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

- 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, mutual information
- 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.
- L17 (3/1):
- L18 (3/6):
- L19 (3/8):

Finals Week: March 12th–16th.
Radio Communications

- Key: If we increase the transmission rate over a noisy channel will the error rate increase?
- Perhaps the only way to achieve error free communication is to have a rate of zero.
- The error profile we might expect to see is the following:

\[
P_e \rightarrow R
\]

- Here, probability of error \( P_e \) goes up linearly with the rate \( R \), with an intercept at zero.
- This was the prevailing wisdom at the time. Shannon was critical in changing that.

A key idea

- If we choose the messages carefully at the sender, then with very high probability, they will be uniquely identifiable at the receiver.
- The idea is that we choose the source messages that (tend to) not have any ambiguity (or have any overlap) at the receiver end. I.e.,

\[
\text{Source Messages} \rightarrow \text{Received Messages}
\]

- This might restrict our possible set of source messages (in some cases severely, and thereby decrease our rate \( R \)), but if any message received in a region corresponds to only one source message, “perfect” communication can be achieved.
Discrete Channels

Definition 3.1 (discrete channel)
A discrete channel is one where there is an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$, and a distribution $p(y|x)$ which is the probability of observing output $y$ after seeing $x$ as input.

Definition 3.2 (memoryless channel)
A discrete channel is memoryless if $y_t$, the output at time $t$, is independent of all previous inputs, given $x_t$. I.e., $y_t \perp \perp x_1:t-1|x_t$.

- We will see many instances of discrete memoryless channels (DMCs).
- Recall back from lecture 1 our general model of communications:

Model of Communication

- **Source** message $W$, one of $M$ messages.
- **Encoder** transforms this into a length-$n$ string of source symbols $X^n$.
- **Noisy channel** distorts this message into a length-$n$ string of receiver symbols $Y^n$.
- **Decoder** attempts to reconstruct original message as best as possible, comes up with $\hat{W}$, one of $M$ possible sent messages.

$$\text{noise } p(y|x)$$

$W \xrightarrow{\text{encoder}} X^n \xrightarrow{\text{channel}} Y^n \xrightarrow{\text{decoder}} \hat{W}$

$n = \log M \text{ bits} \quad n \log |\mathcal{X}| \text{ bits} \quad n \log |\mathcal{Y}| \text{ bits}$
Rates and Capacities

Definition 3.3 (information flow)

The rate of information flow through a channel is given by $I(X; Y)$, the mutual information between $X$ and $Y$, in units of bits per channel use.

Definition 3.4 (capacity)

The information capacity of a channel is the maximum information flow.

$$C \triangleq \max_{p(x) \in \Delta} I(X; Y) \quad (1)$$

where $\Delta$ is the set of all possible probability distributions over source alphabet $\mathcal{X}$. Thus, $C$ is the maximum number of bits sent over the channel per channel use.

Definition 3.5 (rate)

The rate $R$ of a code is measured in the number of bits per channel use.

Fundamental Limits of Data Transmission/Communication

- For communication, lower bound on probability of error becomes bounded away from 0 as the rate of the code $R$ goes above a fundamental quantity $C$. Note, $P_e \propto e^{-nE(R)}$.

  That is,

  $$\log P_e$$

  $R \rightarrow C$

- The only possible way to get low error is if $R < C$. Something funny happens at the point $C$, the capacity of the channel.

- Note that $C$ is not 0, so can still achieve “perfect” communication over a noisy channel as long as $R < C$. 
Consider the channel

\[ A \rightarrow \{A, B\}, \; C \rightarrow \{C, D\}, \; E \rightarrow \{E, F\}, \ldots \]

- 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?
BSC Capacity

\[ I(X; Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x) H(Y|X = x) \]  
\[ = H(Y) - \sum_x p(x) H(p) = H(Y) - H(p) \leq 1 - H(p) \]  

- When is \( H(Y) = 1 \)? Note that
  \[ \Pr(Y = 1) = \Pr(Y = 1|X = 1)\Pr(X = 1) \]  
  \[ + \Pr(Y = 1|X = 0)\Pr(X = 0) \]  
  \[ = (1 - p)\Pr(X = 1) + p\Pr(X = 0) \]  
  \[ = (1 - p)\Pr(X = 1) + p(1 - \Pr(X = 1)) \]

- 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.

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.

$$\Pr(x|y) = \frac{\text{Pr}(y|x) \Pr(x)}{\Pr(y)} = \frac{\Pr(y|x)\Pr(x)}{\sum_{x'} \Pr(y|x')\Pr(x')}$$

(8)

- 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.

Minimum Error Decoding

- Note: Computing quantities such as $\Pr(x|y)$ is a task of probabilistic inference.
- Often this problem is difficult (NP-hard, see Cooper and Herskovitz, 1990). This means that doing minimum error decoding might very well be exponentially expensive (unless $P = NP$).
- Many real world codes are such that computing the exact computation must be approximated (i.e., no known fast algorithm for minimum error or maximum likelihood decoding).
- Instead we do approximate inference algorithms (e.g., loopy belief propagation, message passing, etc.). These algorithms tend still to work very well in practice (achieve close to the capacity $C$).
- But before doing that, we need first to study more channels and the theoretical properties of the capacity $C$. 
Binary Erasure Channel

- $e$ is an erasure symbol, if that happens we don’t have access to the transmitted bit.
- The probability of dropping a bit is then $\alpha$.
- We want to compute capacity. Obviously, $C = 1$ if $\alpha = 0$.

\[
C = \max_p I(X;Y) = \max_p (H(Y) - H(Y|X)) \tag{9}
\]
\[
= \max_p H(Y) - H(\alpha) \tag{10}
\]

- So while $H(Y) \leq \log 3$, we want actual value of the capacity.

Let $E = \{Y = e\}$. Then

\[
H(Y) = H(Y,E) = H(E) + H(Y|E)
\]

Let $\pi = \Pr(X = 1)$. Then

\[
H(Y) = \begin{cases} 
\alpha(1-\pi)(1-\alpha) & \text{if } Y = 0 \\
\alpha & \text{if } Y = e \\
\pi(1-\alpha) & \text{if } Y = 1 
\end{cases}
\]
\[
= H(\alpha) + (1-\alpha)H(\pi) \tag{11}
\]

This last equality follows since $H(E) = H(\alpha)$, and

\[
H(Y|E) = \alpha H(Y|e) + (1-\alpha)H(Y|\neq e) = \alpha \cdot 0 + (1-\alpha)H(\pi)
\]
Then we get

\[ C = \max_{p(x)} H(Y) - H(\alpha) \]  

(13)

\[ = \max_{\pi} (1 - \alpha)H(\pi) + H(\alpha) - H(\alpha) \]  

(14)

\[ = \max_{\pi} (1 - \alpha)H(\pi) = 1 - \alpha \]  

(15)

- Best capacity when \( \pi = 1/2 = \Pr(X = 1) = \Pr(X = 0) \).
- This makes sense, loose \( \alpha \)% of the bits of original capacity.

### Ternary Confusion Channel

\[ P(Y = j|X = ?) = 1/2. \]

- Whenever the symbol “?” is input, the output is random.
- Other inputs are reliable.
- Thus, \( C = 1 \) bit.
**Symmetric Channels**

**Definition 5.1**

A channel is **symmetric** if rows of the channel transmission matrix \( p(y|x) \) are permutations of each other, and columns of this matrix are permutations of each other. A channel is weakly symmetric if every row of the matrix is a permutation of every other row, and all column sums \( \sum_x p(y|x) \) are equal.

**Theorem 5.2**

For weakly symmetric channels, we have that

\[
C = \log |\mathcal{Y}| - H(r) \tag{16}
\]

where \( r \) is the row of the transmission matrix.

- This follows immediately since

\[
I(X; Y) = H(Y) - H(Y|X) = H(Y) - H(r) \leq \log |\mathcal{Y}| - H(r)
\]

**Properties of Channel Capacity \( C \)**

- \( C \geq 0 \) since \( I(X; Y) \geq 0 \).
- \( C \leq \log |\mathcal{X}| \) since \( C = \max_{p(x)} I(X; Y) \leq \max H(X) = \log |\mathcal{X}| \).
- \( C \leq \log |\mathcal{Y}| \) 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, \( I_{\lambda p_1 + (1-\lambda)p_2}(X; Y) \geq \lambda I_{p_1}(X; Y) + (1-\lambda)I_{p_2}(X; Y) \).
- 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 5.3 (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 \( p(A^{(n)}_\epsilon) \approx 1 \), 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

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

(17)

- 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
  \[ \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n(H(Y) - H(Y|X))} = 2^{nI(X;Y)} \]  
  \[ (18) \]

- 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:
  
  \[
  \text{source} \xrightarrow{W} \text{encoder} \xrightarrow{X^n} \text{channel} \xrightarrow{Y^n} \text{decoder} \xrightarrow{\hat{W}} \text{receiver}
  \]

  \[n = \log M \text{ bits} \quad n \log |X| \text{ bits} \quad n \log |Y| \text{ bits}\]

- **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)** \((\mathcal{X}, p(y|x), \mathcal{Y})\)
- **\(n^{th}\) 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 5.4 ((M, n) code)**

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

1. An index set \(\{1, 2, \ldots, M\}\)
2. An encoding function \(X^n : \{1, 2, \ldots, M\} \rightarrow \mathcal{X}^n\) 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., \(g : \mathcal{Y}^n \rightarrow \{1, 2, \ldots, M\}\) 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 5.5 (Probability of Error $\lambda_i$ for message $i \in \{1, \ldots, M\}$)

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

(20)

Definition 5.6 (Max probability of Error $\lambda^{(n)}$ for $(M, n)$ code)

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

(21)

Definition 5.7 (Average probability of error $P_e^{(n)}$)

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

(22)

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

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

(23)

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

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

(24)

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

**Definition 5.9 (Achievability for a given channel)**

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

Capacity

**Definition 5.10 (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 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 5.11 (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:

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

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 sequences: \( |\mathcal{X}^n \times \mathcal{Y}^n| = (|\mathcal{X}| |\mathcal{Y}|)^n \)
Jointly Typical Sequences: Intuition

- So intuitively,

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

(29) \hspace{1cm} (30) \hspace{1cm} (31)

- 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 6.1**

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

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

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

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

(32)

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

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

(33)

- 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,
  \[
  \frac{1}{n} \log \Pr(X^n) \to -E(\log p(X)) = H(X) \quad (34)
  \]
  so $\forall \epsilon > 0$, $\exists m_1$ such that for $n > m_1$
  \[
  \Pr\left(\left|\frac{1}{n} \log \Pr(X^n) - H(X)\right| > \epsilon \right) < \epsilon/3 \quad (35)
  \]
  call this $S_1$
  
  - So, $S_1$ is a non-typical event.

- Also, $\exists m_2, m_3$ such that $\forall n > m_2$, we have
  \[
  \Pr\left(\left|\frac{1}{n} \log \Pr(Y^n) - H(Y)\right| > \epsilon \right) < \epsilon/3 \quad (36)
  \]
  and $\forall n > m_3$, we have
  \[
  \Pr\left(\left|\frac{1}{n} \log \Pr(X^n, Y^n) - H(X, Y)\right| > \epsilon \right) < \epsilon/3 \quad (37)
  \]
  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 $\Pr(\bar{A_{\epsilon}^{(n)}}) \leq \epsilon$ giving $\Pr(A_{\epsilon}^{(n)}) \geq 1 - \epsilon$, as desired. $\Box$ for 1.

Joint AEP proof

Proof of $|A_{\epsilon}^{(n)}| \leq 2^n(H(X,Y)+\epsilon)$.

- We have

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

\[ \Rightarrow |A_{\epsilon}^{(n)}| \leq 2^n(H(X,Y)+\epsilon) \quad (38) \]

\[ \Rightarrow |A_{\epsilon}^{(n)}| \leq 2^n(H(X,Y)+\epsilon) \quad (39) \]

- Also, $\Pr(A_{\epsilon}^{(n)}) \geq 1 - \epsilon$ for big $n$, since

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

\[ \Rightarrow |A_{\epsilon}^{(n)}| \geq (1 - \epsilon)2^n(H(X,Y)-\epsilon) \quad (40) \]

\[ \Rightarrow |A_{\epsilon}^{(n)}| \geq (1 - \epsilon)2^n(H(X,Y)-\epsilon) \quad (41) \]

$\Box$ 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:

\[
\Pr\left( (\tilde{X}^n, \tilde{Y}^n) \right) = \sum_{(x^n, y^n) \in A_{\epsilon}^n} p(x^n)p(y^n) \tag{42}
\]

\[
\leq 2^n(H(X,Y)+\epsilon)2^{-n(H(X)\epsilon)}2^{-n(H(Y)\epsilon)} \tag{43}
\]

\[
= 2^{-n(I(X;Y)\epsilon)} \tag{44}
\]

\[
\Pr\left( (\tilde{X}^n, \tilde{Y}^n) \right) \geq (1 - \epsilon)2^n(H(X,Y)-\epsilon)2^{-n(H(X)+\epsilon)}2^{-n(H(Y)+\epsilon)} \tag{45}
\]

\[
= (1 - \epsilon)2^{-n(I(X;Y)+3\epsilon)} \tag{46}
\]

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:

\[
\frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{n(H(X,Y)-H(X)-H(Y))} = 2^{-nI(X,Y)} \tag{47}
\]

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 before we start using pairs that will be jointly typical and chosen randomly.
- Ex: if $p(x) = 1/M$ then we can choose about $M$ samples before we see a given $x$, 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 $n$. 
Channel Coding Theorem (Shannon 1948): more formally

**Theorem 6.2**

All rates below capacity 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 \( R < C \).