Instructor: Radha Poovendran, Associate Processor, Electrical Engineering Department, EE1, Room 434. Phone: (206) 221-6512. E-mail: rp3@u.washington.edu

Office Hours: Mon, Tue, Wed, Thu: 10:00am-11:00am and by appointment.

Lectures: Mon, Wed 2:30pm-4:20pm, EEB 054

Added Note about class meeting hours: There will be no lectures on the following dates: Jan **, Feb 18 and Feb 20. The makeup classes will be held on Jan * and Feb * from 2:30pm-4:20pm. Meeting room will be announced in class later.

Course Description: This is the first of the two course sequence. (IT-II is scheduled to be taught in spring 2008 by Prof. Jeff Bilmes.) The first course will develop the principal contributions of information theory in modeling communication systems. The main topics are:

1. Measures of Information: entropy, relative entropy and mutual information for probability experiments. Basic inequalities and chain rules among these measures.
2. Data compression by variable-length codes: Kraft inequality and Huffman construction, the relationship of average codeword length to the source entropy.
3. Data compression of discrete memoryless (i.i.d.) sources by means of block codes: the asymptotic equipartition property and the significance of entropy. Data Compression schemes for Markov Sources.
5. Communication over discrete channels: the channel coding theorem and the physical significance of channel capacity.
6. Measures of information for sources with continuous range. Discrete time Gaussian channels, either in series or parallel configuration, extension to continuous time Gaussian Channels.
7. Quantization of discrete time memoryless sources under a fidelity criterion. The mathematical definition of the rate distortion function and its physical significance as revealed in the rate distortion theorem.

Prerequisites: Fundamentals of mathematical analysis (or good knowledge of advanced calculus), some linear algebra (up to diagonalization of square matrices) and a course in probability/random process such as EE 505. Required concepts from probability/random processes are: the calculus of distributions including the multivariate Gaussian, the Chebychev and Jensen inequalities, conditional probability and independence, sequences of independent random variables and laws of large numbers, elements of Markov chains and stationary processes.
Textbook: The required textbook for the course is the 2nd edition of T. M. Cover and J. A. Thomas, “Elements of Information Theory,” John Wiley, 2006. Primary chapters to be covered are: 2, 3, 4, 5, 7, 8, 9, optional chapter 10.

Additional Suggested References


Grading Policy: Homework Assignments 10%; two midterms 20% each; Comprehensive Final 50%

You may bring your text book OR two pages of notes to an exam. Use of any other notes or electronic devices will not be allowed during exam.

Homework Set 1; Due Jan 16th 2008, in class.

Chapter 2, Text, page 44-46
Problems 2.2-2.12 (11 problems)
EE 595

Topics covered today
1. Introduction
2. Start measures of information

Introduction.

Aim of information theory

(1) Definition and measurement of information.

(2) Analysis of optimal systems that generate + reliably xmit information (communication systems).

Two-terminal communication systems.

Fixed components

[Diagram of two-terminal communication systems with components labeled: Source, Source encoder, Channel encoder, Modulation, Channel, Decoder, Source decoder, Demodulation, User.]
Source, channel, user: Fixed entities

To avoid triviality
1. Source output is random (not a priori known to user)
2. Channel is imperfect medium (i.e., channel output is not a deterministic one-to-one function of input)

Q: What does it take to reliably transmit information from the source to user over the channel?

Example

Source message: Single binary digit 0 or 1, with equal probability
Channel: Binary symmetric channel

```
0  \(\rightarrow\) 1-\(p\) \(\rightarrow\) 0
\(\rightarrow\) \(p\) \(\rightarrow\) 1
1  \(\rightarrow\) 1-\(p\) \(\rightarrow\) 1
```

\(0 < p \leq 1/2\)

\(P_r( Y = x | x = 0 ) = P_r( Y = x | x = 1 ) = 1 - p.\)
Qa: If we are allowed to use the channel n times in order to xmit the source message

(a) How do we achieve the smallest probability of error, denoted \( p_n \)?

(b) How does \( p_n \) behave as \( n \to \infty \)?

Answer (b)

<table>
<thead>
<tr>
<th>Source Output</th>
<th>Xmit</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>000000...0</td>
</tr>
<tr>
<td>1</td>
<td>01111...1 Times</td>
</tr>
</tbody>
</table>

Then decode 0 if, in received sequence

# 0's \( \geq \) # 1's.

1 if # 1's \( < \) # 0's.

(b) \( n = 1 \) \( p_1 = p \)

\( n = 2 \) \( p_2 = \frac{1}{2} \text{ Pr (receive 11 | xmit 00) } + \frac{1}{2} \text{ Pr (receive 00, 01, 10 | xmit 11) } \)

\( = \frac{1}{2} \left[ p^2 + p^2 + 2p(1-p) \right] = p^2 + p(1-p) = p = p_1 \)

\( n = 3 \) \( p_3 = p^2 (3 - 2p) < p_1 \)

\( n = 4 \) \( p_4 = p_3 \)

In general (i) \( p_n \) is a decreasing function of \( n \).

(ii) \( p_n > 0 \) for all \( n \).

(iii) \( p_n \to 0 \) as \( n \to \infty \) (Show this).
Thus, in order to have reliable transmission (Prob. error → 0) of source MS6, we need $n \to \infty$ uses of channel. In this case

\[
\frac{\text{# MS6 bits}}{\text{# channel uses}} = \frac{1}{n} \to 0 \text{ as } n \to \infty.
\]

Fundamental result of information theory:

Shannon's channel coding theorem

It is possible to have reliable transmission of messages such that the ratio

\[
\frac{\text{# of message bits}}{\text{# of channel uses}} \sim "\text{rate}" > C - \epsilon
\]

\[
\downarrow \text{arbitrary small}
\]

where $C$ is the so called capacity of the channel.

For binary symmetric channel: $C > 0$ unless $p = \frac{1}{2}$. Thus, even for noisy channels, it is possible to transmit at a rate which is non-zero.
(2) INFORMATION CONTENT OF SOURCE MESSAGE

Suppose we have a m-ary discrete time source so that output is

\[ U_1, U_2, \ldots \text{ where } U_i \in \{1, 2, \ldots, M\} \]

\[ Q_n \text{ How many bits are needed to encode the message } U_k = (U_1, U_2, \ldots, U_k) \text{ in binary form?} \]

Ans: \( M^k \) possible messages so we need \( l(k) \) bits, where

\[ 2^{l(k) - 1} < M^k \leq 2^{l(k)} \]

or

\[ l(k) - 1 < k \log_2 M \leq l(k) \]

\[ \Rightarrow \frac{l(k) - 1}{k} < \frac{\log_2 M}{k} \leq \frac{l(k)}{k} \]

As \( k \to \infty \)

\[ \frac{l(k)}{k} = \frac{\text{ LENGTH OF CODE WORD}}{\text{ LENGTH OF SOURCE MESSAGE}} \to \frac{\log_2 M}{M} \]

\[ \sim \text{ "RATE OF SOURCE CODE"} \]

\[ Q_n \text{ What if we can tolerate a small error in this source code? (i.e., some source messages are mapped onto the same codeword?)} \]
Another Fundamental Result in Information Theory

For a large class of sources (including i.i.d. ones) there is a minimum rate $H$ at which the source can be reliably encoded.

$H$: entropy (or entropy rate) of the source.

Within this class there is only one source for which $H = \log_2 M$.

For all remaining sources $H < \log_2 M$.

What if $M = \infty$?

Discrete Amplitude Sources: Depending on source distribution, it is still possible that $H < \infty$ and result is still valid.

Continuous Amplitude Sources: Quantization is widespread in digital communications. In such cases, error occurs with probability bounded away from 0. (i.e. for all non-trivial cases $P(\text{error}) = 1$)

Question: If we can tolerate a certain average level of quantization error, or distortion, what is the minimum source code (i.e. quantizer) rate required?
Rate-Distortion Theory answers above on.

Results "less" surprising than source and channel coding theorems. However! To achieve low rate, one needs to quantize the entire vector \( u^k \) (vector quantization), which brings in complexity.

Information theory establishes the performance limits of optimal codes for a broad class of com systems.

Yet practical implementation of above optimal codes is in many cases is prohibitively complex.

Note that currently active information theory research includes finding "efficient" codes.
WE WILL CONSIDER \((\Omega, \mathcal{A}, P)\)  
\(X : \Omega \rightarrow \mathcal{X}\) WHERE \(\mathcal{X}\) IS FINITE.

1. **Entropy of Random Variable** \(X\) **is defined as**
\[
H(X) \triangleq - \mathbb{E} \left[ \log_2 P(X) \right] \quad \text{Note } \log^c \text{ will represent } \left(\frac{1}{2}\right)^c \\
\triangleq - \sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x)
\]

2. **Joint Entropy of Two Random Variables**  
\(X : \Omega \rightarrow \mathcal{X}\) \(;\) \(Y : \Omega \rightarrow \mathcal{Y}\) **is defined as**
\[
H(X, Y) \triangleq \mathbb{E} \left[ - \log P_{X,Y}(X,Y) \right] \\
\triangleq - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) \log_2 P_{X,Y}(x,y)
\]

**Think of**  
\(i \{ X = x \} = - \log P_X(x) \) **as the**  
**Information contained in outcome** \(i \{ X = x \} \), **or # of bits needed to represent and convey the outcome.**

**Example. Consider a source with four outcomes**  
\(a_1, a_2, a_3, a_4\) **with probability of appearances**  
\(P(a_1), P(a_2), P(a_3)\) \(P(a_4)\) **if all are equally likely we can do a binary code**
\[a_1 - 00, a_2 - 01, a_3 - 10, a_4 - 11\]

**But if**  
\(P(a_1) = 0.99\) \(P(a_2) > P(a_3) > P(a_4)\) **we can do**  
\(a_1 - 0, a_2 - 01, a_3 - 001, a_4 - 000\)

**This way most of the time we expect to xmit 1 bit (since \(P(a_1) = 0.99\)) instead of 2.**
\( \begin{align*}
\text{IF} \quad & x^2 = x^3 \quad \Rightarrow \quad - \log P(X = x) \\
H(X) &= - \sum_{x \in X} P_X(x) \log P_X(x) \\
&= \sum_{x \in X} P_X(x) \cdot x^3 \\
&= \text{AVERAGE INFORMATION ABOUT THE SOURCE.}
\end{align*} \)

**Note that** \( \log 0 = -\infty; \log 1 = 0 \) \( \forall \text{ finite} \)

\( 0 \leq P_X(x) \leq 1 \Rightarrow 0 \leq - \sum_{x \in X} P_X(x) \log P_X(x) \leq 0 \)

\( \implies H(X) \geq 0 \quad \text{ENTROPY IS NON-NEGATIVE.} \)

**Example of a Binary Source.**

*Coin toss* \( P(H) = p \quad P(T) = 1-p \)

\( H(X) = -P(H) \log P(H) - P(T) \log P(T) \)

\( = -p \log p - (1-p) \log (1-p) \)

For binary \( H(X) = H(p) \); Function of \( p \) only

**Note** \( H(X) \) is symmetric with respect to the point \( p = \frac{1}{2} \).

\( H(X) \big|_{p=0} = H(X) \big|_{p=1} = 0 \quad H(X) \big|_{p=\frac{1}{2}} = 1 \)

On, suppose \( n \) i.i.d. coin tosses take place. What is the entropy of this event?

Ans.

Let \( X_1, \ldots, X_n \) be the \( n \) outcomes.

Tosses i.i.d. 

\( \Rightarrow P(X_1, \ldots, X_n) = P(X_1) \)

\( \Rightarrow \log P(X_1, \ldots, X_n) = \sum_{i=1}^{n} \log P(x_i) \)

\( \Rightarrow H(X_1, \ldots, X_n) = -E \left[ \log P(X_1, \ldots, X_n) \right]_n \)

\( = -E \left[ \sum_{i=1}^{n} \log P(x_i) \right] = \sum_{i=1}^{n} H(x_i) = nH(X), \)
Hence, entropy of $n$ i.i.d. events is $n$ times the entropy of one event.

Another view!

Let $p = \frac{P(H)}{1-p}$ be measured by empirical methods. $p = \frac{N(H|\omega)}{n}$

where $N(H|\omega) = \# \text{ times head appears in } n \text{ coin tosses.}$

Then $P(x_1 = x_1, \ldots, x_n = x_n) = p^N H(1-p)$

But $p = \frac{N(H|\omega)}{n} \Rightarrow N(H|\omega) = np$

$\Rightarrow P(x_1 = x_1, \ldots, x_n = x_n) = p^n H(1-p)$

$\Rightarrow P(x_1^n = x_1^n) = 2^n \left\{ np \log p + n(1-p) \log(1-p) \right\}$

$= \frac{1}{2} \left\{ -n \log H(p) - n \log H(1-p) \right\}$

$= \frac{1}{2} \left\{ -n \log H(p) + (-n \log H(1-p)) \right\}$

$= 2^{-n \log H(1-x)}$

$\Rightarrow$ i.i.d. event then a string of length $n$ has probability $2^{-n H(1-x)}$. But not all strings of length have probability $2^{-n H(x)}$. Very special ones have this property. More later:

**Conditional Entropy $H(y|x)$**

Consider $P_{xy}(x=x, y=y) = P(x=x). P_{y|x}(y=y|x=x)$

Then $H(x, y) = -E \left[ \log P_{xy}(x, y) \right]$  

$= -E \left[ \log P_x(x) \right] - E \left[ \log P_{y|x}(y|x) \right]$  

$\Rightarrow H(x, y) = H(x) + H(y|x)$. 
Q1: WHAT IS THE AVERAGE UNCERTAINTY
IN Y, GIVEN X?

Q2: WHAT IS THE AVERAGE UNCERTAINTY IN Y
GIVEN X = x?

Ans:

\[ H(Y|X=x) = -\sum_{y \in Y} P(y|x) \log P(y|x) \]

CONDITIONAL ENTROPY OF Y GIVEN X = x.

AVERAGE \( H(Y|X=x) \) OVER ALL VALUES OF X TO GET \( H(Y|X) \):

\[ H(Y|X) = \sum_{x \in X} P(x=x) H(Y|X=x) \]

\[ = -\sum_{x \in X} P(x=x, y=y) \log P(y=y|X=x) \]

\[ = -\sum_{x \in X} P(x=x) \sum_{y \in Y} P(y=x) \log P(y=x|x) \]

\[ = - E \left[ \log P(Y|X) \right] \]

NOTE IN PROVING MANY OF THESE START WITH PROBABILITY AND PROCEED.

\[ P_{xy}(x, y) = P_x \cdot P_{y|x} = P_y \cdot P_{x|y} \]

\[ \Rightarrow - E \left[ \log P_{xy} \right] = - E \left[ \log P_x P_{y|x} \right] = - E \left[ \log P_y P_{x|y} \right] \]

\[ \Rightarrow H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \]
Chain Rule for Multivariate

\[ P(x_1, x_n) = P(x_1) \cdot P(x_n | x_1) \cdot P(x_3 | x_1, x_2) \cdot \ldots \cdot P(x_n | x_{n-1}) \]

\[ \Rightarrow -E[\log P(x_n)] = -E[\log P(x_1)P(x_2 | x_1) \ldots P(x_n | x_{n-1})] \]

\[ \Rightarrow H(x_1, \ldots, x_n) = H(x_1) + H(x_2 | x_1) + \ldots + H(x_n | x_{n-1}) \]

\[ = H(x_1) + \sum_{i=2}^{n} H(x_i | x_1, \ldots, x_{i-1}) \]

Note:

1. If \( X \perp Y \) (\( X \) AND \( Y \) independent),
\[ P_{xy} = P_x P_y \Rightarrow E[-\log P_{xy}] = E[-\log P_x P_y] \]
\[ \Rightarrow H(x, y) = H(x) + H(y) \]

2. \( 0 \leq P_{y|x} \leq 1 \Rightarrow 0 \leq -\log P_{y|x} \leq \infty \)
\( x, y \) finite \( \Rightarrow 0 \leq E[-\log P_{y|x}] \leq \infty \)
\[ \Rightarrow H(x) \leq H(x) + E[-\log P_{y|x}] \leq \infty \]
\[ H(x) \leq E[-\log P_x - \log P_{y|x}] \leq \infty \]
\[ H(x) \leq E[-\log P_{x|y}] \leq \infty \]
\[ H(x)^2 \leq H(x, y) \leq \infty \]

Similarly \( H(y) \leq H(x, y) \leq \infty \)

What does \( H(x, y) \) mean? Joint entropy \( xy \)
\[ = \text{entropy of one RV + entropy of the next RV (conditioned based on the fact that 1st one was observed!} \]
Relative Entropy

Given $p(x)$ and $q(x)$ are pmfs on $X$, the relative entropy between $p$ and $q$ is defined as

$$D(p||q) = \mathbb{E}_p \left[ \log \frac{p(x)}{q(x)} \right]$$

Aside log-sum inequality

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq (\sum a_i) \log \frac{\sum a_i}{\sum b_i}$$

with equality iff $\frac{a_i}{b_i} = \frac{\sum a_i}{\sum b_i}$

Using log-sum inequality

$$D(p||q) = \mathbb{E}_p \left[ \log \frac{p(x)}{q(x)} \right] = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

$$\geq \left( \sum_{x \in X} p(x) \right) \log \left( \frac{\sum_{x \in X} p(x)}{\sum_{x \in X} q(x)} \right) = 1 \cdot \log \frac{1}{\sum_{x \in X} q(x)}$$

Note $\sum_{x \in X} p(x) = 1 = \sum_{x \in X} q(x)$

Equality iff $p = q$

$\Rightarrow D(p||q) \geq 0$

Relative entropy is non-negative.

What is this object? Expectation w.r.t. $p(x)$ of log likelihood ratio of $p$ to $q$.

Conventions

$p > 0, q = 0 \quad \log \frac{p}{0} = \infty$

$p = 0, q = 0 \quad 0 \cdot \log \frac{0}{0} = 0$

$p = 0, q > 0 \quad 0 \cdot \log \frac{0}{q} = 0$
CHAIN RULE OF RELATIVE ENTROPY

\( X, Y \) FINITE

\[
\frac{p_{xy}}{q_{xy}} = \frac{p_x}{q_x} \cdot \frac{p_{y|x}}{q_{y|x}}
\]

\[
\Rightarrow E \left[ \log \frac{p_{xy}}{q_{xy}} \right] = E \left[ \log \frac{p_x}{q_x} \cdot \frac{p_{y|x}}{q_{y|x}} \right]
\]

\[
= E_{p_{xy}} \left[ \log \frac{p_x}{q_x} \right] + E_{p_{y|x}} \left[ \log \frac{p_{y|x}}{q_{y|x}} \right]
\]

\[
D(p_{xy} \parallel q_{xy}) = D(p_x \parallel q_x) + D(p_{y|x} \parallel q_{y|x})
\]

MUTUAL INFORMATION.

\( X, Y \) FINITE RV'S \( X, Y \).

\[
I(X, Y) = I(X \land Y) = E_{p_{xy}} \left[ \log \frac{p_{xy}}{p_x p_y} \right]
\]

NOTE \( I(X \land Y) = D(p_{xy} \parallel p_x p_y) \)

\( \geq 0 \) SINCE \( D(\cdot) \geq 0 \)

HENCE MUTUAL INFORMATION BETWEEN TWO RV'S IS NON-NEGATIVE.

\[
I(X \land Y) = E \left[ \log p_{xy} - \log p_x - \log p_y \right]
\]

\[
= - E \left[ - \log p_{xy} \right] - E[\log p_x] - E[\log p_y]
\]

\[
= - H(X \land Y) + H(X) + H(Y)
\]

\[
I(X \land Y) = H(X) + H(Y) - H(X, Y)
\]

\[
= H(X) + H(Y) - \left( H(Y) + H(X|Y) \right)
\]

\[
I(X \land Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \geq 0
\]
\[ I(x; y) = H(x) - H(x|y) = \text{AVERAGE INFORMATION} \]

\[ \text{AVERAGE UNCERTAINTY OF } x \text{ PRIOR TO OBSERVING } y \]

\[ \text{AVERAGE UNCERTAINTY OF } x \text{ AFTER OBSERVING } y \]

**Example**

1. \[ I(x; y) \geq 0 \Rightarrow H(x) - H(x|y) \geq 0 \Rightarrow H(x) \geq H(x|y) \]

\[ \text{CONDITIONING REDUCES ENTROPY} \]

2. \[ \text{DATA} \quad \text{CIPHER} \]

\[ x \quad y \]

**IDEALLY OBTAINING Y SHOULD NOT REVEAL ANY INFORMATION ABOUT X.**

\[ \Rightarrow I(x; y) = 0 \quad \text{PERFECT SECRECY CONDITION OF SHANNON} \]

**Conditional Version of Mutual Information**

\[ I(x; y|z) = E_{p_{y|x,z}} [ \log \frac{p_{y|x,z}}{p_{y|x}} ] \]

\[ = E_{p_{y|x,z}} [ \log \frac{p_{x|z} \cdot p_{y|x,z}}{p_{x|z} \cdot p_{y|z}} ] \]

\[ = H(y|z) - H(y|x,z) \]

\[ = H(x|z) - H(x|y,z) \]

\[ I(x; y|z) = H(y|z) - H(y|x,z) = H(x|z) - H(x|y,z) \]

**Uncertainty in y given z (but not x)**

**Uncertainty in y given x and z**
\[ I (x_1 \ldots x_n \land y) = H (x_1 \ldots x_n) - H (x_1 \ldots x_n \mid y) \]
\[ = H(x_1) + \sum_{i=2}^{n} H(x_i \mid x_i^{i-1}) \]
\[ - H(x_1 \mid y) + \sum_{i=2}^{n} H(x_i \mid x_i^{i-1}, y) \]
\[ = H(x_1) - H(x_1 \mid y) + \sum_{i=2}^{n} H(x_i \mid x_i^{i-1}) - H(x_i \mid x_i^{i-1}, y) \]
\[ \Rightarrow I (x_1 \land y) + \sum_{i=2}^{n} I (x_i \land y \mid x_i^{i-1}) \]

**Examples**

\[ p = 1 - \frac{H(x_2 \mid x_1)}{H(x_1)} = \frac{H(x_1) - H(x_2 \mid x_1)}{H(x_2)} \]
\[ = \frac{H(x_2) - H(x_2 \mid x_1)}{H(x_1)} = \frac{I(x_1 \land x_2)}{H(x_1)} = \frac{I(x_1 \land x_2)}{H(x_2)} \]
\[ H(x_1) = H(x_2) \geq 0 \quad I(x_1 \land x_2) \geq 0 \]

\[ \Rightarrow p \geq 0 \quad \text{and} \quad p = I(x_1 \land x_2) = 1 - \frac{H(x_2 \mid x_1)}{H(x_2)} \]

**Note**

\[ I(x_1 \land x_2) = H(x_2) - H(x_2 \mid x_1) \geq 0 \]
\[ \Rightarrow \frac{H(x_2 \mid x_1)}{H(x_2)} \geq 1 \quad \Rightarrow 1 - \frac{H(x_2 \mid x_1)}{H(x_2)} \geq 0 \]

**Indeed**

\[ H(x_2) \geq H(x_2 \mid x_1) \Rightarrow \frac{I(x_1 \land x_2)}{H(x_2)} \leq \frac{H(x_2)}{H(x_2)} \]
\[ \Rightarrow 0 \leq p \leq 1 \]

\[ \therefore \text{using } I(x_1 \land y) \geq 0 \text{ and Chain Rule,} \]
\[ \text{show } H (x_1 \land y) \leq \sum_{i=1}^{n} H (x_i) \text{ with equality iff } x_i \text{ are independent!} \]
Lemma. Let \( i = 1, \ldots, n \) and \( b_i \geq 1, \ldots, n \) be arbitrary non-negative numbers. Then

\[
\sum_{i=1}^{n} \frac{a_i}{b_i} \log \frac{a_i}{b_i} \geq \sum_{i=1}^{n} a_i \log \left( \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \right)
\]

with equality iff \( \frac{a_i}{b_i} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \) is constant.

Consider \( y = \log t \)

Pick \( t = t' > 0 \).

Tangent to \( y(t) \) at \( t = t' \) is of the form

\[
y = mt + c
\]

\[
m = \frac{dy}{dt} \bigg|_{t=t'} = \frac{1}{t'}
\]

\[
y = \frac{t}{t'} + c
\]

At \( t = t' \), \( y = \log t' \)

\[
c = \log t' - \frac{t}{t'} = \log t' - 1
\]

Tangent is \( y = \frac{t-t'}{t'} + \log t' \)

Tangent is above the curve \( y(t) = \log t \)
\[ \ln t \leq \frac{k - t'}{t'} + \ln t' \quad \text{lead to} \]

\[ \ln \frac{b_i}{b'} \leq \left[ \frac{b_i/a_i - \frac{\sum b_i}{\sum a_i}}{\frac{\sum b_i}{\sum a_i}} \right] + \ln \left( \frac{\sum b_i}{\sum a_i} \right) \]

Multiply both sides by \( a_i \) and sum over \( i \):

\[ \sum_{i=1}^{r} a_i \ln \frac{b_i}{b'} \leq \left( \frac{\sum a_i}{\sum b_i} \right) \left( \sum b_i - \sum b_i \right) \]

\[ + \sum_{i=1}^{r} a_i \ln \left( \frac{\sum b_i}{\sum a_i} \right) \]

\[ \Rightarrow \]

\[ \sum_{i=1}^{r} a_i \ln \frac{b_i}{b'} \leq \left( \frac{\sum a_i}{\sum b_i} \right) \ln \left( \frac{\sum b_i}{\sum a_i} \right) \]

\[ \Rightarrow \]

\[ \sum_{i=1}^{r} a_i \ln \left( \frac{b_i}{b'} \right) \geq \ln \left( \frac{\sum a_i}{\sum b_i} \right) \ln \left( \frac{\sum b_i}{\sum a_i} \right) \]

Equality when the transport function \( \ln \left( \frac{b_i}{b'} \right) \) is

\[ \text{Converse, or } t=t' \Rightarrow \quad \frac{a_i}{b_i} = \frac{\sum a_i}{\sum b_i} \]
LAST CLASS

1. **Entropy** $H(x) = E[- \log P(x)]$
2. **Conditional Entropy** $H(y|x) = E[- \log P(y|x)]$
3. **Differential Entropy** $D(P||Q) = E[- \log \frac{P}{Q}]$
4. **Mutual Information** $I(x; y) = I(x \wedge y) = E[- \log \frac{P(x, y)}{P(x)P(y)}]$
5. **Chain Rule of Entropy** $H(x^n) = H(x_1) + \sum_{i=2}^{n} H(x_i | x_{i-1})$
6. **Chain Rule of Differential Entropy** $D(P_{xy} || Q_{xy}) = D(P_x || Q_x) + D(P_{y|x} || Q_{y|x})$
7. **Chain Rule of Mutual Information** 
   $I(x^n; y) = I(x_1; y) + \sum_{i=2}^{n} I(x_i; y|x_1^n)$
8. Using Log Sum Inequality, 
   \[
   \sum a_i \log \frac{e^{a_i}}{b_i} \geq \sum a_i \log \left( \frac{e^{a_i}}{\sum b_i} \right)
   \]
   And setting $a_i = p(x)$, $b_i = q(x)$, \[\sum p(x) \log \frac{p(x)}{q(x)} = \sum p(x) \log \left( \frac{e^{p(x)}}{\sum q(x)} \right) \geq 0 \]
   $D(P_{xy} || Q_{xy}) \geq 0$ Note $D(\mathcal{Q}_{xy}) = D(P_{xy} || Q_{xy})$ Problem 2.35
9. $I(x; y) = E[- \log \frac{P_{xy}}{P_xP_y}] = D(P_{xy} || P_xP_y) \geq 0$
   Using $\log \frac{P_{xy}}{P_xP_y} = \log \frac{P_{xy}}{P_y} \Rightarrow I(x; y) = H(x) - H(x|y) \geq 0$
   Conditioning reduces entropy on average!
10. $H(x|y) \leq H(x)$
11. $I(x; x) = H(x) - H(x|x) = H(x) - 0 = H(x)$
12. $x \wedge y \Rightarrow I(x; y) = H(x) - H(x|y) = H(x) - H(x) = 0$

TODAY

1. **Examples.**
2. **Data Processing Inequality** $X \rightarrow Y \rightarrow Z$ (Markov)
   \[I(x; y) \geq I(x; z) \text{ and if } Z = g(Y) \]
   This implies processing cannot increase information!
3. **Fano's Inequality**: Given RV's $X, Y$ and $X = g(Y)$
   \[H(x|y) \leq H(x|y) \leq H(P_e) + P_e \log (\frac{1}{P_e} - 1); P_e = P_x \phi \phi \text{.} \]
4. **Jensen and Log Sum**
(4) A STOCHASTIC PROCESS $\{X_i\}_{i=1}^\infty$ is STATIONARY IF
$$P(X_1^n) = P(X_{i+1}^{n+i})$$
(EX: i.i.d. Process)
$$P(X_i) = P(X_1)$$
$$H(X_1^n) = H(X_{i+1}^{n+i})$$

(5) A STOCHASTIC PROCESS IS MARKOV IF
$$P(X_{i+1} | X_i^n) = P(X_{i+1} | X_i^n)$$
$$P(X_i^n) = \prod_{i=2}^{n} P(X_i | X_{i-1}^{i-2}) = P(X_1) \prod_{i=2}^{n} P(X_i | X_{i-1})$$

(C) IF $X_1 \rightarrow X_2 \rightarrow X_3$ OR $X_1 \rightarrow X_2 \rightarrow X_3$ (INDICATION / NOTATION FOR MARKOV)
$$P_{X_1X_2X_3} = \frac{P(x_1x_2x_3)}{P(x_2x_3)} = \frac{P(x_1x_2) P(x_3 | x_1x_2)}{P(x_2) P(x_3 | x_2)} = \frac{P(x_1x_2) P(x_3 | x_2)}{P(x_2) P(x_3 | x_2)} = P_{X_1X_2}

\Rightarrow P_{X_1X_2X_3} = P_{X_3} P_{X_1 | X_3} P_{X_2 | X_3}

IF $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \cdots \rightarrow X_n$ \iff $X_n \rightarrow X_{n-1} \rightarrow \cdots \rightarrow X_2 \rightarrow X_1$.

(d) A MARKOV CHAIN IS TIME IN Variant IF $P(X_{n+1} = j | X_n = i) = P_{ij}$

INDEPENDENT OF $n$. WE CAN USE MARKOV CHAINS TO COMPUTE $P(X_{n+1} = j | X_t = i) = P_{ij}$

IF WE KNOW PMF AT TIME "n" WE CAN COMPUTE PMF AT TIME "n+1" AS
$$P(X_{n+1} = x_{n+1}) = \sum_{x_n \in X} P(X_n = x_n) P(X_{n+1} = x_{n+1} | X_n = x_n)$$
$$= \sum_{x_n \in X} P(X_n = x_n) P_{x_n x_{n+1}}$$

O.NE-STEP TRANSITION MATRIX ELEMENTS.

PROBABILITY DISTRIBUTION IS SAID TO BE STATIONARY IF $\sum_{x_n \in X} P(X_n = x_n) = 1$ WHERE $\sum_{x_n \in X} = 1$
DATA PROCESSING INEQUALITY

PROCESSING CANNOT INCREASE THE INFORMATION

CONSIDER OUTCOME/RV X. ASSUME Y IS OBSERVED.

LET \( \hat{X} \) BE THE ESTIMATE OF X BASED ON Y.

THEN \( X \rightarrow Y \rightarrow \hat{X} \) IF X, Y ARE NOT INDEPENDENT.

ALSO NOTE IF \( X \rightarrow Y \rightarrow Z \) THEN \( P_{Z|X,Y} = P_{Z|Y} \cdot P_{Y|X} \)

DPINES.

1. \( I(X, Y|Z) = 0 \)
2. \( I(X \wedge Y) \geq I(X \wedge Z) \) AS U GO FAR IN CHAIN CORR DECREASES.
3. \( I(X \wedge Y) \geq I(X \wedge Y|Z) \)

PROS

1. FOLLOWS FROM MARKOV CHAIN PROPERTY.

\[ P_{Z|X,Y} = P_{Y|Z} \cdot P_{Z|Y} \]

\[ I(X, Y|Z) = E \left[ \log \frac{P_{X,Y|Z}}{P_{X|Y} \cdot P_{Y|Z}} \right] = E \left[ \log \frac{P_{Y|Z} \cdot P_{Z|Y}}{P_{Z|Y} \cdot P_{Y|Z}} \right] \]

\[ = E \left[ \log 1 \right] = 0 \]

\( \Rightarrow X \rightarrow Y \rightarrow Z \Rightarrow I(X, Y|Z) = 0 \) X AND Z ARE INDEPENDENT CONDITIONED ON Y. NO INFO IS REVEALED ON Y GIVES ALL THAT CAN BE OBTAINED.

2. \( I(X \wedge Y, Z) = E \left[ \log \frac{P_{X,Y|Z}}{P_{X|Y} \cdot P_{Z|Y}} \right] = E \left[ \log \frac{P_{X,Z|Y}}{P_{X|Z} \cdot P_{Y|Z}} \right] \)

\( (X \rightarrow Y \rightarrow Z \Leftrightarrow Z \rightarrow Y \rightarrow X) \Rightarrow P_{X|Y,Z} = P_{X|Y} \cdot P_{Y|Z} \)

\( \Rightarrow I(X \wedge Y, Z) = E \left[ \log \frac{P_{X,Y}}{P_{X} \cdot P_{Y}} \right] = H(X) - H(X \wedge Y) \)

\( = I(X \wedge Y) - (i) \)

BUT \( I(X \wedge Y, Z) = E \left[ \log \frac{P_{X,Y,Z}}{P_{X,Z} \cdot P_{Y|Z}} \right] = E \left[ \log \frac{P_{X,Z|Y}}{P_{X,Z} \cdot P_{Y|Z}} \right] \)

\[ = E \left[ \log \frac{P_{X,Z} + P_{Y|Z}}{P_{X,Z} \cdot P_{Y|Z}} \right] = I(X \wedge Z) + I(X \wedge Y|Z) \]

\( \Rightarrow I(X \wedge Y, Z) = I(X \wedge Y) = I(X \wedge Z) + I(X \wedge Y|Z) \)

Note \( I(X \wedge Y) \geq 0 \) \( I(X \wedge Z) \geq 0 \) \( I(X \wedge Y|Z) \geq 0 \)

\( \Rightarrow I(X \wedge Y) \geq I(X \wedge Z) \) CLOSER => MORE CORRELATED.

\( I(X \wedge Y) \geq I(X \wedge Y|Z) \) KNOWING Z REDUCES WHAT Y REVEALS ABOUT X.
**More Data Processing Inequalities.**

**Lemma** If \( x_1 \circ x_2 \circ x_3 \circ x_4 \), THEN

(i) \( I(x_1 \land x_2) \geq I(x_1 \land x_3) \) \( \Rightarrow \) already know this!

(ii) \( I(x_1 \land x_4) \leq I(x_2 \land x_3) \)

\[
I(x_1 \land x_4) \leq I(x_1, x_2 \land x_4)
\]

**Observation** Incrases Mutual Info.

But \( I(x_1, x_2 \land x_4) = I(x_2 \land x_4) + I(x_1 \land x_4 | x_2), x_1 \circ x_2 \circ x_4 \)

\[
= I(x_2 \land x_4) \quad \text{[Conditioned on } x_2]\]

\[
P_{x_1 x_4 | x_2} = P_{x_1 x_2} P_{x_4 | x_2}
\]

\[
I(x_1 \land x_2 \land x_4) = I(x_1 \land x_4) + I(x_2 \land x_4 | x_1, x_2, x_4, P_{x_1 x_4 | x_2} = P_{x_1 x_2} P_{x_4 | x_2}
\]

\[
I(x_2 \land x_4 | x_1) \geq 0 \Rightarrow I(x_1 \land x_4) + I(x_2 \land x_4 | x_1) = I(x_2 \land x_4)
\]

 Leads to \( I(x_1 \land x_4) \leq I(x_2 \land x_4) \)

Similarly \( x_2 \circ x_3 \circ x_4 \) provides \( I(x_2 \land x_4) \leq I(x_2 \land x_3) \)

\[
\Rightarrow I(x_1 \land x_4) \leq I(x_2 \land x_4) \leq I(x_2 \land x_3)
\]

**Key Point in this Proof is that if**

\[
x_1 \circ x_2 \circ x_3 \cdots \circ x_n \cdots \circ x_i \cdots
\]

Then \( x_1 \circ x_i \circ x_n \) or \( x_1 \cdots x_j \circ x_n \).

**Any sub-sequence is Markov!**
Fano's Inequality

Bounding of Error

In Estimates.

Let \( X \) be RV and \( Y \) be an observed RV that is correlated to \( X \). Let \( \hat{X} \) be the estimate of \( X \). Denote by \( P_e \equiv P(Y \neq \hat{X}) \). Then

Fano's Inequality states

\[
0 \leq H(X|Y) \leq H(X|\hat{X}) \leq H(P_e) + P_e \log \left( \frac{1}{2P_e - 1} \right).
\]

Proof requires an indicator RV \( E \).

Let \( E = \begin{cases} 1 & X \neq \hat{X} \quad P(E=1) = P\{X \neq \hat{X}\} = P_e \\ 0 & X = \hat{X} \quad P(E=0) = P\{X = \hat{X}\} \end{cases} \)

\( P_e, 1 - P_e \) indicate \( E \) is binary RV.

\[
H(E) = H(P_e) = -P_e \log P_e - (1 - P_e) \log (1 - P_e).
\]

Note if we know \( X, \hat{X} \), we know \( E \).

If we know \( X, E \) and \( E \neq 0 \) we only know \( X \in \{\hat{X}, \bar{X}\} \).

Proof.

\[
\begin{align*}
H(E, X|\hat{X}) &= H(E|\hat{X}) + H(X|\hat{X}, E) \\
&= H(X|\hat{X}) + H(E|X, \hat{X})
\end{align*}
\]

\( \Rightarrow H(X|\hat{X}) = H(E|\hat{X}) + H(X|\hat{X}, E) - \log P_e \)

\( I(E;X) \geq 0 \Rightarrow H(E) \geq H(E|\hat{X}) \)

\[
\begin{align*}
H(X|\hat{X}, E) &= P(\hat{X} = \bar{X}|X = X) H(X|\hat{X} = \bar{X}) + P(\hat{X} = \hat{X}|X = X) H(X|\hat{X} = \hat{X}) \\
&= P_e H(X|\hat{X} = \bar{X}) \\
\Rightarrow I(X; E) &= H(X) - H(X|\hat{X}) \geq 0 \Rightarrow H(X) \geq H(X|\hat{X})
\end{align*}
\]

\( \Rightarrow H(X|\hat{X}) = H(E|\hat{X}) + H(X|\hat{X}, E) \leq H(E) + P_e \log \left( \frac{1}{2P_e - 1} \right) \)

or sharply

\( H(X|\hat{X}) \leq H(E) + P_e \log \left( \frac{1}{2P_e - 1} \right) \)

Note \( H(E) = H(P_e) \) and \( H(P_e) \leq 1 \)
HENCE
\[ H(x|\hat{y}) \leq 1 + P_e \log (|x| - 1) \]

Also note from \( X \rightarrow Y \rightarrow \hat{X} \Rightarrow I(X \land Y) \geq I(X \land \hat{X}) \)
\[ I(X \land \hat{X}) \leq I(X \land Y) \]
\[ \Rightarrow H(x) - H(x|\hat{X}) \leq H(x) - H(x|Y) \]
\[ \Rightarrow H(x|Y) \leq H(x|\hat{X}) \]

\[ \Rightarrow H(x|Y) \leq H(x|\hat{X}) \leq H(x) + P_e H(x|\hat{X}, E=1) \]

OR \[ H(x|Y) \leq H(x|\hat{X}) \leq 1 + P_e \log (|x| - 1) \]

A SPECIAL CASE

Let \( X, Y \) be i.i.d. with Pmf \( p(x) \)

\[ P\{X = Y\} = \sum_{x \in X} P\{X = x = Y\} \]
\[ = \sum_{x \in X} P\{X = x\} \cdot P\{Y = x\} \]
\[ = \sum_{x \in X} p(x) \]

Note this is a convex function!

\[ P\{X = Y\} = \sum_{x \in X} p(x) = \sum_{x \in X} p(x) \cdot \log p(x) \]

Jensen's inequality \( E(f(X)) \geq f(E(X)) \)

\[ \Rightarrow P\{X = Y\} = \sum_{x \in X} p(x) = E \left[ \sum_{x \in X} \log p(x) \right] \geq 2 \cdot \frac{E[\log p(x)]}{2} \]

\[ H(x) = -E[\log p(x)] \]
\[ \Rightarrow P\{X = Y\} \geq 2^{-H(x)} \]
Def. \( f(x) \) is said to be convex over interval \((a, b)\) if \( \forall x_1, x_2 \in (a, b) \Delta 0 \leq \lambda \leq 1 \)

\[
\frac{\lambda f(x_1) + (1-\lambda) f(x_2)}{= \lambda f(x_1) + (1-\lambda) f(x_2)}
\]

Strict if equality with \( \lambda = 0/1 \) only. \( f(x) = x^2 \)

\[
x_1 = 1 \quad x_2 = -1
\]

\[
f(1) = 1, \quad f(-1) = 1
\]

\[
A = \frac{1}{2} \quad \frac{1}{2} x_1 + \frac{1}{2} x_2 = 0 \quad f(0) = 0
\]

\[
\frac{f(\lambda x_1 + (1-\lambda) x_2)}{= f\left(\frac{1}{2} x_1 + \frac{1}{2}(-1)\right) = f(0) = 0}
\]

\[
\frac{1}{2} f(1) + \frac{1}{2} f(-1) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1
\]

\[
f(0) < \frac{1}{2} f(1) + \frac{1}{2} f(-1)
\]

If \( f(x) \) is convex \( -f(x) \) is concave

\[
y = mx + c
\]

\[
x_1, x_2
\]

\[
y_1 = mx_1 + c
\]

\[
y_2 = mx_2 + c
\]

\[
\frac{\lambda y_1 + (1-\lambda) y_2}{= \lambda (mx_1 + c) + (1-\lambda)(mx_2 + c) = \lambda mx_1 + (1-\lambda)mx_2 + c}
\]

\[
= \frac{\lambda f(x_1) + (1-\lambda) f(x_2)}{= \lambda f(x_1) + (1-\lambda) f(x_2)}
\]

\[
y_3 = \frac{\lambda}{2} f\left(\lambda x_1 + (1-\lambda) x_2\right) = \frac{\lambda}{2} f(x_1) + (1-\lambda) f(x_2)
\]

\[
= \text{line is convex! with equality!}
\]

Note then \( \text{line is concave! :c} \)

Which is line too!
JENSEN'S INEQUALITY

**THEOREM** IF $f$ IS CONVEX FUNCTION AND $X$ IS A RANDOM VARIABLE

$$E[f(X)] \geq f(E[X])$$

EXPECTED VALUE OF FUNCTION OF RV $\geq$ FUNCTION OF EXPECTED VALUE

IF $f$ IS STRICLY CONVEX, EQUALITY IF $X = E[X]$ W.P. 1 OR $X$ IS A CONSTANT.

**PROOF BY INDUCTION.** FOR PMF

**LET** $p_1$, $p_2$ BE $p_1 + p_2 = 1$

$\Rightarrow p_1 \cdot f(x_1) + p_2 \cdot f(x_2) \geq f(p_1 x_1 + p_2 x_2)$

*Assume it is true for index $i$ up to $k-1$.

$$\sum_{i=1}^{k-1} p_i \cdot f(x_i) \geq f\left(\sum_{i=1}^{k-1} p_i \cdot x_i\right)$$

**LET** $k = n$ **i.e.** $p_1, \ldots, p_{k-1}, p_k$ ARE NEW COMPONENTS

$$\sum_{i=1}^{k} p_i \cdot f(x_i) = p_k \cdot f(x_k) + \sum_{i=1}^{k-1} p_i \cdot f(x_i)$$

**Note** we only know $\sum_{i=1}^{k-1} p_i \cdot f(x_i) \geq f\left(\sum_{i=1}^{k-1} p_i \cdot x_i\right)$

**So set** $p_i = \frac{p_i}{1-p_k}$

$$\Rightarrow \sum_{i=1}^{k} p_i \cdot f(x_i) = p_k \cdot f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p_i \cdot f(x_i)$$

**Induction works up to $k-1$**

**Convexity of $f(*)$**

$$f\left(\sum_{i=1}^{k-1} p_i \cdot x_i\right) \geq f\left(p_k \cdot x_k + (1-p_k) \sum_{i=1}^{k-1} p_i \cdot x_i\right)$$

$$= f\left(\sum_{i=1}^{k} p_i \cdot x_i\right)$$
EE514  WINTER 2008  CLASS #4

LAST CLASS

1. DATA PROCESSING INEQ: (PROCESSING DOES NOT INCREASE INFORMATION) \( X \) (RV); \( Y \) (OBSERVED RV); \( \hat{X} \) (ESTIMATED RX).

Or in general if \( X \rightarrow Y \rightarrow Z \) (Markov), then

\[ P_{Z|Y} = P_{Z|Y} \cdot P_{Y} \cdot P_{X|Y} \cdot P_{X} \] (X AND Z CONDITIONALLY INDEPENDENT GIVEN Y).

a) \( I(X;Y) \geq I(X;Z) \) --- CLOSER VARIABLES IN CHAIN MORE CORRELATED.

b) \( I(X;Y) \geq I(X;Y|Z) \) --- CONDITIONING REDUCES MUTUAL INQA.

c) \( I(X;Z|Y) = 0 \) \[ P_{Z|Y} = P_{XY} \cdot P_{Y} \Rightarrow E \left[ \log P_{Z|Y} \right] = E \left[ \log I \right] \] CAN EXTEND TO MULTIVARIATE CASE.

2. FANO'S INEQUALITY \( X \) (RV); \( Y \) (OBSERVED); \( \hat{X} \) (ESTIMATE RX).

\( P_e = P_{Y \neq \hat{X}} \). BY DEFINING \( E = \begin{cases} 1 & \text{if } X = \hat{X} \\ 0 & \text{if } X \neq \hat{X} \end{cases} \)

3. \( H(X|Y) \leq H(X|\hat{X}) \) DERIVE THIS BY \( I(X;Y|\hat{X}) \geq I(X;Y) \) NOT SURPRISING!

\( H(X|Y) \leq H(X|\hat{X}) \leq H(X) - H(\hat{X}) \leq H(X) \)

4. FANO'S RESULT \( H(X|Y) \leq H(X|\hat{X}) \leq H(X) + P_e \log (1|X|^{-1}) \)

WHERE \( X \in \mathbb{X}, H(\hat{X}) \leq 1 \) BINARY ENTROPY SINGE \( P(0|1) = P_e \).

5. JESEN'S LOG SUM (SELF STUDY)

TODAY: ASYMPTOTIC EQUIPARTITION PROPERTY (AEP) CHAP 3

MAIN IDEA: RECALL IF \( \{X_i\}_{i=1}^{n} \) ARE I.I.D. THEN WEAK LAW OF LARGE # STATES THAT

\[ \frac{1}{n} \sum_{i=1}^{n} X_i \rightarrow E[X] \]

AEP STATES: IF \( \{X_i\}_{i=1}^{n} \) I.I.D. THEN \( \frac{1}{n} \log P(X_i^n) \rightarrow H(X) \)

WHERE \( P(X_i^n) \) IS THE PROBABILITY OF OBSERVING SEQ \( X_i \ldots X_n \)

SET OF SEQUENCES FOR WHICH AEP HOLD ARE CALLED TYPICAL SEQUENCES. A TYPICAL SEQUENCE HAS PROBABILITY \( 2^{-nH(X)} \) WHERE \( n \) IS THE SEQUENCE LENGTH.
RECALL A SEQUENCE $X_1 \ldots X_n$ OF RV'S CONVERGE TO RV $X$ IN PROBABILITY IF $\varepsilon > 0 \quad \Pr \{ |X_n - X| < \varepsilon \} \to 0$

**THEOREM** STATES IF $X_1 \ldots X_n$ i.i.d. w.p. $P(X)$

THEN

\[-\frac{1}{n} \log P(X_1 \ldots X_n) \to H(X) \text{ IN PROBABILITY} \]

**PF:**

$X_i$'s INDEP $\Rightarrow$ $P(X_1 \ldots X_n) = \prod_{i=1}^{n} P(X_i) \quad x_i \sim P \Rightarrow P(x_i) = p_{x_i}$

$\Rightarrow -\frac{1}{n} \log P(X_1 \ldots X_n) = -\frac{1}{n} \sum_{i=1}^{n} \log P(X_i)$

$X_i$'s INDEP $\Rightarrow \log P(X_i)$'s INDEP. HENCE BY WEAK LAW OF LARGE NUMBERS.

\[-\frac{1}{n} \log P(X_1^n) = -\frac{1}{n} \sum_{i=1}^{n} \log P(X_i) \to E \left[ -\log P(X) \right] \]

$= H(X) \text{ IN PROBABILITY.}$

RECALL WLLN $X_1 \ldots X_n$ i.i.d. seq of RV. with MEAN $E[X_i] = \mu$ AND VARIANCE $E[(X_i - \mu)^2] = \sigma^2$. THEN

SAMPLE MEAN, DENOTED BY $\frac{1}{n} \sum_{i=1}^{n} X_i$ SATISFIES

\[\Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right\} \leq \frac{\sigma^2}{\varepsilon^2} \xrightarrow{\varepsilon \to 0} \frac{\sigma^2}{\varepsilon^2} \xrightarrow{n \to \infty} 0 \]

OR $n \to \infty$ \[\Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right\} \xrightarrow{\varepsilon \to 0} 0 \]

THE MOST EXCITING PART IS WE CAN SAY A LOT ABOUT THE SEQUENCE $X_1 \ldots X_n$ OR SUCH "$n$" TUPLES.

LET $A_\varepsilon^n \equiv \left\{ x^n \in X^n : \left| -\frac{1}{n} \log P(X^n) - H(X) \right| < \varepsilon \right\}$

THIS SET $A_\varepsilon^n$ IS CALLED TYPICAL SET.
Properties of Typical Set $A_x^{(n)}$

1. If $x_i^n \in A_x^{(n)}$ then $H(x) - \varepsilon \leq \frac{-1}{n} \log P(x_i^n) \leq H(x) + \varepsilon$

2. If $x_i^n \in A_x^{(n)}$ then $2^{-n(H(x)+\varepsilon)} \leq P(x_i^n) \leq 2^{-n(H(x)-\varepsilon)}$

3. $\Pr \left\{ \frac{1}{n} \sum P(x_i^n) \neq H(x) \right\} > 1 - \delta$

4. $\left| A_x^{(n)} \right| \leq 2^{-n(H(x)+\varepsilon)}$

5. $\left| A_x^{(n)} \right| \geq (1 - \varepsilon) 2^{-n(H(x)-\varepsilon)}$

Proofs

1 and 2 follow by definitions. If $x_i^n \in A_x^{(n)}$ then

$$\left| \frac{-1}{n} \log P(x_i^n) - H(x) \right| \leq \varepsilon$$

$\Rightarrow H(x) - \varepsilon \leq \frac{-1}{n} \log P(x_i^n) \leq H(x) + \varepsilon \quad \{\text{first one}\}$

$\Rightarrow -n(H(x)+\varepsilon) \leq \log P(x_i^n) \leq -n(H(x)-\varepsilon)$

$\Rightarrow 2^{-n(H(x)+\varepsilon)} \leq P(x_i^n) \leq 2^{-n(H(x)-\varepsilon)} \quad \{\text{2nd one}\}$

To get 3rd.

Note that by WLLN $\Pr \left\{ \left| \frac{-1}{n} \log P(x_i^n) - H(x) \right| < \varepsilon \right\} > 1 - \delta$

where $\delta > 0$ is chosen. Then $n_o \geq 1/n > n_o$ above will be computed. Setting $\delta = \varepsilon \geq \delta$ some small constant.

$\Rightarrow \Pr \left\{ \left| \frac{-1}{n} \log P(x_i^n) - H(x) \right| < \varepsilon \right\} > 1 - \varepsilon$

To obtain 4 use the idea $1 = \sum P(x_i^n) = \sum P(x_i^n)_{x_i^n \in X^n, x_i^n \in A_x^{(n)}}$

$\geq \sum 2^{-n(H(x)+\varepsilon)}$

$\Rightarrow \left| A_x^{(n)} \right| \leq 2^{n(H(x)+\varepsilon)}$
Since
\[ \Pr \{ A_\varepsilon^{(n)} \} > 1 - \varepsilon \]
\[ 1 - \varepsilon \leq \Pr \{ x \in A_\varepsilon^{(n)} \} \]
\[ = \sum_{x^n \in A_\varepsilon^{(n)}} 2^{-n(\mathcal{H}(x) - \varepsilon)} = \frac{|A_\varepsilon^{(n)}|}{2} \]
\[ \Rightarrow |A_\varepsilon^{(n)}| \leq 2 \cdot 2^{-n(\mathcal{H}(x) - \varepsilon)} \]

How big is \( A_\varepsilon^{(n)} \)?
\[ \frac{|A_\varepsilon^{(n)}|}{|x|^n} \leq 2 \cdot 2^{-n(\mathcal{H}(x) + \varepsilon - \log |x|)} \]

As \( \varepsilon \to 0 \), if \( \Pr(x) = \frac{1}{|x|} \) (then \( \mathcal{H}(x) < \log |x| \))
Then \( \frac{|A_\varepsilon^{(n)}|}{|x|^n} \to 0 \)

The set \( A_\varepsilon^{(n)} \) is small as \( n \to \infty \). Compared to the set \( x^n \).

1st idea of data compression.

We can on average code \( x^n \) with \( n \mathcal{H}(x) \) bits.

There are \( 2^{n(\mathcal{H}(x) + \varepsilon)} \) elements in \( A_\varepsilon^{(n)} \). We need \( \log |A_\varepsilon^{(n)}| \) or
\[ < n (\mathcal{H}(x) + \varepsilon) + 1 \text{ bits} \]

to uniquely represent each element.

Assuming \( |A_\varepsilon^{(n)}| < |x|^n \), we need \( \log |x|^n \) or
\[ < n \log |x| + 1 \text{ bits to represent elements of } x^n \setminus A_\varepsilon^{(n)} \].

Note we over-count here since we only need \( |x|^n - |A_\varepsilon^{(n)}| + 1 \) at most. But \( |x|^n \gg |A_\varepsilon^{(n)}| \).

So we need \( n (\mathcal{H}(x) + \varepsilon + \log |x| + 1) + 2 \varepsilon \) bits at most to uniquely code \( A_\varepsilon^{(n)} \) and the rest.
Given \( x^n \in A(e^n) \) or \( x^n \in X^n \setminus A(e^n) \),

**How to Differentiate And Uniquely (code/decode)?**

Ans: Add a **Prefix Bit**: \( x^n_i \in A(e^n) \rightarrow 0 \) **Prefix** \( x^n_i \in X^n \setminus A(e^n) \rightarrow 1 \) **Bits Prefix**.

Hence \( x^n_i \in A(e^n) \) requires \( n(\mathbb{H}(x) + \epsilon) + 2 \) **Bits**.
\( x^n_i \in X^n \setminus A(e^n) \) requires \( n(\epsilon g|x|) + 2 \) **Bits**.

**Q:** However, **What is the Average # of Bits Needed to Code** \( x^n \in X^n \) (or an RV \( x^n \))?

Ans: Let \( l(x^n) \) denote **Code Word Length**

\[
E[l(x^n)] = \sum_{x^n} p(x^n) l(x^n)
\]

\[
= \sum_{x^n_i \in A(e^n)} p(x^n) l(x^n) + \sum_{x^n_i \in X^n \setminus A(e^n)} p(x^n) l(x^n)
\]

\[
\leq \sum_{x^n_i \in A(e^n)} p(x^n) (n \mathbb{H}(x) + n \epsilon + 2) + \sum_{x^n_i \in X^n \setminus A(e^n)} p(x^n) (n \epsilon g|x| + 2) \leq \epsilon
\]

**Aside**

\[
P(A(e^n)) + P(X^n \setminus A(e^n)) = 1
\]

\[
\leq n \mathbb{H}(x) + n(\epsilon + 2/n + \epsilon g|x|)
\]

\[
= n(\mathbb{H}(x) + \epsilon') \quad \epsilon' = \epsilon + 2/n + \epsilon g|x|
\]

\[
\Rightarrow \frac{1}{n} E[l(x^n)] \leq \mathbb{H}(x) + \epsilon'
\]

\[
\frac{2}{n} P(A(e^n)) + \frac{2}{n} P(X^n \setminus A(e^n)) = \frac{2}{n}
\]

Hence each RV \( x \) **Requires (on Average)** \( \mathbb{H}(x) \) **Bits**

To **Code It Uniquely!**

A seq \( X^n \) of i.i.d. RVs requires \( n \mathbb{H}(x) \) **Bits**

**On Average to Code!**
WE WANT TO SHOW THAT $A_\epsilon^{(n)}$ IS OF SAME SIZE AS A SET $B_\delta^{(n)} \subseteq \mathcal{X}^n$ WITH $P \{ B_\delta^{(n)} \} \geq 1-\delta$ FOR $n \geq 1$.

A NOTATION $a_n = b_n \iff \lim_{n \to \infty} \frac{\log a_n}{b_n} = 0$ OR $a_n \approx b_n \implies a_n$ AND $b_n$ ARE EQUAL IN FIRST ORDER OF POWER W.R.T A BASE (HERE 2)

THM $X_1, \ldots, X_n$ i.i.d. $\sim P(x)$ FOR $\delta < \frac{1}{2}$ AND ANY $\delta' > 0$ IF $P \{ B_\delta^{(n)} \} > 1-\delta$ THEN,

$$\frac{1}{n} \log |B_\delta^{(n)}| > H - \delta'. \quad \text{FOR} \quad n \uparrow$$

OR

$$|B_\delta^{(n)}| \equiv |A_\epsilon^{(n)}| = 2^{nH(x)} \quad (\text{BECAUSE BASE IS 2})$$

PROOF PROBLEM 3.11

$$P \{ A_\epsilon^{(n)} \} > 1-\epsilon, \quad P \{ B_\delta^{(n)} \} > 1-\delta$$

$$P \{ A_\epsilon^{(n)} \cap B_\delta^{(n)} \} = P \{ A_\epsilon^{(n)} \} + P \{ B_\delta^{(n)} \} - P \{ A_\epsilon^{(n)} \cup B_\delta^{(n)} \}$$

$$\geq 1-\epsilon + 1-\delta + 1 = 1-\epsilon - \delta.$$

$$1-\epsilon - \delta \leq P \{ A_\epsilon^{(n)} \cap B_\delta^{(n)} \}$$

$$= \sum_{x_1^n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} \leq 2^{-n} (H(x)-\epsilon)$$

$$= 2^{-n} (H(x)-\epsilon) \iff A_\epsilon^{(n)} \cap B_\delta^{(n)} \subseteq B_\delta^{(n)}$$

$$\Rightarrow$$

$$|B_\delta^{(n)}| \geq 2^{-n} (H(x)-\epsilon)$$

$$\Rightarrow |B_\delta^{(n)}| \geq (1-\epsilon') \cdot 2^{-n} (H(x)-\epsilon)$$

$$\Rightarrow 1/n \log |B_\delta^{(n)}| \geq H(x) - \epsilon' \cdot 2^{-n}$$
A real valued function \( f(x) \) or \( f(p) \) is convex \( (p \text{ is on } R^v \times) \) if

\[
f(\alpha p_1 + (1-\alpha) p_2) \leq \alpha f(p_1) + (1-\alpha) f(p_2)
\]

for every \( p_1, p_2 \) and \( \alpha \in \{0,1\} \).

Recall log-sum inequality: given \( a_i, b_i \)

\[
\sum_{i=1}^{\hat{a}} a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^{\hat{a}} a_i \right) \log \frac{\sum_{i=1}^{\hat{a}} a_i}{\sum_{i=1}^{\hat{a}} b_i} \quad \text{with equality if} \quad \frac{a_i}{b_i} = \frac{\sum_{i=1}^{\hat{a}} a_i}{\sum_{i=1}^{\hat{a}} b_i}
\]

with log-sum-ineq.

Show (i) \( h(x) \) (or \( h(p) \)) is concave in \( p \)
(ii) \( D(p||q) \) is convex in (w.r.t.) pair \( (p, q) \).
(iii) What can you say about \( I(x, y) \)
\[ A_e^{(n)} = \left\{ x^n \in X^n : \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \varepsilon \right\} \]

1. \( X_i \) s.i.d. \( p(x) \)
2. \( H(X) - \varepsilon \leq -\frac{1}{n} \log p(x^n) \leq H(X) + \varepsilon \)
3. \( 2^{-n(H(X)+\varepsilon)} \leq p(x^n) \leq 2^{-n(H(X)-\varepsilon)} \)
4. \( P \left\{ x^n \in A_e^{(n)} \right\} > 1 - \varepsilon \quad (\exists n_0 \geq n > n_0, \text{we can set } \delta = \varepsilon) \)
5. \( |A_e^{(n)}| \leq 2^{-n(H(X)+\varepsilon)} \)

From this \( \frac{|A_e^{(n)}|}{|X^n|} = 2^{-n(H(X)-\varepsilon)} \rightarrow 0 \text{ as } n \rightarrow \infty. \)

6. **Using methods of types**

\[
p(x^n) = \sum_{i=1}^{N(i|m)} \frac{1}{n} \log \frac{1}{n} p_i = \frac{1}{n} \sum_{i=1}^{k} p_{i}^n \log \frac{1}{n} p_i = -n \left( -\sum_{i=1}^{k} p_i \log p_i \right) = 2^{-nH(X)} \]

\( \frac{N(i|m)}{n} = \frac{p_i}{p_i} = p(X_i = i) \]

\( \frac{n}{\sum_{i=1}^{k} p^n_i} = \frac{n}{\sum_{i=1}^{k} p^n_i} \)

\( \rightarrow \frac{1}{2^n} \quad \text{as } n \rightarrow \infty. \)

\( \frac{N(c|\lambda)}{n} = \frac{1}{2^n} \quad \text{only if } \phi_i = \frac{N(c|\lambda)}{n} \)

\( \text{|sequences } x^n \text{ statistically have probability} \)

\( \frac{1}{2^{nH(X)}} \sum \phi_i = \frac{1}{2^{nH(X)}} \)

\( \text{set size} = \frac{1}{2^{nH(X)}} = \frac{1}{2^{nH(X)}} \)

\( \text{ratio of sets} = \frac{\sum_{i=1}^{n} D(p_i|\lambda)}{\sum_{i=1}^{n} p_i} \quad \frac{\sum_{i=1}^{n} D(p_i|\lambda)}{\sum_{i=1}^{n} p_i} \quad \frac{\sum_{i=1}^{n} D(p_i|\lambda)}{\sum_{i=1}^{n} p_i} \)

\( \text{set size} = \frac{1}{2^{nH(X)}} = \frac{1}{2^{nH(X)}} \)

\( \text{ratio of sets} = \frac{\sum_{i=1}^{n} D(p_i|\lambda)}{\sum_{i=1}^{n} p_i} \quad \frac{\sum_{i=1}^{n} D(p_i|\lambda)}{\sum_{i=1}^{n} p_i} \quad \frac{\sum_{i=1}^{n} D(p_i|\lambda)}{\sum_{i=1}^{n} p_i} \)
Today (i) Entropy Rates. (ii) Stationary Processes & Rate
(iii) Entropy Rate of Markov Chain (Stationary)
(iv) Random Walk on Graph (v) Hidden Markov Models and Rate.

1) Entropy Rate of a Stochastic Process \( \{X_i\}_{i=1}^\infty \) is

\[
H(\mathcal{X}) = \lim_{n \to \infty} \frac{H(X_n, X_{n-1}, \ldots, X_1, X_0)}{n}
\]

when the limit exists.

\( \{X_i\}_{i=1}^\infty \) iid \( \implies H(X_n) = \frac{n}{n} H(X_i) = H(X_i) \)

3) \( H(X) \leq H(X_i) \) as expected.

\( \{X_i\}_{i=1}^\infty \), Not iid but \( X_i \perp X_j \) and all possible comp.

\( X_i \)'s are mutually independent.

\[
H(X_1^n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i) \] "Average" rate of a block

may (not) exist.

Entropy Rate (Alternate)

\[
H'(\mathcal{X}) = \lim_{n \to \infty} H(X_n | X_{n-1})
\]

\( \{X_i\} \) Markov

\[
H'(\mathcal{X}) = \lim_{n \to \infty} H(X_n | X_{n-1}) = \lim_{n \to \infty} H(X_n | X_{n-1})
\]

If Stationary Markov

\[
H'(\mathcal{X}) = \lim_{n \to \infty} H(X_n | X_{n-1}) = H(X_2 | X_1)
\]
WE WANT TO SHOW A FEW THINGS.  FIRST NOTE
\[ H(x) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=2}^{n} H(x_i | x_{i-1}) + H(x_1) \]
(why?)

Look at each of \( H(x_i | x_{i-1}) \). They look similar to

ALTERNATE DEF OR ENTROPY RATE \( H'(x) \)

WE WANT TO KNOW WHEN \( H(x) = H'(x) \) SINCE
WE MAY BE EASIER THAN OTHER TO COMPUTE.

CLAIM! IF \( \{x_i\} \) ARE STATIONARY (\( \forall n \) \( H'(x) \) EXISTS.

NOTE \( H(x) \) or \( H(x_1 | x_{1-n}) \geq 0 \).

\[ H(x_n | x_{1-n}) \leq H(x_n | x_{2-n}) \]  (CONDITIONING REDUCES

\[ = H(x_{n-1} | x_{1-n-2}) \]  (\( x_i \) STATIONARY)

\( \Rightarrow \) \( H(x_n | x_{1-n}) \leq H(x_{n-1} | x_{1-n-2}) \) \( \Rightarrow \) \( H(x_n | x_{1-n}) \) IS A

DECREASING SEQUENCE AT WORST. (AT BEST CONSTANT) \( H(.) \geq 0 \) \( \Rightarrow \) \( \lim \) CONVERGES. \( \exists Bn \)

\( \Rightarrow \) \( \lim \ T(x_n | x_{1-n}) \) EXISTS, NON-NEGATIVE AS WE

\( k \to \infty \)

THIS IS NOT ENOUGH. WE NEED MORE.

WE NEED TO SHOW \( \lim_{n \to \infty} \frac{1}{n} \sum_{i=2}^{n} H(x_i | x_{i-1}) + H(x_1) \]

EXISTS AND IS EQUAL TO \( H'(x) \)!

ASIDE. IF \( a_n \to a \), \( \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} a_i \), THEN \( b_n \to a. \)

\( \text{PROOF.} \quad a_n \to a \Rightarrow \exists N(\varepsilon) \Rightarrow + n \geq N(\varepsilon) \)

\( \Rightarrow \) \( |b_n - a| = \left| \frac{1}{n} \sum_{i=1}^{n} a_i - a \right| \leq \frac{1}{n} \sum_{i=1}^{n} |a_i - a| \)

\( \lim_{n \to \infty} |b_n - a| \leq \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{N(\varepsilon)} |a_i - a| + \left( \frac{n - N(\varepsilon)}{n} \right) \varepsilon \)
Since $N(x)$ is fixed for given $x$, as $n \to \infty$
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} |a_i - a| < \varepsilon \text{ can be achieved.} \tag{3}
\]
\[
\Rightarrow \lim_{n \to \infty} |b_n - a| \leq 2 \varepsilon \Rightarrow b_n \to a, \quad \text{CESARO MEAN}
\]

We are now ready to show:
\[
H(x) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=2}^{n} H(x_i | x_{i-1}) = \lim_{n \to \infty} \frac{n}{n} \sum_{i=2}^{n} H(x_i | x_{i-1}) + H(x_1)
\]
\[
= H'(x^*)
\]

Note:
(i) $a_i = H(x_i | x_{i-1}) \to H'(x) \quad \text{(STATIONARY)}$
(ii) $b_n = \frac{1}{n} \sum_{i=2}^{n} H(x_i | x_{i-1})$; $b_1 = H(x_1) = a_1$

$x_i$'s stationary $\Rightarrow H(x_i | x_{i-1}) \to H'(x) = a$

From earlier proof $b_n \to a = H'(x)$

Now,
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=2}^{n} H(x_i | x_{i-1}) = \lim_{n \to \infty} \frac{n}{n} \sum_{i=2}^{n} H(x_i | x_{i-1}) + H(x_1)
\]
\[
= H'(x)
\]

For stationary processes $H(x) = H'(x)$

For Markov process $\Rightarrow H(x) = H'(x) = H(x_1 | x_1)$
**Example.**

Recall:

\[ P = \begin{bmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{bmatrix} \]

\[ \mu = \mu P \]

\[ \mu_1 = \frac{\beta}{\alpha + \beta}, \quad \mu_2 = \frac{\alpha}{\alpha + \beta} \]

\[ H(X_1) = H(\mu_1, \mu_2) \]

\[ H(X_2 | X_1) = ? \]

\[ H(X_2 | X_1 = 0) = -P(X_2 = 0 | X_1 = 0) \log P(X_2 = 0 | X_1 = 0) \]

\[ -P(X_2 = 1 | X_1 = 0) \log P(X_2 = 1 | X_1 = 0) \]

\[ = H(\alpha) \]

\[ H(X_2 | X_1 = 1) = H(\beta) \]

\[ \Rightarrow H(X_2 | X_1) = H(X_2 | X_1 = 0) \cdot P(X_1 = 0) + H(X_2 | X_1 = 1) \cdot P(X_1 = 1) \]

\[ = \frac{\beta}{\alpha + \beta} H(\alpha) + \frac{\alpha}{\alpha + \beta} H(\beta). \quad \text{Entropy Rate} \]

Generalizing to \( \mu = (\mu_1, \ldots, \mu_n) \):

\[ P = n \times n \]

\[ H(X_2 | X_1) = \sum_{j=1}^{n} \sum_{i=1}^{n} \mu_i P_{ij} \log P_{ij} \]

\[ = \sum_{i=1}^{n} \mu_i \left( -\sum_{j=1}^{n} P_{ij} \log P_{ij} \right) \]

\[ = -\sum_{i=1}^{n} \sum_{j=1}^{n} \mu_i P_{ij} \log P_{ij} \]

\[ P_{ij} = P(X_2 = j | X_1 = i) \]
**HMM.** \( X_i \) is stationary Markov. \( Y_i = \phi(X_i) \) 
\( \Rightarrow Y_i \) is stationary but need not be Markov.

**Claim 2.** \( H(Y_n \mid Y_{i-1}^n, X_i) \leq H(\phi) \leq H(Y_n \mid Y_{i-1}^n) \)

**Proof.** We already know \( Y_i \) is stationary \( \Rightarrow H(Y) \leq H(Y_n \mid Y_{i-1}^n) \)

**Need to show**
\[ H(Y_n \mid Y_{i-1}^n, X_i) \leq H(\phi) \]
\[ H(Y_n \mid Y_{i-1}^n, X_i) = H(Y_n \mid Y_{i-1}^n, X_{i-k}) \quad (\because X_i \text{ is Markov}) \]
\[ = H(Y_n \mid Y_{i-1}^n, X_{i-k}, Y_{i-1}^k) \]
\[ \leq H(Y_n \mid Y_{i-k}) \]
\[ = H(Y_{n+k} \mid Y_{i}^{n+k}) \]
\[ \Rightarrow H(Y_n \mid Y_{i-1}^n, X_i) \leq H(Y_{n+k} \mid Y_{i}^{n+k}) \quad \forall k. \]
\( \Rightarrow \) it is true for \( k \to \infty \) as well.
\[ \Rightarrow H(Y_n \mid Y_{i-1}^n, X_i) \leq \lim_{k \to \infty} H(Y_{n+k} \mid Y_{i}^{n+k}) = H(\phi) \]

\[ \Rightarrow H(Y_n \mid Y_{i-1}^n, X_i) \leq H(\phi) \leq H(Y_n \mid Y_{i-1}^n) \]

**Claim 3.** \[ \lim_{n \to \infty} H(Y_n \mid Y_i^{n-1}) - H(Y_n \mid Y_i^{n-1}, X_i) = 0 \]

**Proof.**
\[ I(Y_n \mid X_i \mid Y_i^{n-1}) = H(Y_n \mid Y_i^{n-1}) - H(Y_n \mid Y_i^{n-1}, X_i) \]
\[ = H(X_i \mid Y_i^{n-1}) - H(X_i \mid Y_i^n) \leq H(X_i) \]
\[
I(Y^n, \Lambda X_1) = I(Y, \Lambda X_1) + \sum_{i=2}^{n} I(X_i, Y_i; X_{i-1}^i)
\]

\[
\Rightarrow \lim_{n \to \infty} I(X_i, Y_i^n) = \lim_{n \to \infty} I(Y_i, X_i) + \sum_{i=2}^{n} I(X_i, Y_i; Y_{i-1}^i)
\]

\[
= I(X_i, Y_i) + \sum_{i=2}^{\infty} I(X_i, Y_i; Y_{i-1}^i)
\]

Each of \(I(X_i, Y_i; Y_{i-1}^i) \geq 0\)

But \[
\lim_{n \to \infty} I(X_i, Y_i^n) \leq \mathbb{H}(X_i)
\]

\[
\Rightarrow \exists \varepsilon > 0 \quad \forall n > N(\varepsilon) \quad I(X_i, Y_i; Y_{i-1}^i) \to 0
\]

\[
\Rightarrow \lim_{i \to \infty} I(X_i, Y_i; Y_{i-1}^i) = 0
\]

or \[
\lim_{i \to \infty} \mathbb{H}(X_i; Y_{i-1}^i) - \mathbb{H}(Y_i; Y_{i-1}^i, X_i) = 0
\]

\[
\Rightarrow \lim_{i \to \infty} \mathbb{H}(Y_i; Y_{i-1}^i) = \lim_{i \to \infty} \mathbb{H}(Y_i; Y_{i-1}^i, X_i)
\]

---

HW #3, Chapter 4, Due Feb 6, 2008

4.1, 4.2, 4.3, 4.4, 4.6, 4.7, 4.11

4.20, 4.21
LAST TIME: CHAPTER 4, ENTROPY RATE
\[ H(X) = \lim_{n \to \infty} \frac{H(X^n)}{n} = \lim_{n \to \infty} H(X^n | X^{n-1}) = H(X) \]
for stationary processes.

TODAY

2. DATA COMPRESSION 3. CODE 4. AVERAGE CODE LENGTH
4. NON-SINGULAR CODE 5. EXTENSION / CONCATENATION CODE
6. UNIQUELY DECodable CODES
7. PREFIX FREE CODES
8. KRAFT INEQUALITY

DEF. A SOURCE CODE OR RV X IS (DENOTECD C)
A MAPPING FROM X TO A SET OF "STRINGS" OF SYMBOLS FROM A D-ARY ALPHABET
\[ C : X \to \mathbb{D} \]
EXAMPLE \( X = \{ H, T \} \)
\[ P(H) = \frac{1}{2}, \quad P(T) = \frac{1}{2} \]

DEF. EXPECTED LENGTH \( L(C) \) OF A SOURCE CODE C
is given by
\[ L(C) = \sum_{x \in X} f(x) \cdot l(x) \]
WHERE \( l(x) \) IS THE # OF SYMBOLS IN C(x).

\[ \Rightarrow \text{FOR EXAMPLE ABOVE} \]

\[ L(C) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1 \]
IF WE DID \( C(H) = 00, \quad C(T) = 11 \)
\[ L(H) = 2, \quad L(T) = 2 \]
\[ L(C) = \frac{1}{2} \cdot 2 + \frac{1}{2} \cdot 2 = 2 \]
We want to have smaller / shorter representation
Example

<table>
<thead>
<tr>
<th>x</th>
<th>P(x)</th>
<th>C(x)</th>
<th>C_2(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/2</td>
<td>000</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1/4</td>
<td>001</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>1/8</td>
<td>01</td>
<td>110</td>
</tr>
<tr>
<td>4</td>
<td>1/8</td>
<td>1</td>
<td>111</td>
</tr>
</tbody>
</table>

\[ H(x) = -\frac{1}{2} \log_2 2 + \frac{1}{4} \log_2 4 + \frac{3}{8} \log_2 8 \]

\[ = \frac{1}{2} + \frac{1}{2} + \frac{3}{4} = 1.75 \text{ bits} \]

\[ L(C_1) = \frac{1}{2} \times 3 + \frac{1}{4} \times 3 + \frac{1}{8} \times 2 + \frac{1}{8} \times 1 \]

\[ = \frac{18}{8} + \frac{3}{8} = \frac{21}{8} = 2.625 > H(x) \]

\[ L(C_2) = \frac{1}{2} + \frac{1}{4} \times 2 + \frac{3}{8} + \frac{3}{8} = \frac{1}{2} + \frac{1}{2} + \frac{3}{4} = 1.75 \]

\[ L(C_2) < L(C_1) \]

Weighted Avg is small if \( p_i > p_j \Rightarrow l_i < l_j \)

\[ p_i l_i + p_j l_j \] is what we seek for.
**NON-SINGULAR CODE**

- $C : X \rightarrow Y$ is NON SINGULAR IF
  
  $x_i \neq y_j, \ x_i, y_j \in X \Rightarrow C(x_i) \neq C(y_j)$

**EXAMPLE**

<table>
<thead>
<tr>
<th>$X$</th>
<th>$C(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>01</td>
</tr>
<tr>
<td>4</td>
<td>010</td>
</tr>
</tbody>
</table>

NO PROBLEM IF WE XMIT ONLY ONE CODE WORD. BUT HOW DOES THE RECEIVER KNOW WHAT IT IS, IS $C(x_4) = 101 = 1, 01 = C(x_1)C(x_3)$

WE NEED MORE RULES!

**EXTENTION CODE** $C^*$ of $C$ is DEFINED AS

$$C^*(x_1, \ldots, x_n) = C(x_1)C(x_2) \cdots C(x_n)$$

$C(x_i)C(x_j)\cdots$ is CONCATENATION OF CODEWORDS

**DEF.** **UNIQUELY DECODABLE CODE** $C$

$C^*$ IS UNIQUELY DECODABLE IF $C^*$ IS NON-SINGULAR
Example:

<table>
<thead>
<tr>
<th>X</th>
<th>C(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>00</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>110</td>
</tr>
</tbody>
</table>

$C(4) = 1110$

Once we get to we know what it is

$C(3) = 1011$  

$C(3) \text{ is prefix of } C(4)$

We may not know immediately but can find and decode once we know the sequence till the end.

Ideally we want to decode as we receive each code word, \( \Rightarrow \) no prefix.

**Def. Prefix (Free) Code.**

A code is prefix (free) or instantaneous if no code word is prefix of other.

We can decode on the fly.

\[ \begin{array}{c|c}
 X & C(X) \\
---&--- \\
 1 & 00 \\
 2 & 01 \\
 3 & 10 \\
 4 & 11 \\
\end{array} \]
The idea behind prefix codes is that if no code word is a prefix of another, each uniquely decodable word can be "thought" of it as an ID. Each ID is unique.

Concretized by Kraft inequality

For any instantaneous code over alphabet of size $D$, the code word lengths $l_1, \ldots, l_{\max}$ must satisfy:

$$\sum_{i=1}^{D-1} l_i \leq 1$$

Conversely, if $l_1, \ldots, l_{\max}$ satisfy (1), there exists a $D$-ary alphabet and codewords such that an instantaneous code with these word lengths exists.

**Proof (Graphical Build a D-ary Tree of Depth $\max \frac{D}{l_i}$)**

Assume $l_1 \leq l_2 \leq \ldots \leq l_{\max}$

Example: $l_{\max} = 3$,

$D = 2$

$l_1 = 2$, $l_2 = 2$, $l_3 = 3$,

Label each child as "0...d-1"

Total leaf = $D^{\max}$

Each leaf is a code word.

Traverse the path from root to leaf.

Read off symbols $C(x_1) = 00$, $C(x_2) = 01$, $C(x_3) = 10$, $C(x_4) = 11$.

Once we go up to depth $l_i$ along a path from root, to choose $C(x_i)$, all children are pruned.
In a depth \( l_i \) we prune \( d^{\text{max} - l_i} \) children.

\[ \Rightarrow \text{each } l_i \text{ leaves removes } d^{\text{max} - l_i} \]

Leaves. No leaves overlap.

\[ \Rightarrow \sum d^{\text{max} - l_i} \text{ is the total leaves pruned.} \]

But this should be less than \( d^{\text{max}} \), the max possible leaves.

\[ \Rightarrow \sum d^{\text{max} - l_i} \leq d^{\text{max}} \]

\[ \Rightarrow \left( \sum d^{-l_i} \right) \leq 1 \]

Problems - 5.1, 5.2, 5.3
LAST WEEK:  (i) STARTED DATA COMPRESSION  (ii) DEFINED YD, PREFIX CODES. (iii) KRAFFT INEQ. \( l_1 \ldots l_{\max} \) ARE CODE WORD LENGTHS OF PREFIX CODE \( \sum d^{-l_i} \leq 1 \). \( (3-\text{ARY ALPHABET}) \). IT IS \textit{LEFT} CASE.

TODAY  (i) RECAP KRAFFT  (ii) DISCUSS COMMON HW PROBLEMS  (iii) RECAP K.E. (iii) OPTIMAL CODE LENGTHS AND ENTRPY.

\( i) \sum P(x,y) \log P(x,y) = 0 . \text{(MOST) OF YOU LOOKED AT PROB.} \)
AND SAID \( P(x) = 0 \) IF \( x \) IS A SOLN. NO! \( \sum P(x) = 1 \) MEANS THIS IS NOT TRUE. WE HAVE \( P_1, P_2 \ldots P_n \) FURTHER VALUES
\( ii) \text{Study } P(y|x) = 0 \text{ in 8.6. } \Rightarrow P(y|x) = 1 \text{ for } x, y. \text{ SEE EXAMPLE.} \)
\( H(y|x) = \sum P(x,y) \log P(y|x) \text{ for } x = x \text{ and } y = y \text{ is 1 mp.} \)

RECALL K.E: FOR PREFIX CODES (ALPHABET SIZE 2)
THE CODE WORD LENGTHS \( l_1 \ldots l_{\max} \) (OR TO ARRANGE \( l_i \leq l_{i+1} \ldots l_{\max} \)) SATISFY \( \sum d^{-l_i} \leq 1 \). MOREOVER IF \( l_1 \ldots l_{\max} \) SATISFY K.E. THERE EXISTS A PREFIX CODE WITH \( l_1 \ldots l_{\max} \) AS LENGTHS.

CONSTRUCTED TREE OF DEGREE-D AND DEPTH \( l_{\max} \).
STARTING WITH \( l_i \), WE CONSTRUCTED UNIQUE PATHS FROM ROOT OF THE TREE TO LEAF AT DEPTH \( l_i \).
PRUNED ALL CHILDREN ONCE A LEAF IS CHOSN.
CODWORD TOTAL OF CHILDREN PRUNED = \( \sum d^{l_{\max} - l_i} \).
THIS SHOULD BE LESS THAN \( d^{l_{\max}} \).

\[ \sum d^{-l_i} \leq 1 \]
Converse: Given \( l_1 \ldots l_m \), construct a tree of max depth \( l_{\text{max}} \). Now construct or choose a path with dept \( l_1 \). Remove all descendants of it \((D_{\text{max}} - l_1)\). This is code word 1.

Then choose another node in the nearest path at depth \( l_2 \) and label it. This is code word 2.

ETC.

**Reading**

**Your Assignment** Theorem 5.2.2

Optimal codes: We know codes that are Optimal codes. We know codes that are instantaneous satisfy KEG. Also known for each symbol \( x \in X \), \( p(x) \), \( c(x) \), \( l(c(x)) \)

Want to look at average code word length and minimize it

**Average Code Word Length**

\[ L = \sum_{x \in X} p(x) c(c(x)) = \sum_{x \in X} p_i \cdot l_i \]

\[ l_i = l(c(x_i)) \]

**Problem:** Minimize \( L \) subject to \( \sum D - l_i \leq 1 \)

Cost function is (Lagrangian formulation)

\[ J = \sum_{x \in X} p_i \cdot l_i + A \left( \sum D - l_i \right) \]

\[ \frac{\partial J}{\partial l_i} = p_i - A \cdot D - l_i \ln D = 0 \]

\[ D - l_i = \frac{p_i}{A \ln D} \]
Using $D_l = 1$ we have

$$\sum_i D_l i = \sum_i \frac{p_i}{A \ln D} = \frac{1}{A \ln D} = 1$$

$$\Rightarrow p_i = D_l i \Rightarrow l_i = -\log D_l i$$

$$=) \text{OPTIMAL AVERAGE CODE WORD LENGTH OF PREFIX CODE } = \sum p_l \log D_l i = H_D(x)$$

D-ARY ENTROPY OF $X$.

Claim! $L = \sum p_l l_i \leq H_D(x)$ for D-ARY CODE WITH $D$ PREFIXES. WITH EQUALITY IFF $D_l i = p_l$.

pf. $L - H_D(x)$

$$= \sum p_l l_i + \sum p_l \log D_l i$$

$$= \sum p_l \log D_l i - \sum p_l \log p_l$$

$$= \sum p_l \log \frac{p_l}{D_l i} = D(\phi || \phi) + \log \frac{1}{m}$$

$$= \sum p_l \frac{p_l}{D_l i} \log \left( \frac{p_l}{D_l i} \right) - \log \left( \sum D_l i \right)$$

$$= D(\phi || \phi) + \log \frac{1}{\sum D_l i} \geq 0 \Rightarrow \boxed{L \geq H_D(x)}$$
Equality holds iff \( \log D - \xi \geq 0 \) then \( \sum D \xi_i = \sum \phi_i = 1 \) \( \frac{g}{\sum D \xi_i} = 0 + 0 = 0 \) \( \frac{g}{\sum D \xi_i} = 0 + 0 = 0 \)

**Bounds on the Optimal Code Length**

**Claim** for an instantaneous code achieving **the optimal value** \( L^* = - \log \pi \)

\[ H(x) \leq L = \sum \phi_i \xi_i \leq H(x) + 1 \]

Consider

\[ L - H_D(x) = D (\phi \log \pi) + \log \frac{1}{\sum D \xi_i} \geq 0 \]

And \( \xi_i = - \log \pi_i \) \( \xi_i \) may not be an integer.

\[ \Rightarrow \left\lceil \log \pi_i \right\rceil \geq \xi_i \] (Smallest integer \( \geq \xi_i \))

\[ 0 \leq \xi_i \leq \left\lceil \log \pi_i \right\rceil \Rightarrow \log \frac{1}{\pi_i} \leq \xi_i < \log \frac{1}{\pi_i} + 1 \]

Hence

\[ \Rightarrow D \left\lceil \log \pi_i \right\rceil \leq D \xi_i \leq D \pi_i \]

\[ \Rightarrow \sum D \left\lceil \log \pi_i \right\rceil \leq \sum D \xi_i = \sum \phi_i = 1 \]

\[ \Rightarrow \pi_i \log \frac{1}{\pi_i} \leq \phi_i \xi_i \leq \pi_i \left( \log \frac{1}{\pi_i} + 1 \right) \]

\[ \Rightarrow \sum \phi_i \log \frac{1}{\pi_i} \leq \sum \phi_i \xi_i \leq \sum \phi_i \left( \log \frac{1}{\pi_i} + 1 \right) \]

\[ \Rightarrow H(x) \leq L \leq H(x) + 1 \]
CLAIM: Since \( h(x) \leq L \leq h(x) + 1 \) for a source distribution \( P \) with \( d \)-ary code words, it is also true if code words were to be optimal.

\[ l_i^* = -\log P_i \]
LAST TIME. KRAFT INEQ FOR PREFIX CODES (Finitel)

(i) \( \sum D \cdot l_i \leq 1 \) \( l_i = l(c(c(x_i))) \) \( x_i \in X \)

(ii) Optimal Code Length. \( \min \{ \sum l_i p_i, l_i \geq 0 \} = \) \( \min J = \sum p_i l_i + H(D, l_i) \) w.r.t. \( l_i \)

\[ l_i^* = -\log_D p_i \]

(iii) Shown \( L = \sum l_i p_i \) \( \Rightarrow L - H_2(x) = D(p || q) + \frac{\log p_0}{\sum p_0 \cdot l_i} \)

\[ p_i = \frac{D \cdot l_i}{\sum D \cdot l_i}, \sum D \cdot l_i \leq 1 \] \( \Rightarrow L - H_2(x) \geq 0 \)

with equality when \( p_i = l_i \) \( \sum x \cdot l_i = 1 \)

\[ \Rightarrow p_i = l_i \text{ or } l_i = -\log_D p_i \]

\( \Rightarrow \) Probabilities are D-Adic; \( H_2(x) \leq L < H_0(x) + 1 \)

TODAY \( l_i = -\log_D p_i \) THEN

(i) \( H_D(x) \leq \sum p_i l_i < H_0(x) + 1 \)

(ii) NF \( l_i = -\log_D p_i \) BUT EXPECTED LENGTH UNDER \( p_i \) THEN \( H(p) + D(p || q) \leq L = \sum p_i l_i \leq D(p || q) + H(p) + 1 \)

(iii) For IID sequences \( H(x, n) = \sum p(x, n) l(x, n) < H(x) + 1 \)

\[ \Rightarrow \text{Bit per symbol} \frac{1}{n} H(x, n) \leq E l_i < H_0(x) + 1/n \]

or \( \frac{1}{n} \leq H(x, n) \rightarrow H(x) \) Entropy Rate.

(iv) TREE FOR IID OR STATIONARY PROCESSES.

(V) KINeq FOR UD \( \sum i \cdot l_i \leq 1 \)

(vi) READ NOTE FOR xi LENGTHS.
\[ l_i = \left\lceil -\log_p q_i \right\rceil \]

\[ \Rightarrow -\log_p q_i \leq l_i < -\log_p q_i + 1 \]

\[ \Rightarrow -\log_p q_i + \log_p l_i \leq l_i < -\log_p q_i + \log_p (l_i + 1) \]

\[ \Rightarrow E\left[-\log_p q_i + \log_p \frac{l_i}{q_i} \right] \leq E[l_i] < E\left[-\log_p q_i + \log_p \frac{l_i}{q_i} + 1 \right] \]

\[ \Rightarrow H_q(x) + D(p || q) \leq E \leq H_q(x) + D(p || q) + 1 \]
(iii) Note \( H = \sum_{i=1}^{l} \frac{1}{l} \) can lead to one bit extra \( \frac{1}{n} \) in coding. Let's now use \( H(x) \) entropy rate to show that block coding will do better!

Consider source with \( x \sim f(x) \), take block of \( n \), code \( C(x; n) \) and length \( l(C(x; n)) = l(x; n) \), note it is non-singular (instantaneous as well!)

\[
\Rightarrow \quad H(x; n) \leq \frac{1}{n} \mathbb{E}[l(x; n)] < \frac{1}{n} H(x; n) + \frac{1}{n}
\]

\( \Rightarrow \quad H(x; n) \leq \frac{1}{n} \mathbb{E}[l(x; n)] < \frac{1}{n} H(x; n) + \frac{1}{n} \)

As \( n \to \infty \), \( \frac{H(x; n)}{n} \to H(x) \) so \( x \) is iid

\[
\Rightarrow \quad \frac{H(x)}{n} \leq \frac{1}{n} \mathbb{E}[l(x; n)] < \frac{1}{n} H(x) + \frac{1}{n}
\]

\( \mathbb{E}[l] = \mathbb{E}[l(x; n)] \to H(x) \) as \( n \to \infty \)

For iid \( H(x) = H(x) = \lim_{n \to \infty} \frac{H(x; n)}{n} \)

\[
\Rightarrow \quad H(x) \leq \frac{1}{n} \mathbb{E}[l(x; n)] < H(x) + \frac{1}{n}
\]

For large blocks, average codeword length approaches entropy rate.
(V) For stationary processes, use of optimal code word length satisfies
\[ H(x^n) \leq E[l(x^n)] \leq H(x^n) + 1 \]
\[ \Rightarrow \frac{1}{n} H(x^n) \leq \frac{1}{n} E[l(x^n)] \leq \frac{1}{n} H(x^n) + \frac{1}{n} \]
\[ \Rightarrow n \rightarrow H(x) \leq E[l(x)] \leq H(x) + \lim_{n \rightarrow \infty} \frac{1}{n} \]

or \[ \underline{\text{or}} \]
\[ E[l(x)] \rightarrow H(x) \]

(VI) Kraft inequality for UD codes.
We know UD codes is a larger set.
But for UD codes with \( l_1, \ldots, l_m \) lengths, inequality holds.
\[ \sum_{i=1}^{m} -l_i \leq 1 \]
Converse holds as well.

Proof due to few observations
(i) Given base \( D \) with \( m \) bits.
\[ \text{max \# sequences of } m \text{ bits is } D^m. \]
(ii) \[ \lim_{k \rightarrow \infty} \left( \frac{k}{c} \right)^{1/k} = 1. \]
**Proof** Consider. Consider \( l(x_i) = l_i \).

\[
\left( \sum_{x_i \in x^k} D - l_i \right)^k = \left( \sum_{x_i \in x^k} D - l(x_i) \right)^k
\]

\[
= \sum_{x_i \in x^k} \sum_{x_j \in x^k} \sum_{x_k \in x^k} D - l(x_i) - l(x_j) - l(x_k)
\]

\[
= \sum_{x_i \in x^k} \sum_{x_i \in x^k} D - l(x_i, k) = \sum_{x_i \in x^k} D - l(x_i, k)
\]

\[
\Rightarrow \left( \sum_{x_i \in x^k} D - l_i \right)^k = \sum_{x_i \in x^k} D - l(x_i, k).
\]

\[
= \sum_{m=1}^{k \cdot \text{lmax}} a(m) \cdot D - m
\]

\( a(m) = \# x_i^k \text{ mapped to code words of length } l(x_i, k) = m, a(m) \leq D^m \)

\[
\Rightarrow \left( \sum_{x_i \in x^k} D - l_i \right)^k \leq \sum_{m=1}^{k \cdot \text{lmax}} D - m = k \cdot \text{lmax}
\]

\[
\Rightarrow \sum_{x_i \in x^k} D - l_i \leq (k \cdot \text{lmax})^{1/k} + k
\]

\[
\Rightarrow \text{true for } k \to \infty \Rightarrow \sum_{x_i \in x^k} D - l_i \leq (k \cdot \text{lmax})^{1/k}
\]

\[
\Rightarrow k \to \infty
\]
LAST WEEK
(i) Kraft Inequality: \( \sum \frac{p_i}{2} \leq \log_2 N \) for UD with codeword lengths \( l_1, l_2, \ldots, l_{\text{max}} \)
(ii) Optimal code-word length \( l_i = -\log_2 p_i \)
(iii) Average length \( E[l_i] = E[-\log_2 p_i] = H_D(X) \)
(iv) If \( l_i = \lceil -\log_2 p_i \rceil \), then \( H_D(X) \leq E[l_i] < H_D(X) + 1 \)
(v) If \( l_i = \lceil -\log_2 p_i \rceil \), then \( H_D(X) + 1 - d(p_i) \leq E[l_i] < H_D(X) + d(p_i) + 1 \)

TODAY (a) Huffman code (b) Properties of optimal code

Huffman Coding (Alg)

Construction of (an) optimal prefix code for given source distribution.

EX: Binary Case

Recursion

1. Merge two least likely symbols to reduce alphabet size by 1.
2. Repeat until alphabet size reduces to 1.
3. Use existing merge lines to form binary tree.

Example: \( |X| = |Y| = 7 \), \( D = 2 \)

Read from root:

```
| Y | X |
+---+---+
| 0 | 0  |
| 1 | 1  |
| 0 | 2  |
| 1 | 3  |
| 0 | 4  |
| 1 | 5  |
| 1 | 6  |
| 1 | 7  |
```

Any same idea but may need dummy symbols of zero probability in order to ensure that last step of merger has exactly \( D \) symbols.
Q: ADD ENOUGH "DUMMY" SYMBOLS?
A: FINAL MERGER LEADS TO 1 SYMBOL AND IF THERE WERE $K$ MERGERS, TOTAL # OF SYMBOLS = $1 + K(D-1)$ SINCE EACH MERGER ABSORBS $D-1$ SYMBOLS.

**Example**

$D = 3$

<table>
<thead>
<tr>
<th>$x_0$</th>
<th>$x_1$</th>
<th>0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_2$</td>
<td>0.20</td>
<td>0</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.15</td>
<td>1</td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.15</td>
<td>0.25</td>
</tr>
<tr>
<td>$x_5$</td>
<td>0.15</td>
<td>1</td>
</tr>
<tr>
<td>$x_6$</td>
<td>0.10</td>
<td>0.25</td>
</tr>
</tbody>
</table>

**Optimality of Huffman Code**

Proving this requires three properties of an optimal code. (Assume $D=2$) $p_1 > p_2 > \ldots > p_m$

**3 Properties**

(i) Every node must have two branches stemming from it.

---

**Optimality of Huffman Code**

Proving this requires three properties of an optimal code. (Assume $D=2$) $p_1 > p_2 > \ldots > p_m$

**3 Properties**

(i) Every node must have two branches stemming from it.

---

**Optimality of Huffman Code**

Proving this requires three properties of an optimal code. (Assume $D=2$) $p_1 > p_2 > \ldots > p_m$

**3 Properties**

(i) Every node must have two branches stemming from it.
2) Nodes at level \( \ell \max \) come in pairs (siblings).

3) \( k = \ell \)

\[ p_i \leq b_i \]

(Contradiction)

\( k = \ell \)

(Contradiction)

\( (h_i - p_i) \leq (h_i + p_i) \)

(Contradiction)

To reduce the average code length, if possible, the path length should be changed.

Optimum code \((b_i, b_i)\) for \( h_i > b_i \).

Better to state \( h_i \) \( \Rightarrow \) \( h_i \) \( \Rightarrow \) \( b_i \).

Locations
(c) By (3) and (b) we can rearrange the two least likely source symbols as two siblings on the highest level of tree. Thus, an optimal code $C_m$ of the form

$$C_m: \text{SUBTREE}$$

$$p_{m-1}$$

$$p_m$$

Two least likely source symbols.

Consider following code $C_{m-1}$ for $(m-1)$-ary distribution.

$$p^{(m-1)} = (p_1, p_2, \ldots, p_{m-2}, \underbrace{p_{m-1} + p_m})$$

$$C_{m-1}: \text{SAME SUBTREE}$$

$\leftarrow$ Single "Super symbol" of prob $p_{m-1} + p_m$

Verify that

$$L(C_m) = L(C_{m-1}) + \underbrace{p_{m-1} + p_m}_{\text{fixed}}$$

Hence, if $L(C_m)$ is optimal for $p^{(m)}$ then $L(C_{m-1})$ is optimal for $p^{(m-1)}$.

Conversely if $C_{m-1}$ is optimal for $p^{(m-1)}$ then

$$L(C_{m-1}) = L(C_{m-1})$$
AND BY TURNING

\[ \begin{array}{c}
p_{m-1} p_m
\end{array} \]

INTO

\[ \begin{array}{c}
p_{m-1}
p_m
\end{array} \]

WE GET A \( C_M' \) FOR \( p_m \) SUCH THAT

\[
L(C_M') = L(C_{m-1}) + p_{m-1} p_m
\]

\[
= L(C_{m-1}) + p_{m-1} + p_m
\]

\[
= L(C_M)
\]

i.e. \( C_M \) IS OPTIMAL.

HENCE HUFFMAN CODING IS OPTIMAL

READ 5 - 7.
CHAPTER 7

DISCRETE MEMORYLESS CHANNELS (DMC)

SIMPLEST FORM OF NON-DETERMINISTIC TRANSMISSION MEDIUM

\[ X \rightarrow DMC \rightarrow Y \in \mathcal{Y} \]

DISCRETE INPUT ALPHABET \( \mathcal{X} \) \& BOTH OUTPUT ALPHABET \( \mathcal{Y} \) DISCRETE

TRANSITION MATRIX \( [P_{xy}] \) WHERE

\[ P_{xy} = P(y|x) \quad \text{not joint prob } P_{x \rightarrow y} \]

MEMORYLESS CHANNEL ACTS ON EACH INPUT SYMBOL INDEPENDENTLY.

EQUIVALENTLY: GIVE THE PRESENT INPUT, THE PRESENT OUTPUT IS INDEPENDENT OF ALL PREVIOUS INPUTS AND OUTPUTS.

\[ X_{i-1}^{n-1}, Y_{i-1}^{n-1} \rightarrow X_n \rightarrow Y_n \]

\[ P_{Y_n|Z_{i-1}^{n-1}X_n} = P_{Y_n|X_n} \]

CHECK ALSO IMPLIES

\[ P\left(y_n|x_n\right) = \prod_{i=1}^{n} P\left(y_i|x_i\right) \]

MAIN QUESTION: HOW MANY MESSAGES CAN WE RELIABLY XMIT IN \( n \) USES OF DMC?

ANS: NUMBER GROWS EXponentially WITH \( n \). - EXPONENT IS CALLED CHANNEL CAPACITY - SHANNON’S CHANNEL CODING THEOREM.
DEFINITION OF CAPACITY

THE CAPACITY OF A DMC (IN BITS/CHANNEL USE) IS DEFINED AS

\[ C = \max_{p(x)} I(x; y) = \max_{p(x)} I(x \wedge y) \]

\[ \text{HERE, THE INPUT DISTRIBUTION } p(x) \text{ IS VARIABLE AND THE CONDITIONAL DISTRIBUTION } p(y|x) \text{ IS FIXED.} \]

COMPUTATION OF CAPACITY

1. \( C \geq 0 \) since \( I(x; y) = D(p_{xy} || p_x p_y) \geq 0 \)
2. \( C \leq \max \left( \log |x|, \log |y| \right) \)

\[ \text{SINCE } I(x; y) = H(x) - H(x|y) \leq H(x) = H(y) - H(y|x) \leq H(y) \]

MAXIMIZATION

OF A CONTINUOUS CONCAVE FUNCTION OF \( p(x) \) OVER A CLOSED SET OF DISTRIBUTION VECTORS \( p \in \mathbb{R}^{|x|} \)

CONCAVITY OF \( I(x; y) \) W.R.T \( p(x) \)

\[ I(x; y) = H(y) - H(y|x) \]
\[ H(y) \text{ IS CONCAVE FUNCTION OF } p(y) \]
\[ p(y) = \sum_{x} p(y|x) p(x) \text{ IS LINEAR FUNCTION OF } p(x) \]
\[ H(y) \text{; CONCAVE FUNCTION of } p(x). \]
\[ H(y|x) = \sum_{x \in \mathcal{X}} p(x) H(x) \text{ (xth row of Xsition Mattrix).} \]

\[ \Rightarrow H(y|x) \text{ is LINEAR FUNCTION of } p(x). \]

\[ \Rightarrow I(x;y) \text{ is CONCAVE FUNCTION of } p(x). \]

Closed form computation of capacity is not always possible.

Start reading ch. 7
LAST TIME (CLASS #8) STARTED CHAPTER 7 ON CAPACITY.

TODAY
(i) CHANNEL + SOURCE MODEL USED
(ii) DEFINITION OF CAPACITY
(iii) TYPES OF SYMMETRIC CHANNELS.
(iv) TYPICAL SETS, (JOINT TYPICALITY)
(v) PRELUDE TO CHANNEL CODING.

MODEL I

\[ E \left[ \ell(x^n) \right] \lim_{n \to \infty} H(x^n) = H(x) \]

SOURCE INFORMATION CAN BE REPRESENTED BY ENCODED CLOSED TO ENTROPY/PARITY USING BLOCK CODES.

@ XMIT WHAT IS THE PROBLEM?
(i) ENCODE AND XMIT X^n THROUGH RANDOM CHANNEL
(ii) OBSERVE Y^n @ RECEIVER END.

GOAL: REPRODUCE FAITHFULLY AS MUCH INFO ABOUT X^n AS POSSIBLE BASED ON Y^n.
Recall \( I (x^n \wedge y^n) = H(x^n) - H(x^n | y^n) \) 
\( = H(y^n) - H(y^n | x^n) \)

Ideally, drive \( H(x^n | y^n) \to 0 \) (if possible)

Then \( I (x^n \wedge y^n) = H(x^n) \)!

Where is the problem? The channel produces \( P_{xy} \neq 1 \). If we know \( P(x | y) \) then we can compute \( P(x^n | y^n) \) and

\[ I (x^n \wedge y^n) = H(x^n) - H(x^n | y^n) \]

is known. We want to maximize \( I (\cdot) \) w.r.t. what we can control. Namely, input distribution.

\[ \max_{p(x)} I (x^n \wedge y^n) \]

Loosely, this is defined as channel capacity

\[ C = \max_{p(x)} I (x^n \wedge y^n) \]

Intuitively, maximizing w.r.t. \( p(x) \) increases correlation between \( x \) and \( y \). \( y \) depends on \( x \) via \( p(y | x) \).

Definition

A discrete channel is a system consisting of an input alphabet \( \mathcal{X} \), output alphabet \( \mathcal{Y} \), and transition probability matrix \( P(y | x) \). \( \{ x, y, P(y | x) \} \)

Discrete memoryless channel (DMC) \((x, y, P(y | x))\)

Is such that \( P(y_n | x_0^n, y_0^{n-1}) = P(y_n | x_n) \)

Output at time \( n \) depends conditionally on input at time \( n \) and is conditionally independent of all past inputs and outputs.
PROPERTY OF DMC

$$P(y^n | x^n) = \frac{1}{n} \sum_{i=1}^{n} P(y_i | x_i)$$

Almost like iid.

BINARY SYMMETRIC CHANNEL (BSC)

$$p(x) \times$$

\[ Y \]

\[ 0 \quad p \]

\[ 0 \quad p \]

\[ 1 \quad 1-p \]

\[ 1-p \quad 1 \]

$$P(y=0 | x=0) = p(Y=1 | x=1) = 1-p$$

$$P_{y|x} = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

Compute $$I(x; y)$$, max $$I(x; y)$$

PROPERTY OF CAPACITY

1. $$C = I(x; y) = D(P_{xy} || P_x P_y) \geq 0$$ with equality @ $$P_{xy} = P_x P_y$$

2. $$C = I(x; y) = H(x) - H(x|y) \leq \log |\mathcal{X}|$$

$$= H(y) - H(y|x) \leq \log |\mathcal{Y}|$$

$$\Rightarrow C \leq \min \{ \log |\mathcal{X}|, \log |\mathcal{Y}| \}$$ (At first cut).

WHAT HAPPENS WITH NOISY CHANNEL?

Eq: $$X \rightarrow \oplus \rightarrow Y$$

$$X \neq Y$$ Here. We need a method to decode.

WHAT IS GOING TO HAPPEN IS TO IMPOSE THAT NO TWO $$X_i(i), X_j(j)$$ PRODUCE SAME OUTPUT $$Y_i, Y_j$$.

SHOW THAT THERE IS A WAY TO SELECT $$X^n$$ SUCH THAT IT IS MOST PROBABLE INPUT TO A GIVEN $$Y_i$$. 
Joint typical set $A^{(n)}_{xy}$ for sequences $\{x^n, y^n\}$ with $p(x^n, y^n)$ is defined as:

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

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

**Properties of $A^{(n)}_{xy}$**

(i) $\Pr \left\{ \frac{\# A^{(n)}_{xy}}{n} \rightarrow 1 \right\} \rightarrow 1$ as $n \rightarrow \infty$

(ii) $|A^{(n)}_{xy}| \leq 2^{n(H(x,y) + \epsilon)}$

(iii) IF $(\tilde{x}^n, \tilde{y}^n) \sim p(\tilde{x}^n) p(\tilde{y}^n)$ i.e. $\tilde{x}^n$ and $\tilde{y}^n$ are independent with marginals of $p(x^n, y^n)$ THEN $p_2 \{ (\tilde{x}^n, \tilde{y}^n) \in A^{(n)}_{xy} \} \geq (1 - \epsilon) 2^{-n(I(x,y) + 3\epsilon)}$

**Proof follows next**

$A^{(n)}_{x^n} = \{ x^n \in X^n : \left| -\frac{1}{n} \log p(x^n) - H(x) \right| < \epsilon \}$

$\Rightarrow$ we can choose $m_1$ such that $p_2 \{ \left| -\frac{1}{n} \log p(x^n) - H(x) \right| > \epsilon \} < \epsilon / 3$
Choose
\[ A_{\varepsilon, y}^{(n)} = \left\{ y^n : \left| -\frac{1}{n} \log \frac{p(y^n)}{p(y)} - H(y) \right| < \varepsilon \right\} \]

\[ \Rightarrow \Pr \left\{ \left| -\frac{1}{n} \log \frac{p(y^n)}{p(y)} - H(y) \right| > \varepsilon \right\} < \frac{\varepsilon}{3} - (1) \]

For \( n = n_2 \) can be chosen for \( n \) large enough.

Choose
\[ A_{\varepsilon, (x, y)}^{(n)} = \left\{ (x^n, y^n) : \left| -\frac{1}{n} \log \frac{p(x^n, y^n)}{p(x, y)} - H(x, y) \right| < \varepsilon \right\} \]

\[ \Rightarrow \Pr \left\{ \left| -\frac{1}{n} \log \frac{p(x^n, y^n)}{p(x, y)} - H(x, y) \right| > \varepsilon \right\} < \frac{\varepsilon}{3} \]

Can be achieved for large \( n = n_3 \).

\[ \Rightarrow \Pr \left\{ (x^n, y^n) \notin A_{\varepsilon, (x, y)}^{(n)} \right\} \]

\[ = \Pr \left\{ x^n \notin A_{\varepsilon, x}^{(n)}, y^n \notin A_{\varepsilon, y}^{(n)} \right\} + \Pr \left\{ (x^n, y^n) \notin A_{\varepsilon, (x, y)}^{(n)} \right\} \]

\[ + \Pr \left\{ (x^n, y^n) \notin A_{\varepsilon, (x, y)}^{(n)} \right\} \]

\[ = \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon \]

\[ \Rightarrow \Pr \left( (x^n, y^n) \in A_{\varepsilon}^{(n)} \right) = 1 \text{ as } n \to \infty \]
\[ 1 = \sum p(x^n, y^n) \geq \sum p(x^n, y^n) \]
\[ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \quad (x^n, y^n) \in \mathcal{A}_e^{(n)} \]
\[ \geq |\mathcal{A}_e^{(n)}| \cdot 2 \quad \forall (x^n, y^n) \in \mathcal{A}_e^{(n)} \]
\[ -n(H(x, y) + \epsilon) \quad \forall (x^n, y^n) \in \mathcal{A}_e^{(n)} \]
\[ \Rightarrow |\mathcal{A}_e^{(n)}| \leq 2^{-n(H(x, y) + \epsilon)} \]

For the last one:
\[ Pr \{ (x^n, y^n) \in \mathcal{A}_e^{(n)} \} = \sum p(x^n) \cdot p(y^n) \]
\[ (x^n, y^n) \in \mathcal{A}_e^{(n)} \]
\[ \leq 2^{-n(H(x^n) - \epsilon)} \cdot 2^{-n(H(y^n) - \epsilon)} \]
\[ = 2^{-n(I(x^n, y^n) + 3\epsilon)} \]
\[ \Rightarrow Pr \{ (x^n, y^n) \in \mathcal{A}_e^{(n)} \} \leq 2^{-n(I(x^n, y^n) + 3\epsilon)} \]

But \[ |\mathcal{A}_e^{(n)}| \geq (1-\epsilon) \cdot 2 \]
\[ \Rightarrow Pr \{ (x^n, y^n) \in \mathcal{A}_e^{(n)} \} = \sum p(x^n) \cdot p(y^n) \]
\[ (x^n, y^n) \in \mathcal{A}_e^{(n)} \]
\[ \geq (1-\epsilon) \cdot 2^{-n(H(x^n) - \epsilon)} \cdot 2^{-n(H(y^n) + \epsilon)} \cdot 2 \]
\[ = (1-\epsilon) \cdot 2^{-n(H(x^n) + H(y^n) - H(x^n, y^n) + 3\epsilon)} \]
\[ = (1-\epsilon) \cdot 2^{-n(I(x^n, y^n) + 3\epsilon)} \]
\[ \Rightarrow Pr \{ (x^n, y^n) \in \mathcal{A}_e^{(n)} \} \geq (1-\epsilon) \cdot 2^{-n(I(x^n, y^n) + 3\epsilon)} \]
Some Strange Behaviors of Weak Typicality.

If $\tilde{x}_i, \tilde{y}_i$ are independent and have property that $\tilde{x}_i \perp \tilde{y}_i$ and $P_{\tilde{x}_i \tilde{y}_i} = \frac{1}{n} \prod_{i=1}^{n} P_{\tilde{x}_i} P_{\tilde{y}_i}$

(Note (1) I(\text{x, y}) is fixed)

Setting $n \to \infty$ makes

$P \{ (\tilde{x}_i, \tilde{y}_i) \in A^{(n)} \} \to 0$ \(\therefore\) I(x, y) is fixed.

i.e. # of such $(\tilde{x}_i, \tilde{y}_i)$ pairs is negligible!

Application to Channel Coding

(i) Size of the set containing $(x^n, y^n)$ pairs such that $x^n$ is $\epsilon$-typical and $y^n$ is $\epsilon$-typical is $2^{n H(x)} 2^{n H(y)} 2^{n (H(x) + H(y))}$

(ii) Size of the set containing $(x^n, y^n)$ pairs that are jointly typical is $2^{n H(x,y)}$

(iii) Every jointly typical pair is marginally typical by definition, i.e. $2^{n (H(x) + H(y))} \geq 2^{n H(x,y)}$

(iv) For a given $x^n$, how many $y^n$ sequences are jointly typical?

For a given $x^n$, there are $2^{n H(x)} x^n \epsilon$-typical sequences. If $\alpha = \# \text{ of } y^n \text{ sequences that are jointly typical then}$

$$\alpha = 2^{n \{H(x,y) - H(x)}$$

$$\alpha = 2^{n \{H(x,y) - H(x)} = 2$$
A \( (m,n) \) code for channel \((X, p(y|x), a)\)
consists of

1. Map \( W \rightarrow \{1, \ldots, m\} \) or message index set.

2. Encoding \( X^n: \{1, \ldots, m\} \rightarrow X^n \). Take message and create a sequence of length \( n \). Set of the sequences \( \{X^n(1), X^n(2), \ldots, X^n(m)\} \) is called codebook.

3. A decoding rule \( g: Y^n \rightarrow \{1, \ldots, m\} \)

4. Probability of error

\[
I(g(y^n) \neq i) = \log \frac{1}{p(y^n)} \sum p(y^n|x^n(i)) \cdot I(g(y^n) \neq i)
\]
MAXIMAL PROBABILITY OF ERROR

\[ y^{(n)} = \max_{i \in \{1, \ldots, M\}} \bar{A}_i \]

AVERAGE PROB OF ERROR

\[ P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \bar{A}_i = P_r \{ I(i \neq g(y^{(n)})) \} \]

RATE OF A CODE \((M, n)\) IS

\[ R = \frac{\log M}{n} \]

\[ \Rightarrow M = \left\lceil 2^{nR} \right\rceil \]

\[ \Rightarrow (M, n) = (2^{nR}, n) \] code.

CAPACITY IS THEIR (LEAST) UPPER BOUND.

So let say \( \varepsilon, R_1, R_2, \ldots, R_i \) are achievable. Arrange to PIC \( R^* \Rightarrow R_i \leq R^* \).
CHANNEL CODING THEOREM

All rates \( R < I(X^n; Y^n) \) below capacity \( C = \max_{P(x)} \frac{1}{n} I(X^n; Y^n) \) are achievable. For every \( R < C \), there exists a sequence of codes \( (2^{nr}, n) \) such that
\[
\lambda^n \to 0 \quad \text{as} \quad n \to \infty.
\]
\( \lambda^n = \max \{ P(W \neq i^n) \} \)

Conversely, for \( (2^{nr}, n) \) codes with \( \lambda^n \to 0 \), all rates \( R < C \) must hold.

A SNEAK PREVIEW. WHEN \( \lambda^n = 0 \) FOR

\( \lambda^n = \max \) probability of error.

\( \Rightarrow \lambda^n = 0 \Rightarrow \) ? \[ I (Y^n \wedge W) = H(W). \]

\[ W \to X^n(w) \to Y^n \to \hat{W} \]

MARKOV CHAIN - DATA PROCESSING INEQUALITY

NOTE \( I(W \wedge Y^n) \leq I(Y^n \wedge X^n) = H(Y^n) \)

\[ C = \max_{P(x)} I(X^n \wedge Y^n) \]

\[ \begin{align*}
& \geq H(Y^n) - H(Y^n | X^n) \\
& \leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i | X_i) \\
& = \sum_{i=1}^{n} I(X_i \wedge Y_i) \\
& \leq n \cdot C
\end{align*} \]

But \( I(W \wedge Y^n) = H(W) = nR \leq nC \Rightarrow [R < C] \]
Today

- Revisit joint typicality
- Definitions
- Applications to channel coding.

Jointly Typical Sequences.

**Def:** For $\varepsilon > 0$, the jointly $\varepsilon$-typical set or jointly typical set or set of jointly typical sequences for an i.i.d. pmf on $\mathbb{R}^n \times \mathbb{R}^n$ which is $p_{xy}(x,y)$ is defined as:

$$A_{\varepsilon}^{(n)} = \left\{ (x^n, y^n) \in \mathbb{R}^n \times \mathbb{R}^n : 
\begin{align*}
&\left| -\frac{1}{n} \log p_{x^n}(x^n) - H(X) \right| < \varepsilon, \\
&\left| -\frac{1}{n} \log p_{y^n}(y^n) - H(Y) \right| < \varepsilon, \\
&\left| -\frac{1}{n} \log p_{x^n y^n}(x^n, y^n) - H(X,Y) \right| < \varepsilon
\end{align*}
\right\}$$

where $p_{x^n y^n} = \prod_{i=1}^{n} p_{xy}(x_i, y_i)$

**Theorem (JACEP)**

Let $p_{xy}$ be i.i.d. pmf of sequence $\{X_t, Y_t\}_{t=1}^\infty$ on $\mathbb{R} \times \mathbb{R}$. Then

(i) $\lim_{n \to \infty} P\left( (x^n, y^n) \in A_{\varepsilon}^{(n)} \right) \to 1$

(ii) $|A_{\varepsilon}^{(n)}| \leq 2^n (H(X,Y) + \varepsilon) \neq n$.

(iii) $|A_{\varepsilon}^{(n)}| \geq (1 - \varepsilon) 2^n (H(X,Y) - \varepsilon)$ as $n \to \infty$

(iv) If $(\tilde{x}^n, \tilde{y}^n) \sim p_x [\tilde{x}^n] p_y [\tilde{y}^n]$ i.e. same marginals as $(x^n, y^n) \in A_{\varepsilon}^{(n)}$ then
(a) \( \mathbb{P} \left( \left( \tilde{X}_n, \tilde{Y}_n \right) \in A^{(n)}_\varepsilon \right) \leq \frac{-n (I(X \land Y) - 3\varepsilon)}{2} \)

(b) \( \mathbb{P} \left( \left( \check{X}_n, \check{Y}_n \right) \in A^{(n)}_\varepsilon \right) \geq (-\varepsilon) \frac{2}{2} \) as \( n \to \infty \).

**Proofs for (i), (ii), (iii) in class #10 & book.**

For (iv) \( P_{X \to Y} (\tilde{x}, \tilde{y}) = P_{X \to Y} (\check{x}, \check{y}) = P_X (\tilde{x}) P_Y (\tilde{y}) \) is given.

\( \mathbb{P} \left( \left( \tilde{X}_n, \tilde{Y}_n \right) \in A^{(n)}_\varepsilon \right) = \sum \mathbb{P}_{X \to Y} (\tilde{x}_n) \mathbb{P}_{Y \to Y} (\tilde{y}_n) \left( \tilde{x}_n, \tilde{y}_n \right) \in A^{(n)}_\varepsilon \)

(Note: \( \mathbb{P}_{X \to Y} (\tilde{x}_n) \Rightarrow \text{it is } \varepsilon \text{-typical in marginals too!} \)

\( \Rightarrow \mathbb{P} \left( \left( \tilde{X}_n, \tilde{Y}_n \right) \in A^{(n)}_\varepsilon \right) \leq \sum \frac{-n (H(X) - \varepsilon) - n (H(Y) - \varepsilon)}{2} \)

\( \Rightarrow \mathbb{P} \left( \left( \tilde{X}_n, \tilde{Y}_n \right) \in A^{(n)}_\varepsilon \right) \leq \frac{2}{2} \left( \frac{-n (I(X \land Y) - 3\varepsilon)}{2} \right) \)

Also \( \mathbb{P} \left( \left( \check{X}_n, \check{Y}_n \right) \in A^{(n)}_\varepsilon \right) \)

\( = \sum \mathbb{P}_{X \to Y} (\check{x}_n) \mathbb{P}_{Y \to Y} (\check{y}_n) \left( \check{x}_n, \check{y}_n \right) \in A^{(n)}_\varepsilon \)

\( \Rightarrow \sum \frac{-n (H(X) + \varepsilon) - n (H(Y) + \varepsilon)}{2} \)

\( \Rightarrow \mathbb{P} \left( \left( \check{X}_n, \check{Y}_n \right) \in A^{(n)}_\varepsilon \right) \geq \frac{2}{2} \left( \frac{n (H(X, Y) - 3\varepsilon)}{2} \right) \)

\( \Rightarrow \mathbb{P} \left( \left( \check{X}_n, \check{Y}_n \right) \in A^{(n)}_\varepsilon \right) \geq \frac{2}{2} \left( \frac{n (H(X, Y) + 3\varepsilon)}{2} \right) \)

\( \Rightarrow \mathbb{P} \left( \left( \tilde{X}_n, \tilde{Y}_n \right) \in A^{(n)}_\varepsilon \right) \geq (-\varepsilon) \frac{2}{2} \)

\( \Rightarrow \mathbb{P} \left( \left( \check{X}_n, \check{Y}_n \right) \in A^{(n)}_\varepsilon \right) \geq (-\varepsilon) \frac{2}{2} \)

\( \Rightarrow \mathbb{P} \left( \left( \tilde{X}_n, \tilde{Y}_n \right) \in A^{(n)}_\varepsilon \right) \geq (-\varepsilon) \frac{2}{2} \)

\( \Rightarrow \mathbb{P} \left( \left( \check{X}_n, \check{Y}_n \right) \in A^{(n)}_\varepsilon \right) \geq (-\varepsilon) \frac{2}{2} \)

\( \Rightarrow \mathbb{P} \left( \left( \tilde{X}_n, \tilde{Y}_n \right) \in A^{(n)}_\varepsilon \right) \geq (-\varepsilon) \frac{2}{2} \)

\( \Rightarrow \mathbb{P} \left( \left( \check{X}_n, \check{Y}_n \right) \in A^{(n)}_\varepsilon \right) \geq (-\varepsilon) \frac{2}{2} \)
\[ P \left( (\tilde{x}^n, \tilde{y}^n) \in A_\epsilon^{(n)} \right) \geq (1-\epsilon) 2^{-n(I(x; y) + 2\epsilon)} \]

WHENEVER \( \tilde{x}^n \) AND \( \tilde{y}^n \) ARE INDEPENDENT
AND HAVE THE PROPERTY THAT \( P_{\tilde{x}^n \tilde{y}^n} = \prod \frac{P_x(x^n) P_y(y^n)}{2^n} \)

BY SETTING \( n \to \infty \), WE CAN SET

\[ P \left( (\tilde{x}^n, \tilde{y}^n) \in A_\epsilon^{(n)} \right) \to 0 \]

NOTE IF \( (\tilde{x}^n, \tilde{y}^n) \sim P_{\tilde{x}^n}(\tilde{x}^n) P_{\tilde{y}^n}(\tilde{y}^n) \) THEN

\[ (\tilde{x}, \tilde{y}) \sim P_x(x^n) P_y(y^n), \quad (\tilde{x}^2, \tilde{y}^2) \sim P_{x^2}(\tilde{x}^2) P_{y^2}(\tilde{y}^2) \]

\[ \ldots \]

\[ \text{NOTE 1. THERE ARE } 2^n H(x) \text{ E-TYPICAL } x^n \text{ SEQUENCES} \]

\[ \text{2. THERE ARE } 2^n H(y) \text{ E-TYPICAL } y^n \text{ SEQUENCES}. \]

\[ \text{3. SIZE OF } (\tilde{x}^n, \tilde{y}^n) \text{ WHERE EACH IS E-TYPICAL} \]

\[ = 2^n H(x) \cdot 2^n H(y) = 2^n (H(x) + H(y)) \]

\[ \text{4. SIZE OF JOINTLY TYPICAL } n H(x, y) \]

\[ (x^n, y^n) \text{ PAIR IS EQUAL TO } 2 \]

BY DEFINITION OF \( A_\epsilon^{(n)} \) EVERY JOINTLY

TYPICAL PAIR IS MARGINAL TYPICAL.

\[ \text{i.e. } n (H(x) + H(y)) \geq n H(x \| y), \quad P_{xy} = P_x P_y \]

\[ \text{EQUALITY IF } x \perp y \]

Q1: HOW MANY JOINTLY TYPICAL SEQUENCES ARE THERE FOR A GIVEN \( x^n \)?

IF \( \alpha \) IS THE \# OF (JOINTLY TYPICAL) \( y^n \)

THAT ARE JOINTLY TYPICAL WITH \( x^n \) THEN

\[ \alpha \cdot 2^{n H(x)} \cdot 2^{n H(x, y)} \]

\[ = 3 \]

\[ \Rightarrow \alpha = 2^{n H(x | y) - H(x)} = 2^{n H(y | x)} \]

So FOR EACH \( x^n \), THERE ARE \( 2^n H(y | x) \) JOINTLY TYPICAL \( y^n \) SEQUENCES.
Another way to look at it:

Given \( x^n \), the \( y^n \) is produced with PMF

\[
P(y_i | x^n) = \frac{1}{m} P(y_i | x_i)
\]

Hence there are \( 2^n H(y | x) \)

\[
= 2^n \cdot \frac{1}{n} \sum_{i=1}^{n} H(y_i | x_i)
\]

Hence there are at most \( 2^n H(y | x) \) typical \( y^n \)

Note: Given set of sequences \((x^n, y^n)\), the size of the set where either \( x^n \) or \( y^n \) is typical is \( 2^n (H(x) + H(y)) \). There are \( 2^n H(x, y) \) elements that have \((x^n, y^n)\) jointly typical. Hence the relative frequency or the probability of picking a jointly typical pair \((x^n, y^n)\) is

\[
\frac{n H(x, y)}{2^n (H(x) + H(y))} = 2^n I(x, y)
\]

\[
= \frac{1}{2^n (H(x) + H(y))}
\]

Main idea of channel coding is that given \( x^n \) typical, only those \( y^n \) that are jointly typical with \( x^n \) will occur with high probability.

In general, a given \( y^n \) may be jointly typical with more than one \( x^n \). This leads to decoding error! Need to be avoided.

As long as \( B_1 \cap B_1 = \emptyset \), we can decode \( x^n \).

** Aim:** Pick \( x^n \)'s such that \( B_1 \cap B_2 = \emptyset \).
What is the maximal number of typical sequences that we can pick with no overlap in Bi's?

An: For each \( x^n \), there are \( 2^{n H(y|x)} y^n \) that are jointly typical with that \( x^n \). Totally \( 2^{n H(Y)} y^n \) sequences are \( \epsilon \)-typical.

So we can have \( \frac{2^n H(Y)}{2^n H(Y|x)} = 2^{n I(X;Y)} \)

\[ 2^n I(X;Y) \leq 2^{n I(X;Y)} \]

\( \{ \text{that are jointly typical (with } y^n) \} \) that are disjoint.

Note for \( 2^n I(X;Y) \) we need \( \log 2^{n I(X;Y)} \) or \( n I(X;Y) \) bits for representation. We use channel \( n \) times. So the rate of the code representing \( x^n \) is \( \frac{n I(X;Y)}{n} = I(X;Y) \)

Some Definitions

(4) An \((M, n)\) code for channel \((X, P(y|x), M)\) consists of

(4(i) An (message) index set \( \{1, 2, \ldots, M\} \)

(4(ii) Encoding function \( X^n : \{1, 2, \ldots, M\} \rightarrow X^n \)

Message is mapped into sequence of length \( n \)

Set of the sequences \( \{ X^n(1), X^n(2), \ldots, X^n(M) \} \)

is called codebook. \( X^n(1) \) is a codeword.

(3) A decoding function \( \hat{q} : X^n \rightarrow \{1, 2, \ldots, M\} \)

A rule for estimation/assignment of known form.
(b) DEFINITION OF PROBABILITY OF ERROR.

**Define** \( \delta (g(y^n) \neq i) = \sum_{i \neq g(y^n)} p_i \)

**Let** \( \lambda_i = P_r (g(y^n) \neq i \mid x^n = x^n(i)) \)

\[
\lambda_i = \sum_{y^n} p(y^n \mid x^n(i)) \delta (g(y^n) \neq i)
\]

(c) **MAXIMAL PROBABILITY OF ERROR** \( \lambda^{(n)} \) **FOR** \((M, n)\) **CODE** **IS DEFINED AS**:

\[
\lambda^{(n)} = \max_{i \in \{1, \ldots, M\}} \lambda_i.
\]

(d) **AVERAGE PROBABILITY OF ERROR** \( P_{e}^{(n)} \) **FOR AN** \((M, n)\) **CODE** **IS DEFINED AS**

\[
P_{e}^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i = P_r (g(y^n) \neq i) \quad \text{if} \quad p(i) = 1/M.
\]

\[
P_{e}^{(n)} \leq \lambda^{(n)}
\]

(e) **RATE** \( R \) **OF A CODE** \((M, n)\) **IS**

\[
R = \frac{\log M}{n} \quad \text{BITS PER XMISSION}.
\]

(f) A RATE \( R \) IS **SAID TO BE ACHIEVABLE IF** **THERE EXISTS A SEQUENCE OF** \( (\left\lceil 2^{nR}, n \right\rceil) \) **CODING** **SUCH THAT** **MAXIMAL** **PROBABILITY** **OF ERROR** \( \lambda^{(n)} \downarrow 0 \) **AS** \( n \to \infty \).
Note that if there are $2^nR$ messages, then there are $2^nR$ codewords. Decoder is in error if there is no $x^n$ that is jointly typical with $y^n$ or more than one $x^n$ is jointly typical with $y^n$.

\[ 2^nR \leq -nI(x^n;y^n) \]

\[
\text{Probability of error} = \sum_{m_1 \neq m_l} \mathbb{P}\left(\left(\frac{x^n(m_1) \sim y^n}{x^n(m_l) \sim y^n} \in \mathcal{A}_n^{(n)}\right) \mid x^n(m) \text{ is sent}\right)
\]

\[
= \sum_{m_1 \neq m_l} 2^{-nI(x^n;y^n)} \leq 2^{-n(I(x^n;y^n) - R)} \rightarrow 0 \text{ as } n \rightarrow \infty, \text{ if } R \leq I(x^n;y^n)
\]

If $R < I(x^n;y^n)$, then probability of error $\rightarrow 0$.

**Channel Coding Theorem**

All rates $R \leq \frac{\log M}{n}$ below capacity ($= \max_{p(x)} I(x^n;y^n)$) are achievable. For every $R < C$, there exists sequence of codes $(2^n, n)$ such that $A^n \rightarrow 0$ as $n \rightarrow \infty$.

Recall $A^n = \max_{\hat{W} \in \hat{W}^n} \left\{ x^n : x^n \in \hat{W} \right\}$

Conversely, for $(2^n, n)$ codes with $A^n \rightarrow 0$, all rates $R \leq C$ must hold.
ZERO ERROR CASE $A^{(n)} = 0$ (DMC)

Converse:

\[
W \xrightarrow{\text{Source}} E \xrightarrow{\text{Encode}} X^n \xrightarrow{\text{Channel}} Y^n \xrightarrow{\text{Decode}} \hat{W}
\]

Recall data processing inequality:

\[
n_R = I(W \wedge Y^n) \leq I(X^n \wedge Y^n)
\]

\[
= H(Y^n) - H(Y^n | X^n)
\]

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

\[
= \sum_{i=1}^{n} I(X_i ; Y_i)
\]

\[
\leq n C
\]

\[
\Rightarrow R \leq C
\]
Proof of converse when $P_e^{(n)} \neq 0$

We need to recall some old inequalities (Fano's)

From Chapter 2. Note the chain below.

\[ W \xrightarrow{\alpha} X^n \xrightarrow{\beta} Y^n \xrightarrow{\gamma} W \]

Note $X^n$ is a function of $W$.

Hence

\[ H(X^n, W) = H(W) + H(X^n | W) \]

\[ = H(X^n) + H(W | X^n) \]

\[ \Rightarrow H(W) \geq H(X^n) \quad \text{with Eq 11} \quad \text{iff} \quad X^n \text{ uniquely determines} \quad W. \]

Also

\[ I(W, X^n, Y^n) \leq I(W, Y^n) + I(X^n | Y^n | W) \]

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

\[ \Rightarrow I(X^n | Y^n) \leq I(W | Y^n) \]

\[ \Rightarrow H(X^n) - H(X^n | Y^n) \leq H(W) - H(W | Y^n) \]

\[ \Rightarrow H(W) - H(X^n) \leq H(W | Y^n) - H(X^n | Y^n) \]

\[ \Rightarrow H(W) - H(X^n) \geq 0 \]

\[ \Rightarrow 0 \leq H(W | Y^n) - H(X^n | Y^n) \]

\[ \Rightarrow H(X^n | Y^n) \leq H(W | Y^n) \]

Now for error part & Fano's Ineq.

Let

\[ E = \begin{cases} 1 & \text{if} \; W \neq \hat{W} \\ 0 & \text{if} \; W = \hat{W} \end{cases} \]

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

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

\[ \exists H(W | Y^n) \leq H(E) + P(E = 0) \cdot H(E = 0 | Y^n) + P(E = 1) \cdot H(W | Y^n, E = 1) \]

\[ \leq 1 + P_e^{(n)} \frac{1}{n} \log \frac{1}{P_e^{(n)}}(1 - P_e^{(n)}) \leq 1 + P_e^{(n)} \cdot nR. \]
Hence

\[ H(X^n | Y^n) \leq H(W | Y^n) \leq 1 + P_e^{(n)} \cdot R \]

Now the proof of converse.

For codes \((2^{nR}, n)\) with \(A^{(n)} \to 0\) \(R \leq C\).

**Main Idea**: 
\[ A^{(n)} = \max_{i \in \{1, \ldots, m\}} a_i^n \Rightarrow A^{(n)} \to 0 \Rightarrow A_i \to 0 \]

\[ P_e^{(n)} = P_r \{ W \neq \hat{W} \} \to 0 \text{ as } A^{(n)} \to 0 \]

\[ I(W \wedge Y^n) \leq I(X^n \wedge Y^n) \]

\[ I(W \wedge Y^n) = H(W) - H(W | Y^n) \]

\[ I(X^n \wedge Y^n) = H(W) - H(W | Y^n) \]

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

\[ \leq H(W | Y^n) + I(X^n \wedge Y^n) \]

\[ \leq 1 + P_e^{(n)} \cdot nR + nC \]

\[ \Rightarrow nR \leq 1 + P_e^{(n)} \cdot nR + nC \]

\[ \Rightarrow R \leq \frac{1}{n} + P_e^{(n)} R + C \]

Now, if \(A^{(n)} \to 0\) as \(n \to \infty\), \(P_e^{(n)} \to 0\), \(R\) is finite.

\[ \Rightarrow R \leq C \]

Also, \(P_e^{(n)} \geq 1 - \frac{1}{nR} - \frac{C}{R} \)

\[ P_e \]

We do not know how \(P_e^{(n)}\) grows.

\[ C \]
LAST TIME:
A QUESTION AROSE IN DEF OF THEP & MOTIVATION.
WILL RESOLVE THIS POINT ON 25TH. HENCE FORWARD
PART OF CHANNEL CODING WILL BE DONE LATER.
TODAY CONVERSE.
(i) RECALL DEFINITIONS OF PROBABILITY OF ERROR
(ii) ZERO ERROR CASE.
(iii) FANO'S INEQUALITY IN THIS CONTEXT.
(iv) PROOF OF CONVERSE.
(v) CAPACITY WITH FEEDBACK. CFB ≥ C; C ≥ CFB
   LEADING TO CFB = C FOR DMC.

---

\[ \begin{array}{c}
 W \xrightarrow{\text{ENCODING}} X^n \xrightarrow{\text{CHANNEL}} Y^n \xrightarrow{\text{DECODING}} W
\end{array} \]

AN \((M, n)\) CODE FOR CHANNEL \((X, \mathcal{Y}, p(y|x))\) CONSISTS OF
(i) A MESSAGE INDEX SET \(\{1, 2, \ldots, M\}\), W.G.S.
(ii) ENCODING FUNCTION \(X^n: \{1, \ldots, M\} \rightarrow \mathcal{X}^n\)
\[ X^n(i) = \{X_n^i, \ldots, X_1^i\} \]
WITH \(X^n(i)\) A CODE-WORD, \(\{X_n^i, \ldots, X_1^i\}\) CODE-WORD.
(iii) A DECODING FUNCTION (PRIOR RULE KNOWN TO RECEIVER)
\[ q: \mathcal{Y}^n \rightarrow \{1, \ldots, M\} \]

PROBABILITY OF ERROR BASED ON MESSAGE XMISSION.
\[ \lambda_i = \Pr \{ q(y^n) \neq i \mid X^n = X^n(i) \} \]
i.e. PROBABILITY THAT DECODING LEADS TO INDEX OTHER THAN \(i\),
GIVEN THAT XMITTED INDEX WAS \(i\). \(i \in \{1, \ldots, M\}\)

\[ \lambda_i^{(n)} = \max_{i \in \{1, \ldots, M\}} \lambda_i \]
MAX PROBABILITY OF ERROR FOR \((M, n)\) CODE.
AVERAGE PROBABILITY OF ERROR FOR (MIN) CODE IS DEFINED AS

\[ P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i \]

WHEN ALL INPUTS ARE EQUALLY LIKELY,

\[ P_e^{(n)} = P_i \{ g(y^n) \neq w \} = \sum_{i=1}^{M} P_i \{ g(y^n) \neq i | x^n = x^n(i) \} \times P_i \{ x^n = x^n(i) \} \]

ALSO NOTE \( \min_{i \in \{1, \ldots, M\}} \lambda_i \leq P_e^{(n)} \leq \max_{i \in \{1, \ldots, M\}} \lambda_i = \lambda_i \)

TRIVIAL BUT WILL USE LATER FOR FORWARD PROOF TO!

RATE OF AN \((M,n)\) CHANNEL CODE IS DEFINED AS

\[ R = \frac{\log M}{n} \text{ BITS PER XMISSION.} \]

A RATE \( R \) IS SAID TO BE ACHIEVABLE IF THERE EXISTS A SEQUENCE OF \( (2^{nR}, n) \) CHANNEL CODE FOR WHICH \( J(n) \to 0 \) AS \( n \to \infty \)

CHANNEL CODE & 2^{nd} THEOREM.

ALL RATES \( R (\leq \log M/n) \) THAT ARE BELOW CAPACITY \( C (= \max_{p(x)} I(X;Y)) \) ARE ACHIEVABLE.

FOR EVERY RATE \( R < C \), THERE EXISTS \( (2^{nR}, n) \) CODE SEQUENCE FOR WHICH \( J(n) \to 0 \) AS \( n \to \infty \).

CONVERSELY, ANY SEQUENCE OF \( (2^{nR}, n) \) CODES WITH \( J(n) \to 0 \) (AS \( n \to \infty \)) MUST HAVE \( R \leq C \).
\[ P_x (\text{ERR}) = P_x \{ W \neq \hat{W} \} \xrightarrow{\Delta} 0 \text{ as } n \to \infty. \text{ Then } R \leq C \]

\[
\begin{align*}
W & \xrightarrow{E} X^n(W) \xrightarrow{p(y|x) \text{ DMC}} Y^n \xrightarrow{D} \hat{W} \quad \text{DATA PROCESSING IN "\textit{q}".} \\
\text{NOTE } & W \rightarrow X^n(W) \rightarrow Y^n \Rightarrow I (W \wedge Y^n) \leq I (X^n(W) \wedge Y^n)
\end{align*}
\]

Mutual information between \( W \) and \( Y^n \) is

\[
I (W \wedge Y^n) = H(W) - H(W | Y^n) \quad \frac{M = 2^n R}{\delta R} \Rightarrow R = \frac{\log M}{n}
\]

Simple case \( P_x \{ W \neq \hat{W} \} = 0 \), \( \hat{W} = q(Y^n) \text{ DECODING RULE} \)

\[
\Rightarrow H(W | Y^n) = 0
\]

\[
\Rightarrow H(W) = I(W \wedge Y^n) \leq I(X^n(W) \wedge Y^n)
\]

\[
I(W \wedge Y^n) = H(Y^n) - \sum_{i=1}^{M} H(Y_i | Y_i^{-1}, X^n(W))
\]

\[
= H(Y^n) - \sum_{i=1}^{M} H(Y_i | Y_i^{-1}, X_i(W)) \quad \text{DMC}
\]

\[
\leq \sum H(Y_i) - \sum H(Y_i | X_i(W)) \quad \text{INDEPENDENCE BOUND ON } Y_i
\]

\[
\leq \sum I(Y_i \wedge X_i(W)) \quad c = \max \frac{I(X \wedge Y_i)}{p(x)}
\]

\[
I(W \wedge Y^n) \leq n C \quad \text{--- (2)}
\]

By but \( H(W) \leq \log 2^{n R} = n R \). Set \( P_w \sim \frac{1}{2^{n R}} \) to get

\[
\Rightarrow H(W) = n R \leq n C \quad \text{OR} \quad R \leq C \quad \text{for zero error case.}
\]
Case \( P_r \{ W = \hat{W} \} \downarrow 0 \) as \( n \to \infty \). Set \( P_w = \frac{1}{2} n \Rightarrow H(w) = nR \) 4

\[ (\exists) \quad H(w) = H(w|Y^n) + I(w \wedge Y^n) \]

\[ (\exists) \quad H(w) \leq H(w|Y^n) + nC \] — 3

\[ \Rightarrow \quad nR \leq H(w|Y^n) + nC \] — 4

We need to bound \( H(w|Y^n) \)

**Fano's Inequality, Main Idea: Define Error Variable**  
\[ E = \begin{cases} 1 & w \neq \hat{w} \\ 0 & w = \hat{w} \end{cases} \quad E \text{ is Binary RV} \]

\[ P_r(E) = P_r(w \neq \hat{w}) \]

\[ H(E, w|Y^n) = H(E|Y^n) + H(w|E, Y^n) \] — 5

\[ = H(w|Y^n) + H(E|w, Y^n) \] — 6

\[ \Rightarrow \quad \text{given } w, Y^n \text{ we know } w, \hat{w} = f(Y^n) \]

\[ \Rightarrow \quad \text{know } E \Rightarrow H(E|w, Y^n) = 0 \]

\[ \Rightarrow \quad H(w|Y^n) = H(E, w|Y^n) = H(E|Y^n) + H(w|E, Y^n) \]

\[ = H(E) \leq 1 \leq 1 + H(w|E, Y^n) \] — 7

\[ H(w|Y^n) = 1 + H(w|E = 0, Y^n) P_r(E = 0) + H(w|E = 1, Y^n) P_r(E = 1) \]

\[ E = 0 \Rightarrow w = \hat{w} \Rightarrow H(w|E = 0, Y^n) = H(w|w, Y^n) = 0 \] — 8

\[ E = 1 \Rightarrow w \neq \hat{w} \Rightarrow H(w|E = 1, Y^n) \leq \log_2(2^n - 1) \leq nR \]

When \( E = 1 \): You know \( w \neq \hat{w} \) but you still have \( 2^n - 1 \) values from which you need to choose.
\( H(W|Y^n) \leq 1 + n \cdot Pr \{ \hat{Y} = 0 \} + nR \cdot Pr \{ \hat{Y} = 1 \} \)

\[
H(W|Y^n) = 1 + nR \cdot Pr \{ W \neq W^\hat{Y} \}
\]

Substituting 9 into 7 gives:

\[
H(W) = nR \leq 1 + nR \cdot Pr \{ W \neq W^\hat{Y} \} + nC
\]

Or

\[
R \leq 1 + \frac{nC}{n} R \cdot Pr \{ W \neq W^\hat{Y} \} + C
\]

As \( n \to \infty \), \( Pr \{ W \neq W^\hat{Y} \} \to 0 \), \( \frac{1}{n} \to 0 \)

\[
\Rightarrow \lim_{n \to \infty} R \leq C
\]

(NOT \( \Rightarrow R \leq C \)) Actually only shows \( R \leq C + C\) in the long run.

Read section 8.10 for case of equality (similar to zero error case)

Feedback capacity \( C_{FB} \)

Model:

\[
\begin{align*}
  W & \xrightarrow{E} X_i = f(W, Y_{i-1}) \\
  Y_i, X_i & \xrightarrow{\text{Noiseless channel}} Y^n \\
  D & \xrightarrow{\hat{W}} W
\end{align*}
\]

1. \( X_i \) is a function of \( W, Y_{i-1}, \ldots, Y_{i-1} \)
2. \( Y_i \) depends on \( X_i \) given \( W, Y_{i-1}, \ldots, Y_{i-1}, X_i, Y_i \)
PROOF TECHNIQUE: SHOW $C_{FB} \geq C$ AND $C_{FB} \leq C$ AND CONCLUDE $C_{FB} = C$

To show $C_{FB} \geq C$, since channel without feedback is a special case of channel with feedback, any rate achieved without feedback (no further info) should be achievable with feedback.

$\Rightarrow [C_{FB} \geq C]$

To show $C_{FB} \leq C$, assume $W \sim \frac{1}{2^{nr}}$

RECALL

$I(\bar{W} \wedge \bar{Y}) = H(W) - H(W | Y)$

AND $H(W) = \underbrace{H(W | Y)}_{nR} + I(\bar{W} \wedge \bar{Y})$

$\Rightarrow nR \leq 1 + nR P_{Y} P_{W} + I(\bar{W} \wedge \bar{Y})$

$I(\bar{W} \wedge \bar{Y}) = H(Y) - H(Y | W)$

$= H(Y) - \sum H(Y_i | W, Y_{i-1})$

$= H(Y) - \sum H(Y_i | W, Y_{i-1}, X_i)$

$\leq \sum H(Y_i) - \sum H(Y_i | X_i)$

$I(\bar{W} \wedge \bar{Y}) = \sum I(X_i \wedge Y_i) \leq nC$
\[ nR \leq 1 + nR \mathbb{P}\{W = \hat{W}\} + nc \]

or

\[ R \leq \frac{1}{n} + R \mathbb{P}\{W = \hat{W}\} + c \]

if \( R \mathbb{P}\{W = \hat{W}\} \downarrow 0 \) as \( n \uparrow \infty \), \( R \leq c \)

\( \Rightarrow \) **WE CANNOT GET A HIGHER RATE THAN C.**

\( \Rightarrow \) \( CF_B \leq c \). **BUT WE SHOWED CF_B \geq c**

\( \Rightarrow \) \( \boxed{CF_B = c} \)
LAST TIME:
1. PROVED CONVERSE TO CHANNEL CODING (DNC)
   i.e. FOR SEQUENCE OF CODES $(2^n, n)$ WITH $\varepsilon(n) \to 0$
   $(\text{MAX PROB OF ERROR})$ MUST HAVE $R < C$.
   
   \[ R = \frac{\log M}{n}, \quad C = \max I(x; y) \]

   MAIN IDEA WAS
   
   \[ I(w^n, \hat{w}) \leq I(w^n, y^n) \] & (FANO'S INEQ)
   
   \[ H(w) \leq H(w, \hat{w}) + I(w^n, \hat{w}) \]
   
   \[ \leq 1 + \varepsilon(n) nR + I(w^n, \hat{y}^n) \leq 1 + \varepsilon(n) nR + nC. \]

   BY CHOOSING
   
   \[ W \sim 2^{-nR} \text{ at max val of } H(W) = n R \]
   
   \[ nR \leq 1 + nR \varepsilon(n) + nC \]
   
   \[ R \leq \frac{1}{\ln M} + R \varepsilon(n) + C \]

   IF $\varepsilon(n) \to 0$ AS $n \to \infty$
   
   THEN $X_R \to 0$, $\ln M \to 0$
   
   \[ \Rightarrow \quad R \leq C + n \varepsilon(n) \to 0 \] AS $n \to \infty$.

   ALSO SHOWN FEEDBACK DOES NOT INCREASE CAPACITY.
   FOR DNC.

---

TODAY: ALL RATES BELOW CAPACITY ARE ACHIEVABLE.

SPECIFICALLY FOR EVERY $R < C$, THERE EXISTS A

SEQUENCE OF $(2^n, n)$ CODES FOR WHICH $\varepsilon(n) \to 0$ AS

\[ n \to \infty. \]

IN PROVING THIS, NO SPECIFIC CODE IS GENERATED

WE ASSUME ONE CAN GENERATE A CODE OF LENGTH $n$ CODEWORDS.

* NOTE THE FOLLOWING IF $X^n(i) \perp X^n(j)$ AND $Y^n$

IS $Y^n = f(X^n(i))$, THEN $\text{i.i.d.}$

THE FOLLOWING HELDS

\[ \text{i.i.d. } \Rightarrow \text{(TRIVIALLY MARKOV)} \]

\[ X^n(i) \perp X^n(j) \perp Y^n \]

\[ I(X^n(i^n), Y^n | X^n(i^n)) = H(X^n(j^n) | X^n(i^n)) - H(X^n(i^n) | Y^n, X^n(i^n)) \]

\[ = H(Y^n | X^n(i^n)) - H(Y^n | X^n(i^n), X^n(j^n)) \]

\[ = 0 \text{, ideally if no errors i.e. } P_{\text{err}}(x) = 0. \]
\[ I(\mathcal{X}^n(i) \land \mathcal{Y}^n, \mathcal{X}^n(j)) \]
\[ = I(\mathcal{X}^n(i) \land \mathcal{X}^n(j)) + I(\mathcal{X}^n(i) \land \mathcal{Y}^n | \mathcal{X}^n(j)) \]
\[ = I(\mathcal{X}^n(i) \land \mathcal{Y}^n) + I(\mathcal{X}^n(i) \land \mathcal{Y}^n | \mathcal{Y}^n) \]
\[ \geq 0 \]
\[ \Rightarrow \text{each} \quad I(\mathcal{X}^n(i) \land \mathcal{Y}^n) = 0 \]
\[ H(\mathcal{X}^n(i)) - H(\mathcal{X}^n(i) | \mathcal{Y}^n) = 0 \]
\[ \text{or} \quad H(\mathcal{X}^n(j)) = H(\mathcal{X}^n(j) | \mathcal{Y}^n) \]

Recall
\[ I(\mathcal{X}^n(j) \land \mathcal{Y}^n) = E \left[ \log \frac{P_{\mathcal{X}^n(j) \mathcal{Y}^n}}{P_{\mathcal{X}^n(j)} P_{\mathcal{Y}^n}} \right] = 0 \]
with the equality if
\[ P_{\mathcal{X}^n(j) \mathcal{Y}^n} = P_{\mathcal{X}^n(j)} P_{\mathcal{Y}^n} \]

So what we have is \( \mathcal{X}^n(i) \) transmitted to gain \( \mathcal{Y}^n \)

User may do something to decode. But \( \mathcal{Y}^n = f(\mathcal{X}^n) \)

If \( \mathcal{X}^n(i) \perp \mathcal{X}^n(j) \), we have \( \mathcal{X}^n(i) \perp \mathcal{Y}^n \)

Leading to \( P(\mathcal{X}^n(j), \mathcal{Y}^n) = P_{\mathcal{X}^n(j)} P_{\mathcal{Y}^n} = P_{\mathcal{X}^n(j)} P_{\mathcal{Y}^n} \).

Again, proof does not generate specific code.

Steps:
1. Construct a codebook \( C \) of \( (m=2^n, n) \) code,
   generating codewords in \( \mathcal{X}^n \) with distribution
   \[ p(\mathcal{X}^n) = \prod_{i=1}^{n} p(\mathcal{X}_i) \] (i.e. each \( \mathcal{X}, \mathcal{X}_i \sim p(\mathcal{X}) \)).
   \[ \mathcal{X}(i) = (\mathcal{X}_1(i), \mathcal{X}_2(i), \ldots, \mathcal{X}_n(i)) \quad i = 1, \ldots, m = 2^n \]
2. Both sender and receiver know \( C \).
3. Pick a message \( \mathcal{W} \), uniformly \( \sim 1_M = 1/2^{nk} \).
4. Xmit over channel \( \mathcal{X}^n(\mathcal{W}) = (\mathcal{X}_1(\mathcal{W}), \mathcal{X}_2(\mathcal{W}), \ldots, \mathcal{X}_n(\mathcal{W})) \).
$$P(Y^n = y^n | X^n(w) = x^n) = \prod_{i=1}^{n} P(y_i | x_i) \quad \text{DMC}.$$ 

5. **Decoding.** If \((X^n(\hat{w}), Y^n) \in A_{\epsilon}^{(n)}\) and there is no other \(w' \neq \hat{w}\) with \((X^n(w'), Y^n) \in A_{\epsilon}^{(n)}\) then declare \(\hat{g}(y^n) = \hat{w}\) is the transmitted msg.

If \((X^n(\hat{w}), Y^n) \notin A_{\epsilon}^{(n)}\) or there are more than one \(\hat{w}\) such that \(\hat{w}\) pairs belong to typical set, declare error.

Let \(E_{\epsilon} = \{ \hat{w} \neq \hat{w} \}_{\epsilon}\)

\[E_{\epsilon} = \left\{ (X^n(w), Y^n) \in A_{\epsilon}^{(n)} \right\}\]

\[P_r(E_{\epsilon}) = \sum_{w=1}^{M} \left[ P_r(e_{\epsilon}^{(n)} | W = w) \cdot P(W = w) \right] \]

\[= \sum_{\epsilon \in \mathcal{C}} P_r(\hat{w} \neq \hat{w} | C_{\epsilon}) \cdot P_{r}(C_{\epsilon}) \]

\[= \frac{1}{M} \sum_{W = 1}^{M} \sum_{\epsilon \in \mathcal{C}} P_r(\hat{w} \neq w | W = w, C_{\epsilon} \neq P_r(\epsilon)) \]

\[= \frac{1}{M} \sum_{\epsilon \in \mathcal{C}} \left( \sum_{W = 1}^{M} P_{r}(\hat{w} \neq w | W = w, C_{\epsilon} \neq P_r(\epsilon)) \right) \]
\[ P_r \in \mathcal{E} | \mathcal{F} = \frac{1}{M} \sum_{w=1}^{M} \sum_{\mathcal{F}} P_r \in \mathcal{E} \wedge \mathcal{W} \wedge \mathcal{W} = w, \mathcal{W} = \mathcal{W}, \mathcal{F} \]

\[ P_r(\mathcal{C}) = \prod_{i=1}^{n} \prod_{w=1}^{M} P(X_i(w)) \uparrow \downarrow \text{INDEP OF } \mathcal{W} \]

\[ P_r \in \mathcal{E} | \mathcal{F} = \frac{1}{M} \sum_{w=1}^{M} \left( \sum_{\mathcal{F}} P_r \in \mathcal{E} \wedge \mathcal{W} \wedge \mathcal{W} = w, \mathcal{W} = \mathcal{W}, \mathcal{F} \right) \]

\[ = P_r \in \mathcal{E} | \mathcal{W} = 1 \]

So we have averaged probability of error over all codebooks.

Let's compute decoding error terms.

\( W = 1 \) was sent.

1. \( E_i^C = \mathcal{E}^C(X^n(i), Y^n) \notin \mathcal{A}^{(n)C} \)

Then \( E_i^C = \mathcal{E}^C(X^n(i), Y^n) \notin \mathcal{A}^{(n)C} \).

Error occurs if \( E_i^C \) occurs or \( Y^n \) is not jointly typical with \( X^n(i) \), \( i \neq 1 \), \( i \leq 2, 3, \ldots, M \).

Hence

\[ P_r \in \mathcal{E} | \mathcal{W} = 1 \]

\[ \leq \sum_{i=2}^{M} P_r \in \mathcal{E} | \mathcal{F} \]

\[ \leq \varepsilon. \]

\[ = \prod_{i=2}^{M} \left( \frac{1}{M} \right) \]

\[ \leq \varepsilon + \frac{1}{M} \left( \mathbb{H}(X^n\cap Y^n) - 3\varepsilon \right) \]

\[ \leq \varepsilon + \frac{1}{M} \left( \mathbb{H}(X^n\cap Y^n) - 3\varepsilon \right) \]

\[ \leq \varepsilon + \frac{1}{M} \left( \mathbb{H}(X^n\cap Y^n) - 3\varepsilon \right) \]

\[ \leq \varepsilon + \frac{1}{M} \left( \mathbb{H}(X^n\cap Y^n) - 3\varepsilon \right) \]

\[ \leq \varepsilon + \frac{1}{M} \left( \mathbb{H}(X^n\cap Y^n) - 3\varepsilon \right) \]

\[ \leq \varepsilon + \frac{1}{M} \left( \mathbb{H}(X^n\cap Y^n) - 3\varepsilon \right) \]

\[ \leq \varepsilon + \frac{1}{M} \left( \mathbb{H}(X^n\cap Y^n) - 3\varepsilon \right) \]

\[ \leq \varepsilon + \frac{1}{M} \left( \mathbb{H}(X^n\cap Y^n) - 3\varepsilon \right) \]

\[ \leq \varepsilon + \frac{1}{M} \left( \mathbb{H}(X^n\cap Y^n) - 3\varepsilon \right) \]
\[ P_{x} \in \mathbb{E}_{2} | W=1 \]

\[ = P_{x} \sum_{(x^n(2), y^n) \in A^{(n)}} P_{z} | W=1 \]

\[ = \sum_{(x^n(2), y^n) \in A^{(n)}} P_{z} \cdot P(y^n) \]

\[ \ln \left( \frac{1}{2^n} \right) = -n (H(x, y) + \epsilon) - n (H(x) - \epsilon) - n (H(y) - \epsilon) \]

\[ = \frac{1}{2} \cdot 2 \cdot 2 \]

\[ = -n (H(x) + H(y) - 3 \epsilon) - \frac{n (I(x; y) - 3 \epsilon)}{2} = 2 \]

\[ P_{x} \in \mathbb{E} | W=1 \] = \[ P(\mathbb{E}) \leq \epsilon + (M-1) \frac{1}{2} \]

\[ \leq \epsilon + \frac{n (I(x; y) - R - 3 \epsilon)}{2} \]

As \( n \to \infty \), if \( R < I(x; y) - 3 \epsilon \), then

\[ \frac{n (I(x; y) - R - 3 \epsilon)}{2} \to 0 \] or \( \frac{1}{2} \leq \epsilon \)

\[ \Rightarrow P_{x} \in \mathbb{E} | W=1 \] = \[ P(\mathbb{E}) \to 0 \) as \( n \to \infty \).

So far we showed average probability of error \( \to 0 \) over all possible codebooks.

Need to show \( \gamma^{(n)} \to 0 \) as \( n \to \infty \).

Also want \( R < I(x; y) \) to be \( R < C \)

1. Set \( p(x) = p(x^n) \) where \( C = I(x^n; y) \)
   i.e., \( p(x^n) \) achieves capacity.
   Then we have \( P_{x} \in \mathbb{E} | W=1 \] = \[ P(\mathbb{E}) \to 0 \) as \( n \to \infty \)
   when \( R < C \).

2. Pick the codebook \( C^* \) such that
   \[ P_{x} \in \mathbb{E} | C^* \] \( \leq 2 \epsilon \). Find it by Exhaustive Search...
For \( C^* \), we have

\[
P_y \leq \epsilon | C^* | = \frac{1}{M} \sum_{w=1}^{M} P_{y}^{(w)} w | w \neq w, C^* y \leq 2 \epsilon.
\]

\[
\frac{\text{BEST HALF + WORST HALF}}{2} \leq 2 \epsilon \Rightarrow \text{BEST HALF} \leq 4 \epsilon
\]

\[
\Rightarrow \text{EVEN IF WE THROW AWAY WORST HALF OF}
\]

\[
\text{CODEWORDS, RESULTING CODE WILL HAVE}
\]

\[
\text{SUM OF PROBABILITIES OF ERRORS} < 4 \epsilon.
\]

\[
\Rightarrow \text{MAXIMAL PROBABILITY OF ERROR ON BEST}
\]

\[
\text{HALF OF CODEWORDS} \leq 4 \epsilon.
\]

\[
\Rightarrow \lambda^{(n)}(C^*) \leq 4 \epsilon.
\]

\[
\#\text{OF CODEWORDS} = BM/2 = \frac{nR}{2} = \frac{n(R-1/n)}{2} = 2^n R^\prime = R-1/n.
\]

Hence we have new code with \( \lambda^{(n)} \to 0 \) as \( n \to \infty \),

and rate \( R^\prime = R-1/n < C \).

We don't know what \( C^* \) is! But we know it is OUT THERE.
LAST TIME

Forward part of the Channel Coding Theorem was proved.

\[ W \xrightarrow{x^n} Y^n \quad \hat{W} \]

Statement

All rates below capacity are achievable. Specifically, for \( R < C \), there exists code sequences for which \( x^n \to 0 \text{ as } n \to \infty \).

Some definitions needed in proof.

\[ \mathbb{E} [ error ] = \sum_{w \neq \hat{w}} \mathbb{P} \]

\[ P(\epsilon) = \mathbb{P} \left( \frac{\mathbb{E} [ error ]}{w} \right) \]

\[ \lambda_w = \mathbb{P} \left( \frac{\mathbb{E} [ error ]}{w} \right) \mid w = w_0 \]

\[ \lambda^{(n)} = \max_{w \in \mathcal{C}} \lambda_w \]

\[ P_e^{(n)} = \frac{1}{m} \sum_{i \in \mathcal{C}} \lambda_i = \mathbb{P} \left( \frac{\mathbb{E} [ error ]}{w} \right) \text{ if } w \sim \frac{1}{m} \]

\[ E_w = \sum_{w \in \mathcal{C}} (x^n(w), y^n) \in A^{(n)} \]

\[ E_w^c = \sum_{w \in \mathcal{C}} (x^n(w), y^n) \notin A^{(n)} \]

For DMC, we also showed that if \( y^n \) was output when \( x^n(i) \) was sent \& \( x^n(i) \equiv x^n(j) \) then \( y^n \equiv x^n(j) \)

Proof of forward part looks to construct random code, compute average probability of error, averaged over all codebooks, pick the best codebook and shows that max prob of error for such/least one \( n \to 0 \quad i.e. \quad A^{(n)} \to 0 \text{ as } n \to \infty \)

For at least one such book!

Key point in this argument is Joint Typical Decoding, which picks of codebooks from atypical set.
**Proof.**

Fix $p(x)$. Generate 2nd sequence of length $Mn$, where $M = 2^n R$.

$$X_1^{Mn} = (x_1 x_2 \ldots x_n | x_{n+1} \ldots x_{2n} | x_{2n+1} \ldots x_{Mn})$$

Generate $M$, $n$-length sequences.

$$
\begin{bmatrix}
X_1^n \\
X_2^n \\
\vdots \\
X_M^n
\end{bmatrix} =
\begin{bmatrix}
x_1 x_2 \ldots x_n \\
x_{n+1} \ldots x_{2n} \\
\vdots \\
x_{(M-1)n+1} \ldots x_{Mn}
\end{bmatrix}
$$

Assign/call each $n$-length sequence as codeword and $Mn$ sequence as codebook $C = (X_1^n, X_2^n, X_M^n)^T$.

$$p(C) = \prod_{i=1}^{Mn} p(x_i) = \prod_{j=1}^{M} p_{X_j^n}(x_j^n)$$

Note that we can generate $(Mn)!$ codebooks all having equal probability.

Q: How many ways in which we can arrange $X_1^{Mn}$?

A: $(Mn)!$ ways.

Q: For all possible codebooks, each has $(Mn)!$ equal probability ones.

Pick one partition from there. This is what we have above.

Now for codewords.

Pick a message index $i$ we have $M$ different assignments that are possible if we fix $n$-length sequences.
For illustration, consider $M$, $n$-length sequences $(X_1^n, X_2^n, \ldots, X_M^n)^T = C$. There are $M!$ codebooks that can be generated by permutation, all of which have equal probability. How do they look?

\[
\begin{bmatrix}
X_1^n & X_2^n & \ldots & X_M^n \\
X_2^n & X_1^n & \ldots & X_M^n \\
\vdots & \vdots & \ddots & \vdots \\
X_M^n & X_1^n & \ldots & X_{M-1}^n
\end{bmatrix}
\]

There are $(M-1)!$ codebooks in which any one row of elements remain fixed. Since each row is assigned an index to encode, there are $(M-1)!$ codebooks in which index $i$ is assigned the same codeword say $X_j^n = X_i^n(i)$.

Now we can compute the $P(E)$ based on this computation as:

\[
P(E) = \frac{1}{M} \sum_{C \in \mathcal{C}} \sum_{i=1}^{M} P(C) P \{ S(Y^n) \neq X_j^n \Rightarrow X_i^n(i) \Rightarrow i \mid X_j^n = X_i^n(i) \}
\]

\[
= \frac{1}{M} \sum_{i=1}^{M} \sum_{C \in \mathcal{C}} P(C) \frac{((M-1)n)!}{(Mn)!} \left\{ \sum_{i=1}^{M} P_r \{ S(Y^n) = \hat{W} \neq X_j^n \Rightarrow X_i^n(i) \Rightarrow i \mid X_j^n = X_i^n(i) \} \right\}
\]

\[
= \frac{1}{M} \sum_{i=1}^{M} \sum_{C \in \mathcal{C}} P(C) \lambda_i(C)
\]

\[
= \sum_{C \in \mathcal{C}} P(C) \lambda_i(C) = P_r(E \mid W=1)
\]
RECALL \( W = 1 \) WAS SENT AND \( Y^n \) WAS OBSERVED.

\[ E_i = \{ (X^n(i), Y^n) \in A^{(n)}_e \} \]

AN ERROR OCCURS IF \( \mathcal{G}(Y^n) = \hat{W} \) LEADS TO \( \{ (X^n(\hat{W}), Y^n) \notin A^{(n)}_e \} \) OR MORE THAN ONE \( Wi \)'s ARE COMPUTED PROVIDING

\[ (X^n(\hat{W}_i), Y^n) \in A^{(n)}_e \quad i \in \{1, 2, \ldots, M\} \]

\( \Rightarrow \) NO ERROR \( \{ \mathcal{E}^C \mid W = 1 \} = \left\{ \bigcup_{i=1}^{M} \left( E_i \cap E_j^c \right) \mid W = 1 \right\} \)

\( \Rightarrow \)

\[ p(\mathcal{E}^C \mid W = 1) = p \left\{ \bigcup_{i=1}^{M} \left( E_i \cap E_j^c \right) \mid W = 1 \right\} \]

\[ = p \left\{ E_j^c \cap \bigcup_{i=1}^{M} E_i \mid W = 1 \right\} \]

\[ \Rightarrow p_{s} \{ \mathcal{E} \mid W = 1 \} = 1 - p_{s} \{ \mathcal{E}^C \mid W = 1 \} \]

\[ \leq 1 - p_{s} \left\{ \bigcup_{i=1}^{M} E_i \cap E_j^c \mid W = 1 \right\} \]

\( E_i \)

\[ = p_{s} \left\{ (E_i \cap E_j^c)^c \mid W = 1 \right\} \]

\[ = p_{s} \left\{ E_i^c \cup E_j \mid W = 1 \right\} \]

\[ \Rightarrow \]

\[ \left( p_{s} \{ \mathcal{E} \mid W = 1 \} \right) \leq p_{s} \left\{ \bigcup_{j=2}^{M} E_i^c \cup E_j \mid W = 1 \right\} \]

\[ \leq p_{s} \{ E_i \mid W = 1 \} \]

\[ \leq p_{s} \{ E_j \mid W = 1 \} \]

\[ \leq p_{s} \{ E_j \mid W = 1 \} \]

\[ \leq p_{s} \{ E_j \mid W = 1 \} \]

\[ \leq p_{s} \{ E_j \mid W = 1 \} \]
\[
Pr \left\{ \sum_{i=2}^{M} E_i \mid W=1 \right\} \leq Pr \left\{ \sum_{i=2}^{M} E_i \mid W=1 \right\} + \sum_{i=2}^{M} Pr \left\{ E_i \mid W=1 \right\}
\]

1. \[
Pr \left\{ \sum_{i=2}^{M} E_i \mid W=1 \right\} = Pr \left\{ \sum_{i=2}^{M} (X^n(i), Y^n) \notin A^n_e \mid W=1 \right\} \leq \epsilon.
\]

2. **CHANNEL INPUT WAS X^n(i) GIVING OUTPUT Y^n.**

ALSO X^n(i) \(\equiv\) X^n(i), i \(\neq\) 1. **CHANNEL DMC.**

RECALL FROM LAST CLASS X^n(i) \(\equiv\) Y^n.

HENCE \[
Pr \left\{ \sum_{i=2}^{M} E_i \mid W=1 \right\} = Pr \left\{ (X^n(i), Y^n) \notin A^n_e \mid W=1 \right\}
\]

\[
= \sum_{i=2}^{M} Pr \left\{ (X^n(i), Y^n) \in A^n_e \right\} = \sum_{i=2}^{M} P(x^n(i)) P(y^n) (X^n(i), Y^n) \in A^n_e (X^n(i), Y^n) \in A^n_e
\]

\[
\leq \sum_{i=2}^{M} \left( -nH(x) - nH(y) + n2\epsilon \right)
\]

\[
\leq \sum_{i=2}^{M} \left( n(H(x,y) + \epsilon) - n(H(x) + H(y)) - 2\epsilon \right)
\]

\[
\leq \sum_{i=2}^{M} \left( n(H(x) + H(y) - H(x,y)) - 3\epsilon \right)
\]

\[
= \sum_{i=2}^{M} \left( n(I(x,y) - 3\epsilon) \right)
\]

\[
\Rightarrow \sum_{i=2}^{M} Pr \left\{ E_i \mid W=1 \right\} \leq (M-1) 2^n \leq 2^{nR}
\]

\[
Pr \left\{ \sum_{i=2}^{M} E_i \mid W=1 \right\} \leq \epsilon + 2^{nR}
\]

1. IF \( R < I(x,y) - 3\epsilon \), \( R < I(x,y) \) AND
\[ n(I(x; y) - R - 3\epsilon) = 2 \implies 0 \quad \text{as} \quad n \to \infty. \]

So, for \( n \) large enough

\[ P(\epsilon | W = 1) \leq \epsilon + \epsilon = 2\epsilon \to 0 \quad \text{as} \quad n \to \infty. \]

So far we only showed \( P(\epsilon) \to 0 \) averaged over all codebooks. This means, there must be at least one codebook \( c^* \) for which

\[ P(\epsilon | c^*) \leq 2\epsilon. \]

\[ P(\epsilon | c^*) = P\{ \tilde{W} \neq W | c^* \} = \frac{1}{2^n} \sum_{M=2}^{M=2^n} \lambda_{\tilde{W}}(c^*) \leq 2\epsilon. \]

Note if \( \sum_{M=2}^{M=2^n} \lambda_{\tilde{W}}(c^*) \leq 2\epsilon \), at least the best half of codewords in \( c^* \) must have total error \( \leq 4\epsilon \). Get rid of worst half of codewords in \( c^* \) and retain best half. After reindexing, we have

\[ \frac{1}{2^{n+1}} \sum_{\tilde{W} = 1}^{2^n} \lambda_{\tilde{W}}(c^*) \leq 4\epsilon. \]

\[ \Rightarrow \quad \hat{\lambda}_n^{(n)}(c^*) = \max_{\tilde{W} \in \{1, \ldots, 2^n-1\}} \hat{\lambda}_{\tilde{W}}(c^*) \leq 4\epsilon \to 0 \quad \text{as} \quad n \to \infty. \]

\( \Rightarrow \) we have at least one codebook \( c^* \) for which we have max probability of error \( \hat{\lambda}_n \to 0 \) as \( n \to \infty \). For this codebook, raise = \( R - 1/n = R' \). But \( R < I(x; y) \leq C \)

\[ \Rightarrow \quad R' < R < I(x; y) \leq C. \]

Hence absurd.
Joint Source-Channel Coding Theorem (Read 7.13)

Only sketch of proof stated.

\[ I (V^n \wedge \hat{V}^n) \leq I (X^n \wedge Y^n) \]

If \( V^n = \{ V_1, \ldots, V_n \} \) is generated from a source with finite alphabet \( \mathcal{A} \) satisfying AEP with \( H(Y) < C \), there exists a joint source-channel code with

\[ P_{\hat{V}^n} \leq \sum \hat{V}^n \neq V^n \rightarrow 0 \]

Conversely, if \( \lim_{n \to \infty} P_{\hat{V}^n} \leq V^n = 0 \) then \( H(V_n) \leq C \).

Recall \( H(Y) = \frac{H(Y^n)}{n} \) or

\[ \frac{1}{n} \sum_{i=1}^{n} H(Y_i) = H(Y) \]

\[ H(V) = 2^{-n} H(V) \]

Also \( V^n \in A_{\varepsilon}^{(n)} \Rightarrow |A_{\varepsilon}^{(n)}| \leq 2^{n(H(V) + \varepsilon)} \)

\[ 1 \geq \sum_{\omega^n \in V^n} \sum_{\hat{V}^{(n)}} p(\omega^n, \hat{V}^{(n)}) = \sum_{\omega^n \in A_{\varepsilon}^{(n)}} \sum_{\hat{V}^{(n)}} p(\omega^n, \hat{V}^{(n)}) \geq 2^{-n(H(V) + \varepsilon)} \]

\[ \Rightarrow |A_{\varepsilon}^{(n)}| \leq 2^{n(H(V) + \varepsilon)} \]

We need at most \( n(H(V) + \varepsilon) \) bits to code source. \( \Rightarrow \) Each symbol needs at most \( (H(V) + \varepsilon) \) bits.

When we decode \( I \{ \hat{V}^n \neq V^n \} = \sum_{\omega^n \in V^n} \hat{V}^n \neq V^n \)

\[ E[I(\hat{V}^n \neq V^n)] \]

\[ = \sum_{(\omega^n, \hat{\omega}^n)} p(\omega^n, \hat{\omega}^n) I(\omega^n \neq \hat{\omega}^n) \]
\[
E[1(\hat{V}^n \neq V^n)] = \sum_{(x^n, \hat{x}^n) : \hat{V}^n \neq V^n} p(x^n, \hat{x}^n) = p(\hat{V}^n \neq V^n)
\]
\[
P_e^{(n)} = E[1(\hat{V}^n \neq V^n)]
\]

**FORWARD PART OF THEM**

\[
P_e^{(n)} = Pr\{\hat{V}^n \neq V^n\} \leq \epsilon
\]

\[
= \epsilon \leq \epsilon + \epsilon = 2\epsilon \rightarrow 0 \text{ as } n \rightarrow \infty.
\]

**Hence, if** \(H(V) + \epsilon < R \leq C\) **then** \(P_e^{(n)} \rightarrow 0\) **as** \(n \rightarrow \infty\).

**Converse:** \(P_e^{(n)} \rightarrow 0\) **as** \(n \rightarrow \infty\). **Then** \(H(V^n) \leq C\)

\[
H(V^n) = H(V^n|\hat{V}^n) + I(V^n;\hat{V}^n)
\]

\[
\leq H(V^n|\hat{V}^n) + I(X^n;Y^n)
\]

\[
\leq 1 + nPr\{\hat{V}^n \neq V^n\} \log |2| + nC
\]

\[
\Rightarrow \frac{H(V^n)}{n} \leq \frac{1}{n} + Pr\{\hat{V}^n \neq V^n\} \log |2| + C
\]

**If** \(Pr\{\hat{V}^n \neq V^n\} \rightarrow 0\) **as** \(n \rightarrow \infty\) **then**

\[
\frac{H(V^n)}{n} \rightarrow H(\mathcal{Y}) \leq C
\]