Logistics

Class Road Map - IT-I

- L1 (9/26): Overview, Communications, Information, Entropy
- L2 (10/1): Props. Entropy, Mutual Information
- L5 (10/10): AEP, Compression
- L6 (10/15): Compression, Method of Types
- L8 (10/22): Entropy rates, HMMs, Coding, Kraft
- L9 (10/24): Kraft, Shannon Codes, Huffman, Shannon/Fano/Elias
- L10 (10/28): Huffman, Shannon/Fano/Elias
- L11 (10/29): Shannon Games
- LXX (10/31): Midterm, in class.
- L12 (11/7): Arith. Coding, Channel Capacity
- L13 (11/12): Channel Capacity
- L14 (11/14): Channel Capacity
- L15
- L16
- L17
- L18
- L19

Finals Week: December 12th–16th.
Cumulative Outstanding Reading

- Read chapters 1 and 2 in our book (Cover & Thomas, “Information Theory”) (including Fano’s inequality).
- Chapter 3 in our book (Cover & Thomas, “Information Theory”).
- Section 11.1 (method of types).
- Chapter 4 and 5 in our book (Cover & Thomas, “Information Theory”)
- Read stream code chapter 6 in “Information Theory, Inference, and Learning Algorithms” by David J.C. MacKay (available online at http://www.inference.phy.cam.ac.uk/mackay/itila/)
- Read Chapter 7 in our book (Cover & Thomas, “Information Theory”).

Homework

- Homework 5 on our web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/), due Thursday, Nov 14th, at 11:45pm.
Announcements

- Office hours, every week, **now Thursdays 4:30-5:30pm**. Can also reach me at that time via a canvas conference.
- Final assignment, need to upload pdf scan we send you.

Model of Communication

- **Source** message $W$, one of $M$ messages.
- **Encoder** transforms this into a length-$n$ string of source symbols $X^n$ (we might call them “channel input symbols”).
- **Noisy channel** distorts this message into a length-$n$ string of receiver symbols $Y^n$ (we might call them “channel output symbols”).
- **Decoder** attempts to reconstruct original message as best as possible, comes up with $\hat{W}$, one of $M$ possible sent messages.
- Note, $p(x, y) = p(x)p(y|z)$. $p(y|x)$ will model our channel and is fixed in most cases (we can’t control the channel). $p(x)$ is our source distribution, and we get to determine (and optimize over) it.
Rates and Capacities

**Definition 14.2.3 (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 14.2.4 (capacity)**

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

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

(14.3)

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 14.2.5 (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, we have a “dual” situation to entropy compression, i.e.,

![Error Exponent](image)

- We will show: only way to get low error is $R < C$. Something funny happens at the point $C$, the channel capacity.
- Note that $C$ is not 0, so can still achieve “perfect” communication over a noisy channel as long as $R < C$. 
For the moment, all we known about $C \triangleq \max_{p(x) \in \Delta} I(X; Y)$ is its definition, which is the result of an optimization problem.

- We will do this. For now, think of $C$ has being measured in units of “bits per channel use”.

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.,

- 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.
Consider the channel A → B → C → D → E → F → G → H → I → J → K → L → M → N → O → P → Q → R → S → T → U → V → W → X → Y → Z

- 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 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, ...
- 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 \(1/2\)?

### BSC Capacity

\[
I(X; Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|X = x) \quad (14.6)
\]

\[
= H(Y) - \sum_x p(x)H(p) = H(Y) - H(p) \leq 1 - H(p) \quad (14.7)
\]

- To achieve the upper bound, need \(H(Y) = 1\). Note that \(H(Y) = 1\) if \(H(X) = 1\) (i.e., if \(\Pr(X = 1) = 1/2\)).
- Thus, we get that \(C = 1 - H(p)\) which happens when \(X\) is uniform.
Ternary Confusion Channel

\[
\begin{array}{ccc}
X & ? & Y \\
0 & \frac{\sqrt{2}}{} & 0 \\
1 & \frac{\sqrt{2}}{} & 1 \\
\end{array}
\]

- \( P(Y = j|X = ?) = \frac{1}{2} \).
- Whenever the symbol “?” is input, the output is random.
- Other inputs are reliable.
- Thus, \( C = 1 \) bit.

Symmetric Channels

**Definition 14.4.1**
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 14.4.2**
For weakly symmetric channels, we have that
\[
C = \log |\mathcal{Y}| - H(r) \tag{14.1}
\]
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 can 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 14.4.3** (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.
**Shannon’s 2nd Theorem**

- 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”

**Shannon’s 2nd Theorem: Bin packing intuition**

- Bin packing: 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

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

(14.2)

- 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
  \[
  \leq \frac{2^n H(Y)}{2^n H(Y|X)} = 2^n (H(Y) - H(Y|X)) = 2^n I(X;Y)
  \]  
  \[(14.3)\]

- 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.

  \[
  \frac{2^n H(Y)}{2^n H(Y|X)}
  \]

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:

  \[ \text{noise } p(y | x) \]

  \[
  \begin{array}{c}
  \text{source} \\
  W \\
  n = \log M \text{ bits}
  \end{array}
  \quad
  \begin{array}{c}
  \text{encoder} \\
  X^n \\
  n \log |\mathcal{X}| \text{ bits}
  \end{array}
  \quad
  \begin{array}{c}
  \text{channel} \\
  Y^n \\
  n \log |\mathcal{Y}| \text{ bits}
  \end{array}
  \quad
  \begin{array}{c}
  \text{decoder} \\
  \hat{W}
  \end{array}
  \quad
  \begin{array}{c}
  \text{receiver}
  \end{array}
  \]

  \[
  g(Y^n)
  \]

  \[
  \hat{W} = g(Y^n)
  \]

- 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) \( (\mathcal{X}, p(y | x), \mathcal{Y}) \)
- \( n^{\text{th}} \) 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.

\[ \text{(M, n) code} \]

\[ \text{Definition 14.4.4 ((M, n) code)} \]

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

- An index set \( \{1, 2, \ldots, M\} \)
- 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.
- 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 14.4.5 (Probability of Error $\lambda_i$ for message $i \in \{1, \ldots, M\}$)**

$$\lambda_i \triangleq \Pr(g(Y^n) \neq i|X^n = X^n(i)) = \sum_{y^n \in Y^n} p(y^n|X^n(i))1(g(y^n) \neq i)$$  \hspace{1cm} (14.5)

**Definition 14.4.6 (Max probability of Error $\lambda^{(n)}$ for $(M, n)$ code)**

$$\lambda^{(n)} \triangleq \max_{i \in \{1,2,\ldots,M\}} \lambda_i$$  \hspace{1cm} (14.6)

**Definition 14.4.7 (Average probability of error $P_e^{(n)}$)**

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i = \Pr(I \neq g(Y^n))$$  \hspace{1cm} (14.7)

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} (14.8)

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

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

(14.9)

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

**Definition 14.4.9 (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 14.4.10 (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 any longer go to zero with increasing $n$.
- Note: this is a different notion of capacity that we encountered before.
- Before we defined $C = \max_{p(x)} I(X; Y)$ as the “information capacity”
- 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 14.4.11 (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 \( (\in A^{(n)}_\epsilon) \) as per the following definition:

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

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

- So intuitively,

\[
\frac{\text{num. jointly typical seqs.}}{\text{num ind. chosen typical seqs.}} = \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{n(H(X,Y) - H(X) - H(Y))} = 2^{-nI(X;Y)} \tag{14.14}
\]

- 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 14.5.1**

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

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

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 \((\hat{X}^n, \hat{Y}^n) \sim p(x^n)p(y^n)\) are drawn independently, then

\[
\Pr\left( (\hat{X}^n, \hat{Y}^n) \in A_{\epsilon}^{(n)} \right) \leq 2^{-n(I(X;Y)-3\epsilon)} \tag{14.17}
\]

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

\[
\Pr\left( (\hat{X}^n, \hat{Y}^n) \in A_{\epsilon}^{(n)} \right) \geq (1 - \epsilon)2^{-n(I(X;Y)+3\epsilon)} \tag{14.18}
\]

- 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 \( \Pr \left( (X^n, Y^n) \in A^{(n)}_\varepsilon \right) \rightarrow 1 \).

- We have, by the w.l.l.n.s,
  \[
  -\frac{1}{n} \log \Pr(X^n) \rightarrow -E(\log p(X)) = H(X)
  \]
  \( (14.19) \)
  so \( \forall \varepsilon > 0, \exists m_1 \) such that for \( n > m_1 \)
  \[
  \Pr \left( \left| -\frac{1}{n} \log \Pr(X^n) - H(X) \right| > \varepsilon \right) < \varepsilon/3
  \]
  call this \( S_1 \)

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

- Also, \( \exists m_2, m_3 \) such that \( \forall n > m_2 \), we have
  \[
  \Pr \left( \left| -\frac{1}{n} \log \Pr(Y^n) - H(Y) \right| > \varepsilon \right) < \varepsilon/3
  \]
  call this \( S_2 \)

  and \( \forall n > m_3 \), we have
  \[
  \Pr \left( \left| -\frac{1}{n} \log \Pr(X^n, Y^n) - H(X, Y) \right| > \varepsilon \right) < \varepsilon/3
  \]
  call this \( S_3 \)

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

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

- 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. $\square$ for 1.

\[ 1 = \sum_{x^n, y^n} p(x^n, y^n) \geq \sum_{(x^n, y^n) \in A_{\epsilon}^{(n)}} p(x^n, y^n) \geq |A_{\epsilon}^{(n)}|2^{-n(H(X,Y)+\epsilon)} \]
\[ \Rightarrow |A_{\epsilon}^{(n)}| \leq 2^n(H(X,Y)+\epsilon) \]  \hspace{1cm} (14.23)
\[ \Rightarrow |A_{\epsilon}^{(n)}| \leq 2^n(H(X,Y)+\epsilon) \]  \hspace{1cm} (14.24)

- Also, from before, $\Pr(A_{\epsilon}^{(n)}) \geq 1 - \epsilon$ for big $n$, giving:
\[ 1 - \epsilon \leq \sum_{(x^n, y^n) \in A_{\epsilon}^{(n)}} p(x^n, y^n) \leq |A_{\epsilon}^{(n)}|2^{-n(H(X,Y) - \epsilon)} \]  \hspace{1cm} (14.25)
\[ \Rightarrow |A_{\epsilon}^{(n)}| \geq (1 - \epsilon)2^n(H(X,Y) - \epsilon) \]  \hspace{1cm} (14.26)

$\square$ 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:

\[
\Pr \left( (\tilde{X}^n, \tilde{Y}^n) \in A_{\epsilon}^{(n)} \right) = \sum_{(x^n, y^n) \in A_{\epsilon}^{(n)}} p(x^n)p(y^n) \tag{14.27}
\]

\[
\leq 2^n(H(X,Y)+\epsilon)2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)} \tag{14.28}
\]

\[
= 2^{-n(I(X;Y)-3\epsilon)} \tag{14.29}
\]

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

\[
= (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \tag{14.30}
\]

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:

\[
\frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{n(H(X,Y)-H(X)-H(Y))} = 2^{-nI(X,Y)} \tag{14.31}
\]

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 14.6.1**

*All rates below* $C \triangleq \max_{p(x)} I(X;Y)$ *are achievable. Specifically,*

\[ \forall R < C, \text{ there exists a sequence of } (2^{nR}, n) \text{ codes with maximum} \]

\[ \lambda^{(n)} \rightarrow 0 \text{ as } n \rightarrow \infty. \text{ Conversely, any } (2^{nR}, n) \]

\[ \text{sequence of codes with } \lambda^{(n)} \rightarrow 0 \text{ as } n \rightarrow \infty \text{ must have that } R < C. \]

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

$$C = \begin{bmatrix} x_1(1) & x_2(1) & x_3(1) & \ldots & x_n(1) \\ x_1(2) & x_2(2) & x_3(2) & \ldots & x_n(2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & x_3(2^{nR}) & \ldots & x_n(2^{nR}) \end{bmatrix}$$  \hfill (14.32)

- Since $C$ is random, $\Pr(C)$ is sensible.
- 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 $x_{1:n}(\omega) = (x_1(\omega), x_2(\omega), \ldots, x_n(\omega))$.

All rates $R < C$ are achievable.

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

- Can compute probabilities of a given codeword for $\omega \in \{1, 2, \ldots, M\}$

$$p(x^n(\omega)) = \prod_{i=1}^{n} p(x_i(\omega)), \; \omega \in \{1, \ldots, M\}$$  \hfill (14.33)

- \ldots or even the entire codebook:

$$p(C) = \prod_{\omega=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(\omega))$$  \hfill (14.34)
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
   
   $Y^n \sim p(y^n|x^n(\omega)) = \prod_{i=1}^{n} p(y_i|x_i(\omega))$  \hfill (14.35)

6. The signal is decoded using typical set decoding (to be described).

Three types of errors might occur (type A, B, or C).

A: \exists k \neq \hat{\omega} \text{ s.t. } (x^n(k), y^n) \in A^{(n)}(\epsilon) \text{ (i.e., } \hat{\omega} \text{ is unique)}

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

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

where (as a reminder)

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

but we would like something easier to analyze.

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

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

(14.37)

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

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

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

$$\Pr(\mathcal{E}) = \sum_C \Pr(C) P_e^{(n)}(C) = \sum_C \Pr(C) \frac{1}{2^n R} \sum_{\omega=1}^{2^n R} \lambda_\omega(C) \tag{14.39}$$

$$= \frac{1}{2^n R} \sum_{\omega=1}^{2^n R} \sum_C \Pr(C) \lambda_\omega(C) \tag{14.40}$$

but

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

$$= \sum_{x^n(1), x^n(2), \ldots, x^n(2^n R)} \text{stuff} \tag{14.41}$$

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

$$\sum_C \Pr(C) \lambda_1(C) = \beta \tag{14.42}$$

$$\sum_{x^n(\omega)} \Pr(g(Y^n) \neq \omega | X^n = x^n(\omega)) \Pr(x^n(\omega)) = 1 \tag{14.43}$$

$$= \sum_{x^n(\omega)} \Pr(g(Y^n) \neq 1 | X^n = x^n(1)) \Pr(x^n(1)) = \sum_C \Pr(C) \lambda_1(C) = \beta \tag{14.44}$$
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:

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

(14.45)

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

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

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

(14.47)

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

Then the no error event is the same as: $E_1 \cap \neg(E_2 \cup E_3 \cup \ldots \cup E_M)$.

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

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

Various flavors of error

- $E_1^n$ 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
  - 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.

Vertical axis is lexicographic order of possible codewords

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

Dots are the jointly typical sequences
All rates $R < C$ are achievable.

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

Goal: bound the probability of error:

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

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

We have that

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

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

$$\Pr(E_1^c) \leq \epsilon, \ \forall n > n_0$$

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

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

This gives, for $i \neq 1$,

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

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:

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

(14.54)

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

(14.55)

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

(14.56)

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

(14.57)

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

(14.58)

$$\leq 2\epsilon \quad \text{for large enough } n$$

(14.59)

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.

- Lets break apart the this error probability.

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

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

  \[ \leq 2\epsilon \]

- 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:

  \[ \frac{1}{2^{nR}} \sum_{i: \lambda_i \geq 4\epsilon} \lambda_i \geq \frac{1}{2^{nR}} \sum_{i: \lambda_i \geq 4\epsilon} 4\epsilon = \frac{1}{2^{nR}} |\{i : \lambda_i \geq 4\epsilon\}| 4\epsilon > \frac{1}{2} 4\epsilon = 2\epsilon \]

  \[ \Rightarrow |\{i : \lambda_i \geq 4\epsilon\}| > \frac{2^{nR}}{2} \]

  \[ 14.63 \]

- 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

  \[ |\{i : \lambda_i \geq 4\epsilon\}| \leq \frac{2^{nR}}{2} \Rightarrow |\{i : \lambda_i < 4\epsilon\}| \geq \frac{2^{nR}}{2} \]

  \[ 14.64 \]

- 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 \( (2^n R, n) \) 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 prove the converse . . .

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

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