exactly one codeword $x_i$, set

$$\text{Decode}(y) = x_i.$$ 

• Otherwise set

$$\text{Decode}(y) = \text{Failed}$$
Shannon’s Decoding Strategy Continued

- The decoder searches $S(y)$ for codewords. If there is exactly one codeword in $S(y)$, that is the decoder’s decision.
- Otherwise, the decoder reports Failed.
Analyzing the Decoder Error Probability

• If $x_i$ is transmitted, define

$$P_{E}^{(i)} \overset{\text{def}}{=} \Pr \{ \text{Decode}(y) \neq x_i \}.$$ 

• By the union bound,

$$P_{E}^{(i)} \leq \Pr \{ x_i \notin S(y) \} + \sum_{\substack{j=1 \atop j \neq i}}^{M} \Pr \{ x_j \in S(y) \}.$$
Interpretation

\[
P_E^{(i)} \leq \Pr \{x_i \notin S(y)\} + \sum_{\substack{j=1 \atop j \neq i}}^{M} \Pr \{x_j \in S(y)\}.
\]
Interpretation

\[ P_{E}^{(i)} \leq \Pr \{ x_i \notin S(y) \} + \sum_{\substack{j=1 \\ j \neq i}}^{M} \Pr \{ x_j \in S(y) \} . \]

• “Bad Noise” Term (The causal codeword is not a “reasonable cause” of \( y \).)
Interpretation

\[ P_{E}^{(i)} \leq \Pr \{ x_i \notin S(y) \} + \sum_{\substack{j=1 \atop j \neq i}}^{M} \Pr \{ x_j \in S(y) \}. \]

- “Bad Code” Term (A noncausal codeword is a “reasonable cause” of \( y \).)
An Alternative Expression for $P_E^{(i)}$

- We had

$$P_E^{(i)} \leq \Pr \{x_i \not\in S(y)\} + \sum_{\substack{j=1 \atop j \neq i}}^{M} \Pr \{x_j \in S(y)\}.$$

- Define indicator functions as follows

$$\Delta(x, y) = \begin{cases} 
1 & \text{if } (x, y) \in T_n \\
0 & \text{if } (x, y) \not\in T_n
\end{cases}$$

$$\overline{\Delta}(x, y) = \begin{cases} 
0 & \text{if } (x, y) \in T_n \\
1 & \text{if } (x, y) \not\in T_n
\end{cases}$$
An Alternative Expression for $P_E^{(i)}$

- We had

$$P_E^{(i)} \leq \Pr \{x_i \not\in S(y)\} + \sum_{\substack{j=1 \atop j \neq i}}^{M} \Pr \{x_j \in S(y)\}.$$  

- Which can now be written as

$$P_E^{(i)} \leq \sum_{y \in B^n} \Delta(x_i, y)Q(y|x_i) + \sum_{j=1}^{M} \sum_{\substack{y \in B^n \atop j \neq i}} \Delta(x_j, y)Q(y|x_i)$$

$$= Q_i(x_1, \ldots, x_M).$$
Random Coding

- $Q_i(\mathcal{C})$ is an extremely complicated expression that is impossible to compute except in a few trivial cases.

- Shannon’s brilliant idea was to average $Q_i(\mathbf{x}_1, \ldots, \mathbf{x}_M)$ over all codes!

- The averaging is with respect to the model in which each letter in each codeword is chosen i.i.d. with probability $p(x)$. 

![Image of Claude Shannon]
\begin{align*}
E(Q_i(x_1, \ldots, x_M)) &= E \left( \sum_{y \in B^n} \Delta(x_i, y) Q(y \mid x_i) \right) \\
&\quad + \sum_{j=1 \atop j \neq i}^M E \left( \sum_{y \in B^n} \Delta(x_j, y) Q(y \mid x_i) \right) \\
&= E_1(n) + \sum_{j=1 \atop j \neq i}^M E_2^{(j)}(n)
\end{align*}
First We Estimate $E_1(n)$

$$E_1(n) = E \left( \sum_{y \in B^n} \Delta(x_i, y)Q(y|x_i) \right)$$
First We Estimate $E_1(n)$

$$E_1(n) = E \left( \sum_{y \in B^n} \bar{\Delta}(x_i, y) Q(y|x_i) \right)$$

$$= \sum_{x_i} p(x_i) \left( \sum_{y \in B^n} \bar{\Delta}(x_i, y) Q(y|x_i) \right)$$
First We Estimate $E_1(n)$

\[ E_1(n) = E \left( \sum_{y \in B^n} \overline{\Delta}(x_i, y)Q(y|x_i) \right) \]

\[ = \sum_{x_i} p(x_i) \left( \sum_{y \in B^n} \overline{\Delta}(x_i, y)Q(y|x_i) \right) \]

\[ = \sum_{x, y} p(x)Q(y|x)\overline{\Delta}(x, y) \]
First We Estimate $E_1(n)$

$$E_1(n) = E \left( \sum_{y \in B^n} \Delta(x_i, y) Q(y|x_i) \right)$$

$$= \sum_{x_i} p(x_i) \left( \sum_{y \in B^n} \Delta(x_i, y) Q(y|x_i) \right)$$

$$= \sum_{x, y} p(x) Q(y|x) \Delta(x, y)$$

$$= \sum_{x, y} p_1(x, y) \Delta(x, y)$$
First We Estimate $E_1(n)$

$$E_1(n) = E \left( \sum_{y \in B^n} \Delta(x_i, y) Q(y|x_i) \right)$$

$$= \sum_{x_i} p(x_i) \left( \sum_{y \in B^n} \Delta(x_i, y) Q(y|x_i) \right)$$

$$= \sum_{x, y} p(x) Q(y|x) \Delta(x, y)$$

$$= \sum_{x, y} p_1(x, y) \Delta(x, y)$$

$$= \Pr_1 \{ (x, y) \notin T_n \} \to 0$$
Now We Estimate $E_2^{(j)}(n)$

$$E_2^{(j)}(n) = E \left( \sum_{y \in B^n} \Delta(x_j, y) Q(y|x_i) \right)$$
Now We Estimate $E_2^{(j)}(n)$

$$E_2^{(j)}(n) = E \left( \sum_{y \in B^n} \Delta(x_j, y) Q(y | x_i) \right)$$

$$= \sum_{x_i, x_j} p(x_i) p(x_j) \left( \sum_{y \in B^n} \Delta(x_j, y) Q(y | x_i) \right)$$
Now We Estimate $E_2^{(j)}(n)$

\[
E_2^{(j)}(n) = E \left( \sum_{y \in B^n} \Delta(x_j, y) Q(y|x_i) \right)
\]

\[
= \sum_{x_i, x_j} p(x_i)p(x_j) \left( \sum_{y \in B^n} \Delta(x_j, y) Q(y|x_i) \right)
\]

\[
= \sum_{x_j, y} p(x_j) \Delta(x_j, y) \sum_{x_i} p(x_i) Q(y|x_i)
\]
Now We Estimate $E_2^{(j)}(n)$

$$E_2^{(j)}(n) = E \left( \sum_{y \in B^n} \Delta(x_j, y) Q(y|x_i) \right)$$

$$= \sum_{x_i, x_j} p(x_i)p(x_j) \left( \sum_{y \in B^n} \Delta(x_j, y) Q(y|x_i) \right)$$

$$= \sum_{x_j, y} p(x_j) \Delta(x_j, y) \sum_{x_i} p(x_i) Q(y|x_i)$$

$$= \sum_{x, y} p(x) \Delta(x, y) q(y) = \sum_{(x, y) \in T_n} p(x) q(y)$$

46
Now We Estimate $E_2^{(j)}(n)$

$$E_2^{(j)}(n) = E \left( \sum_{y \in B^n} \Delta(x_j, y) Q(y|x_i) \right)$$

$$= \sum_{x_i, x_j} p(x_i)p(x_j) \left( \sum_{y \in B^n} \Delta(x_j, y) Q(y|x_i) \right)$$

$$= \sum_{x_j, y} p(x_j) \Delta(x_j, y) \sum_{x_i} p(x_i) Q(y|x_i)$$

$$= \sum_{x, y} p(x) \Delta(x, y) q(y) = \sum_{(x, y) \in T_n} p(x)q(y)$$

$$= \sum_{(x, y) \in T_n} p_2(x, y) \leq 2^{-nR'}$$
Recap:

\[ E(P_E^{(i)}) \leq E(Q_i(x_1, \ldots, x_M)) \]

\[ = E_1(n) + \sum_{j=1}^{2^n R} E_2^{(j)}(n) \]

\[ \leq o(1)^* + 2^{Rn} 2^{-nR'} \]

\[ = o(1) + 2^{-n(R' - R)} \]

* The notation \( o(1) \) stands for a quantity that approaches 0 as \( n \to \infty \).
Thus for any $i \in \{1, \ldots, M\}$ and $\epsilon > 0$ there exists a value $n_0$ such that for $n \geq n_0$, $E(P_E^{(i)}) < \epsilon$, which is almost what we need.
The Last Step

• Define

\[ P_E(x_1, \ldots, x_M) = \frac{1}{M} \sum_{i=1}^{M} P_E^{(i)}, \]

which is a kind of overall error probability for the code \( C = \{x_1, \ldots, x_M\} \), and its average (under \( p_1(x, y) \)), being the sum of the averages, is \( < \epsilon \) for all \( n \geq n_0 \).

• But not everyone can be above average!
We Have Therefore Proved

**Theorem.** Given a DMC with capacity $C$: For any $R < C$, and $\epsilon > 0$, for all sufficiently large values of $n$ there exists a channel code $C = \{x_1, x_2, \ldots, x_M\}$ of length $n$ and rate $R$, and a suitable decoding rule, such that $P_E < \epsilon$. 
But We Promised

**Theorem.** Given a DMC with capacity $C$: For any $R < C$, and $\epsilon > 0$, for all sufficiently large values of $n$ there exists a channel code $\mathcal{C} = \{x_1, x_2, \ldots, x_M\}$ of length $n$ and rate $R$, and a suitable decoding rule, such that $P_E^{(i)} < \epsilon$ for all $i = 1, \ldots, M$. 
An Easy Fix: Expurgation

• Choose $n_0$ so that $E(P_E) < \epsilon/2$ for $n \geq n_0$. Then there exists a particular code with $P_E < \epsilon/2$.

* At most half of the class can be twice the average height.
An Easy Fix: Expurgation

• Choose $n_0$ so that $E(P_E) < \epsilon/2$ for $n \geq n_0$. Then there exists a particular code with $P_E < \epsilon/2$.

• At most $M/2$ of the corresponding $P_E^{(i)}$’s can be $\geq \epsilon^*$. (bad codewords)

* At most half of the class can be twice the average height.
An Easy Fix: Expurgation

• Choose $n_0$ so that $E(P_E) < \epsilon/2$ for $n \geq n_0$. Then there exists a particular code with $P_E < \epsilon/2$.

• At most $M/2$ of the corresponding $P_E^{(i)}$’s can be $\geq \epsilon^*$. (bad codewords)

• Expurgate these bad codewords.

* At most half of the class can be twice the average height.
An Easy Fix: Expurgation

- Choose $n_0$ so that $E(P_E) < \epsilon/2$ for $n \geq n_0$. Then there exists a particular code with $P_E < \epsilon/2$.

- At most $M/2$ of the corresponding $P_E^{(i)}$'s can be $\geq \epsilon^*$. (bad codewords)

- Expurgate these bad codewords.

- What Remains? A code with $\geq 2^{Rn - 1}$ codewords and $P_e^{(i)} < \epsilon$ for all $i$.

* At most half of the class can be twice the average height.
Now We Have Proved

**Theorem.** Given a DMC with capacity $C$: For any $R < C$, and $\epsilon > 0$, for all sufficiently large values of $n$ there exists a channel code $C = \{x_1, x_2, \ldots, x_M\}$ of length $n$ and rate $R$, and a suitable decoding rule, such that $P_{E}^{(i)} < \epsilon$ for all $i = 1, \ldots, M$. 