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 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.
So 26 input symbols, and each symbol maps probabilistically to itself or its lexicographic neighbor.

- i.e., $p(A \rightarrow A) = p(A \rightarrow B) = 1/2$, etc.
- Each symbol always has some chance of error, so how can we communicate without error?
- Choose subset of symbols that can be uniquely disambiguated on receiver side. Choose every other source symbol, A, C, E, etc.
- Thus $A \rightarrow \{A, B\}$, $C \rightarrow \{C, D\}$, $E \rightarrow \{E, F\}$, etc. so that each received symbols has only one unique source symbol.
- Capacity $C = \log 13$
- Q: what happens to $C$ when probabilities are not all 1/2?

### Binary Symmetric Channel (BSC)

- A bit that is sent will be flipped with probability $p$.
- $p(Y = 1|X = 0) = p = 1 - p(Y = 0|X = 0)$. $p(Y = 0|X = 1) = p = p(Y = 1|X = 1)$.
- The BSC is an important channel since it is a simple model but at the same time captures some of the difficulties of more complicated channels.
- Q: can we still achieve reliable (“guaranteed” error free) communication with this channel? A: Yes, if $p < 1/2$ and if we do not ask for too high a transmission rate (which would be $R > C$), then we can. Actually, any $p \neq 1/2$ is sufficient.
- Intuition: think about AEP and/or block coding.
- But how to compute $C$ the capacity?
$I(X; Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X = x)$  

$(1)$

$= H(Y) - \sum_x p(x)H(p) = H(Y) - H(p) \leq 1 - H(p)$  

$(2)$

- When is $H(Y) = 1$? Note that

  $\Pr(Y = 1) = \Pr(Y = 1|X = 1)\Pr(X = 1)$  

  $(3)$

  $+ \Pr(Y = 1|X = 0)\Pr(X = 0)$  

  $(4)$

  $= (1 - p)\Pr(X = 1) + p\Pr(X = 0)$  

  $(5)$

  $= (1 - p)\Pr(X = 1) + p(1 - \Pr(X = 1))$  

  $(6)$

- So $H(Y) = 1$ if $H(X) = 1$ (i.e., $\Pr(X = 1) = 1/2$)
- Thus, we get that $C = 1 - H(p)$ which happens when $X$ is uniform.

- If $p = 1/2$ then $C = 0$, so if it randomly flips bits, then no information can be sent.
- If $p \neq 1/2$, then we can communicate, albeit potentially slowly. E.g., if $p = 0.499$ then $C = 2.8854 \times 10^{-6}$ bits per channel use. So to send one bit, need to use the channel quite a bit.
- If $p = 0$ or $p = 1$, then $C = 1$ and we can get maximum possible rate (i.e., the capacity is one bit per channel use).
Decoding

- Lets temporarily look ahead to address this problem.
- We can “decode” the source using the received string, source distribution, and the channel model \( p(y|x) \) via Bayes rule. I.e.

\[
\Pr(x|y) = \frac{\Pr(y|x) \Pr(x)}{\Pr(y)} = \frac{\Pr(y|x)\Pr(x)}{\sum_{x'} \Pr(y|x')\Pr(x')}
\]  

(7)

- If we get a particular \( y \), we can compute \( p(x|y) \) and make a decision based on that. I.e., \( \hat{x} = \arg\max_x p(x|y) \) (maximum likelihood decoding).
- This is optimal decoding in that it minimizes the error.
- Error if \( x \neq \hat{x} \), and \( \Pr(\text{error}) = \Pr(x \neq \hat{x}) \).
- This is minimal if we chose \( \arg\max_x p(x|y) \) since the error \( 1 - \Pr(\hat{x}|y) \) is minimal.

Properties of Channel Capacity \( C \)

- \( C \geq 0 \) since \( I(X;Y) \geq 0 \).
- \( C \leq \log |\mathcal{X}| \) since \( C = \max_p I(X;Y) \leq \max H(X) = \log |\mathcal{X}| \).
- \( C \leq \log |\mathcal{Y}| \) for same reason. Thus, the alphabet sizes limit the transmission rate.
- \( I(X;Y) = I_{p(x)}(X;Y) \) is a continuous function of \( p(x) \).
- Recall, \( I(X;Y) \) is a concave function of \( p(x) \) for fixed \( p(y|x) \). Thus, \( I_{\lambda_1 p_1 + (1-\lambda) p_2}(X;Y) \geq \lambda I_{p_1}(X;Y) + (1-\lambda) I_{p_2}(X;Y) \).
- Interestingly, since concave, this makes computing something like the capacity easier. I.e., a local maximum is a global maximum, and computing the capacity for a general channel model is a convex optimization procedure.
- Recall also, \( I(X;Y) \) is a convex function of \( p(y|x) \) for fixed \( p(x) \).
Shannon’s 2nd Theorem

- One of the most important theorems of the last century.
- We’ll see it in various forms, but we state it here somewhat informally to start acquiring intuition.

**Theorem 2.1 (Shannon’s 2nd Theorem)**

\( C \) is the maximum number of bits (on average, per channel use) that we can transmit over a channel reliably.

- Here, “reliably” means with vanishingly small and exponentially decreasing probability of error as the block length gets longer. We can easily make this probability essentially zero.
- Conversely, if we try to push \( > C \) bits through the channel, error quickly goes to 1.

Intuition of this we’ve already seen in the noisy typewriter and the region partitioning.
- Slightly more precisely, this is a sort of bin packing problem.
- We’ve got a region of possible codewords, and we pack as many smaller non-overlapping bins into the region as possible.
- The smaller bins correspond to the noise in the channel, and the packing problem depends on the underlying “shape”
- Not really a partition, since there might be wasted space, also depending on the bin and region shapes.
Shannon’s 2nd Theorem

- Intuitive idea: use typicality argument, like in chapter 3.
- There are $\approx 2^{nH(X)}$ typical sequences, each with probability $2^{-nH(X)}$ and with $p(A^{(n)}_c) \approx 1$, so the only thing with “any” probability is the typical set and it has all the probability.
- The same thing is true for conditional entropy.
- That is, for a typical input $X$, there are $\approx 2^{nH(Y|X)}$ output sequences.
- Overall, there are $2^{nH(Y)}$ typical output sequences, and we know that $2^{nH(Y)} \geq 2^{nH(Y|X)}$.

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

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

(8)

- 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
  \[
  \leq \frac{2^nH(Y)}{2^nH(Y|X)} = 2^n(H(Y) - H(Y|X)) = 2^nI(X;Y) \tag{9}
  \]

- 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.

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:
  
  \[
  \begin{align*}
  &\text{noise} \quad p(y|x) \\
  &\text{source} \quad W \\
  &\text{encoder} \quad X^n \\
  &\text{channel} \quad Y^n \\
  &\text{decoder} \quad \hat{W} \\
  &\text{receiver}
  \end{align*}
  \]

  \( n = \log M \) bits \quad \( n \log |X| \) bits \quad \( n \log |Y| \) bits

- **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)** \( (X, p(y|x), Y) \)
- \( n^{th} \) extension to channel is \( (X^n, p(y^n|x^n), Y^n) \)
- **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 2.2 ((\(M, n\)) 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 3.1 (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} (11)

Definition 3.2 (Max probability of Error $\lambda^{(n)}$ for $(M, n)$ code)

$$\lambda^{(n)} \triangleq \max_{i \in \{1, 2, \ldots, M\}} \lambda_i$$  \hspace{1cm} (12)

Definition 3.3 (Average probability of error $P_{e}^{(n)}$)

$$P_{e}^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i = \Pr(I \neq g(Y^n))$$  \hspace{1cm} (13)

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

$$= E(1(I \neq g(Y^n))) = \sum_{i=1}^{M} \Pr(g(Y^n) \neq i|X^n = X^n(i))p(i)$$  \hspace{1cm} (14)

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!
**Rate**

**Definition 3.4 (Rate $R$ of an $(M,n)$ code)**

\[
R = \frac{\log M}{n} = \frac{\text{total num. of bits in a source message}}{\text{total num. of channel uses needed to send a message}}
\]  
(15)

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

**Definition 3.5 (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 $n \to \infty$.

**Capacity**

**Definition 3.6 (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 3.7 (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:

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

with \( p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i) \).

Jointly Typical Sequences: Picture

- Set of all jointly typical pairs of sequences \( 2^{nH(X,Y)} \)
- Set of all pairs of marginally typical sequences
- \( 2^{nH(Y|X)} \)
- \( 2^{nH(Y)} \)
- \( 2^{nH(X)} \)
Jointly Typical Sequences: Intuition

- So intuitively,

\[
\frac{\text{num. jointly typical seqs.}}{\text{num ind. chosen typical seqs.}} = \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{n(H(X,Y) - H(X) - H(Y))} = 2^{-nI(X;Y)} \tag{21}
\]

- 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 \(2^{-nC}\).

Joint AEP

**Theorem 4.1**

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

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

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

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

\[
\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_c^{(n)}\right) \leq 2^{-n(I(X;Y)-3\epsilon)} \tag{23}
\]

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

\[
\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_c^{(n)}\right) \geq (1 - \epsilon)2^{-n(I(X;Y)+3\epsilon)} \tag{24}
\]

- Key property: we have bound on the probability of independently drawn sequences being jointly typical, falls off exponentially fast with \(n\), if \(I(X;Y) > 0\).
Joint AEP proof

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

- We have, by the w.l.l.n.s,
  \[
  -\frac{1}{n} \log \Pr(X^n) \to -E(\log p(X)) = H(X)
  \]
  so \( \forall \epsilon > 0, \exists m_1 \) such that for \( n > m_1 \)
  \[
  \Pr \left( \left| -\frac{1}{n} \log \Pr(X^n) - H(X) \right| > \epsilon \right) < \epsilon/3
  \]
  call this \( S_1 \)

- So, \( S_1 \) is a non-typical event.

Also, \( \exists m_2, m_3 \) such that \( \forall n > m_2, m_3 \), we have

- \[
  \Pr \left( \left| -\frac{1}{n} \log \Pr(Y^n) - H(Y) \right| > \epsilon \right) < \epsilon/3
  \]
  call this \( S_2 \)

and \( \forall n > m_3 \), we have

- \[
  \Pr \left( \left| -\frac{1}{n} \log \Pr(X^n, Y^n) - H(X,Y) \right| > \epsilon \right) < \epsilon/3
  \]
  call this \( S_3 \)

- So all events \( S_1, S_2 \) and \( S_3 \) are non-typical events.
Joint AEP proof

Proof of $\Pr\left( (X^n, Y^n) \in A_{\epsilon}^{(n)} \right) \rightarrow 1$.

- For $n > \max(m_1, m_2, m_3)$, we have that $p(S_1 \cup S_2 \cup S_3) \leq \epsilon = 3\epsilon/3$ by the union bound.
- So, non-typicality has probability $< \epsilon$, meaning $\Pr(A_{\epsilon}^{(n)^c}) \leq \epsilon$ giving $\Pr(A_{\epsilon}^{(n)}) \geq 1 - \epsilon$, as desired. □ for 1.

\[
1 = \sum_{x^n, y^n} p(x^n, y^n) \geq \sum_{(x^n, y^n) \in A_{\epsilon}^{(n)}} p(x^n, y^n) \geq |A_{\epsilon}^{(n)}| 2^{-n(H(X,Y)+\epsilon)} \quad (29)
\]

\[
|A_{\epsilon}^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)} \quad (30)
\]

- Also, from before, $\Pr(A_{\epsilon}^{(n)}) \geq 1 - \epsilon$ for big $n$, giving:

\[
1 - \epsilon \leq \sum_{(x^n, y^n) \in A_{\epsilon}^{(n)}} p(x^n, y^n) \leq |A_{\epsilon}^{(n)}| 2^{-n(H(X,Y)-\epsilon)} \quad (31)
\]

\[
|A_{\epsilon}^{(n)}| \geq (1 - \epsilon) 2^{n(H(X,Y)-\epsilon)} \quad (32)
\]

□ for 2.
Joint AEP proof

Proof of two indep. sequences are likely not jointly typical.
- Let $\tilde{X}^n, \tilde{Y}^n$ be independent $\sim p(x^n)p(y^n)$, i.e. the two sequences are independent of each other.
- Then we have the following two derivations:

$$
\Pr\left( (\tilde{X}^n, \tilde{Y}^n) \in A_{\epsilon}^{(n)} \right) = \sum_{(x^n,y^n) \in A_{\epsilon}^{(n)}} p(x^n)p(y^n) \tag{33}
\leq 2^n(H(X,Y)+\epsilon)2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)} \tag{34}
= 2^{-n(I(X;Y)-3\epsilon)} \tag{35}
$$

$$
\Pr\left( (\tilde{X}^n, \tilde{Y}^n) \in A_{\epsilon}^{(n)} \right) \geq (1-\epsilon)2^n(H(X,Y)-\epsilon)2^{-n(H(X)+\epsilon)}2^{-n(H(Y)+\epsilon)}
= (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \tag{36}
$$

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:

$$
\frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{n(H(X,Y)-H(X)-H(Y))} = 2^{-nI(X,Y)} \tag{37}
$$

and this is essentially the probability that a randomly chosen pair of (marginally) typical sequences is jointly typical.
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.

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 $y^n$.
- 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 5.1

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

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: \( X^n : \{1, \ldots, M\} \to \mathcal{X}^n \) maps to codewords \( X^n(i) \)
  3. Decoder: \( g : \mathcal{Y}^n \to \{1, \ldots, M\} \).
- 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 $p(x^n) = \prod_{i=1}^{n} p(x_i)$.
- 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:

\[
C = \begin{bmatrix}
  x_1(1) & x_2(1) & x_3(1) & \cdots & x_n(1) \\
  x_1(2) & x_2(2) & x_3(2) & \cdots & x_n(2) \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  x_1(2^{nR}) & x_2(2^{nR}) & x_3(2^{nR}) & \cdots & x_n(2^{nR})
\end{bmatrix}
\] (38)

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

... or even the entire codebook:

\[
p(C) = \prod_{\omega=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(\omega))
\] (40)

All rates $R < C$ are achievable.

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

- Can compute probabilities of a given codeword for $\omega$ ...
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
\[
Y^n \sim p(y^n|x^n(\omega)) = \prod_{i=1}^{n} p(y_i|x_{i}(\omega))
\]  
(41)
6. The signal is decoded using typical set decoding (to be described). . . .

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. $(x^n(k), y^n) \in A_c(n)$ (i.e., $\hat{\omega}$ is unique)

Otherwise output special invalid integer “0” (error). Three types of errors might occur.

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

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.

1. Code specific error

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

(42)

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.)

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

(43)

Surprisingly, this is much easier to analyze than $P_e$.

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

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

$$\Pr(\mathcal{E}) = \sum_C \Pr(C) P_e^{(n)}(C) = \sum_C \Pr(C) \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \lambda_\omega(C)$$  \hspace{1cm} (45)

$$= \frac{1}{2^{nR}} \sum_{\omega=1}^{2^{nR}} \sum_C \Pr(C) \lambda_\omega(C)$$  \hspace{1cm} (46)

but

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

$$= \sum_{x^n(1), x^n(2), \ldots, x^n(2^{nR})} \text{stuff}$$  \hspace{1cm} (47)

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.

So we get

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

(51)

Next, define the random events:

$$E_i \triangleq \{ (x^n(i), y^n) \in A_e^{(n)} \} \quad \text{for} \quad i = 1, \ldots, 2^{nR}$$

(52)

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.

Vertical axis is lexicographic order of possible codewords

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

Dots are the jointly typical sequences

Dots are the jointly typical sequences

Vertical axis is lexicographic order of possible codewords

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

Dots are the jointly typical sequences

Dots are the jointly typical sequences
Proof that all rates $R < C$ are achievable.

Goal: bound the probability of error:

$$\Pr(\mathcal{E}|W = 1) = \Pr(E_1^c \cup E_2 \cup E_3 \ldots)$$  \hspace{1cm} (53)

$$\leq \Pr(E_1^c) + \sum_{i=2}^{2^n R} \Pr(E_i) \text{ by the union bound} \hspace{1cm} (54)$$

We have that

$$\Pr(E_1^c) = \Pr(A_{\epsilon^n}^c) \to 0 \text{ as } n \to \infty$$  \hspace{1cm} (55)

So, $\forall \epsilon, \exists n_0$ s.t.

$$\Pr(E_1^c) \leq \epsilon, \ \forall n > n_0$$  \hspace{1cm} (56)
All rates $R < C$ are achievable.

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

Also, because of random code generation process (and recall, $W = 1$)

\[ X^n(1) \perp \perp X^n(i) \Rightarrow Y^n \perp \perp X^n(i), \text{ for } i \neq 1 \]  (57)

this gives

\[ \Pr((X^n(i), Y^n) \in A^{(n)}_{\epsilon}) \leq 2^{-n(I(X,Y) - 3\epsilon)} \]  (58)

by the joint AEP.

So we get:

\[
\Pr(\mathcal{E}) = \Pr(\mathcal{E}|W = 1) \leq \Pr(E_{1}^{c}) + \sum_{i=2}^{2^{nR}} \Pr(E_{i}) \\
\leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X,Y) - 3\epsilon)} \\
= \epsilon + (2^{nR} - 1)2^{-n(I(X,Y) - 3\epsilon)} \\
\leq \epsilon + 2^{3n\epsilon}2^{-n(I(X,Y) - R)} \\
= \epsilon + 2^{-n((I(X,Y) - 3\epsilon) - R)} \\
\leq 2\epsilon \quad \text{for large enough } n 
\]  (59-64)

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

- 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.

$$P_e(n) (C^*) \leq 2\epsilon$$

(65)
All rates \( R < C \) are achievable.

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