Lecture 16 - Feb 30th, 2012
Outstanding Reading

- Read chapters 1, and 2 in C&T.
- Read chapter 3 in C&T.
- Read section 11.1, 11.3, method of types and universal source coding.
- Read chapter 4.
- Read chapter 5.
- Read stream code chapter 6 in “Information Theory, Inference, and Learning Algorithms” by David J.C. MacKay (available online http://www.inference.phy.cam.ac.uk/mackay/itila/)
- Read chapter 7 in Cover and Thomas, channel capacity
Announcements, Assignments, and Reminders

- Homework 5 out, due Thursday (tonight) March 1st, 11:45pm via our dropbox
  (https://catalyst.uw.edu/collectit/dropbox/karna/19164)
- Homework 6 is likely to be available on Friday, March 2nd.
- Late policy: 10% every 24 hour period that you are late, and no more than 3 days late accepted.
- Lowest grade out of all HW grades is not counted towards final grade (so you can skip one HW with impunity).
- Please do use our discussion board (https://catalyst.uw.edu/gopost/board/karna/25503/) for all questions, so that all will benefit from them being answered.
Class Road Map - IT-I

- L1 (1/3): Overview, Entropy
- L2 (1/5): Props. Entropy, Mutual Information, KL-Divergence
- L3 (1/10): KL-Divergence, Jensen, properties, Data Proc. Inequality
- L5 (1/17): Fano, AEP
- L6 (1/19): snow
- L6 (1/24): AEP, source coding
- L7 (1/26): Method of Types
- L9 (2/2): HMMs, coding
- L10 (2/7): Coding, Kraft,
- L11 (2/9): Huffman, midterm
- L12 (2/14): Midterm
- L13 (2/16): Shannon Games, Arithmetic
- L14 (2/21): Channel Capacity
- L16 (2/28): Shannon’s 2nd theorem.
- L18 (3/6):
- L19 (3/8):

Finals Week: March 12th–16th.
Shannon’s 2nd Theorem: Intuition

- Goal: find a non-confusable subset of the inputs that produce disjoint output sequences (as in picture).
Shannon’s 2nd Theorem: Intuition

- Goal: find a non-confusable subset of the inputs that produce disjoint output sequences (as in picture).
- There are $\approx 2^{nH(Y)}$ (typical) outputs (i.e., the marginally typical $Y$ sequences).
Shannon’s 2nd Theorem: Intuition

- **Goal**: find a non-confusable subset of the inputs that produce disjoint output sequences (as in picture).
- There are \( \approx 2^{nH(Y)} \) (typical) outputs (i.e., the marginally typical \( Y \) sequences).
- There are \( \approx 2^{nH(Y|X)} \) (\( X \)-conditionally typical \( Y \) sequences) outputs.

Note, in non-ideal case, there could be overlap of the typical \( Y \)-given-\( X \) sequences, but the best we can do (in terms of maximizing the number of non-confusable inputs) is when there is no overlap on the output. This is assumed in the above.
Shannon’s 2nd Theorem: Intuition

- Goal: find a non-confusable subset of the inputs that produce disjoint output sequences (as in picture).
- There are \( \approx 2^{nH(Y)} \) (typical) outputs (i.e., the marginally typical \( Y \) sequences).
- There are \( \approx 2^{nH(Y|X)} \) (\( X \)-conditionally typical \( Y \) sequences) outputs. \( \equiv \) the average possible number of outputs for a possible input, so this many could be confused with each other.
Goal: find a non-confusable subset of the inputs that produce disjoint output sequences (as in picture).

- There are \( \approx 2^{nH(Y)} \) (typical) outputs (i.e., the marginally typical \( Y \) sequences).
- There are \( \approx 2^{nH(Y|X)} \) (\( X \)-conditionally typical \( Y \) sequences) outputs. \( \equiv \) the average possible number of outputs for a possible input, so this many could be confused with each other. I.e., on average, for a given \( X = x \), this is approximately how many outputs there might be.
Shannon’s 2nd Theorem: Intuition

- **Goal:** find a non-confusable subset of the inputs that produce disjoint output sequences (as in picture).
- There are $\approx 2^{nH(Y)}$ (typical) outputs (i.e., the marginally typical $Y$ sequences).
- There are $\approx 2^{nH(Y|X)}$ ($X$-conditionally typical $Y$ sequences) outputs. \(\equiv\) the average possible number of outputs for a possible input, so this many could be confused with each other. I.e., on average, for a given $X = x$, this is approximately how many outputs there might be.
- **So the number of non-confusable inputs is**

$$\leq \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^n(H(Y) - H(Y|X)) = 2^{nI(X;Y)} \quad (1)$$

\[\text{Note, in non-ideal case, there could be overlap of the typical } Y - \text{given-} X \text{ sequences, but the best we can do (in terms of maximizing the number of non-confusable inputs) is when there is no overlap on the output. This is assumed in the above.}\]
Shannon’s 2nd Theorem: Intuition

- Goal: find a non-confusable subset of the inputs that produce disjoint output sequences (as in picture).
- There are $\approx 2^{nH(Y)}$ (typical) outputs (i.e., the marginally typical $Y$ sequences).
- There are $\approx 2^{nH(Y|X)}$ ($X$-conditionally typical $Y$ sequences) outputs. $\equiv$ the average possible number of outputs for a possible input, so this many could be confused with each other. I.e., on average, for a given $X = x$, this is approximately how many outputs there might be.
- So the number of non-confusable inputs is

$$\leq \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^n(H(Y) - H(Y|X)) = 2^nI(X;Y) \quad (1)$$

- Note, in non-ideal case, there could be overlap of the typical $Y$-given-$X$ sequences, but the best we can do (in terms of maximizing the number of non-confusable inputs) is when there is no overlap on the output. This is assumed in the above.
Shannon’s 2nd Theorem: Intuition

- The number of non-confusable inputs is

\[
\leq \frac{2^nH(Y)}{2^nH(Y|X)} = 2^n(H(Y) - H(Y|X)) = 2^nI(X;Y)
\]  (2)
Shannon’s 2nd Theorem: Intuition

- The number of non-confusable inputs is
  \[
  \leq \frac{2^n H(Y)}{2^n H(Y|X)} = 2^n (H(Y) - H(Y|X)) = 2^n I(X;Y) \tag{2}
  \]

- We can view this as a volume. $2^n H(Y)$ is the total number of possible slots, while $2^n H(Y|X)$ is the number of slots taken up (on average) for a given source word. Thus, the number of source words that can be used is the ratio.
Shannon’s 2nd Theorem: Intuition

- The number of non-confusable inputs is

\[
\leq \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n(H(Y) - H(Y|X))} = 2^{nI(X;Y)}
\]  

- We can view this as a volume. $2^{nH(Y)}$ is the total number of possible slots, while $2^{nH(Y|X)}$ is the number of slots taken up (on average) for a given source word. Thus, the number of source words that can be used is the ratio.
Shannon’s 2nd Theorem: Intuition

- The number of non-confusable inputs is

\[ \leq \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^n(H(Y) - H(Y|X)) = 2^nI(X;Y) \] (3)
Shannon’s 2nd Theorem: Intuition

- The number of non-confusable inputs is

\[
\leq 2^{nH(Y)} \leq 2^{n(H(Y) - H(Y|X))} = 2^{nI(X;Y)} \tag{3}
\]

- Now of course, to maximize this number, for a fixed channel \( p(y|x) \), we find the best \( p(x) \) which gives \( I(X;Y) = C \), which is the log of the maximum number of inputs possible to use.
Shannon’s 2nd Theorem: Intuition

- The number of non-confusable inputs is

$$\leq \frac{2^n H(Y)}{2^n H(Y|X)} = 2^n (H(Y) - H(Y|X)) = 2^n I(X;Y) \quad (3)$$

- Now of course, to maximize this number, for a fixed channel $p(y|x)$, we find the best $p(x)$ which gives $I(X;Y) = C$, which is the log of the maximum number of inputs possible to use.

- This is the capacity.
Some Definitions

- Reminder: model of communication:
  
  **noise** \( p(y|x) \)

- **Message** \( W \in \{1, \ldots, M\} \) requiring \( \log M \) bits per message.
- **Signal** sent through channel \( X^n(W) \), a random codeword.
- **Received signal** from channel \( Y^n \sim p(y^n|x^n) \)
- **Decoding** via guess \( \hat{W} = g(Y^n) \).
- **Discrete memoryless channel (DMC)** \( (\mathcal{X}, p(y|x), \mathcal{Y}) \)
- \( n^{th} \) extension to channel is \( (\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n) \)
- **Feedback** if \( x_k \) can use both previous inputs and outputs.
- **No feedback** if \( p(x_k|x_{1:k-1}, y_{1:k-1}) = p(x_k|x_{1:k-1}) \). We’ll analyze feedback a bit later.
Definition 2.1 \((M, n)\) code

An \((M, n)\) code for channel \((X, p(y|x), Y)\) is:

1. An index set \(\{1, 2, \ldots, M\}\)

2. An encoding function \(X^n : \{1, 2, \ldots, M\} \rightarrow X^n\) yielding codewords \(X^n(1), X^n(2), X^n(3), \ldots, X^n(M)\). Each source message has a codeword, and each codeword is \(n\) code symbols.

3. Decoding function, i.e., \(g : Y^n \rightarrow \{1, 2, \ldots, M\}\) which makes a "guess" about original message given channel output.

- In an \((M, n)\) code, \(M = \) the number of possible messages to be sent, and \(n = \) number of channel uses by the codewords of the code.
Definition 2.2 (Probability of Error $\lambda_i$ for message $i \in \{1, \ldots, M\}$)

$$
\lambda_i \triangleq \Pr(g(Y^n) \neq i | X^n = X^n(i)) = \sum_{y^n \in Y^n} p(y^n | X^n(i))1(g(y^n) \neq i)
$$
Definition 2.2 (Probability of Error $\lambda_i$ for message $i \in \{1, \ldots, M\}$)

$$\lambda_i \triangleq \Pr(g(Y^n) \neq i | X^n = X^n(i)) = \sum_{y^n \in Y^n} p(y^n | X^n(i)) 1(g(y^n) \neq i)$$  \hspace{1cm} (4)

Definition 2.3 (Max probability of Error $\lambda^{(n)}$ for $(M, n)$ code)

$$\lambda^{(n)} \triangleq \max_{i \in \{1, 2, \ldots, M\}} \lambda_i$$  \hspace{1cm} (5)
Definition 2.4 (Average probability of error $P_{e}^{(n)}$)

$$P_{e}^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_{i} = \Pr(I \neq g(Y^{n}))$$

(6)

where $I$ is a r.v. with probability $\Pr(I = i)$ according to a uniform source distribution . . .

$$= E(1(I \neq g(Y^{n}))) = \sum_{i=1}^{M} \Pr(g(Y^{n}) \neq i | X^{n} = X^{n}(i))p(i)$$

(7)

with $p(i) = 1/M$. 

Definition 2.4 (Average probability of error $P_e^{(n)}$)

$$
P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i = \Pr(I \neq g(Y^n))
$$

where $I$ is a r.v. with probability $\Pr(I = i)$ according to a uniform source distribution . . .

$$
= E(\mathbf{1}(I \neq g(Y^n))) = \sum_{i=1}^{M} \Pr(g(Y^n) \neq i | X^n = X^n(i)) p(i)
$$

with $p(i) = 1/M$.

A key Shannon’s result is that a small average probability of error means we must have a small maximum probability of error!
Definition 2.5 (Rate $R$ of an $(M, n)$ code)

$$R = \frac{\log M}{n} = \frac{\text{total num. of bits in a source message}}{\text{total num. of channel uses needed to send a message}}$$  \hspace{1cm} (8)
Definition 2.5 (Rate $R$ of an $(M, n)$ code)

$$R = \frac{\log M}{n} = \frac{\text{total num. of bits in a source message}}{\text{total num. of channel uses needed to send a message}}$$ (8)

- The rate $R$ is in units of bits per channel use, or bits per transmission.
**Rate**

**Definition 2.5 (Rate $R$ of an $(M,n)$ code)**

\[ R = \frac{\log M}{n} = \frac{\text{total num. of bits in a source message}}{\text{total num. of channel uses needed to send a message}} \]

The rate $R$ is in units of bits per channel use, or bits per transmission.

**Definition 2.6 (Achievability for a given channel)**

A given rate $R$ is achievable for a given channel if $\exists$ a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the maximal probability of error $\lambda^{(n)} \to 0$ as $n \to \infty$. 
Definition 2.7 (Capacity of a DMC)

The capacity of a DMC is the largest possible achievable rate.

- So the capacity of a DMC is the rate beyond which the error won’t any longer go to zero with increasing $n$. 
Capacity

Definition 2.7 (Capacity of a DMC)

The capacity of a DMC is the largest possible achievable rate.

- So the capacity of a DMC is the rate beyond which the error won’t any longer go to zero with increasing $n$.
- Note: this is a different notion of capacity that we encountered before.
Definition 2.7 (Capacity of a DMC)

The capacity of a DMC is the largest possible achievable rate.

- So the capacity of a DMC is the rate beyond which the error won’t any longer go to zero with increasing $n$.
- Note: this is a different notion of capacity that we encountered before.
- Before we defined $C = \max_{p(x)} I(X; Y)$. 
Definition 2.7 (Capacity of a DMC)

The **capacity** of a DMC is the largest possible achievable rate.

- So the capacity of a DMC is the rate beyond which the error won’t any longer go to zero with increasing $n$.
- Note: this is a different notion of capacity that we encountered before.
- Before we defined $C = \max_{p(x)} I(X; Y)$.
- Here we are defining something called the “capacity of a DMC”.
Definition 2.7 (Capacity of a DMC)

The capacity of a DMC is the largest possible achievable rate.

- So the capacity of a DMC is the rate beyond which the error won’t any longer go to zero with increasing $n$.
- Note: this is a different notion of capacity that we encountered before.
- Before we defined $C = \max_{p(x)} I(X; Y)$.
- Here we are defining something called the “capacity of a DMC”.
- We have not yet compared the two
Definition 2.7 (Capacity of a DMC)

The **capacity** of a DMC is the largest possible achievable rate.

- So the capacity of a DMC is the rate beyond which the error won’t any longer go to zero with increasing $n$.
- Note: this is a different notion of capacity that we encountered before.
- Before we defined $C = \max_{p(x)} I(X; Y)$.
- Here we are defining something called the “capacity of a DMC”.
- We have not yet compared the two (but of course we will 😊).
Joint Typicality

Definition 2.8 (Joint typicality of a set of sequences)

A set of sequences \( \{(x_1:n, y_1:n)\} \) w.r.t. \( p(x, y) \) is jointly typical \( (\in A^{(n)}_\epsilon) \) as per the following definition:

\[
A^{(n)}_\epsilon = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \right. \\
\left. \left| \left| \frac{1}{n} \log p(x^n) - H(X) \right| \right| < \epsilon, \text{ } x\text{-typical} \right) \\
\left. \left| \left| \frac{1}{n} \log p(y^n) - H(Y) \right| \right| < \epsilon, \text{ } y\text{-typical} \right) \\
\left. \left| \left| \frac{1}{n} \log p(x^n, y^n) - H(X,Y) \right| \right| < \epsilon, \text{ } (x,y)\text{-typical} \right) 
\]

(9)

\[
A^{(n)}_\epsilon = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \right. \\
\left. \left| \left| \frac{1}{n} \log p(x^n, y^n) - \frac{1}{n} \sum_i \log p(x_i, y_i) \right| \right| < \epsilon, \text{ } (x,y)\text{-typical} \right) 
\]

(12)

with \( p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i) \).
Definition 2.8 (Joint typicality of a set of sequences)

A set of sequences \( \{(x_{1:n}, y_{1:n})\} \) w.r.t. \( p(x, y) \) is jointly typical (\( \in A^{(n)}_\epsilon \)) as per the following definition:

\[
A^{(n)}_\epsilon = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \right. \\
\left. \begin{array}{l}
a) \quad \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \quad \text{x-typical} \\
b) \quad \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \quad \text{y-typical} \\
c) \quad \left| -\frac{1}{n} \log p(x^n, y^n) - H(X,Y) \right| < \epsilon, \quad \text{(x,y)-typical}
\end{array} \right\}
\]

with \( p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i) \).
Definition 2.8 (Joint typicality of a set of sequences)

A set of sequences \( \{(x_1^n, y_1^n)\} \) w.r.t. \( p(x, y) \) is jointly typical \((\in A_{\epsilon}^{(n)})\) as per the following definition:

\[
A_{\epsilon}^{(n)} = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \right. \\
\left. a) \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \quad x\text{-typical} \right. \\
\left. b) \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \quad y\text{-typical} \right. \\
\left. c) \left| -\frac{1}{n} \log p(x^n, y^n) - H(X,Y) \right| < \epsilon, \quad (x,y)\text{-typical} \right\} 
\]

with \( p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i) \).
Definition 2.8 (Joint typicality of a set of sequences)

A set of sequences \( \{(x_{1:n}, y_{1:n})\} \) w.r.t. \( p(x, y) \) is jointly typical \((\in A^{(n)}_\epsilon)\) as per the following definition:

\[
A^{(n)}_\epsilon = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \right. \\
\begin{aligned}
& a) \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \quad x\text{-typical} \\
& b) \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \quad y\text{-typical} \\
& c) \left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon, \quad (x, y)\text{-typical}
\end{aligned}
\]  

(9)

(10)

(11)

(12)

with \( p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i) \).
Jointly Typical Sequences: Picture

- Set of all jointly typical pairs of sequences: $2^{nH(X,Y)}$
- Set of all pairs of sequences: $|\mathcal{X}^n \times \mathcal{Y}^n| = (|\mathcal{X}|^n |\mathcal{Y}|^n)$

$\mathcal{X}^n \rightarrow 2^{nH(X)}$

$\mathcal{Y}^n \rightarrow 2^{nH(Y)}$

$2^{nH(Y|X)}$

$2^{nH(X,Y)}$

$2^{nH(Y)}$
So intuitively,

\[
\frac{\text{num. jointly typical seqs.}}{\text{num ind. chosen typical seqs.}} = \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{n(H(X,Y) - H(X) - H(Y))} = 2^{-nI(X;Y)}
\]

(13) \hspace{1cm} (14) \hspace{1cm} (15)
So intuitively,

\[
\frac{\text{num. jointly typical seqs.}}{\text{num ind. chosen typical seqs.}} = \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}}
\]

\[
= 2^{n(H(X,Y) - H(X) - H(Y))}
\]

\[
= 2^{-nI(X;Y)}
\]

So if we independently at random choose two (singly) typical sequences for $X$ and $Y$, then the chance that it will be an $(X, Y)$ jointly typical sequence decreases exponentially with $n$, as long as $I(X;Y) > 0$. 
So intuitively,

\[
\frac{\text{num. jointly typical seqs.}}{\text{num ind. chosen typical seqs.}} = \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^n(H(X,Y) - H(X) - H(Y)) = 2^{-nI(X;Y)}
\]  (13) 

(14) 

(15)

So if we independently at random choose two (singly) typical sequences for \(X\) and \(Y\), then the chance that it will be an \((X, Y)\) jointly typical sequence decreases exponentially with \(n\), as long as \(I(X; Y) > 0\).

To decrease this chance as much as possible, it can become \(2^{-nC}\).
Theorem 2.9

Let \((X^n, Y^n) \sim p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)\). Then
Joint AEP

Theorem 2.9

Let \((X^n, Y^n) \sim p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)\). Then

1. \(\Pr((X^n, Y^n) \in A^{(n)}_c) \rightarrow 1\) as \(n \rightarrow \infty\).
Joint AEP

Theorem 2.9

Let \((X^n, Y^n) \sim p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)\). Then

1. \(\Pr\left( (X^n, Y^n) \in A_{\epsilon}^{(n)} \right) \to 1 \text{ as } n \to \infty.\)

2. \(|A_{\epsilon}^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)} \text{ and } (1 - \epsilon)2^{n(H(X,Y)-\epsilon)} \leq |A_{\epsilon}^{(n)}|\).
Joint AEP

Theorem 2.9

Let $(X^n, Y^n) \sim p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$. Then

1. $Pr \left( (X^n, Y^n) \in A^{(n)}_\epsilon \right) \to 1$ as $n \to \infty$.

2. $|A^{(n)}_\epsilon| \leq 2^n(H(X,Y)+\epsilon)$ and $(1 - \epsilon)2^n(H(X,Y)-\epsilon) \leq |A^{(n)}_\epsilon|$.

3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$ are drawn independently, then

   $Pr \left( (\tilde{X}^n, \tilde{Y}^n) \in A^{(n)}_\epsilon \right) \leq 2^{-n(I(X;Y)-3\epsilon)}$ \hspace{1cm} (16)

and for sufficiently large $n$, we have

   $Pr \left( (\tilde{X}^n, \tilde{Y}^n) \in A^{(n)}_\epsilon \right) \geq (1 - \epsilon)2^{-n(I(X;Y)+3\epsilon)}$ \hspace{1cm} (17)
Joint AEP

Theorem 2.9

Let \((X^n, Y^n) \sim p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)\). Then

1. \(\Pr \left( (X^n, Y^n) \in A_{\epsilon}^{(n)} \right) \to 1 \text{ as } n \to \infty.\)

2. \(|A_{\epsilon}^{(n)}| \leq 2^n(H(X,Y)+\epsilon) \text{ and } (1-\epsilon)2^n(H(X,Y)-\epsilon) \leq |A_{\epsilon}^{(n)}|\).

3. If \((\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)\) are drawn independently, then

\[
\Pr \left( (\tilde{X}^n, \tilde{Y}^n) \in A_{\epsilon}^{(n)} \right) \leq 2^{-n(I(X;Y)-3\epsilon)} \tag{16}
\]

and for sufficiently large \(n\), we have

\[
\Pr \left( (\tilde{X}^n, \tilde{Y}^n) \in A_{\epsilon}^{(n)} \right) \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \tag{17}
\]

Key property: we have bound on the probability of independently drawn sequences being jointly typical, falls off exponentially fast with \(n\), if \(I(X;Y) > 0\).
Another Intuitive (and somewhat redundant) Reprieve

- There are $\approx 2^{nH(X)}$ typical $X$ sequences
Another Intuitive (and somewhat redundant) Reprieve

- There are $\approx 2^{nH(X)}$ typical $X$ sequences
- There are $\approx 2^{nH(Y)}$ typical $Y$ sequences.
Another Intuitive (and somewhat redundant) Reprieve

- There are $\approx 2^{nH(X)}$ typical $X$ sequences.
- There are $\approx 2^{nH(Y)}$ typical $Y$ sequences.
- The total number of independent typical pairs is $\approx 2^{nH(X)}2^{nH(Y)}$. 
Another Intuitive (and somewhat redundant) Reprieve

- There are $\approx 2^{nH(X)}$ typical $X$ sequences.
- There are $\approx 2^{nH(Y)}$ typical $Y$ sequences.
- The total number of independent typical pairs is $\approx 2^{nH(X)}2^{nH(Y)}$, but not all of them are jointly typical. Rather only $\approx 2^{nH(X,Y)}$ of them are jointly typical.
Another Intuitive (and somewhat redundant) Reprieve

- There are $\approx 2^{nH(X)}$ typical $X$ sequences.
- There are $\approx 2^{nH(Y)}$ typical $Y$ sequences.
- The total number of independent typical pairs is $\approx 2^{nH(X)}2^{nH(Y)}$, but not all of them are jointly typical. Rather only $\approx 2^{nH(X,Y)}$ of them are jointly typical.
- The fraction of independent typical sequences that are jointly typical is:

$$\frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{n(H(X,Y) - H(X) - H(Y))} = 2^{-nI(X,Y)} \quad (18)$$
Another Intuitive (and somewhat redundant) Reprieve

- There are $\approx 2^{nH(X)}$ typical $X$ sequences.
- There are $\approx 2^{nH(Y)}$ typical $Y$ sequences.
- The total number of independent typical pairs is $\approx 2^{nH(X)}2^{nH(Y)}$, but not all of them are jointly typical. Rather only $\approx 2^{nH(X,Y)}$ of them are jointly typical.
- The fraction of independent typical sequences that are jointly typical is:

$$\frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{n(H(X,Y) - H(X) - H(Y))} = 2^{-nI(X,Y)}$$

(18)

and this is essentially the probability that a randomly chosen pair of (marginally) typical sequences is jointly typical.
Jointly Typical Sequences: Picture

- Set of all jointly typical pairs of sequences: $2^{nH(X,Y)}$
- Set of all pairs of sequences: $|X^n \times Y^n| = (|X||Y|)^n$

Diagram:
- $X^n$ and $Y^n$ with corresponding entropies $2^{nH(X)}$ and $2^{nH(Y)}$
- $2^{nH(Y|X)}$ and $2^{nH(X,Y)}$
More Intuition

- So if we use typicality to decode (which we will) then there are about $2^{nI(X;Y)}$ pairs of sequences available before we start needing to use pairs that would be jointly typical if chosen randomly.
More Intuition

- So if we use typicality to decode (which we will) then there are about $2^{nI(X;Y)}$ pairs of sequences available before we start needing to use pairs that would be jointly typical if chosen randomly.

- Ex: if $p(x) = 1/M$ then we can choose about $M$ samples before we see a given particular $x$, on average.
The basic idea is to use joint typicality.
Channel Coding Theorem (Shannon 1948)

- The basic idea is to use joint typicality.
- Given a received codeword $y^n$, find an $x^n$ that is jointly typical with $y^n$. 
The basic idea is to use joint typicality.

Given a received codeword $y^n$, find an $x^n$ that is jointly typical with $y^n$.

This $x^n$ will occur jointly with $y^n$ with probability 1, for large enough $n$. 
The basic idea is to use joint typicality.

Given a received codeword $y^n$, find an $x^n$ that is jointly typical with $y^n$. This $x^n$ will occur jointly with $y^n$ with probability 1, for large enough $n$.

Also, the probability that some other $\hat{x}^n$ is jointly typical with $y^n$ is about $2^{-nI(X;Y)}$. 
The basic idea is to use joint typicality.

Given a received codeword $y^n$, find an $x^n$ that is jointly typical with $y^n$.

This $x^n$ will occur jointly with $y^n$ with probability 1, for large enough $n$.

Also, the probability that some other $\hat{x}^n$ is jointly typical with $y^n$ is about $2^{-nI(X;Y)}$,

so if we use $< 2^{nI(X;Y)}$ codewords, then some other sequence being jointly typical will occur with vanishingly small probability for large $n$. 
Theorem 3.1

Channel Coding Theorem (Shannon 1948): more formally

All rates below $C$ are achievable. Specifically, $\forall R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda(n) \to 0$ as $n \to \infty$.

Conversely, any $(2^{nR}, n)$ sequence of codes with $\lambda(n) \to 0$ as $n \to \infty$ must have that $R < C$.

Implications: as long as we do not code above capacity we can, for all intents and purposes, code with zero error. This is true for all noisy channels representable under this model. We're talking about discrete channels now, but we generalize this to continuous channels in the coming weeks.
Theorem 3.1

All rates below \( C \) are achievable. Specifically, \( \forall R < C \), there exists a sequence of \( (2^{nR}, n) \) codes with maximum probability of error \( \lambda^{(n)} \to 0 \) as \( n \to \infty \).
Theorem 3.1

All rates below $C$ are achievable. Specifically, $\forall R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda(n) \to 0$ as $n \to \infty$. Conversely, any $(2^{nR}, n)$ sequence of codes with $\lambda(n) \to 0$ as $n \to \infty$ must have that $R < C$. 

Implications: as long as we do not code above capacity we can, for all intents and purposes, code with zero error. This is true for all noisy channels representable under this model. We’re talking about discrete channels now, but we generalize this to continuous channels in the coming weeks.
Channel Coding Theorem (Shannon 1948): more formally

**Theorem 3.1**

All rates below $C$ are achievable. Specifically, $\forall R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$ as $n \to \infty$. Conversely, any $(2^{nR}, n)$ sequence of codes with $\lambda^{(n)} \to 0$ as $n \to \infty$ must have that $R < C$.

- Implications: as long as we do not code above capacity we can, for all intents and purposes, code with zero error.
Theorem 3.1

All rates below $C$ are achievable. Specifically, $\forall R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$ as $n \to \infty$. Conversely, any $(2^{nR}, n)$ sequence of codes with $\lambda^{(n)} \to 0$ as $n \to \infty$ must have that $R < C$.

- Implications: as long as we do not code above capacity we can, for all intents and purposes, code with zero error.
- This is true for all noisy channels representable under this model.
Theorem 3.1

All rates below $C$ are achievable. Specifically, $\forall R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \rightarrow 0$ as $n \rightarrow \infty$. Conversely, any $(2^{nR}, n)$ sequence of codes with $\lambda^{(n)} \rightarrow 0$ as $n \rightarrow \infty$ must have that $R < C$.

- Implications: as long as we do not code above capacity we can, for all intents and purposes, code with zero error.
- This is true for all noisy channels representable under this model.
- We’re talking about discrete channels now, but we generalize this to continuous channels in the coming weeks.
Channel Theorem

- We could look at error for a particular code and bound its errors.
Channel Theorem

- We could look at error for a particular code and bound its errors.
- Instead, we look at the average probability of errors of all codes generated randomly.
Channel Theorem

- We could look at error for a particular code and bound its errors.
- Instead, we look at the average probability of errors of all codes generated randomly.
- We then prove that this average error is small.
Channel Theorem

- We could look at error for a particular code and bound its errors.
- Instead, we look at the average probability of errors of all codes generated randomly.
- We then prove that this average error is small.
- This implies \( \exists \) many good codes to make the average small.
Channel Theorem

- We could look at error for a particular code and bound its errors.
- Instead, we look at the average probability of errors of all codes generated randomly.
- We then prove that this average error is small.
- This implies ∃ many good codes to make the average small.
- To show that the maximum probability of error also small, we throw away the worst 50% of the codes.
Channel Theorem

- We could look at error for a particular code and bound its errors.
- Instead, we look at the average probability of errors of all codes generated randomly.
- We then prove that this average error is small.
- This implies $\exists$ many good codes to make the average small.
- To show that the maximum probability of error also small, we throw away the worst 50% of the codes.
- Recall: idea is, for a given channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ come up with a $(2^{nR}, n)$ code of rate $R$ which means we need:
Channel Theorem

- We could look at error for a particular code and bound its errors.
- Instead, we look at the average probability of errors of all codes generated randomly.
- We then prove that this average error is small.
- This implies $\exists$ many good codes to make the average small.
- To show that the maximum probability of error also small, we throw away the worst 50% of the codes.
- Recall: idea is, for a given channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ come up with a $(2^{nR}, n)$ code of rate $R$ which means we need:
  - Index set $\{1, \ldots, M\}$
Channel Theorem

- We could look at error for a particular code and bound its errors.
- Instead, we look at the average probability of errors of all codes generated randomly.
- We then prove that this average error is small.
- This implies \( \exists \) many good codes to make the average small.
- To show that the maximum probability of error also small, we throw away the worst 50% of the codes.
- Recall: idea is, for a given channel \( (\mathcal{X}, p(y|x), \mathcal{Y}) \) come up with a \( (2^{nR}, n) \) code of rate \( R \) which means we need:
  1. Index set \( \{1, \ldots, M\} \)
  2. Encoder: \( X^n : \{1, \ldots, M\} \rightarrow \mathcal{X}^n \) maps to codewords \( X^n(i) \)
We could look at error for a particular code and bound its errors. Instead, we look at the average probability of errors of all codes generated randomly. We then prove that this average error is small. This implies \( \exists \) many good codes to make the average small. To show that the maximum probability of error also small, we throw away the worst 50% of the codes.

Recall: idea is, for a given channel \((\mathcal{X}, p(y|x), \mathcal{Y})\) come up with a \((2^{nR}, n)\) code of rate \(R\) which means we need:

1. Index set \(\{1, \ldots, M\}\)
2. Encoder: \(X^n : \{1, \ldots, M\} \rightarrow \mathcal{X}^n\) maps to codewords \(X^n(i)\)
3. Decoder: \(g : \mathcal{Y}^n \rightarrow \{1, \ldots, M\}\).
Channel Theorem

- We could look at error for a particular code and bound its errors.
- Instead, we look at the average probability of errors of all codes generated randomly.
- We then prove that this average error is small.
- This implies \( \exists \) many good codes to make the average small.
- To show that the maximum probability of error also small, we throw away the worst 50% of the codes.

Recall: idea is, for a given channel \((\mathcal{X}, p(y|x), \mathcal{Y})\) come up with a \((2^{nR}, n)\) code of rate \(R\) which means we need:

1. Index set \(\{1, \ldots, M\}\)
2. Encoder: \(X^n : \{1, \ldots, M\} \rightarrow \mathcal{X}^n\) maps to codewords \(X^n(i)\)
3. Decoder: \(g : \mathcal{Y}^n \rightarrow \{1, \ldots, M\}\).

Two parts to prove: 1) all rates \(R < C\) are achievable (exists a code with vanishing error). Conversely, 2) if the error goes to zero, then must have \(R < C\).
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Given $R < C$, assume use of $p(x)$ and generate $2^{nR}$ random codewords using $p(x^n) = \prod_{i=1}^{n} p(x_i)$. 
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Given $R < C$, assume use of $p(x)$ and generate $2^{nR}$ random codewords using $p(x^n) = \prod_{i=1}^{n} p(x_i)$.

- Choose $p(x)$ arbitrarily for now, and then change it later to get $C$.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Given $R < C$, assume use of $p(x)$ and generate $2^{nR}$ random codewords using $p(x^n) = \prod_{i=1}^{n} p(x_i)$.
- Choose $p(x)$ arbitrarily for now, and then change it later to get $C$.
- Set of random codewords (the codebook) can be seen as a matrix:

$$C = \begin{bmatrix}
    x_1(1) & x_2(1) & x_3(1) & \ldots & x_n(1) \\
    x_1(2) & x_2(2) & x_3(2) & \ldots & x_n(2) \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    x_1(2^{nR}) & x_2(2^{nR}) & x_3(2^{nR}) & \ldots & x_n(2^{NR})
\end{bmatrix} \quad (19)$$
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Given $R < C$, assume use of $p(x)$ and generate $2^{nR}$ random codewords using $p(x^n) = \prod_{i=1}^{n} p(x_i)$.
- Choose $p(x)$ arbitrarily for now, and then change it later to get $C$.
- Set of random codewords (the codebook) can be seen as a matrix:

$$
\mathcal{C} = 
\begin{bmatrix}
x_1(1) & x_2(1) & x_3(1) & \cdots & x_n(1) \\
x_1(2) & x_2(2) & x_3(2) & \cdots & x_n(2) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
x_1(2^{nR}) & x_2(2^{nR}) & x_3(2^{nR}) & \cdots & x_n(2^{nR})
\end{bmatrix}
$$

- So, there are $2^{nR}$ codes each of length $n$ generated via $p(x)$. 

...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Given $R < C$, assume use of $p(x)$ and generate $2^{nR}$ random codewords using $p(x^n) = \prod_{i=1}^{n} p(x_i)$.
- Choose $p(x)$ arbitrarily for now, and then change it later to get $C$.
- Set of random codewords (the codebook) can be seen as a matrix:

$$C = \begin{bmatrix}
x_1(1) & x_2(1) & x_3(1) & \ldots & x_n(1) \\
x_1(2) & x_2(2) & x_3(2) & \ldots & x_n(2) \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
x_1(2^{nR}) & x_2(2^{nR}) & x_3(2^{nR}) & \ldots & x_n(2^{NR})
\end{bmatrix} \quad (19)$$

- So, there are $2^{nR}$ codes each of length $n$ generated via $p(x)$.
- To send any message $i \in \{1, 2, \ldots, M = 2^{nR}\}$, we send codeword $x_{1:n}(i) = (x_1(i), x_2(i), \ldots, x_n(i))$. 
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Can compute probabilities of a given codeword for $\omega$ …

$$p(x_n(\omega)) = \prod_{i=1}^{n} p(x_i(\omega)), \omega \in \{1, \ldots, M\} \tag{20}$$
All rates $R < C$ are achievable.

**Proof that all rates $R < C$ are achievable.**

- Can compute probabilities of a given codeword for $\omega$ ...

$$p(x^n(\omega)) = \prod_{i=1}^{n} p(x_i(\omega)), \quad \omega \in \{1, \ldots, M\}$$  \hspace{1cm} (20)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

1. Can compute probabilities of a given codeword for $\omega$ ...

$$p(x^n(\omega)) = \prod_{i=1}^{n} p(x_i(\omega)), \quad \omega \in \{1, \ldots, M\}$$ (20)

2. ...or even the entire codebook:

$$p(C) = 2^{nR} \prod_{\omega=1}^{M} \prod_{i=1}^{n} p(x_i(\omega))$$ (21)

...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Can compute probabilities of a given codeword for $\omega$ ...

\[ p(x^n(\omega)) = \prod_{i=1}^{n} p(x_i(\omega)), \; \omega \in \{1, \ldots, M\} \]  \hspace{1cm} (20)

- ... or even the entire codebook:

\[ p(C) = \prod_{\omega=1}^{2^nR} \prod_{i=1}^{n} p(x_i(\omega)) \]  \hspace{1cm} (21)

...
Proof that all rates $R < C$ are achievable.

Consider the following encoding/decoding scheme:
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Consider the following encoding/decoding scheme:

1. Generate a random codebook as above according to $p(x)$
Proof that all rates $R < C$ are achievable.

Consider the following encoding/decoding scheme:

1. Generate a random codebook as above according to $p(x)$
2. Codebook known to both sender/receiver (who also knows $p(y|x)$.

...
Proof that all rates $R < C$ are achievable.

Consider the following encoding/decoding scheme:

1. Generate a random codebook as above according to $p(x)$
2. Codebook known to both sender/receiver (who also knows $p(y|x)$).
3. Generate messages $W$ according to the uniform distribution (we’ll see why shortly), $p(W = \omega) = 2^{-nR}$, for $\omega = 1, \ldots, 2^{nR}$. 

...
Proof that all rates $R < C$ are achievable.

Consider the following encoding/decoding scheme:

1. Generate a random codebook as above according to $p(x)$
2. Codebook known to both sender/receiver (who also knows $p(y|x)$).
3. Generate messages $W$ according to the uniform distribution (we’ll see why shortly), $p(W = \omega) = 2^{-nR}$, for $\omega = 1, \ldots, 2^{nR}$.
4. Send $x^n(\omega)$ over the channel.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Consider the following encoding/decoding scheme:

1. Generate a random codebook as above according to $p(x)$
2. Codebook known to both sender/receiver (who also knows $p(y|x)$).
3. Generate messages $W$ according to the uniform distribution (we’ll see why shortly), $p(W = \omega) = 2^{-nR}$, for $\omega = 1, \ldots, 2^{nR}$.
4. Send $x^n(\omega)$ over the channel.
5. Receiver receives $Y^n$ according to distribution

$$Y^n \sim p(y^n|x^n(\omega)) = \prod_{i=1}^{n} p(y_i|x_i(\omega))$$ (22)

...
Proof that all rates $R < C$ are achievable.

Consider the following encoding/decoding scheme:

1. Generate a random codebook as above according to $p(x)$
2. Codebook known to both sender/receiver (who also knows $p(y|x)$).
3. Generate messages $W$ according to the uniform distribution (we’ll see why shortly), $p(W = \omega) = 2^{-nR}$, for $\omega = 1, \ldots, 2^{nR}$.
4. Send $x^n(\omega)$ over the channel.
5. Receiver receives $Y^n$ according to distribution

$$Y^n \sim p(y^n|x^n(\omega)) = \prod_{i=1}^{n} p(y_i|x_i(\omega))$$ (22)

6. The signal is decoded using typical set decoding (to be described).
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Typical set decoding: Decode message as $\hat{\omega}$ if

...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Typical set decoding: Decode message as $\hat{\omega}$ if

1. $(x^n(\hat{\omega}), y^n)$ is jointly typical
Proof that all rates $R < C$ are achievable.

Typical set decoding: Decode message as $\hat{\omega}$ if

1. $(x^n(\hat{\omega}), y^n)$ is jointly typical
2. $\exists$ no other $k$ s.t. $(x^n(k), y^n) \in A^n_\epsilon$ (i.e., $\hat{\omega}$ is unique)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Typical set decoding: Decode message as $\hat{\omega}$ if

1. $(x^n(\hat{\omega}), y^n)$ is jointly typical
2. $\exists$ no other $k$ s.t. $(x^n(k), y^n) \in A_e^n$ (i.e., $\hat{\omega}$ is unique)

Otherwise output special invalid integer “0” (error).
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Typical set decoding: Decode message as $\hat{\omega}$ if

1. $(x^n(\hat{\omega}), y^n)$ is jointly typical
2. $\exists$ no other $k$ s.t. $(x^n(k), y^n) \in A_e^{(n)}$ (i.e., $\hat{\omega}$ is unique)

Otherwise output special invalid integer “0” (error). Three types of errors might occur (type A, B, or C).
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Typical set decoding: Decode message as $\hat{\omega}$ if

1. $(x^n(\hat{\omega}), y^n)$ is jointly typical
2. $\exists$ no other $k$ s.t. $(x^n(k), y^n) \in A_\epsilon^n$ (i.e., $\hat{\omega}$ is unique)

Otherwise output special invalid integer “0” (error). Three types of errors might occur (type A, B, or C).

A: $\exists k \neq \hat{\omega}$ s.t. $(x^n(k), y^n) \in A_\epsilon^n$ (i.e., $> 1$ possible typical message).
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Typical set decoding: Decode message as $\hat{\omega}$ if

1. $(x^n(\hat{\omega}), y^n)$ is jointly typical
2. $\exists$ no other $k$ s.t. $(x^n(k), y^n) \in A_\epsilon^{(n)}$ (i.e., $\hat{\omega}$ is unique)

Otherwise output special invalid integer “0” (error). Three types of errors might occur (type A, B, or C).

A: $\exists k \neq \hat{\omega}$ s.t. $(x^n(k), y^n) \in A_\epsilon^{(n)}$ (i.e., $> 1$ possible typical message).

B: no $\hat{\omega}$ s.t. $(x^n(\hat{\omega}), y^n)$ is jointly typical.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Typical set decoding: Decode message as $\hat{\omega}$ if

1. $(x^n(\hat{\omega}), y^n)$ is jointly typical
2. $\exists$ no other $k$ s.t. $(x^n(k), y^n) \in A_\epsilon^n$ (i.e., $\hat{\omega}$ is unique)

Otherwise output special invalid integer "0" (error). Three types of errors might occur (type A, B, or C).

A: $\exists k \neq \hat{\omega}$ s.t. $(x^n(k), y^n) \in A_\epsilon^n$ (i.e., $> 1$ possible typical message).

B: no $\hat{\omega}$ s.t. $(x^n(\hat{\omega}), y^n)$ is jointly typical.

C: if $\hat{\omega} \neq \omega$, i.e., wrong codeword is jointly typical.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Typical set decoding: Decode message as $\hat{\omega}$ if

1. $(x^n(\hat{\omega}), y^n)$ is jointly typical
2. $\exists$ no other $k$ s.t. $(x^n(k), y^n) \in A_\varepsilon^{(n)}$ (i.e., $\hat{\omega}$ is unique)

Otherwise output special invalid integer “0” (error). Three types of errors might occur (type A, B, or C).

A: $\exists k \neq \hat{\omega}$ s.t. $(x^n(k), y^n) \in A_\varepsilon^{(n)}$ (i.e., $> 1$ possible typical message).

B: no $\hat{\omega}$ s.t. $(x^n(\hat{\omega}), y^n)$ is jointly typical.

C: if $\hat{\omega} \neq \omega$, i.e., wrong codeword is jointly typical.

Note: maximum likelihood decoding is optimal, but typical set decoding is not, but this will be good enough to show the result.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

(Also) three types of quality measures we might be interested in.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

(also) three types of quality measures we might be interested in.

1. Code specific error

$$P_e^{(n)}(C) = \Pr(\hat{\omega} \neq \omega|C) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i$$

(23)

where (as a reminder)

$$\lambda_i = \Pr(g(y^n) \neq i|X^n = x^n(i)) = \sum_{y^n} p(y^n|x^n(i)) \mathbf{1}\{g(y^n) \neq i\}$$
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

(Also) three types of quality measures we might be interested in.

1. Code specific error

$$P_e(n)(C) = \Pr(\hat{\omega} \neq \omega|C) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i$$

where (as a reminder)

$$\lambda_i = \Pr(g(y^n) \neq i|X^n = x^n(i)) = \sum_{y^n} p(y^n|x^n(i)) 1\{g(y^n) \neq i\}$$

but we would like something easier to analyze.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

(also) three types of quality measures we might be interested in.

1. **Code specific error**

   \[
   P_e^{(n)}(C) = \Pr(\hat{\omega} \neq \omega | C) = \frac{1}{2nR} \sum_{i=1}^{2nR} \lambda_i
   \]  

   where (as a reminder)

   \[
   \lambda_i = \Pr(g(y^n) \neq i | X^n = x^n(i)) = \sum_{y^n} p(y^n|x^n(i)) 1\{g(y^n) \neq i\}
   \]

   but we would like something easier to analyze.

2. **Average error over all randomly generated codes (avg. of avg.)**

   \[
   \Pr(\mathcal{E}) = \sum_C \Pr(C) \Pr(\hat{W} \neq W | C) = \sum_C \Pr(C) P_e(C)
   \]  

   \[
   \text{...}
   \]
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

(also) three types of quality measures we might be interested in.

1. Code specific error

$$P_e^{(n)}(C) = \Pr(\hat{\omega} \neq \omega | C) = \frac{1}{2nR} \sum_{i=1}^{2nR} \lambda_i$$

(23)

where (as a reminder)

$$\lambda_i = \Pr(g(y^n) \neq i | X^n = x^n(i)) = \sum_{y^n} p(y^n | x^n(i))1\{g(y^n) \neq i\}$$

but we would like something easier to analyze.

2. Average error over all randomly generated codes (avg. of avg.)

$$\Pr(\mathcal{E}) = \sum_C \Pr(C) \Pr(\hat{W} \neq W | C) = \sum_C \Pr(C) P_e(C)$$

(24)

Surprisingly, this is much easier to analyze than $P_e$
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

(also) three types of quality measures we might be interested in.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

(also) three types of quality measures we might be interested in.

3. Max error of the code, ultimately what we want to use

$$P_{C,\text{max}}(C) = \max_{i \in \{1, 2, \ldots, M\}} \lambda_i$$ (25)

We want to show that if $R < C$, then exists a codebook $C$ s.t. this error $\to 0$ (and that if $R > C$ error must $\to 1$).
Proof that all rates $R < C$ are achievable.

(also) three types of quality measures we might be interested in.

3. Max error of the code, ultimately what we want to use

$$P_{c, \text{max}}(C) = \max_{i \in \{1,2,\ldots,M\}} \lambda_i$$

(25)

We want to show that if $R < C$, then exists a codebook $C$ s.t. this error $\to 0$ (and that if $R > C$ error must $\to 1$).
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

(also) three types of quality measures we might be interested in.

3. Max error of the code, ultimately what we want to use

$$P_{C, \text{max}}(C) = \max_{i \in \{1, 2, ..., M\}} \lambda_i$$ (25)

We want to show that if $R < C$, then exists a codebook $C$ s.t. this error $\rightarrow 0$ (and that if $R > C$ error must $\rightarrow 1$).

Our method is to:

1. Expand average error (bullet 2 above) and show that it is small.

...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

(Also) three types of quality measures we might be interested in.

1. Max error of the code, ultimately what we want to use

$$P_{C,\text{max}}(C) = \max_{i \in \{1,2,...,M\}} \lambda_i$$ (25)

We want to show that if $R < C$, then exists a codebook $C$ s.t. this error $\rightarrow 0$ (and that if $R > C$ error must $\rightarrow 1$).

Our method is to:

1. Expand average error (bullet 2 above) and show that it is small.
2. Deduce that $\exists$ at least 1 code with small error

...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

(also) three types of quality measures we might be interested in.

3. Max error of the code, ultimately what we want to use

$$P_{C,\text{max}}(C) = \max_{i \in \{1,2,\ldots,M\}} \lambda_i$$  \hspace{1cm} (25)

We want to show that if $R < C$, then exists a codebook $C$ s.t. this error $\to 0$ (and that if $R > C$ error must $\to 1$).

Our method is to:

1. Expand average error (bullet 2 above) and show that it is small.
2. deduce that $\exists$ at least 1 code with small error
3. show that this can be modified to have small maximum probability of error.

...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

$$\Pr(\mathcal{E}) = \sum_{C} \Pr(C) P_{e}^{(n)}(C)$$

(27)

$$
\sum_{C} \Pr(C) \lambda_{\omega}(C) = \sum_{C} \Pr(C) \lambda_{\omega}(g(Y_{n}) | X_{n} = x_{n}(\omega))
$$

(28)

\[ \text{stuff} \]
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

\[
Pr(\mathcal{E}) = \sum_{C} Pr(C) P_{e}^{(n)}(C) = \sum_{C} Pr(C) \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \lambda_{\omega}(C)
\]  (26)

\[
Pr(x_{n}(1), x_{n}(2), \ldots, x_{n}(2^{nR}) \mid x_{n}) = \sum_{x_{n}} \prod_{i=1}^{2^{nR}} Pr(x_{n}(i))
\]  (28)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

\[
\Pr(\mathcal{E}) = \sum_C \Pr(C) P_e^n(C) = \sum_C \Pr(C) \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \lambda_\omega(C)
\]  

(26)

\[
= \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \sum_C \Pr(C) \lambda_\omega(C)
\]  

(27)

(28)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

\[
\Pr(\mathcal{E}) = \sum_{C} \Pr(C) P_e^{(n)}(C) = \sum_{C} \Pr(C) \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \lambda_{\omega}(C) \quad (26)
\]

\[
= \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \sum_{C} \Pr(C) \lambda_{\omega}(C) \quad (27)
\]

but

\[
(28)
\]
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

\[ \Pr(\mathcal{E}) = \sum_C \Pr(C) P_e^n(C) = \sum_C \Pr(C) \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \lambda_\omega(C) \]  

(26)

\[ = \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \sum_C \Pr(C) \lambda_\omega(C) \]  

(27)

but

\[ \sum_C \Pr(C) \lambda_\omega(C) \]  

(28)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

\[ \Pr(\mathcal{E}) = \sum_C \Pr(C) P_{e}^{(n)}(C) = \sum_C \Pr(C) \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \lambda_{\omega}(C) \]  

(26)

\[ = \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \sum_C \Pr(C) \lambda_{\omega}(C) \]  

(27)

but

\[ \sum_C \Pr(C) \lambda_{\omega}(C) = \sum_C \Pr(g(Y^n) \neq \omega | X^n = x^n(\omega)) \Pr(x^n(1), \ldots, x^n(2^{nR})) \]  

(28)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

$$\Pr(\mathcal{E}) = \sum_C \Pr(C) P_e^{(n)}(C) = \sum_C \Pr(C) \frac{1}{2nR} \sum_{\omega=1}^{2^{nR}} \lambda_\omega(C)$$ (26)

$$= \frac{1}{2nR} \sum_{\omega=1}^{2^{nR}} \sum_C \Pr(C) \lambda_\omega(C)$$ (27)

but

$$\sum_C \Pr(C) \lambda_\omega(C) = \sum_C \Pr(g(Y^n) \neq \omega | X^n = x^n(\omega)) \Pr(x^n(1), \ldots, x^n(2^{nR}))$$ (28)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

$$\Pr(\mathcal{E}) = \sum_C \Pr(C) P_{e}^{(n)}(C) = \sum_C \Pr(C) \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \lambda_{\omega}(C) \quad (26)$$

$$= \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \sum_C \Pr(C) \lambda_{\omega}(C) \quad (27)$$

but

$$\sum_C \Pr(C) \lambda_{\omega}(C) = \sum_C \Pr(g(Y^n) \neq \omega | X^n = x^n(\omega)) \Pr(x^n(1), \ldots, x^n(2^{nR}))$$

$$\prod_{i=1}^{2^{nR}} \Pr(x^n(i))$$

$$= \prod_{i=1}^{2^{nR}} \Pr(x^n(i))$$

stuff

$$\quad (28)$$
All rates \( R < C \) are achievable.

Proof that all rates \( R < C \) are achievable.

\[
\Pr(\mathcal{E}) = \sum_C \Pr(C) P_e^{(n)}(C) = \sum_C \Pr(C) \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \lambda_\omega(C) \\
= \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \sum_C \Pr(C) \lambda_\omega(C)
\] (26)

but

\[
\sum_C \Pr(C) \lambda_\omega(C) = \sum_C \Pr(g(Y^n) \neq \omega | X^n = x^n(\omega)) \Pr(x^n(1), \ldots, x^n(2^{nR}))
\]

\[
= \sum_{x^n(1), x^n(2), \ldots, x^n(2^{nR})} \text{stuff}
\] (27)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

$$\sum_C \Pr(C) \lambda_\omega(C)$$  \hspace{1cm} (29)

Last sum is same regardless of $\omega$, call it $\beta$. Thus, we can arbitrarily assume that $\omega = 1$. ...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

\[ \sum_{C} \text{Pr}(C) \lambda_{\omega}(C) \]  

\[ = \sum_{x^n(1), \ldots, x^n(\omega-1), x^n(\omega+1), \ldots, x^n(2nR)} p \left( \frac{x^n(1), \ldots, x^n(\omega-1), x^n(\omega+1), \ldots, x^n(2nR)}{x^n(\omega)} \right) \sum_{x^n(\omega)} \text{Pr}(g(Y^n) \neq \omega | X^n = x^n(\omega)) \text{Pr}(x^n(\omega)) \]  

\[ = \sum_{C} \text{Pr}(C) \lambda_{\omega}(C) \beta \]  

(29)  

(30)  

(31)
Proof that all rates $R < C$ are achievable.

\[
\sum_C \Pr(C) \lambda_\omega(C)
\]

\[
= \sum_{x^n(1), \ldots, x^n(\omega-1), x^n(\omega+1), \ldots, x^n(2^nR)} \prod_{i \neq \omega} \Pr(x^n(i)) \sum_{x^n(\omega)} \Pr(g(Y^n) \neq \omega|X^n = x^n(\omega)) \Pr(x^n(\omega))
\]

Last sum is same regardless of $\omega$, call it $\beta$. Thus, we can assume that $\omega = 1$. 

\[
\text{(29)}
\]

\[
\text{(31)}
\]
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

\[ \sum_C \Pr(C) \lambda_\omega(C) \]

\[ = \sum_{x^n(1), \ldots, x^n(\omega-1), x^n(\omega+1), \ldots, x^n(2nR)} \prod_{i \neq \omega} \Pr(x^n(i)) \]

\[ = \sum_{x^n(\omega)} \Pr(g(Y^n) \neq \omega | X^n = x^n(\omega)) \Pr(x^n(\omega)) \]

\[ = 1 \]
Proof that all rates $R < C$ are achievable.

$$\sum_{C} \Pr(C) \lambda_{\omega}(C)$$

$$= \sum_{x^n(1), \ldots, x^n(\omega-1), x^n(\omega+1), \ldots, x^n(2^nR)} \prod_{i \neq \omega} \Pr(x^n(i)) \cdot p\left(\begin{array}{c} x^n(1), \ldots, x^n(\omega-1), \\
 x^n(\omega+1), \ldots, x^n(2^nR) \end{array}\right) \sum_{x^n(\omega)} \Pr(g(Y^n) \neq \omega \mid X^n = x^n(\omega)) \Pr(x^n(\omega))$$

$$= \sum_{x^n(\omega)} \Pr(g(Y^n) \neq \omega \mid X^n = x^n(\omega)) \Pr(x^n(\omega))$$

$$= \sum_{C} \Pr(C) \lambda_{\omega}(C)$$

(29)  (30)  (31)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

\[ \sum_{C} \Pr(C) \lambda_\omega(C) \]

\[ = \sum_{x^n(1),\ldots,x^n(\omega-1),x^n(\omega+1),\ldots,x^n(2nR)} \prod_{i \neq \omega} \Pr(x^n(i)) \cdot p\left(\frac{x^n(1),\ldots,x^n(\omega-1),x^n(\omega+1),\ldots,x^n(2nR)}{x^n(\omega)}\right) \sum_{x^n(\omega)} \Pr(g(Y^n) \neq \omega | X^n = x^n(\omega)) \Pr(x^n(\omega)) \]

\[ = 1 \]

\[ = \sum_{x^n(\omega)} \Pr(g(Y^n) \neq \omega | X^n = x^n(\omega)) \Pr(x^n(\omega)) \]

\[ = \sum_{x^n} \Pr(g(Y^n) \neq 1 | X^n = x^n(1)) \Pr(x^n(1)) \]

(30)

(31)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

\[
\sum_{C} \Pr(C) \lambda_{\omega}(C)
\]

\[
= \sum_{x^n(1),\ldots,x^n(\omega-1),\atop x^n(\omega+1),\ldots,x^n(2nR)} \prod_{i \neq \omega} \Pr(x^n(i)) \left[ \frac{x^n(1),\ldots,x^n(\omega-1),\atop x^n(\omega+1),\ldots,x^n(2nR)}{x^n(\omega)} \right]
\]

\[
= \sum_{x^n(\omega)} \Pr(g(Y^n) \neq \omega | X^n = x^n(\omega)) \Pr(x^n(\omega))
\]

\[
= \sum_{x^n} \Pr(g(Y^n) \neq 1 | X^n = x^n(1)) \Pr(x^n(1)) = \sum_{C} \Pr(C) \lambda_1(C)
\]
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

\[ \sum_{C} \Pr(C) \lambda_{\omega}(C) = \sum_{x^{n}(1),...,x^{n}(\omega-1),x^{n}(\omega+1),...,x^{n}(2nR)} \left( \prod_{i \neq \omega} \Pr(x^{n}(i)) \right) \sum_{x^{n}(\omega)} \Pr(g(Y^{n}) \neq \omega | X^{n} = x^{n}(\omega)) \Pr(x^{n}(\omega)) \]

\[ = \sum_{x^{n}(\omega)} \Pr(g(Y^{n}) \neq \omega | X^{n} = x^{n}(\omega)) \Pr(x^{n}(\omega)) \]

\[ = \sum_{x^{n}} \Pr(g(Y^{n}) \neq 1 | X^{n} = x^{n}(1)) \Pr(x^{n}(1)) = \sum_{C} \Pr(C) \lambda_{1}(C) = \beta \]
Proof that all rates $R < C$ are achievable.

\[
\sum_C \Pr(C) \lambda_\omega(C) = \sum_{x^n(1), \ldots, x^n(\omega-1), x^n(\omega+1), \ldots, x^n(2nR)} \prod_{i \neq \omega} \Pr(x^n(i)) \sum_{x^n(\omega)} \Pr(g(Y^n) \neq \omega | X^n = x^n(\omega)) \Pr(x^n(\omega)) = 1
\]

\[
= \sum_{x^n(\omega)} \Pr(g(Y^n) \neq \omega | X^n = x^n(\omega)) \Pr(x^n(\omega)) = \sum_{x^n} \Pr(g(Y^n) \neq 1 | X^n = x^n(1)) \Pr(x^n(1)) = \sum_C \Pr(C) \lambda_1(C) = \beta
\]

Last sum is same regardless of $\omega$, call it $\beta$. Thus, we can can arbitrarily assume that $\omega = 1$. ...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Example: intuition as to how this becomes $\beta$. 

\[ \text{prob. of choosing } x_1 \text{ for } \omega \text{ and not choosing } y_1 + \text{prob. of choosing } x_2 \text{ for } \omega \text{ and not choosing } y_2 + \ldots \]

\[ \text{this is just the same for all } \omega \in \{1, 2, \ldots, M\} \]

\[ \text{so we may just pick } \omega = 1. \]
All rates \( R < C \) are achievable.

Proof that all rates \( R < C \) are achievable.

Example: intuition as to how this becomes \( \beta \).
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Example: intuition as to how this becomes $\beta$.

4 possible codewords

These are the possible associations between $\omega$ and one of the codewords. Considering all associations, we have the same average error for each $\omega$. Thus, we just choose $\omega = 1$. 

...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Example: intuition as to how this becomes $\beta$.

So error is equal to:

\[
\text{prob. of choosing } x_1 \text{ for } \omega \text{ and not choosing } y_1 +
\text{prob. of choosing } x_2 \text{ for } \omega \text{ and not choosing } y_2 + 
\ldots
\]

(32)

These are the possible associations between $\omega$ and one of the codewords. Considering all associations, we have the same average error for each $\omega$. Thus, we just choose $\omega = 1$. 

\[
\omega \ x_1 \ x_2 \ x_3 \ x_4 \\
\downarrow \downarrow \downarrow \downarrow \\
y_1 \ n_2 \ n_3 \ n_4
\]

4 possible codewords
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Example: intuition as to how this becomes $\beta$.

So error is equal to:

$$
\text{prob. of choosing } x_1 \text{ for } \omega \text{ and not choosing } y_1 \\
+ \text{prob. of choosing } x_2 \text{ for } \omega \text{ and not choosing } y_2 \\
+ \ldots \\
(32)
$$

this is just the same for all $\omega \in \{1, 2, \ldots, M\}$ so we may just pick $\omega = 1$. 

These are the possible associations between $\omega$ and one of the codewords. Considering all associations, we have the same average error for each $\omega$. Thus, we just choose $\omega = 1$. 

All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

So we get

\[
\Pr(\mathcal{E}) = \sum_C \Pr(C) P_e^{(n)}(C) = \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \beta = \sum_C \Pr(C) \lambda_1(C) = \Pr(\mathcal{E} | W = 1)
\]

(33)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

So we get

$$\Pr(E) = \sum_C \Pr(C) P_{\epsilon}^{(n)}(C) = \frac{1}{2^n R} \sum_{\omega=1}^{2^n R} \beta = = \sum_C \Pr(C) \lambda_1(C) = \Pr(E|W = 1)$$

(33)

Next, define the random events (again considering $\omega = 1$):

$$E_i \triangleq \left\{(x^n(i), y^n) \in A_{\epsilon}^{(n)}\right\} \text{ for } i = 1, \ldots, 2^n R$$

(34)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

So we get

$$\Pr(\mathcal{E}) = \sum_{C} \Pr(C) P_{e}^{(n)}(C) = \frac{1}{2nR} \sum_{\omega=1}^{2^{nR}} \beta = \sum_{C} \Pr(C) \lambda_1(C) = \Pr(\mathcal{E}|W = 1)$$

(33)

- Next, define the random events (again considering $\omega = 1$):

$$E_i \triangleq \left\{ (x^n(i), y^n) \in A_{\epsilon}^{(n)} \right\} \text{ for } i = 1, \ldots, 2^{nR}$$

(34)

- Assume that input is $x^n(1)$ (i.e., first message sent).

...
Proof that all rates $R < C$ are achievable.

So we get

$$\Pr(\mathcal{E}) = \sum_C \Pr(C) P_{e}^{(n)}(C) = \frac{1}{2nR} \sum_{\omega=1}^{2^{nR}} \beta = \sum_C \Pr(C) \lambda_1(C) = \Pr(\mathcal{E}|W = 1)$$

(33)

- Next, define the random events (again considering $\omega = 1$):

$$E_i \triangleq \left\{ (x^n(i), y^n) \in A_{e}^{(n)} \right\} \text{ for } i = 1, \ldots, 2^{nR}$$

(34)

- Assume that input is $x^n(1)$ (i.e., first message sent).
- Then the no error event is the same as: $E_1 \cap \neg(E_2 \cup E_3 \cup \cdots \cup E_M)$. 

...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Various flavors of error

- $E_1^c$ means that the transmitted and received codeword are not jointly typical (this is error type B from before).
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Various flavors of error

- $E_1^c$ means that the transmitted and received codeword are not jointly typical (this is error type B from before).
- $E_2 \cup E_3 \cup \cdots \cup E_{2nR}$. This is either:
All rates $R < C$ are achievable.

**Proof that all rates $R < C$ are achievable.**

Various flavors of error

- $E_1^C$ means that the transmitted and received codeword are not jointly typical (this is error type B from before).
- $E_2 \cup E_3 \cup \cdots \cup E_{2^n R}$. This is either:
  - **Type C**: wrong codeword is jointly typical with received sequence.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Various flavors of error

- $E_1^c$ means that the transmitted and received codeword are not jointly typical (this is error type B from before).
- $E_2 \cup E_3 \cup \cdots \cup E_{2nR}$. This is either:
  - Type C: wrong codeword is jointly typical with received sequence
  - Type A: greater than 1 codeword is jointly typical with received sequence
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Various flavors of error

- $E_1^c$ means that the transmitted and received codeword are not jointly typical (this is error type B from before).
- $E_2 \cup E_3 \cup \cdots \cup E_{2nR}$. This is either:
  - Type C: wrong codeword is jointly typical with received sequence
  - Type A: greater than 1 codeword is jointly typical with received sequence

so this is type $C$ and $A$ both.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Various flavors of error

- $E_1^c$ means that the transmitted and received codeword are not jointly typical (this is error type B from before).
- $E_2 \cup E_3 \cup \cdots \cup E_{2nR}$. This is either:
  - Type C: wrong codeword is jointly typical with received sequence
  - Type A: greater than 1 codeword is jointly typical with received sequence

so this is type $C$ and $A$ both.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Various flavors of error

- $E_1^c$ means that the transmitted and received codeword are not jointly typical (this is error type B from before).
- $E_2 \cup E_3 \cup \cdots \cup E_{2nR}$. This is either:
  - Type C: wrong codeword is jointly typical with received sequence
  - Type A: greater than 1 codeword is jointly typical with received sequence

so this is type $C$ and $A$ both.

Our goal is to bound the probability of error, but let's look at some figures first.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

All rates $\begin{align*} R < C \end{align*}$ are achievable.

Proof that all rates $\begin{align*} R < C \end{align*}$ are achievable.

Set of all jointly typical pairs of sequences $\begin{align*} 2^nH(X,Y) \end{align*}$

Set of all pairs of sequences $\begin{align*} |\mathcal{X}^n \times \mathcal{Y}^n| = (|\mathcal{X}| |\mathcal{Y}|)^n \end{align*}$
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Vertical axis is lexicographic order of possible codewords

Subset selection of the $2^{nR}$ random $\mathcal{X}^n$ codewords (chosen by the random selection procedure) for $i = 1, 2, \ldots, M$. Here, $2^{nR} = M = 4$.

Dots are the jointly typical sequences

$\mathcal{X}^n$

$\mathcal{Y}^n$

$2^{nH(X)}$

$2^{nH(Y)}$
Proof that all rates $R < C$ are achievable.

$X_n$ $x(1)$ $x(3)$ $x(4)$ $x(2)$

$y(a)$, on sending $x(\ast)$
$y(d)$, on sending $x(1)$
$y(b)$, on sending $x(4)$
$y(c)$, on sending $x(1)$

$\lambda(y(a)) = 0$
$\lambda(y(b)) = 4$
$\lambda(y(c)) = 0$

$E_1^c$
$E_2 \cup E_3 \cup \ldots$

$y(a)$ not jointly typical with any of the sent codewords.
Error type B

$g(y(a)) = 0$

$g(y(d)) = 4$
$y(d)$ should not be jointly typical with $x(4)$ but it is. Wrong jointly typical sequence.
Error type B

$g(y(b)) = 4$
$y(b)$ is jointly typical only with $x(4)$, so no error

$E_2 \cup E_3 \cup \ldots$

$y(d)$ should not be jointly typical with $x(4)$ but it is. Wrong jointly typical sequence.
Error type B

$g(y(d)) = 4$

$y(c)$ is jointly typical with both $x(1)$ and $x(3)$, so
Error type A

$g(y(c)) = 0$
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

Goal: bound the probability of error:

$$\Pr(E | W = 1) = \Pr(E_1^c \cup E_2 \cup E_3 \ldots)$$

$$\leq \Pr(E_1^c) + \sum_{i=2}^{2^nR} \Pr(E_i) \text{ by the union bound}$$

We have that

$$\Pr(E_1^c) = \Pr(A_\epsilon^{(n)c}) \to 0 \text{ as } n \to \infty$$

So, $\forall \epsilon, \exists n_0 \text{ s.t.}$

$$\Pr(E_1^c) \leq \epsilon, \forall n > n_0$$
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Also, because of random code generation process (and recall, $\omega = 1$)

$$X^n(1) \perp\!\!\!\!\!\!\perp X^n(i) \Rightarrow Y^n \perp\!\!\!\!\!\!\perp X^n(i), \text{ for } i \neq 1$$ (39)

This gives, for $i \neq 1$,

$$\Pr((X^n(i), Y^n) \in A(n)) \leq 2^{-n(I(X; Y) - 3\epsilon)}$$ (40)

by the joint AEP.

This will allow us to bound the error, as long as $I(X; Y) > 3\epsilon$. 

...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Also, because of random code generation process (and recall, $\omega = 1$)

\[ X^n(1) \perp X^n(i) \Rightarrow Y^n \perp X^n(i), \text{ for } i \neq 1 \]  \hspace{1cm} (39)

- This gives, for $i \neq 1$,

\[
\Pr((X^n(i), Y^n) \in A_{\epsilon}^{(n)}) \leq 2^{-n(I(X;Y) - 3\epsilon)}
\]  \hspace{1cm} (40)

by the joint AEP.

\[
\text{indep. events}
\]

\[ \ldots \]
All rates \( R < C \) are achievable.

Proof that all rates \( R < C \) are achievable.

- Also, because of random code generation process (and recall, \( \omega = 1 \))

\[
X^n(1) \perp \perp X^n(i) \Rightarrow Y^n \perp \perp X^n(i), \text{ for } i \neq 1
\]  

(39)

- This gives, for \( i \neq 1 \),

\[
\Pr\left( (X^n(i), Y^n) \in A^{(n)}_{\epsilon} \right) \leq 2^{-n\left(I(X;Y) - 3\epsilon\right)}
\]  

(40)

by the joint AEP.

- This will allow us to bound the error, as long as \( I(X;Y) > 3\epsilon \)....
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

So we get:

$$\Pr(\mathcal{E})$$

(46)

...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

So we get:

$$\Pr(\mathcal{E}) = \Pr(\mathcal{E}|W = 1)$$

(46)
All rates \( R < C \) are achievable.

Proof that all rates \( R < C \) are achievable.

So we get:

\[
\Pr(\mathcal{E}) = \Pr(\mathcal{E}|W = 1) \leq \Pr(E_1^c) + \sum_{i=2}^{2nR} \Pr(E_i) \tag{41}
\]

The last statement is true only if

\[
I(X;Y) - 3\epsilon > R.
\]

(46)
Proof that all rates $R < C$ are achievable.

So we get:

$$\Pr(\mathcal{E}) = \Pr(\mathcal{E}|W = 1) \leq \Pr(\mathcal{E}_1^c) + \sum_{i=2}^{2nR} \Pr(\mathcal{E}_i)$$  \hspace{1cm} (41)

$$\leq \epsilon + \sum_{i=2}^{2nR} 2^{-n(I(X;Y)-3\epsilon)}$$  \hspace{1cm} (42)

$$\leq \epsilon + 2^n \sum_{i=2}^{2nR} 2^{-n(I(X;Y)-3\epsilon)}$$  \hspace{1cm} (46)

...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

So we get:

$$\Pr(\mathcal{E}) = \Pr(\mathcal{E} | W = 1) \leq \Pr(\mathcal{E}_1^c) + \sum_{i=2}^{2^nR} \Pr(\mathcal{E}_i)$$  \hspace{1cm} (41)

$$\leq \epsilon + \sum_{i=2}^{2^nR} 2^{-n(I(X;Y)-3\epsilon)}$$  \hspace{1cm} (42)

$$= \epsilon + (2^{nR} - 1)2^{-n(I(X;Y)-3\epsilon)}$$  \hspace{1cm} (43)

$$\leq 2\epsilon$$  \hspace{1cm} (46)

The last statement is true only if $I(X;Y) - 3\epsilon > R$. 

...
Proof that all rates $R < C$ are achievable.

So we get:

$$\Pr(\mathcal{E}) = \Pr(\mathcal{E}|W = 1) \leq \Pr(E_1^c) + \sum_{i=2}^{2nR} \Pr(E_i)$$

$$\leq \epsilon + \sum_{i=2}^{2nR} 2^{-n(I(X;Y)-3\epsilon)}$$

$$= \epsilon + (2^{nR} - 1)2^{-n(I(X;Y)-3\epsilon)}$$

$$\leq \epsilon + 2^{3n\epsilon}2^{-n(I(X;Y)-R)}$$

(41)

(42)

(43)

(44)

(46)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

So we get:

$$\Pr(\mathcal{E}) = \Pr(\mathcal{E} | W = 1) \leq \Pr(\mathcal{E}_{1}^{c}) + \sum_{i=2}^{2nR} \Pr(E_i)$$

(41)

$$\leq \epsilon + \sum_{i=2}^{2nR} 2^{-n(I(X;Y) - 3\epsilon)}$$

(42)

$$= \epsilon + (2^{nR} - 1)2^{-n(I(X;Y) - 3\epsilon)}$$

(43)

$$\leq \epsilon + 2^{3n\epsilon}2^{-n(I(X;Y) - R)}$$

(44)

$$= \epsilon + 2^{-n((I(X;Y) - 3\epsilon) - R)}$$

(45)

$$\leq \epsilon + 2^{-n(3\epsilon + (R - I(X;Y)))}$$

(46)

...
Proof that all rates $R < C$ are achievable.

So we get:

$$
\Pr(\mathcal{E}) = \Pr(\mathcal{E}|W = 1) \leq \Pr(E_1^c) + \sum_{i=2}^{2^n} \Pr(E_i) \tag{41}
$$

$$
\leq \epsilon + \sum_{i=2}^{2^n} 2^{-n(I(X;Y)-3\epsilon)} \tag{42}
$$

$$
= \epsilon + (2^{nR} - 1)2^{-n(I(X;Y)-3\epsilon)} \tag{43}
$$

$$
\leq \epsilon + 2^{3n\epsilon}2^{-n(I(X;Y)-R)} \tag{44}
$$

$$
= \epsilon + 2^{-n((I(X;Y)-3\epsilon)-R)} \tag{45}
$$

$$
\leq 2\epsilon \quad \text{for large enough } n \tag{46}
$$
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

So we get:

$$\Pr(\mathcal{E}) = \Pr(\mathcal{E}|W = 1) \leq \Pr(E_1^c) + \sum_{i=2}^{2nR} \Pr(E_i)$$  \hspace{1cm} (41)

$$\leq \epsilon + \sum_{i=2}^{2nR} 2^{-n(I(X;Y) - 3\epsilon)}$$  \hspace{1cm} (42)

$$= \epsilon + (2^{nR} - 1)2^{-n(I(X;Y) - 3\epsilon)}$$  \hspace{1cm} (43)

$$\leq \epsilon + 2^{3n\epsilon}2^{-n(I(X;Y) - R)}$$  \hspace{1cm} (44)

$$= \epsilon + 2^{-n((I(X;Y) - 3\epsilon) - R)}$$  \hspace{1cm} (45)

$$\leq 2\epsilon \quad \text{for large enough } n$$  \hspace{1cm} (46)

The last statement is true only if $I(X;Y) - 3\epsilon > R$.  \hspace{1cm} ...
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- So if we chose $R < I(X; Y)$ (strictly), we can find an $\epsilon$ and $n$ so that the average probability of error $\Pr(\mathcal{E}) \leq 2\epsilon$, can be made as small as we want.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- So if we chose $R < I(X; Y)$ (strictly), we can find an $\epsilon$ and $n$ so that the average probability of error $Pr(\mathcal{E}) \leq 2\epsilon$, can be made as small as we want.

- But, we need to get from an average to a max probability of error, and bound that.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- So if we chose $R < I(X; Y)$ (strictly), we can find an $\epsilon$ and $n$ so that the average probability of error $Pr(\mathcal{E}) \leq 2\epsilon$, can be made as small as we want.

- But, we need to get from an average to a max probability of error, and bound that.

- First, choose $p^*(x) = \arg\max_{p(x)} I(X; Y)$ rather than uniform $p(x)$, to change the condition from $R < I(X; Y)$ to $R < C$. Thus, this gives us higher rate limit.
Proof that all rates $R < C$ are achievable.

- So if we chose $R < I(X; Y)$ (strictly), we can find an $\epsilon$ and $n$ so that the average probability of error $Pr(\mathcal{E}) \leq 2\epsilon$, can be made as small as we want.
- But, we need to get from an average to a max probability of error, and bound that.
- First, choose $p^*(x) = \arg\max_{p(x)} I(X; Y)$ rather than uniform $p(x)$, to change the condition from $R < I(X; Y)$ to $R < C$. Thus, this gives us higher rate limit.
- If $Pr(\mathcal{E}) \leq 2\epsilon$, the bound on the average error is small, so there must exist some specific code, say $C^*$ s.t.

$$P_e^{(n)}(C^*) \leq 2\epsilon$$  (47)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Let's break apart this error probability.

$$P_e^{(n)}(C^*)$$

(50)
Proof that all rates $R < C$ are achievable.

- Lets break apart the this error probability.

\[ P_e(n)(C^*) = \frac{1}{2nR} \sum_{i=1}^{2^n R} \lambda_i(C^*) \]  \hspace{1cm} (48)

\[ \leq 2\epsilon \]  \hspace{1cm} (50)

\[ \ldots \]
Proof that all rates $R < C$ are achievable.

- Let's break apart the this error probability.

\[
P_{e}^{(n)}(C^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(C^*)
\]

\[
= \frac{1}{2^{nR}} \sum_{i: \lambda_i < 4\epsilon} \lambda_i(C^*) + \frac{1}{2^{nR}} \sum_{i: \lambda_i \geq 4\epsilon} \lambda_i(C^*)
\]
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Lets break apart the this error probability.

$$P_e(n)(C^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(C^*)$$

$$= \frac{1}{2^{nR}} \sum_{i: \lambda_i < 4\epsilon} \lambda_i(C^*) + \frac{1}{2^{nR}} \sum_{i: \lambda_i \geq 4\epsilon} \lambda_i(C^*)$$

$$\leq 2\epsilon$$
Proof that all rates $R < C$ are achievable.

- Let's break apart the error probability.

$$P_e^{(n)}(C^*) = \frac{1}{2nR} \sum_{i=1}^{2^{nR}} \lambda_i(C^*)$$

$$= \frac{1}{2nR} \sum_{i: \lambda_i < 4\epsilon} \lambda_i(C^*) + \frac{1}{2nR} \sum_{i: \lambda_i \geq 4\epsilon} \lambda_i(C^*)$$

$$\leq 2\epsilon$$

- Now suppose more than half of the indices had error $\geq 4\epsilon$ (i.e., suppose $|\{i : \lambda_i \geq 4\epsilon\}| > 2^{nR}/2$). Under this assumption:

$$\frac{1}{2nR} \sum_{i: \lambda_i \geq 4\epsilon} \lambda_i \geq \frac{1}{2nR} \sum_{i: \lambda_i \geq 4\epsilon} 4\epsilon = \frac{1}{2nR} |\{i : \lambda_i \geq 4\epsilon\}| 4\epsilon > \frac{1}{2} 4\epsilon = 2\epsilon$$
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Can’t be since these alone would be more than our $2\epsilon$ upper bound.
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Can’t be since these alone would be more than our $2\epsilon$ upper bound.
- Hence, at most half the codewords can have error $\geq 4\epsilon$, and we get

$$\left| \left\{ i : \lambda_i \geq 4\epsilon \right\} \right| \leq \frac{2^{nR}}{2} \quad \Rightarrow \quad \left| \left\{ i : \lambda_i < 4\epsilon \right\} \right| \geq \frac{2^{nR}}{2}$$

(51)
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Can’t be since these alone would be more than our $2\varepsilon$ upper bound.
- Hence, at most half the codewords can have error $\geq 4\varepsilon$, and we get

$$|\{i : \lambda_i \geq 4\varepsilon\}| \leq \frac{2^nR}{2} \implies |\{i : \lambda_i < 4\varepsilon\}| \geq \frac{2^nR}{2} \quad (51)$$

- Create a new codebook that eliminates all bad codewords (i.e., those in with index $\{i : \lambda_i \geq 4\varepsilon\}$). There are at most half of them.
Proof that all rates $R < C$ are achievable.

- Can’t be since these alone would be more than our $2\epsilon$ upper bound.
- Hence, at most half the codewords can have error $\geq 4\epsilon$, and we get

$$|\{i : \lambda_i \geq 4\epsilon\}| \leq \frac{2^{nR}}{2} \Rightarrow |\{i : \lambda_i < 4\epsilon\}| \geq \frac{2^{nR}}{2} \quad (51)$$

- Create a new codebook that eliminates all bad codewords (i.e., those in with index $\{i : \lambda_i \geq 4\epsilon\}$). There are at most half of them.
- The remaining codewords are of size $\geq \frac{2^{nR}}{2} = 2^{nR-1} = 2^n\left(R-1/n\right)$ (at least half of them).
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Can’t be since these alone would be more than our $2\epsilon$ upper bound.
- Hence, at most half the codewords can have error $\geq 4\epsilon$, and we get

$$|\{i : \lambda_i \geq 4\epsilon\}| \leq \frac{2^{nR}}{2} \quad \Rightarrow \quad |\{i : \lambda_i < 4\epsilon\}| \geq \frac{2^{nR}}{2} \quad (51)$$

- Create a new codebook that eliminates all bad codewords (i.e., those in with index $\{i : \lambda_i \geq 4\epsilon\}$). There are at most half of them.
- The remaining codewords are of size $\geq \frac{2^{nR}}{2} = 2^{nR-1} = 2^n(R-1/n)$ (at least half of them). They all have max probability $\leq 4\epsilon$. 
All rates $R < C$ are achievable.

Proof that all rates $R < C$ are achievable.

- Can’t be since these alone would be more than our $2\epsilon$ upper bound.
- Hence, at most half the codewords can have error $\geq 4\epsilon$, and we get

$$|\{i : \lambda_i \geq 4\epsilon\}| \leq \frac{2^{nR}}{2} \quad \Rightarrow \quad |\{i : \lambda_i < 4\epsilon\}| \geq \frac{2^{nR}}{2} \quad (51)$$

- Create a new codebook that eliminates all bad codewords (i.e., those in with index $\{i : \lambda_i \geq 4\epsilon\}$). There are at most half of them.
- The remaining codewords are of size $\geq 2^{nR}/2 = 2^{nR-1} = 2^n(R-1/n)$ (at least half of them). They all have max probability $\leq 4\epsilon$.
- We now code with rate $R' = R - 1/n \to R$ as $n \to \infty$, but for this new sequence of codes, the max error probability $\lambda^{(n)} \leq 4\epsilon$, which can be made as small as we wish.
To summarize, random coding is the method of proof to show that if $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ as $n \to \infty$. 
To summarize, random coding is the method of proof to show that if $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with $\lambda(n) \to 0$ as $n \to \infty$.

This might not be the best code, but it is sufficient. It is an existence proof.
To summarize, random coding is the method of proof to show that if $R < C$, there exists a sequence of $({2^nR}, n)$ codes with $\lambda(n) \to 0$ as $n \to \infty$.

This might not be the best code, but it is sufficient. It is an existence proof.

Huge literature on coding theory. We’ll discuss Hamming codes.
To summarize, random coding is the method of proof to show that if $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with $\lambda(n) \to 0$ as $n \to \infty$.

This might not be the best code, but it is sufficient. It is an existence proof.

Huge literature on coding theory. We’ll discuss Hamming codes.

But many good codes exist today: Turbo codes, Gallager (or low-density-parity-check) codes, and new ones are being proposed often.
To summarize, random coding is the method of proof to show that if $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with $\lambda(n) \to 0$ as $n \to \infty$.

This might not be the best code, but it is sufficient. It is an existence proof.

Huge literature on coding theory. We’ll discuss Hamming codes.

But many good codes exist today: Turbo codes, Gallager (or low-density-parity-check) codes, and new ones are being proposed often.

Perhaps if there is enough demand, we’ll have a quarter class just on coding theory.
To summarize, random coding is the method of proof to show that if \( R < C \), there exists a sequence of \( (2^{nR}, n) \) codes with \( \lambda^{(n)} \to 0 \) as \( n \to \infty \).

This might not be the best code, but it is sufficient. It is an existence proof.

Huge literature on coding theory. We’ll discuss Hamming codes.

But many good codes exist today: Turbo codes, Gallager (or low-density-parity-check) codes, and new ones are being proposed often.

Perhaps if there is enough demand, we’ll have a quarter class just on coding theory.

But we have yet to proof the converse . . .
We next need to show that any sequence of \((2^{nR}, n)\) codes with \(\lambda^{(n)} \to 0\) must have that \(R \leq C\).
We next need to show that any sequence of \((2^{nR}, n)\) codes with 
\(\lambda^{(n)} \to 0\) must have that \(R \leq C\).

First lets consider the case if \(P_e^{(n)} = 0\), in such case it is easy to 
show that \(R \leq C\).
Zero Error Codes

- If $P_e^n = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
Zero Error Codes

- If $P_e^n = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$.)
Zero Error Codes

- If $P_e^{(n)} = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$).
- First lets consider the case if $P_e^{(n)} = 0$, in such case it is easy to show that $R \leq C$.  

\[ nR = H(W) = H(W|Y^n) + I(W;Y^n) \]

\[ I(W;Y^n) = \sum_{i=1}^{n} I(Y_i;X_{i-1}) \leq \sum_{i=1}^{n} H(Y_i;X_i) - H(Y_i;X_{i-1}) \]

\[ n \sum_{i=1}^{n} I(Y_i;X_i) \leq nC \]
Zero Error Codes

- If $P_e(n) = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$).
- First let's consider the case if $P_e(n) = 0$, in such case it is easy to show that $R \leq C$. Then we get

\[ (56) \]
Zero Error Codes

- If $P_e^{(n)} = 0$, then $H(W|Y^n) = 0$ (no uncertainty).
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$).
- First let's consider the case if $P_e^{(n)} = 0$, in such case it is easy to show that $R \leq C$. Then we get

  $$nR$$
Zero Error Codes

- If $P_e(n) = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$).
- First lets consider the case if $P_e(n) = 0$, in such case it is easy to show that $R \leq C$. Then we get

\[ nR = H(W) \]
Zero Error Codes

- If $P_e(n) = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over \{1, 2, \ldots, M\}).
- First lets consider the case if $P_e(n) = 0$, in such case it is easy to show that $R \leq C$. Then we get

\[
nR = H(W) = H(W|Y^n) + I(W; Y^n)
\]
Zero Error Codes

- If $P_e(n) = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over \{1, 2, \ldots, M\}).
- First let's consider the case if $P_e(n) = 0$, in such case it is easy to show that $R \leq C$. Then we get

$$nR = H(W) = H(W|Y^n) + I(W; Y^n) = I(W; Y^n)$$

(52)
Zero Error Codes

- If $P_e^n = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$).
- First let's consider the case if $P_e^n = 0$, in such case it is easy to show that $R \leq C$. Then we get
  
  $$nR = H(W) = H(W|Y^n) + I(W; Y^n) = I(W; Y^n) \leq I(X^n; Y^n)$$  

  //Since $W \to X^n \to Y^n$ and data proc. ineq. (53)
Zero Error Codes

- If $P_e^n = 0$, then $H(W|Y^n) = 0$ (no uncertainty).
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$).
- First let's consider the case if $P_e^n = 0$, in such case it is easy to show that $R \leq C$. Then we get

$$nR = H(W) = H(W|Y^n) + I(W;Y^n) = I(W;Y^n)$$  \hspace{1cm} (52)

$$\leq I(X^n;Y^n) \quad \text{//Since } W \rightarrow X^n \rightarrow Y^n \text{ and data proc. ineq.}$$  \hspace{1cm} (53)

$$= H(Y^n) - H(Y^n|X^n)$$  \hspace{1cm} (56)
Zero Error Codes

- If $P_e^{(n)} = 0$, then $H(W|Y^n) = 0$ (no uncertainty).
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$).
- First let’s consider the case if $P_e^{(n)} = 0$, in such case it is easy to show that $R \leq C$. Then we get

\[
nR = H(W) = H(W|Y^n) + I(W; Y^n) = I(W; Y^n)
\]

\[
\leq I(X^n; Y^n) \quad \text{//Since } W \rightarrow X^n \rightarrow Y^n \text{ and data proc. ineq.} \quad (53)
\]

\[
= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, X^n)
\]

(56)
Zero Error Codes

- If $P_e^{(n)} = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$).
- First let's consider the case if $P_e^{(n)} = 0$, in such case it is easy to show that $R \leq C$. Then we get

$$nR = H(W) = H(W|Y^n) + I(W; Y^n) = I(W; Y^n)$$

$$\leq I(X^n; Y^n) \quad /\text{Since } W \rightarrow X^n \rightarrow Y^n \text{ and data proc. ineq.}$$

$$= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, X^n)$$

But $Y_i \perp \perp \{Y_{1:i-1}, X_{1:i-1}, X_{i+1:n}\}|X_i$,
Zero Error Codes

- If $P_e^{(n)} = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$.
- First let's consider the case if $P_e^{(n)} = 0$, in such case it is easy to show that $R \leq C$. Then we get

$$nR = H(W) = H(W|Y^n) + I(W; Y^n) = I(W; Y^n)$$  \hspace{1cm} (52)

$$\leq I(X^n; Y^n) \quad \text{//Since } W \rightarrow X^n \rightarrow Y^n \text{ and data proc. ineq.} \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad (53)$$

$$= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, X^n)$$  \hspace{1cm} (54)

But $Y_i \perp \perp \{Y_{1:i-1}, X_{1:i-1}, X_{i+1:n}\}|X_i$, so we can continue as

(56)
Zero Error Codes

- If \( P_e(n) = 0 \), then \( H(W|Y^n) = 0 \) (no uncertainty)
- For the sake of an easy proof, assume \( H(W) = nR = \log M \) (i.e., uniform distribution over \( \{1, 2, \ldots, M\} \)).
- First lets consider the case if \( P_e(n) = 0 \), in such case it is easy to show that \( R \leq C \). Then we get

\[
nR = H(W) = H(W|Y^n) + I(W;Y^n) = I(W;Y^n) \tag{52}
\]

\[
\leq I(X^n;Y^n) \quad \text{//Since } W \rightarrow X^n \rightarrow Y^n \text{ and data proc. ineq.} \tag{53}
\]

\[
= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1},X^n) \tag{54}
\]

But \( Y_i \perp \perp \{Y_{1:i-1},X_{1:i-1},X_{i+1:n}\}|X_i \), so we can continue as

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \tag{55}
\]

\[
\leq \sum_{i=1}^{n} I(Y_i;X_i) \leq nC \tag{56}
\]
Zero Error Codes

- If $P_e^n = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$).
- First let's consider the case if $P_e^n = 0$, in such case it is easy to show that $R \leq C$. Then we get

$$nR = H(W) = H(W|Y^n) + I(W; Y^n) = I(W; Y^n) \quad (52)$$

$$\leq I(X^n; Y^n) \quad //\text{Since } W \to X^n \to Y^n \text{ and data proc. ineq.} \quad (53)$$

$$= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, X^n) \quad (54)$$

But $Y_i \perp \perp \{Y_{1:i-1}, X_{1:i-1}, X_{i+1:n}\}|X_i$, so we can continue as

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \leq \sum_{i} \left[ H(Y_i) - H(Y_i|X_i) \right] \quad (55)$$

(56)
Zero Error Codes

- If \( P_e(n) = 0 \), then \( H(W|Y^n) = 0 \) (no uncertainty)
- For the sake of an easy proof, assume \( H(W) = nR = \log M \) (i.e., uniform distribution over \( \{1, 2, \ldots, M\} \)).
- First, let's consider the case if \( P_e(n) = 0 \), in such case it is easy to show that \( R \leq C \). Then we get

\[
nR = H(W) = H(W|Y^n) + I(W;Y^n) = I(W;Y^n) \tag{52}
\]

\[
\leq I(X^n;Y^n) \quad \text{//Since } W \rightarrow X^n \rightarrow Y^n \text{ and data proc. ineq.} \tag{53}
\]

\[
= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1},X^n) \tag{54}
\]

But \( Y_i \perp \{Y_{1:i-1},X_{1:i-1},X_{i+1:n}\}|X_i \), so we can continue as

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \leq \sum_{i} \left[ H(Y_i) - H(Y_i|X_i) \right] \tag{55}
\]

\[
= \sum_{i=1}^{n} I(Y_i;X_i) \tag{56}
\]
Zero Error Codes

- If $P_e^{(n)} = 0$, then $H(W|Y^n) = 0$ (no uncertainty)
- For the sake of an easy proof, assume $H(W) = nR = \log M$ (i.e., uniform distribution over $\{1, 2, \ldots, M\}$).
- First let's consider the case if $P_e^{(n)} = 0$, in such case it is easy to show that $R \leq C$. Then we get

$$nR = H(W) = H(W|Y^n) + I(W; Y^n) = I(W; Y^n) \quad (52)$$

$$\leq I(X^n; Y^n) \quad /\text{Since } W \to X^n \to Y^n \text{ and data proc. ineq.} \quad (53)$$

$$= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, X^n) \quad (54)$$

But $Y_i \perp \perp \{Y_{1:i-1}, X_{1:i-1}, X_{i+1:n}\} | X_i$, so we can continue as

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \leq \sum_{i} \left[ H(Y_i) - H(Y_i|X_i) \right] \quad (55)$$

$$= \sum_{i=1}^{n} I(Y_i; X_i) \leq nC \quad (56)$$
Zero Error Codes

- Thus, $nR \leq nC$ and $R \leq C$ when $P_e^{(n)} = 0$. 
Zero Error Codes

• Thus, \( nR \leq nC \) and \( R \leq C \) when \( P_e(n) = 0 \).

• In fact, the proof shows \( H(W) \leq nC \), which means that \( \max_p H_p(W) \leq nC \) implying that

\[
H(W) \leq \max_p H_p(W) = nR \leq nC \quad (57)
\]

so we get \( R \leq C \) regardless of the source distribution.
Zero Error Codes

- Thus, $nR \leq nC$ and $R \leq C$ when $P_e^{(n)} = 0$.
- In fact, the proof shows $H(W) \leq nC$, which means that
  $\max_p H_p(W) \leq nC$ implying that

\[ H(W) \leq \max_p H_p(W) = nR \leq nC \tag{57} \]

so we get $R \leq C$ regardless of the source distribution.
- It also shows a sub-lemma, namely that $I(X^n; Y^n) \leq nC$ that we'll use later. Let's name it:
Thus, $nR \leq nC$ and $R \leq C$ when $P_e^n = 0$.

In fact, the proof shows $H(W) \leq nC$, which means that $\max_p H_p(W) \leq nC$ implying that

$$H(W) \leq \max_p H_p(W) = nR \leq nC$$

so we get $R \leq C$ regardless of the source distribution.

It also shows a sub-lemma, namely that $I(X^n; Y^n) \leq nC$ that we’ll use later. Let’s name it:

**Lemma 4.1**

$$I(X^n; Y^n) \leq nC$$
Zero Error Codes

- Thus, $nR \leq nC$ and $R \leq C$ when $P_e^n = 0$.
- In fact, the proof shows $H(W) \leq nC$, which means that
  $\max_p H_p(W) \leq nC$ implying that

  $$H(W) \leq \max_p H_p(W) = nR \leq nC$$  \hspace{1cm} (57)

  so we get $R \leq C$ regardless of the source distribution.
- It also shows a sub-lemma, namely that $I(X^n; Y^n) \leq nC$ that we’ll use later. Let’s name it:

  **Lemma 4.1**

  $$I(X^n; Y^n) \leq nC$$  \hspace{1cm} (58)

- We also need Fano’s inequality. Recall, before it took the form

  $$H(X|Y) \leq 1 + P_e \log \mathcal{X}$$  \hspace{1cm} (59)
Fano’s Lemma (needed for proof)

**Theorem 4.2 (Fano)**

For a DMC with codebook $C$ and uniformly distributed input messages ($H(W) = nR$) and $P_e(n) = \Pr(W \neq g(Y^n))$, then

$$H(X^n|Y^n) \leq 1 + P_e(n)nR$$

(60)
Fano’s Lemma (needed for proof)

Theorem 4.2 (Fano)

For a DMC with codebook $C$ and uniformly distributed input messages $(H(W) = nR)$ and $P_{e}^{(n)} = Pr(W \neq g(Y^n))$, then

$$H(X^n|Y^n) \leq 1 + P_{e}^{(n)}nR$$

(60)

Proof.

Let $E \triangleq 1\{W \neq \hat{W}\}$. 

...
Fano’s Lemma (needed for proof)

**Theorem 4.2 (Fano)**

For a DMC with codebook $C$ and uniformly distributed input messages ($H(W) = nR$) and $P_e^{(n)} = \Pr(W \neq g(Y^n))$, then

$$H(X^n|Y^n) \leq 1 + P_e^{(n)}nR \tag{60}$$

**Proof.**

Let $E \triangleq 1\{W \neq \hat{W}\}$. Then we get:

$$H(E, W|Y^n) \tag{62}$$
Fano’s Lemma (needed for proof)

Theorem 4.2 (Fano)

*For a DMC with codebook \( C \) and uniformly distributed input messages \( (H(W) = nR) \) and \( P_{e}^{(n)} = Pr(W \neq g(Y^{n})) \), then*

\[
H(X^{n}|Y^{n}) \leq 1 + P_{e}^{(n)}nR
\]  
(60)

Proof.

Let \( E \triangleq 1\{W \neq \hat{W}\} \). Then we get:

\[
H(E, W|Y^{n}) = H(W|Y^{n}) + H(E|Y^{n}, W) = 0
\]  
(61)

\[
H(X^{n}|Y^{n}) \leq 1 + P_{e}^{(n)}nR
\]  
(62)
Fano’s Lemma (needed for proof)

Theorem 4.2 (Fano)

For a DMC with codebook $C$ and uniformly distributed input messages ($H(W) = nR$) and $P_{e}^{(n)} = Pr(W \neq g(Y^{n}))$, then

\[ H(X^{n}|Y^{n}) \leq 1 + P_{e}^{(n)} nR \]  \hspace{1cm} (60)

Proof.

Let $E \triangleq 1\{W \neq \hat{W}\}$. Then we get:

\[ H(E,W|Y^{n}) = H(W|Y^{n}) + H(E|Y^{n},W) \]

\[ = H(E|Y^{n}) + H(W|Y^{n},E) \]

\[ \leq 1 \]  \hspace{1cm} (61)

\[ = 0 \]  \hspace{1cm} (62)

...
Fano’s Lemma (needed for proof)

proof continued.

\[ H(W | Y^n, E) = \Pr(E = 0) \leq 1 - \frac{\Pr(n)}{e} H(W | Y^n, E = 0) \] (63)

\[ + \Pr(E = 1) \leq \frac{\Pr(n)}{e} nR \Rightarrow H(W | Y^n) \leq 1 + \frac{\Pr(n)}{e} nR \] (64)
Fano’s Lemma (needed for proof)

proof continued.

and

\[
H(W|Y^n, E) = \Pr(E = 0) H(W|Y^n, E = 0) \]

\[
= 0 \]

\[
+ \Pr(E = 1) H(W|Y^n, E = 1) \]

\[
\leq 1 + P_e n \log(2^{nR} - 1) \]

(63)
proof continued.

and

\[
H(W|Y^n, E) = \Pr(E = 0) H(W|Y^n, E = 0)
\]

\[
= \frac{1}{1-P_e(n)} \left( 1 - P_e(n) \right) + \Pr(E = 1) H(W|Y^n, E = 1)
\]

\[
= P_e(n) nR
\]
Fano’s Lemma (needed for proof)

proof continued.

and

\[
H(W|Y^n, E) = \underbrace{\Pr(E = 0) H(W|Y^n, E = 0)}_{1 - P_e^{(n)}} + \underbrace{\Pr(E = 1) H(W|Y^n, E = 1)}_{P_e^{(n)} \log(2^{nR} - 1)}
\]

\[
= P_e^{(n)} nR \quad \Rightarrow \quad H(W|Y^n) \leq 1 + P_e^{(n)} nR
\]
proof continued.

\[
H(W|Y^n, E) = \begin{cases} 
1 - P_e(n) & H(W|Y^n, E = 0) \\
Pr(E = 1) & H(W|Y^n, E = 1)
\end{cases} \\
= P_e(n) n R 
\Rightarrow \quad H(W|Y^n) \leq 1 + P_e(n) n R 
\]

but $X^n = X^n(W)$ and functions of random variables can only reduce entropy.
proof continued.

and

\[ H(W|Y^n, E) = Pr(E = 0) H(W|Y^n, E = 0) \]

\[ + Pr(E = 1) H(W|Y^n, E = 1) \]

\[ = P_e^{(n)} nR \quad \Rightarrow \quad H(W|Y^n) \leq 1 + P_e^{(n)} nR \quad (64) \]

but \( X^n = X^n(W) \) and functions of random variables can only reduce entropy. So we get:

\[ H(X^n|Y^n) \leq H(W|Y^n) \leq 1 + P_e^{(n)} nR \quad (65) \]
Sequence of codes w. vanishing error must have $R < C$.

The converse states: any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have that $R \leq C$.

Proof that $\lambda^{(n)} \to 0$ as $n \to \infty \Rightarrow R < C$.

- Average prob. goes to zero if max probability does: $\lambda^{(n)} \to 0 \Rightarrow P_e^{(n)} \to 0$, where $P_e^{(n)} = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i$

- Lets set $H(W) = nR$ for now (i.e., $W$ uniform on $\{1, 2, \ldots, M = 2^{nR}\}$). Again, makes the proof a bit easier and doesn't affect relationship between $R$ and $C$.

- So, $\Pr(W = \hat{W}) = P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$ as we saw in last lecture. ...
Sequence of codes w. vanishing error must have $R < C$.  

Proof that $\lambda^{(n)} \to 0$ as $n \to \infty \Rightarrow R < C$.

$nR$
Sequence of codes w. vanishing error must have $R < C$. 

Proof that $\lambda^{(n)} \to 0$ as $n \to \infty \Rightarrow R < C$.

\[ nR = H(W) \]
Sequence of codes w. vanishing error must have $R < C$.

Proof that $\lambda^{(n)} \rightarrow 0$ as $n \rightarrow \infty \Rightarrow R < C$.

\[ nR = H(W) = H(W|Y^n) + I(W;Y^n) \quad (66) \]

\[ \Rightarrow R \leq P(n) + 1/n + C \quad (69) \]
Sequence of codes w. vanishing error must have $R < C$.

Proof that $\lambda^n \to 0$ as $n \to \infty \Rightarrow R < C$.

\[
\begin{align*}
  nR &= H(W) = H(W|Y^n) + I(W;Y^n) \\
  &\leq H(W|Y^n) + I(X^n(W);Y^n) \quad \text{//Since } W \to X^n \to Y^n \\
  &\leq 1 + P(n)e^{nR} + I(X^n(W);Y^n) \quad \text{//by Fano} \\
  &\leq 1 + P(n)e^{nR} + nC \quad \text{//by lemma 4.1} \\
  \Rightarrow R &\leq P(n)e^{R} + \frac{1}{n} + C
\end{align*}
\]
Proof that $\lambda^{(n)} \to 0$ as $n \to \infty \Rightarrow R < C$.

\[ nR = H(W) = H(W|Y^n) + I(W;Y^n) \]  \hspace{1cm} (66)

\[ \leq H(W|Y^n) + I(X^n(W);Y^n) \quad \text{\texttt{//Since } W \to X^n \to Y^n} \]  \hspace{1cm} (67)

\[ \leq 1 + P_{e}^{(n)}nR + I(X^n(W);Y^n) \quad \text{\texttt{//by Fano}} \]  \hspace{1cm} (68)

\[ \Rightarrow R \leq P_{e}^{(n)}R + \frac{1}{n} + C \]  \hspace{1cm} (70)
Proof that $\lambda^{(n)} \to 0$ as $n \to \infty \Rightarrow R < C.$

$nR = H(W) = H(W|Y^n) + I(W; Y^n)$

$\leq H(W|Y^n) + I(X^n(W); Y^n)$  \quad \text{//Since } W \to X^n \to Y^n

$\leq 1 + P_e^{(n)} nR + I(X^n(W); Y^n)$  \quad \text{//by Fano}

$\leq 1 + P_e^{(n)} nR + nC$  \quad \text{//by lemma 4.1}
Proof that $\lambda^{(n)} \to 0$ as $n \to \infty \Rightarrow R < C$.

\begin{align*}
nR &= H(W) = H(W|Y^n) + I(W;Y^n) \\
&\leq H(W|Y^n) + I(X^n(W);Y^n) \quad \text{//Since } W \to X^n \to Y^n \\
&\leq 1 + P_e^{(n)} nR + I(X^n(W);Y^n) \quad \text{//by Fano} \\
&\leq 1 + P_e^{(n)} nR + nC \quad \text{//by lemma 4.1} \\
\Rightarrow R &\leq P_e^{(n)} R + 1/n + C
\end{align*}
Sequence of codes w. vanishing error must have $R < C$.

Proof that $\lambda^{(n)} \to 0$ as $n \to \infty \Rightarrow R < C$.

\[
nR = H(W) = H(W|Y^n) + I(W; Y^n) \tag{66}
\]
\[
\leq H(W|Y^n) + I(X^n(W); Y^n) \quad //\text{Since } W \to X^n \to Y^n \tag{67}
\]
\[
\leq 1 + P_e^{(n)} nR + I(X^n(W); Y^n) \quad //\text{by Fano} \tag{68}
\]
\[
\leq 1 + P_e^{(n)} nR + nC \quad //\text{by lemma 4.1} \tag{69}
\]
\[
\Rightarrow R \leq P_e^{(n)} R + 1/n + C \tag{70}
\]

Now as $n \to \infty$, $P_e^{(n)} \to 0$, and $1/n \to 0$ as well.
Sequence of codes w. vanishing error must have $R < C$.

Proof that $\lambda^{(n)} \to 0$ as $n \to \infty \Rightarrow R < C$.

\[
nR = H(W) = H(W|Y^n) + I(W; Y^n) \\
\leq H(W|Y^n) + I(X^n(W); Y^n) \quad //\text{Since } W \to X^n \to Y^n \\
\leq 1 + P_e^{(n)}nR + I(X^n(W); Y^n) \quad //\text{by Fano} \\
\leq 1 + P_e^{(n)}nR + nC \quad //\text{by lemma 4.1} \\
\Rightarrow R \leq P_e^{(n)} R + 1/n + C
\]

Now as $n \to \infty$, $P_e^{(n)} \to 0$, and $1/n \to 0$ as well. Thus

\[
\Rightarrow R < C
\]
Sequence of codes w. vanishing error must have $R < C$.

Also,

$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR} \quad (72)$$

This means that:
Sequence of codes w. vanishing error must have $R < C$.

Also,

$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR} \quad (72)$$

This means that:

- if $n \to \infty$ and $R > C$, then error lower bound is strictly positive, and depends on $1 - C/R$. 
Sequence of codes w. vanishing error must have $R < C$.

Also,

$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR} \quad (72)$$

This means that:

- if $n \to \infty$ and $R > C$, then error lower bound is strictly positive, and depends on $1 - C/R$.

- Even for small $n$, $P_e^{(n)} > 0$, since otherwise, if $P_e^{(n_0)} = 0$ for some code, we can concatenate code to get large $n$ same rate code, contradicting $P_e > 0$. 
Sequence of codes w. vanishing error must have $R < C$.

Lower bound on error:

$$P_{e(n)} \geq 1 - \frac{C}{R}$$

(73)
Sequence of codes w. vanishing error must have $R < C$. 

Lower bound on error:

$$P_e^{(n)} \geq 1 - \frac{C}{R}$$  \hspace{1cm} (73)
Sequence of codes w. vanishing error must have $R < C$.

Lower bound on error:

$$P_e^n \geq 1 - \frac{C}{R} \quad (74)$$
Sequence of codes w. vanishing error must have $R < C$.

Lower bound on error:

$$P_e^{(n)} \geq 1 - \frac{C}{R}$$  \hspace{1cm} (74)

generates this plot:

Meaning

$$P_e \propto e^{-nE(R)}$$  \hspace{1cm} (75)
Zero-error capacity

What if we insist on $R = C$ and $P_e = 0$. In such case, what are the requirements of any such code.

$nR$
Zero-error capacity

What if we insist on $R = C$ and $P_e = 0$. In such case, what are the requirements of any such code.

$$nR = H(W)$$
Zero-error capacity

What if we insist on $R = C$ and $P_e = 0$. In such case, what are the requirements of any such code.

$$nR = H(W) = H(X^n(W))$$  //if codewords distinct  (76)
Zero-error capacity

What if we insist on \( R = C \) and \( P_e = 0 \). In such case, what are the requirements of any such code.

\[
nR = H(W) = H(X^n(W)) \quad //\text{if codewords distinct} \tag{76}
\]

\[
= H(X^n|Y^n) + I(X^n;Y^n)
\]

\[
= 0 \quad \text{since } P_e=0
\]
Zero-error capacity

What if we insist on $R = C$ and $P_e = 0$. In such case, what are the requirements of any such code.

\[
\begin{align*}
nR &= H(W) = H(X^n(W)) \quad \text{if codewords distinct} \\
    &= H(X^n|Y^n) + I(X^n;Y^n) = I^n(Y^n) \\
    &= 0 \text{ since } P_e = 0
\end{align*}
\]
Zero-error capacity

What if we insist on $R = C$ and $P_e = 0$. In such case, what are the requirements of any such code.

$$nR = H(W) = H(X^n(W)) \quad // \text{if codewords distinct}$$  \hspace{1cm} (76)

$$= H(X^n|Y^n) + I(X^n; Y^n) = I^n(Y^n)$$  \hspace{1cm} (77)

$$= 0 \text{ since } P_e = 0$$

$$= H(Y^n) - H(Y^n|X^n)$$  \hspace{1cm} (78)
What if we insist on $R = C$ and $P_e = 0$. In such case, what are the requirements of any such code.

\[ nR = H(W) = H(X^n(W)) \quad \text{//if codewords distinct} \quad (76) \]
\[ = H(X^n|Y^n) + I(X^n;Y^n) = I^n(Y^n) \quad (77) \]
\[ = 0 \quad \text{since } P_e = 0 \]
\[ = H(Y^n) - H(Y^n|X^n) \quad (78) \]
\[ = H(Y^n) - \sum_{i=1}^{n} H(Y^i|X_i) \quad (79) \]
Zero-error capacity

What if we insist on $R = C$ and $P_e = 0$. In such case, what are the requirements of any such code.

\[
nR = H(W) = H(X^n(W)) \text{  \footnotesize{\text{//if codewords distinct}} (76)}
\]

\[
= H(X^n|Y^n) + I(X^n;Y^n) = I(n;Y^n) \text{  \footnotesize{(77)}}
\]

\[
= 0 \text{ since } P_e=0
\]

\[
= H(Y^n) - H(Y^n|X^n) \text{  \footnotesize{(78)}}
\]

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y^i|X_i) \text{  \footnotesize{(79)}}
\]

\[
= \sum_{i} H(Y_i) - \sum_{i} H(Y_i|X_i) \text{  \footnotesize{\text{//if all } Y_i \text{'s are indep}} (80)}
\]
Zero-error capacity

What if we insist on $R = C$ and $P_e = 0$. In such case, what are the requirements of any such code.

$$nR = H(W) = H(X^n(W)) \quad \text{//if codewords distinct}$$

$$= H(X^n|Y^n) + I(X^n; Y^n) = I^n(Y^n)$$

$$= 0 \quad \text{since } P_e = 0$$

$$= H(Y^n) - H(Y^n|X^n)$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y^i|X_i)$$

$$= \sum_i H(Y_i) - \sum_i H(Y_i|X_i) \quad \text{//if all } Y_i's \text{'s are indep}$$

$$= \sum_i I(X_i; Y_i)$$

(82)
Zero-error capacity

What if we insist on $R = C$ and $P_e = 0$. In such case, what are the requirements of any such code.

\[ nR = H(W) = H(X^n(W)) \quad // \text{if codewords distinct} \quad (76) \]

\[ = H(X^n|Y^n) + I(X^n; Y^n) = I^n(Y^n) \quad (77) \]

\[ = 0 \text{ since } P_e = 0 \]

\[ = H(Y^n) - H(Y^n|X^n) \quad (78) \]

\[ = H(Y^n) - \sum_{i=1}^{n} H(Y^i|X^i) \quad (79) \]

\[ = \sum_{i} H(Y_i) - \sum_{i} H(Y_i|X_i) \quad // \text{if all } Y_i \text{'s are indep} \quad (80) \]

\[ = \sum_{i} I(X_i; Y_i) \quad (81) \]

\[ = nC \quad // \text{if we choose } p^*(x) \in \arg\max_{p(x)} I(X; Y) \quad (82) \]
Zero-error capacity

So there are 3 conditions for equality, $R = C$, namely

1. all codewords must be distinct
Zero-error capacity

So there are 3 conditions for equality, $R = C$, namely

1. all codewords must be distinct
2. $Y_i$'s are independent
Zero-error capacity

So there are 3 conditions for equality, \( R = C \), namely

1. all codewords must be distinct
2. \( Y_i \)'s are independent
3. distribution on \( x \) is \( p^*(x) \), a capacity achieving distribution.
Does feedback help for DMC

Consider a sequence of channel uses.
Does feedback help for DMC

Consider a sequence of channel uses.

Without Feedback

\[ X_1 \rightarrow Y_1 \]

\[ X_2 \rightarrow Y_2 \]

\[ X_3 \rightarrow Y_3 \]

\[ \vdots \]

\[ X_n \rightarrow Y_n \]
Does feedback help for DMC

Consider a sequence of channel uses.

Without Feedback

\[
X_1 \rightarrow Y_1 \\
X_2 \rightarrow Y_2 \\
X_3 \rightarrow Y_3 \\
\vdots \\
X_n \rightarrow Y_n
\]

With Feedback

\[
Y_i \perp \perp \{\text{all else}\} | X_i
\]
Consider a sequence of channel uses.

Without Feedback

\[ X_1 \rightarrow Y_1 \]
\[ X_2 \rightarrow Y_2 \]
\[ X_3 \rightarrow Y_3 \]
\[ \vdots \]
\[ X_n \rightarrow Y_n \]

With Feedback

\[ X_1 \rightarrow Y_1 \]
\[ X_2 \rightarrow Y_2 \]
\[ X_3 \rightarrow Y_3 \]
\[ \vdots \]
\[ X_n \rightarrow Y_n \]
\[ Y_i \perp \{ \text{all else} \} | X_i \]

Another way of looking at it is:

Encoder \( W \rightarrow X_i(W, Y_{1:i-1}) \)

Channel \( p(y|x) \)

Decoder \( Y_i \rightarrow \hat{W} \)

Error free feedback

Can this help? I.e., can this increase \( C \)?
Consider a sequence of channel uses.

Without Feedback
\[ X_1 \rightarrow Y_1 \]
\[ X_2 \rightarrow Y_2 \]
\[ X_3 \rightarrow Y_3 \]
\[ \vdots \]
\[ X_n \rightarrow Y_n \]

With Feedback
\[ X_1 \rightarrow Y_1 \]
\[ X_2 \rightarrow Y_2 \]
\[ X_3 \rightarrow Y_3 \]
\[ \vdots \]
\[ X_n \rightarrow Y_n \]
\[ Y_i \perp \{ \text{all else} \} | X_i \]

Another way of looking at it is:

Can this help? I.e., can this increase \( C \)?
Does feedback help for DMC

A:

No. Intuition: without memory, feedback tells us nothing more than what we already know, namely $p(y|x)$.

Can feedback made decoding easier? Yes, consider binary erasure channel, when we get $Y = e$ we just re-transmit.

In general, yes.
Does feedback help for DMC

- A: No.
Does feedback help for DMC

- A: No.
- Intuition: w/o memory, feedback tells us nothing more than what we already know, namely $p(y|x)$. 
Does feedback help for DMC

- A: No.
- Intuition: w/o memory, feedback tells us nothing more than what we already know, namely $p(y|x)$.
- Can feedback made decoding easier? Yes, consider binary erasure channel, when we get $Y = e$ we just re-transmit.
A: No.

Intuition: w/o memory, feedback tells us nothing more than what we already know, namely $p(y|x)$.

Can feedback made decoding easier? Yes, consider binary erasure channel, when we get $Y = e$ we just re-transmit.

Can feedback help for channels with memory?
Does feedback help for DMC

- A: No.
- Intuition: w/o memory, feedback tells us nothing more than what we already know, namely $p(y|x)$.
- Can feedback made decoding easier? Yes, consider binary erasure channel, when we get $Y = e$ we just re-transmit.
- Can feedback help for channels with memory? In general, yes.
Feedback for DMC

Definition 6.1 \((2^{nR}, n)\) feedback code

Such a code is the encoder \(X_i(W, Y_{1:i-1})\), a decoder \(g : Y^n \rightarrow \{1, 2, \ldots, 2^{nR}\}\), and \(P_e^{(n)} = \Pr(g(Y^n) \neq W)\) for \(H(W) = nR\) (uniform).

Definition 6.2 (Capacity)

The capacity with feedback \(C_{FB}\) of a DMC is the max of all rates achievable by feedback codes.

Theorem 6.3

\[
C_{FB} = C = \max_{p(x)} I(X; Y) \quad \text{for a DMC} \tag{84}
\]
Feedback codes for DMC

Proof.

- Clearly, $C_{\text{FB}} \geq C$, since FB code is a generalization.
- Next, we use $W$ instead of $X$ and bound $R$.
- We have

\[
H(W) = H(W|Y^n) + I(W; Y^n) \leq 1 + P_e(n) nR + I(W; Y^n) \quad \text{//by fano}
\]

We next bound $I(W; Y^n)$
Feedback codes from DMC

... proof continued.

\[ I(W; Y^n) = H(Y) - H(Y|W) \]
... proof continued.

\[
I(W; Y^n) = H(Y) - H(Y|W) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1:i-1, W) \tag{87}
\]

\[
\sum_{i=1}^{n} I(X_i; Y_i) \leq nC \tag{90}
\]
\[
I(W; Y^n) = H(Y) - H(Y|W) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, W) \quad (87)
\]

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, W, X_i) \quad \text{//note } X_i = f(W, Y_{1:i-1})
\]

\[
\leq \sum_{i=1}^{n} I(X_i; Y_i) \quad (90)
\]
Feedback codes from DMC

... proof continued.

\[
I(W; Y^n) = H(Y) - H(Y|W) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, W) \tag{87}
\]

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, W, X_i) \quad \text{//note } X_i = f(W, Y_{1:i-1})
\tag{88}
\]

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i)
\tag{90}
\]
Feedback codes fro DMC

... proof continued.

\[
I(W; Y^n) = H(Y) - H(Y|W) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1:i-1, W) \quad (87)
\]

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1:i-1, W, X_i) \quad /\text{note } X_i = f(W, Y_1:i-1) \quad (88)
\]

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \leq \sum_{i} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \quad (89)
\]

(90)
Feedback codes from DMC

... proof continued.

\[
I(W; Y^n) = H(Y) - H(Y|W) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, W) \quad (87)
\]

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, W, X_i) \quad \text{//note } X_i = f(W, Y_{1:i-1})
\]

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \leq \sum_{i} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \quad (89)
\]

\[
= \sum_{i} I(X_i; Y_i) \quad (90)
\]
Feedback codes fro DMC

. . . proof continued.

\[
I(W; Y^n) = H(Y) - H(Y|W) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1:i-1, W) \quad (87)
\]

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1:i-1, W, X_i) \quad /\!\!\!/ \text{note } X_i = f(W, Y_1:i-1)
\]

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \quad (89)
\]

\[
= \sum_{i} I(X_i; Y_i) \leq nC \quad (90)
\]

...
Thus we have

\[ H(W) \leq 1 + P_e^{(n)} nR + nC \]  \hspace{1cm} (91)

\[ \Rightarrow nR \leq 1 + P_e^{(n)} nR + nC \]  \hspace{1cm} (92)
Thus we have

\[ H(W) \leq 1 + P_e^{(n)} nR + nC \] \hspace{1cm} (91)

\[ \Rightarrow nR \leq 1 + P_e^{(n)} nR + nC \] \hspace{1cm} (92)

The implication follows since the first inequality holds for all \( H(W) \) including the maximum case at which \( H(W) = nR \).
Thus we have

\[ H(W) \leq 1 + P_e^{(n)} nR + nC \]  \hspace{1cm} (91)

\[ \Rightarrow nR \leq 1 + P_e^{(n)} nR + nC \]  \hspace{1cm} (92)

- The implication follows since the first inequality holds for all \( H(W) \) including the maximum case at which \( H(W) = nR \).
- This gives \( R \leq \frac{1}{n} P_e^{(n)} R + C \) or \( R \leq C \) as \( n \to \infty \).
Feedback codes from DMC

... proof continued.

Thus we have

\[ H(W) \leq 1 + P_e^{(n)} nR + nC \]  
(91)

\[ \Rightarrow nR \leq 1 + P_e^{(n)} nR + nC \]  
(92)

- The implication follows since the first inequality holds for all \( H(W) \) including the maximum case at which \( H(W) = nR \).
- This gives \( R \leq \frac{1}{n} P_e^{(n)} R + C \) or \( R \leq C \) as \( n \to \infty \).
- Thus feedback doesn’t help
Data compression: We now know that it is possible to achieve error free compression if our average rate of compression, $R$, measured in units of bits per source symbol, is such that $R > H$ where $H$ is the entropy of the generating source distribution.
Joint Source/Channel Theorem

- **Data compression:** We now know that it is possible to achieve error-free compression if our average rate of compression, $R$, measured in units of bits per source symbol, is such that $R > H$ where $H$ is the entropy of the generating source distribution.

- **Data Transmission:** We now know that it is possible to achieve error-free communication and transmission of information if $R < C$, where $R$ is the average rate of information sent (units of bits per channel use), and $C$ is the capacity of the channel.
Data compression: We now know that it is possible to achieve error free compression if our average rate of compression, $R$, measured in units of bits per source symbol, is such that $R > H$ where $H$ is the entropy of the generating source distribution.

Data Transmission: We now know that it is possible to achieve error free communication and transmission of information if $R < C$, where $R$ is the average rate of information sent (units of bits per channel use), and $C$ is the capacity of the channel.

Q: Does this mean that if $H < C$, we can reliably send a source of entropy $H$ over a channel of capacity $C$?
Data compression: We now know that it is possible to achieve error free compression if our average rate of compression, $R$, measured in units of bits per source symbol, is such that $R > H$ where $H$ is the entropy of the generating source distribution.

Data Transmission: We now know that it is possible to achieve error free communication and transmission of information if $R < C$, where $R$ is the average rate of information sent (units of bits per channel use), and $C$ is the capacity of the channel.

Q: Does this mean that if $H < C$, we can reliably send a source of entropy $H$ over a channel of capacity $C$?

This seems intuitively reasonable.
Joint Source/Channel Theorem: process

The process would go something as follows:

1. Compress a source down to its entropy, using Huffman, LZ, arithmetic coding, etc.
Joint Source/Channel Theorem: process

The process would go something as follows:

1. Compress a source down to its entropy, using Huffman, LZ, arithmetic coding, etc.
2. Transmit it over a channel.
Joint Source/Channel Theorem: process

The process would go something as follows:

1. Compress a source down to its entropy, using Huffman, LZ, arithmetic coding, etc.
2. Transmit it over a channel.
3. If all sources could share the same channel, would be very useful.
Joint Source/Channel Theorem: process

The process would go something as follows:

1. Compress a source down to its entropy, using Huffman, LZ, arithmetic coding, etc.
2. Transmit it over a channel.
3. If all sources could share the same channel, would be very useful.
4. I.e., perhaps the same channel coding scheme could be used regardless of the source, if the source is first compressed down to the entropy. The channel encoder/decoder need not know anything about the original source (or how to encode it).
Joint Source/Channel Theorem: process

The process would go something as follows:

1. Compress a source down to its entropy, using Huffman, LZ, arithmetic coding, etc.
2. Transmit it over a channel.
3. If all sources could share the same channel, would be very useful.
4. I.e., perhaps the same channel coding scheme could be used regardless of the source, if the source is first compressed down to the entropy. The channel encoder/decoder need not know anything about the original source (or how to encode it).
5. Joint source/channel decoding as in the following figure:
Joint Source/Channel Theorem: process

The process would go something as follows:

1. Compress a source down to its entropy, using Huffman, LZ, arithmetic coding, etc.
2. Transmit it over a channel.
3. If all sources could share the same channel, would be very useful.
4. I.e., perhaps the same channel coding scheme could be used regardless of the source, if the source is first compressed down to the entropy. The channel encoder/decoder need not know anything about the original source (or how to encode it).
5. Joint source/channel decoding as in the following figure:

   ![Joint Source/Channel Diagram]

6. Maybe obvious now, but at the time (1940s) it was a revolutionary idea!
Joint Source/Channel Theorem

- **Source**: $V \in \mathcal{V}$ that satisfies AEP (e.g., stationary ergodic).
Joint Source/Channel Theorem

- **Source:** \( V \in \mathcal{V} \) that satisfies AEP (e.g., stationary ergodic).
- **Send** \( V_1:n = V_1, V_2, \ldots, V_n \) over channel, entropy rate \( H(\mathcal{V}) \) of stochastic process (if i.i.d., \( H(\mathcal{V}) = H(V_i), \forall i \)).
Joint Source/Channel Theorem

- Source: $V \in \mathcal{V}$ that satisfies AEP (e.g., stationary ergodic).
- Send $V_{1:n} = V_1, V_2, \ldots, V_n$ over channel, entropy rate $H(\mathcal{V})$ of stochastic process (if i.i.d., $H(\mathcal{V}) = H(V_i), \forall i$).
- $V_{1:n} \rightarrow \text{Encoder} \rightarrow X^n \rightarrow \text{Channel} \rightarrow Y^n \rightarrow \text{Decoder} \rightarrow \hat{V}_{1:n}$
Source: $V \in \mathcal{V}$ that satisfies AEP (e.g., stationary ergodic).

Send $V_{1:n} = V_1, V_2, \ldots, V_n$ over channel, entropy rate $H(\mathcal{V})$ of stochastic process (if i.i.d., $H(\mathcal{V}) = H(V_i), \forall i$).

$V_{1:n} \rightarrow \text{Encoder} \rightarrow X^n \rightarrow \text{Channel} \rightarrow Y^n \rightarrow \text{Decoder} \rightarrow \hat{V}_{1:n}$

Error probability and setup:

$$P_e^{(n)} = P(V_{1:n} \neq \hat{V}_{1:n})$$

$$= \sum_{y_{1:n}, v_{1:n}} \Pr(v_{1:n}) \Pr(y_{1:n}|X^n(v_{1:n})) \mathbf{1}\{g(y_{1:n}) \neq v_{1:n}\}$$

(93)  
(94)
Joint Source/Channel Theorem

Theorem 7.1 (Source/Channel Coding Theorem)

If $V_{1:n}$ satisfies AEP, then there exists a sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$ if $H(\mathcal{V}) < C$.
Theorem 7.1 (Source/Channel Coding Theorem)

If \( V_1:n \) satisfies AEP, then there exists a sequence of \( (2^{nR}, n) \) codes with \( P_e^{(n)} \to 0 \) if \( H(V) < C \). Conversely, if \( H(V) > C \), then \( P_e^{(n)} > 0 \) for all \( n \) and cannot send with arbitrarily low probability of error.