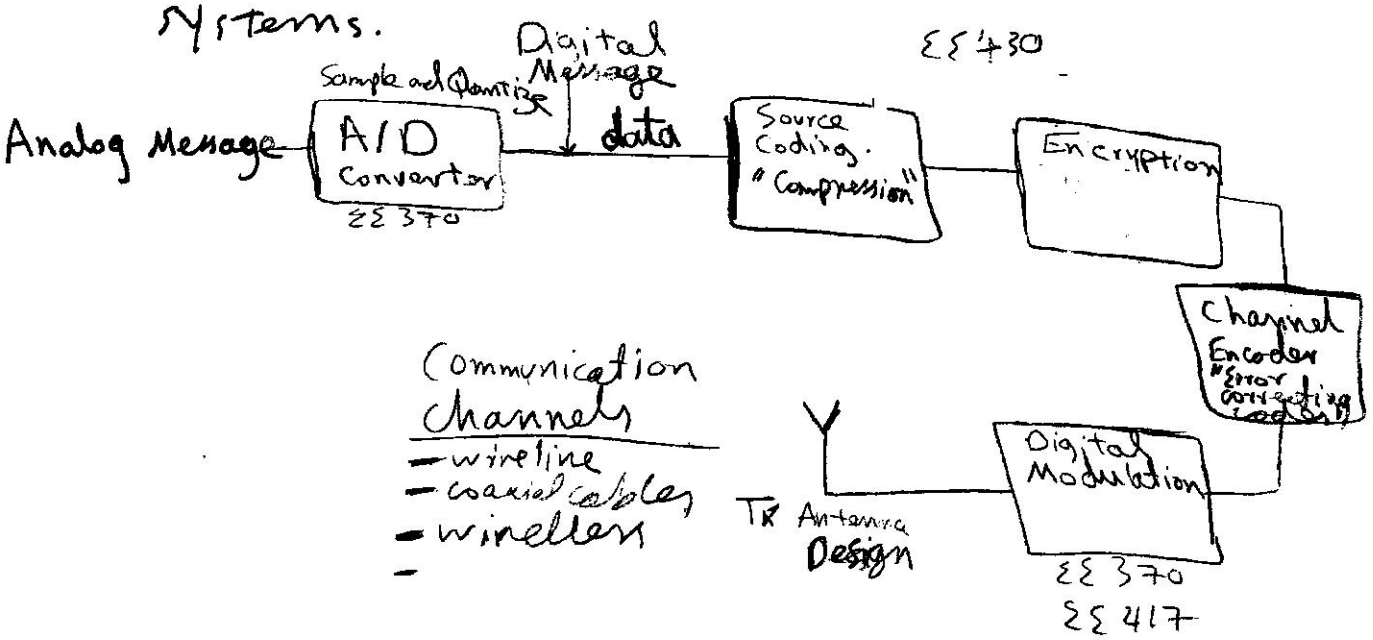
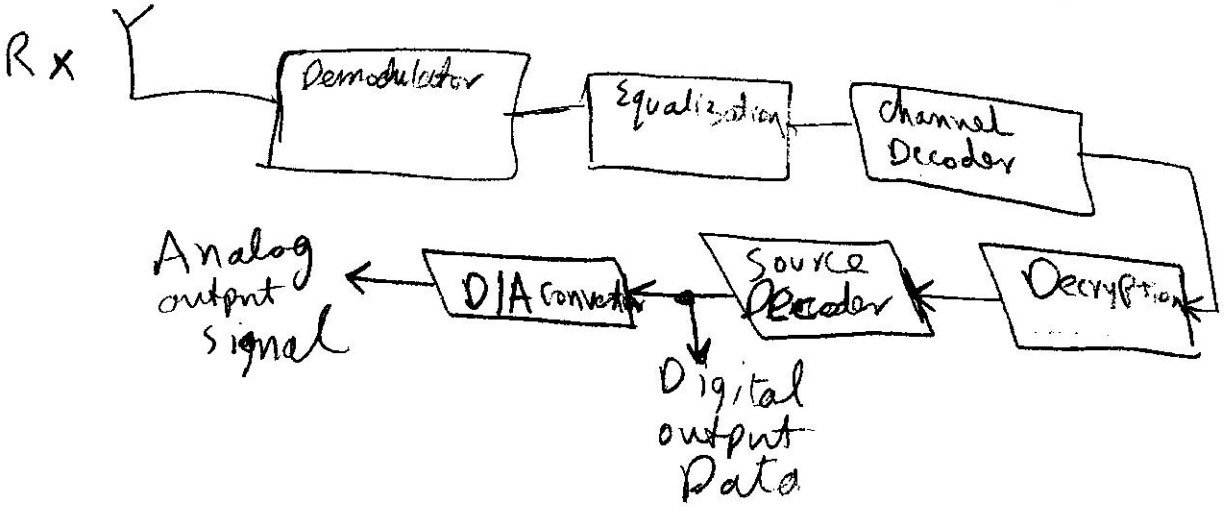
Introduction class

— Let take a quick Review of Digital Communication systems.

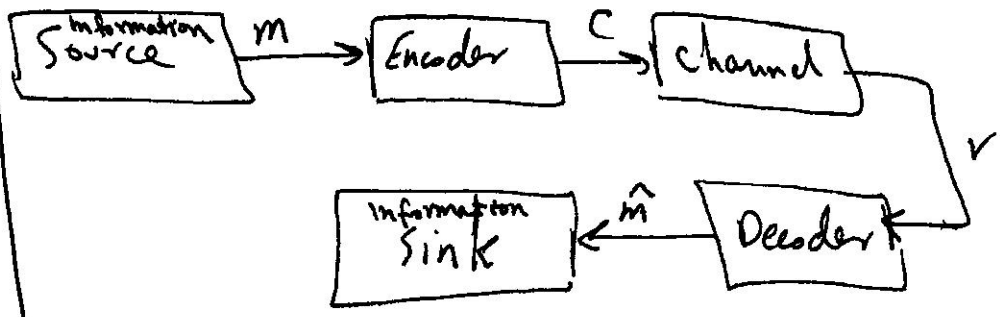




Chapter 6 Proakis

Material
6.2
6.5

Discrete Sources and Entropy



1.2 Source Alphabets and Entropy.

Def. — An information source outputs a set of symbols.

— A finite discrete source ~~output~~ output belongs to a finite discrete set of symbols. The symbol set is called the source alphabet.

— $A = \{a_0, a_1, \cdots, a_{M-1}\}$ is a source alphabet (symbol set) of cardinality $M = |A|$. (size) or (number of elements)

— The source outputs symbols in a time sequence represented by the notation $\bar{a} = (S_0 S_1 \cdots S_t \cdots)$ where $S_t \in A$

— The probability that the source emits symbol a_m is written as $P_m = Pr(a_m)$.

— The set of probabilities for the source alphabet is

$$P_A = \{P_0, P_1, \cdots, P_{M-1}\}$$

— Information Theory makes an important distinction between data and information. Not all data carries information. Information is the part that adds knowledge.

~~Given a set of data~~

Example — Assume that we have a source that emits only one symbol. A set of data from this source will have zero information content since the receiver knows that the source only send one unique symbol all the time. Thus, we have a set of data but no information content.

— From the above example, we can see that the information content of the data increases with uncertainty of transmitted symbols. —

— Entropy

Information theory provides a measure of the average amount of information conveyed per source symbol. This measure is called the entropy of the source and is defined as:

$$H(A)=\sum_{m=0}^{M-1} P_m \log_2\left(\frac{1}{P_m}\right)$$

; where P_m is the probability that symbol m was transmitted. Notice that the Logarithm is to the base 2.

the units of $H(A)$ is bits. It tells us how much information carried by each symbol.

To help in calculation, Recall that $\log_2(x)=\frac{\ln(x)}{\ln(2)}$

Also, $\lim_{x\to 0} x\log(x)=0$

Notice that the information per symbol is $I(A)=\log_2\left(\frac{1}{P_m}\right)$ and Entropy is the average information per

Example 1.2.1

What is the average amount of information conveyed per source symbol of a 4-ary source having probabilities

$$P_A = \{0.5, 0.3, 0.15, 0.05\} \; ?$$

Solution:

$$H(A) = 0.5\log_2(2) + 0.3\log_2(10/3) + 0.15\log_2(100/15) + 0.05\log_2(20)$$

$$= 1.6477 \text{ bits.}$$

So, each symbol carries 1.6477 bits of information. while it carries 2 bits of data.

Therefore, it is possible to use some data compression techniques to compress the above source and use fewer bits on average.

Def: Information efficiency = $\dfrac{\text{The entropy of the source}}{\text{Average number of bits used to represent the source data.}}$

For the above example,

the information efficiency is $= \dfrac{1.6477}{2} = 82.387\%$

this means that approximately 17.6% of the bits are redundant and carry no information.

Example: Find the entropy of the 4-ary signal if all symbols are equally probable. i.e $P_m = \frac{1}{4}$

Lemma: Consider an M-ary source A. The maximum entropy of this source is $\log_2 M$ and it happens when all symbols are equally probable $\Rightarrow$ $p_m = \frac{1}{M}$ for all $m \in A$

See example 1.2.2 for proof.

This result makes sense intuitively. If every symbol in A is equally probable, an observer would have no idea what symbol will be emitted next by the source. Thus, each symbol carries the maximum surprise value and the average amount of information is maximized.

1.2.2 Joint and Conditional Entropy

Most communication systems are designed to be used by a large number of users. The designers of such a system are concerned with maximizing the total information carrying capacity of the system.

Information theory gives us also tools to measure the joint and conditional entropy for more than one source.

Consider a situation where we have two information sources, A and B. Let $|A| = M_A$ and $|B| = M_B$.

The joint probability that A sends symbol a_i and B sends symbol b_j is $p_{ij} = Pr(a_i, b_j)$

If the two symbol are statistically independent, then

Statistically independent case

If sources A and B are statistically independent, the total entropy of this system will be

$$H(A,B) = H(A) + H(B)$$

$H(A,B)$: joint entropy; $H(A)$: Entropy for A; $H(B)$: Entropy for B

If the two sources are dependent, what do you expect will happen to the joint entropy? will it increase or decrease?
The answer is that it will decrease. There is a fundamental theorem in information theory that says "Side information never increases entropy". The joint entropy

$$H(A,B) \leq H(A) + H(B)$$

with equality if and only if A and B are statistically independent.

Statiscally dependent Case

Denote the combined emission of a_i and b_j as a compound symbol $c_{ij} \equiv \langle a_i, b_j \rangle$. let the probability of emiting $c_{ij} = P_{ij}$.

The entropy of C is

joint entropy $\rightarrow H(C) = \sum_{c_{ij} \in C} P_{ij} \log_2\left(\frac{1}{P_{ij}}\right)$

$$= \sum_{i=0}^{M-1} \sum_{j=0}^{M_B-1} P_{ij} \log_2\left(\frac{1}{P_{ij}}\right)$$

The joint probability P_{ij} may be written in terms of a conditional probability $\Rightarrow \left[P_{j|i} = Pr(b_j | a_i)\right]$

as $P_{ij} \equiv P_{j|i} \cdot P_i$

$$\therefore H(C) = \sum_i \sum_j P_{ij} \log_2\left(\frac{1}{P_{j|i} P_i}\right)$$

$$= \sum_i \sum_j P_{ij} \log_2\left(\frac{1}{P_i}\right) + \sum_i \sum_j P_{ij} \log_2\left(\frac{1}{P_{j|i}}\right)$$

Notice that, $\sum_i \sum_j P_{ij} \log_2\left(\frac{1}{P_i}\right) = \sum_i \sum_j P_{j|i} P_i \log_2\left(\frac{1}{P_i}\right)$

$$= \sum_i P_i \log_2\left(\frac{1}{P_i}\right) \underbrace{\sum_j P_{j|i}}_{=1}$$

$$\therefore H(C) = \sum_i P_i \log_2\left(\frac{1}{P_i}\right) + \sum_i \sum_j P_{ij} \log_2\left(\frac{1}{P_{j|i}}\right)$$

$$H(C) = H(A,B) = H(A) + H(B|A)$$

Similarly, $H(A,B) = H(B) + H(A|B)$

$H(B|A)$ and $H(A|B)$ are Conditional Entropy.

$$H(B|A) = \sum_i \sum_j P_{ij} \log\left(\frac{1}{P_{j|i}}\right)$$

$$H(A|B) = \sum_i \sum_j P_{ij} \log\left(\frac{1}{P_{i|j}}\right)$$

Also, $H(B|A) < H(B)$

and $H(A|B) < H(A)$

The end result is that the joint entropy of two sources A and B is

$$H(A,B) = H(A) + H(B|A)$$
$$= H(B) + H(A|B)$$

and

$$H(A,B) \leq H(A) + H(B)$$

with equality if and only if A and B are independent.

Continue 1.2.2

— Joint and Conditional Entropy.

$$H(c) = H(A,B) = H(A) + H(B|A) = H(B) + H(A|B)$$

($H(A,B)$: joint entropy; $H(B|A)$: conditional entropy.)

$$H(A,B) \leq H(A) + H(B)$$

and equality happens when A and B are independent.

Example 1.2.3

Error detection using parity bits

Parity bits are added to the transmitted bits to detect errors. Let A be an information source with alphabet $A = \{0,1,2,3\}$. Let each symbol a be equally probable and Let $B = \{0,1\}$ be a parity generator with

$$b_j = \begin{cases} 0 & \text{if } a = 0 \text{ or } a = 3 \\ 1 & \text{if } a = 1 \text{ or } a = 2 \end{cases}$$

What are $H(A)$, $H(B)$ and $H(A,B)$?

$H(A) = 4 \cdot \frac{1}{4} \log_2(4) = 2$ bit

Likewise $H(B) = 2 \cdot \frac{1}{2} \log_2(2) = 1$ bit

} This is because each symbol is equally probable.

The conditional probabilities $Pr(b/a)$ are

$Pr(0|0) = 1$, $Pr(1|0) = 0$ $\rightarrow \sum_j Pr(b_j|0) = 1$

$Pr(0|1) = 0$, $Pr(1|1) = 1$

$Pr(0|2) = 0$, $Pr(1|2) = 1$

$Pr(0|3) = 1$, $Pr(1|3) = 0$

Therefore,

$$H(B|A) = \sum_{i=0}^{3} P_i \sum_{j=0}^{1} P_{j|i} \log_2\left(\frac{1}{P_{j|i}}\right)$$

$$= 4 \cdot \frac{1}{4} (1 \log_2 1 - 0 \log_2 0)$$

This says that B is completely determined by A.

$$H(A,B) = H(A) + H(B|A) = 2 + 0 = 2$$

∴ source B contributes no information to the compound signal.

1.7.3 Entropy of Symbol Blocks and the Chain Rule.

What is the information content of a block of symbols?

A block of symbols is a sequence of n symbols $(s_0, s_1, \cdots, s_{n-1})$ produced by source A. $\{s_i \in A\}$.

The entropy of this block is denoted as

Joint Entropy → $H(A_0, A_1, \cdots, A_{n-1})$.

We can use the joint entropy result to express the entropy as

$$H(A_0, A_1, \cdots, A_{n-1}) = H(A_0) + H(A_1, A_2, \cdots, A_{n-1}|A_0)$$

This term can also be written as → $H(A_1, A_2, \cdots, A_{n-1}|A_0) = H(A_1|A_0) + H(A_2, \cdots, A_{n-1}|A_0, A_1)$

Repeating this argument inductively, we get

$$H(A_0, A_1, \cdots, A_{n-1}) = H(A_0) + H(A_1|A_0) + H(A_2|A_0, A_1) + \cdots H(A_{n-1}|A_0, \cdots, A_{n-2})$$

This is called the chain rule for entropy.

Since $H(B|A) \leq H(B)$

The upper bound on the Entropy of symbol blocks is

$$H(A_0, \cdots, A_{n-1}) \leq \sum_{i=0}^{n-1} H(A_i)$$

With equality if and only if all the symbols in the equence are statistically independent.

Memoryless Source

A source that emits statistically independent symbols is called a memoryless source. That is because it doesn't remember from one time to the next what symbols it has previously emitted. So, there is no correlation between symbols.

→ If the information source is the same all the time and the probabilities of the source don't change over time, then we have the joint entropy to be.

$$H(A_0, \cdots, A_{n-1}) \leq n \cdot H(A)$$

~~See~~ Examples 1.2.4

Suppose a memoryless source with $A = \{0, 1\}$ having equal symbol probabilities emits a sequence of six symbols. Following the sixth symbol, a seven symbol is added which is the sum modulo 2 of the six previous symbols. What is the entropy of this sequence?

Solution

Let $b = \overset{5}{\underset{t=0}{\bigoplus}} s_t$, where $\bigoplus$ is just the summation Modulo 2 and $s_t \in A$.

$H(A_0, A_1, \cdots, A_5, b) = H(A_0) + H(A_1|A_0) + \cdots + H(b|A_0, \cdots, A_5)$

Since the First six symbols are statistically independent we ~~get~~

$H(A_0, A_1, \cdots, A_5, b) = 6\,H(A) + H(b|A_0, \cdots, A_5)$

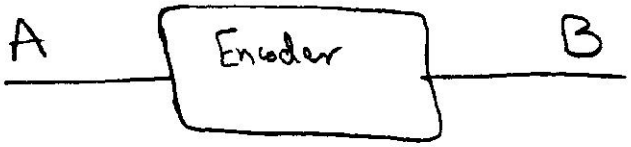
since b is completly determined by b, it has no new information and $H(b/A_0, \cdots, A_5) = 0$

$\Rightarrow$ The ~~point~~ entropy of the ~~s~~block in

$H(A_0, A_1, \cdots, A_5, b) = 6\,H(A)$

$= 6$ since the symbols of A are equally probable.

1.3.2 Mutual Information



The Mutual information $I(B;A)$ is the amount of reduction in uncertainty about B when we observe A.

$$I(B;A) = H(B) - H(B/A)$$

$$= \sum_{b\in B}\sum_{a\in A} P_{ba}\log_2\left(\frac{P_{ba}}{P_b P_a}\right)$$

Also, $I(A;B) = H(A) - H(A/B)$

$$= \sum_{b\in B}\sum_{a\in A} P_{ba}\log_2\left(\frac{P_{ba}}{P_b P_a}\right)$$

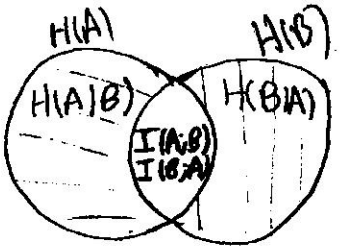
Notice that $I(A;B) = I(B;A)$

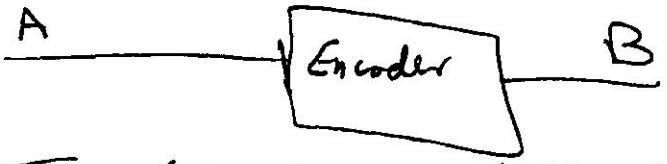
In other words, the Mutual information $I(B;A)$ the amount of information we can tell about B from observing A.

Graphical Illustration

~~Left~~ H(A) • H(B)

Now





The relation between H(A) and H(B) defines three kinds of Encoders.

1- Lossless Encoder

No information is lost. In this case,

$$H(A) = H(B) \quad \text{and} \quad H(B|A) = 0$$

and $$I(B;A) = H(B)$$

$\Rightarrow$ The two circles are on top of each other and the information content of A and B are the same.

2- Lossy Encoder

Some information is Lost.

$$\Rightarrow \quad H(B) < H(A)$$

Also, $I(A;B) < H(A)$

$\Rightarrow I(B;A) < H(A)$

3- Encryption or Channel Coding

In this case, the information content in B is increased In order to improve performance over noisy channels "channel coding" or to increase security "Encryption".

$$H(B) > H(A)$$

The extra amount of information is redundant. It doesn't add more knowledge to A.

This added information makes it hard for

Ch. 9 [Cover]

Example [Entropy of a multivariate normal distribution]

Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with mean μ and covariance matrix K. $[\mathcal{N}_n(\mu, K)]$

Then,

$$h(X_1, X_2, \ldots, X_n) = h(\mathcal{N}_n(\mu, K)) = \frac{1}{2}\log(2\pi e)^n |K| \text{ bits}$$

Proof:

$$h(f) = -\int f(\vec{x}) \ln f(\vec{x})\, d\vec{x}$$

where $f(\vec{x}) = \frac{1}{(\sqrt{2\pi})^n |K|^{1/2}} e^{-\frac{1}{2}(\vec{x}-\mu)^T K^{-1}(\vec{x}-\mu)}$

$\therefore$

$$h(f) = -\int f(\vec{x}) \left[-\tfrac{1}{2}(\vec{x}-\mu)^T K^{-1}(\vec{x}-\mu) - \ln(\sqrt{2\pi})^n |K|^{1/2}\right] d\vec{x}$$

$$= \frac{1}{2} E\left[\sum_{i,j} (x_i - \mu_i)(K^{-1})_{ij}(x_j - \mu_j)\right] + \frac{1}{2}\ln(2\pi)^n |K|$$

$$= \frac{1}{2} E\left[\sum_{i,j} (x_i - \mu_i)(x_j - \mu_j)(K^{-1})_{ij}\right] + \frac{1}{2}\ln(2\pi)^n |K|$$

$$= \frac{1}{2}\sum_{i,j} E[(x_j - \mu_j)(x_i - \mu_i)](K^{-1})_{ij} + \frac{1}{2}\ln(2\pi)^n |K|$$

$$= \frac{1}{2}\sum_j \sum_i K_{ji}(K^{-1})_{ij} + \frac{1}{2}\ln(2\pi)^n |K|$$

$$= \frac{1}{2}\sum_j I_{jj} + \frac{1}{2}\ln(2\pi)^n |K|$$

$$= \frac{n}{2} + \frac{1}{2}\ln(2\pi)^n |K|$$

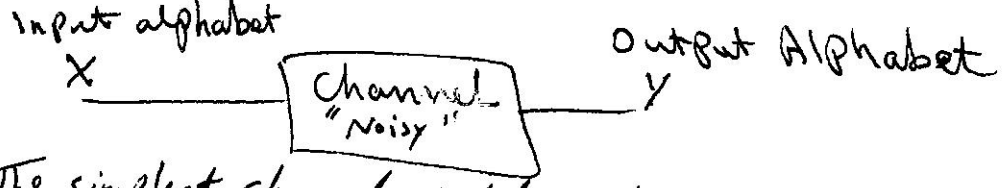
$$= \frac{1}{2}\ln(2\pi e)^n |K| \text{ nats}$$

$$= \frac{1}{2}\log_2(2\pi e)^n |K| \text{ bits}.$$

Chapter 2

Lecture 7

Channels and channel Capacity.



The simplest channel Model used in communication is the AWGN channel.

$y = x + n$; where n is a zero-mean Gaussian Random Variable.

Channels in general have memory. That means that the output sequence can be correlated. A special case that is used a lot in analyzing communication systems is the Memoryless channels.

2.1 Discrete Memoryless Channel Model

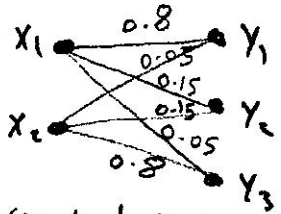
Since the channel is noisy, errors in transmission will occure, we can model a Memoryless channel as a function with transition probabilities that maps the input sequence X to output sequence Y. This mapping is not one-to-one and reversing the function is not possible.

Example 2X3 Memoryless channel Model

Let P_x be the Prob. that symbol x is transmitted.

$\Rightarrow P_y = \sum_{x \in X} P_{y/x} P_x$

P_y: Prob. that the received symbol is y.



the forward transition Prob. can be represented by a matrix

$$P_{Y/X} = \begin{bmatrix} 0.8 & 0.05 \\ 0.15 & 0.15 \\ 0.05 & 0.8 \end{bmatrix} = \begin{bmatrix} P_{Y_1/X_1} & P_{Y_2/X_2} \\ P_{Y_2/X_1} & P_{Y_2/X_2} \\ P_{Y_3/X_1} & P_{Y_3/X_2} \end{bmatrix}$$

In Matrix Form:

the bar denotes a Matrix or a vector.

$$\overline{P}_Y = \overline{P}_{Y/X}\,\overline{P}_X$$

$$\begin{bmatrix} P_{Y_1} \\ P_{Y_2} \\ P_{Y_3} \end{bmatrix} = \begin{bmatrix} P_{Y_1/X_1} & P_{Y_1/X_2} \\ P_{Y_2/X_1} & P_{Y_2/X_2} \\ P_{Y_3/X_1} & P_{Y_3/X_2} \end{bmatrix} \begin{bmatrix} P_{X_1} \\ P_{X_2} \end{bmatrix}$$

Notice that

$$\sum_{y\in Y} P_{Y/X_1} = \sum_{y\in Y} P_{Y/X_2} = 1$$

$\Rightarrow$ columns of $P_{Y/X}$ sum to unity.

For the previous example

$$\overline{P}_Y = \begin{bmatrix} 0.425 \\ 0.15 \\ 0.425 \end{bmatrix} \text{ if } P_X = [0.5 \;\; 0.5] \rightarrow \text{equally probable.}$$

— Notice that the calculations of the transition probabilities are part of the communication systems performance evaluations covered in Digital Communication Courses.

— Notice that $|Y| \neq |X|$ in the above example.

if $|Y| = |X| \rightarrow$ Hard-decision Decoding

if $|Y| > |X| \rightarrow$ Soft-decision Decoding.

2.1.2 Output Entropy and Mutual Information

What is the Mutual information between the input to the channel and the output?

First, examine the previous example,

Entropy $\rightarrow H(Y) = \sum_{y\in Y} P_Y \log_2\left(\frac{1}{P_Y}\right) = 1.4598$ bits

while $H(X) = 1$ bits since P_X are equally Prob.

Notice that $H(Y) > H(X) \Rightarrow$ The channel adds information to the source. More randomness

this is not always the case, sometimes we get $H(Y) < H(X)$ $\Rightarrow$ information loss.

Assume we have a 2x2 ~~Discrete~~ DMC [Discrete Memoryless channel]

$$P_{Y|X} = \begin{bmatrix} 0.98 & 0.05 \\ 0.02 & 0.95 \end{bmatrix} \text{ and } P_X = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

$$\Rightarrow P_Y = \begin{bmatrix} 0.515 \\ 0.485 \end{bmatrix} \Rightarrow H(Y) = 0.99935 < H(X)$$

$\Rightarrow$ we have information Loss.

The information in the above example was lost during the transmission process.

Information Lossy channels $\Rightarrow H(Y) < H(X)$

Mutual Information

The receiver observes Y, How much can it tell us about the transmitted information sequence?

This is the Mutual information of X observing Y.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} P_{xy} \log_2 \left(\frac{P_{xy}}{P_y P_x} \right)$$

$$I(X;Y) = H(X) - H(X/Y)$$

- If Y and X are independent $\Rightarrow I(X;Y) = 0$

  $\Rightarrow$ Y tells us nothing at all about X.

- Upper bound on Mutual information is

$$I(X;Y) \leq H(X)$$

with equality if and only if $H(X/Y) = 0$.

that means that there is no information Loss, and Y contains sufficient information to tell us what the transmitted sequence was.

Example 2.1.4

For the 2X3 DMC in Example 2.1.1

$I(X;Y) = 0.57566$ bits

$\Longrightarrow$ since I is much less than $|\mathcal{X}|=1$, the channel has a high level of information loss.

Example 2.1.5

In this case, (2X2 DMC) $I(X;Y) = 0.78543$

Also, this 2x2 DMC causes information loss.

2.2 Channel Capacity and the Binary Symmetric Channel.

2.2.1 Channel Capacity

Channel capacity is the maximum average amount of information that can be sent per channel use. Each time the transmitter sends a symbol, it is said to use the channel.

Channel capacity is

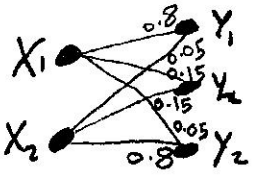
$$C_x = \max_{P_x} I(X;Y)$$

$\Rightarrow$ is the maximum Mutual information achieved for a given channel.
Maximization is done over the input probabilities.

In otherwords, the channel capacity is the maximum information rate that can be supported by the channel.

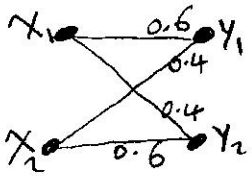
2.2.2 Symmetric Channels

Example



$$P_{Y|X} = \begin{bmatrix} 0.8 & 0.05 \\ 0.15 & 0.15 \\ 0.05 & 0.8 \end{bmatrix}$$

Example



$\Rightarrow$ Binary symmetric channel

$$P_{Y|X} = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$$

Lecture 8 P:

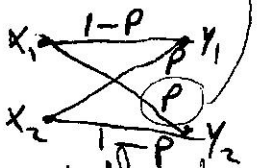
* The capacity of Binary symmetric channels is (BSC)

$$C_x = 1 - H(P)$$
$$= 1 + (1-p)\log_2(1-p) + p\log_2(p)$$

where p is the crossover probability (prob. of error)

The BSC ~~Prob~~ transition Prob. is

$$P_{Y|X} = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$



— The above capacity is the maximum Mutual information And that happens when the the input distribution (P_x) is uniform [equally likely].

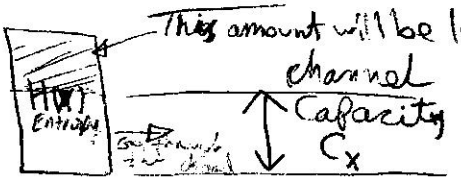
— $0 \leq C_x \leq 1$ ($H(x)$) for BSC

— the upper bound is achieved only if $p = 0$ or $p = 1$. At these values, $C_x = 1$

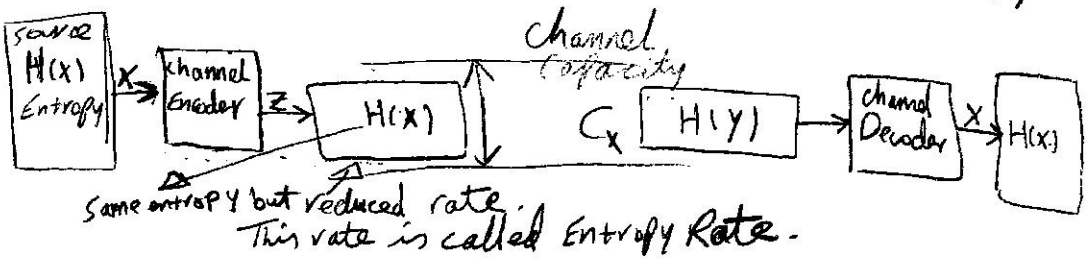
— the lower bound, $C_x = 0$, is achieved when $p = 0.5 \Rightarrow$ 50% chance to get the output correct. Then, Error and correct transmition are equally likely. And the information loss is total.

- If we send data with information $H(x) > C_x$, then for sure we will loose some information over the channel. However, we can reduce the entropy of the data by adding redundant symbols.

Thus, this the key idea of coding. We added redundant bits using a coding algorithm so that we reduce the information of the source and make it able to pass the channel with very low prob. of lost information.



But, Add more redundant bits $\Rightarrow$ Lower Entropy



2.3 Block Coding and Shannon's Second Theorem:

2.3 Block Coding and Shannon's Second Theorem.

The channel capacity is

$$C_X = \max_{P_X} I(X;Y)$$
$$= \max_{P_X} \left(H(X) - H(X|Y) \right)$$

The maximum capacity is $C_X = H(X)$,

thus, the term $H(X|Y)$ is the information loss from the Maximum. This conditional entropy is called Equivocation. Recall that

$$H(X|Y) \leq H(X).$$

2.3.2 Entropy Rate and Channel Coding Theorem.

- The entropy of a block of n symbols is

$$H(X_0, X_1, X_2, \ldots, X_{n-1}) \leq n H(X)$$

if they are drawn from the same source and the prob. doesn't change with time.

- So, this is the information content of the block.
- The average information per channel use is the Rate

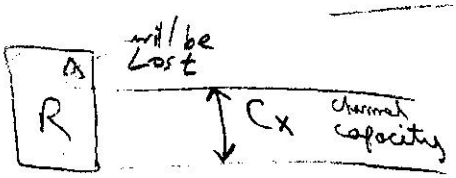
$$R = \frac{H(X_0, X_1, \ldots, X_{n-1})}{n} \leq H(X)$$

- Taking the limit $n \to \infty$, we get Entropy Rate

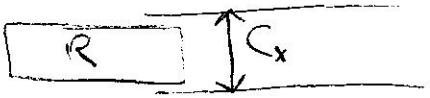
Entropy Rate $\longrightarrow$ $R = \lim_{n\to\infty} \frac{H(X_0, X_1, \ldots, X_{n-1})}{n} \leq H(X)$

What is the relation between R and C_x?
Can we control R in order to have almost zero information loss?

Theoritically, the answer is Yes. We can reduce R by adding redundant symbols such that $R \leq C_x$. This process is called channel coding.



However, if we let $R \leq C_x$, we can pass the information with Low Prob. of error.



Shannon's Theorem

Suppose $R < C_x$, where C_x is the capacity of a memoryless channel. Then for any $\epsilon > 0$, there exists a block length n and a code of block length n and rate R whose probability of block decoding error P_e satisfies $P_e \leq \epsilon$ when the code is used on this channel.

Cutoff rate, R_0

Another bound for practical error-correcting codes is called the cutoff rate (R_0). For a binary symmetric channel, the cutoff rate (R_0) is

$$R_0 = -\log_2\left(0.5 + \sqrt{p(1-p)}\right)$$

This bound applies for most of the codes, However, new codes, such as, Turbo codes or LDPC codes can operate very close to channel capacity.

Example 2.3.2

Find C_x and R_0 for the BSC with

a) $p=0.1$ b) $p=0.01$ c) $p=0.001$ d) $p=0.0001$

a) $C_x=0.531$ $R_0=0.322$ b) $C_x=0.919$ $R_0=0.738$

c) $C_x=0.988$ $R_0=0.911$ d) $C_x=0.998$ $R_0=0.971$

Notice that the cutoff rate approaches the capacity as the crossover prob. p grows smaller.

Differential Entropy

Def $h(X) = -\int_S f(x) \log_2 f(x)\, dx$

where S is the support set of the random variable.

Example uniform distribution

$$h(X) = -\int_0^a \frac{1}{a} \log_2 \frac{1}{a}\, dx = \log_2 a$$

Note that for $a < 1$, $\log_2 a < 0$

hence, differential entropy can be negative.

Example 9.1.2 [Normal Distribution]

Let $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$

$$h(\phi) = -\int \phi \ln \phi$$

$$= -\int \phi(x) \left[-\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right]$$

$$= \frac{1}{2\sigma^2} E[X^2] + \ln \sqrt{2\pi\sigma^2} \int \phi(x)\, dx$$

$$= \frac{\sigma^2}{2\sigma^2} + \frac{1}{2} \ln 2\pi\sigma^2$$

$$= \frac{1}{2} \ln e + \frac{1}{2} \ln 2\pi\sigma^2$$

$$= \frac{1}{2} \ln 2\pi e \sigma^2 \text{ nats}$$

~~$= \frac{1}{2} \log_2 2\pi e \sigma^2$~~

changing the base of the logarithm

$$h(\phi) = \frac{1}{2} \log_2 2\pi e \sigma^2 \text{ bits}$$

Def: Joint Differential Entropy

$$h(X_1, X_2, \cdots, X_n) = -\int f(x_1, x_2, \cdots, x_n) \log_2 f(x_1, x_2, \cdots, x_n)\, dx_1\, dx_2 \cdots dx_n$$

Def: Conditional differential Entropy

$$h(X|Y) = -\int f(x,y) \log_2 f(x|y)\, dx\, dy$$

Since $f(x|y) = \frac{f(x,y)}{f(y)}$

$$\Rightarrow h(X|Y) = h(X,Y) - h(Y)$$

Def: Relative Entropy

The relative Entropy between two densities f and g is:

$$D(f||g) = \int f \log \frac{f}{g}$$

Def: Mutual information

$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)}\, dx\, dy$$

$$I(X;Y) = h(X) - h(X|Y)$$
$$= h(Y) - h(Y|X)$$

Also $I(X;Y) = D[f(x,y)||f(x)f(y)]$

Properties

1. $D(f||g) \geq 0$ with equality iff $f = g$
2. $I(X;Y) \geq 0$ with equality iff X and Y are independent
3. $h(X|Y) \leq h(X)$ with equality if X and Y are independent
4. $h(X_1, X_2, \cdots, X_n) \leq \sum h(X_i)$ with equality iff $X_1, X_2, \cdots, X_n$ are independent.
5. $h(X+c) = h(X)$

"Translation does not change the differential Entropy"

6. $h(aX) = h(X) + \log_2 |a|$
7. Let A be a Matrix, then $h(AX) = h(X) + \log_2 |A|$

~~9.9 [Cover]~~

Theorem 9.6.5:

Let $X \in R^n$ be a R.V. with zero mean and covariance $K = E[XX^T]$,

then $h(X) \leq \frac{1}{2}\log(2\pi e)^n |K|$

with equality iff $X \sim \mathcal{N}(0, K)$

$\Rightarrow$ multivariate normal distribution maximizes the entropy over all distributions with the same covariance.

Proof: Let $g(x)$ be any density satisfying

$$\int g(x) x_i x_j dx = K_{ij} \rightarrow \text{covariance}$$

Let ϕ_k be the density of a $\mathcal{N}(0, K)$.

Note that $\log \phi_k(x)$ is a quadratic form

and $\int x_i x_j \phi_k(x) dx = K_{ij}$

then $D(g \| \phi_k) \geq 0$

$$\int g \log\left(\frac{g}{\phi_k}\right) \geq 0$$

$$-h(g) - \int g \log \phi_k \geq 0$$

since $\log \phi_k(x)$ is a quadratic form and

$$\int g(x) x_i x_j dx = \int \phi_k(x) x_i x_j dx = K_{ij}$$

$\Rightarrow$

$$-h(g) - \int g \log \phi_k$$

$$= -h(g) - \int \phi_k \log(\phi_k)$$

$$= -h(g) + h(\phi_k)$$

$$\therefore h(g) \leq h(\phi_k)$$

Ch. 10 [Cover]

Lecture 9 P.1

The Gaussian Channel

$$Y_i = X_i + Z_i \qquad Z_i \sim \mathcal{N}(0, N)$$

The noise Z_i is an i.i.d R.V. with Gaussian distribution with variance N.

Special Cases:

① If the noise variance is zero
⇒ Infinite capacity
⇒ No transmission errors

② No constraint on the input.
⇒ Choose infinite subset of inputs arbitrarily apart
⇒ Infinite capacity

However, there are always constraints on the inputs in terms of power or energy in addition to limited alphabet size. For any codeword $(x_1, x_2, \dots, x_n)$ transmitted over the channel, we require $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le P$

Gaussian Channel Capacity

Def: The information capacity of the Gaussian channel with power constraint P is

~~$C = \max_{p(x)\ \int} I(X;Y)$~~

$$C = \max_{p(x):\, E[X^2] \le P} I(X;Y)$$

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(X+Z|X) \\ &= h(Y) - h(Z|X) \\ &= h(Y) - h(Z) \end{aligned}$$

since Z is independent of X.

Also, $h(Z) = \frac{1}{2}\log_2 2\pi e N$

and $E[Y^2] = E[X+Z]^2$

$$= E[X^2] + 2E[X]E[Z] + E[Z^2] = P + N$$

∴ The entropy of Y is bounded by $\frac{1}{2}\log_2 2\pi e(P+N)$

Applying this to $I(X;Y)$

$$I(X;Y) = h(Y) - h(Z)$$
$$\leq \frac{1}{2}\log 2\pi e(P+N) - \frac{1}{2}\log 2\pi eN$$
$$= \frac{1}{2}\log\left(1+\frac{P}{N}\right)$$

So

$$C = \max_{E(X^2)=P} I(X;Y) = \frac{1}{2}\log\left(1+\frac{P}{N}\right)$$

and the maximum is attained when $X \sim \mathcal{N}(0,P)$

Theorem:

The capacity of a Gaussian channel with power constraint P and noise variance N is

$$C = \frac{1}{2}\log\left(1+\frac{P}{N}\right) \text{ bits per transmission}$$

Def: A (M,n) code for the Gaussian channel with power constraint P consists of the following:

1. An index set $\{1,2,\dots,M\}$
2. An encoding function
   $X: \{1,2,\dots,M\} \to \mathcal{X}^n$
   yielding codewords $x^n(1), x^n(2), \dots, x^n(M)$, satisfying the power constraint P
   $$\sum_{i=1}^{n} x_i^2(w) \leq nP$$
   where $w = 1,2,\dots,M$
3. A decoding function
   $g: \mathcal{Y}^n \to \{1,2,\dots,M\}$

Definition:

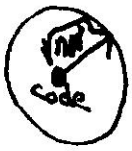
A rate R is said to be achievable for a Gaussian channel with a power constraint P if there exists a sequence of $(2^{nR}, n)$ codes with codewords satisfying the power constraint such that the maximal prob. of error $\lambda^{(n)}$ tends to zero.

The capacity of the channel is the supremum of the achievable rates.

Sphere Packing Argument.

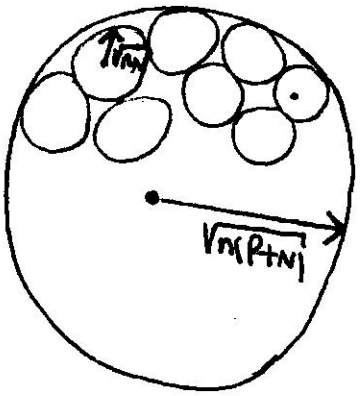
Why we may construct $(2^{nC}, n)$ codes with low Prob. of Error.?

- Consider any codeword of length n. the received vector is normally distributed with mean equal to the true codeword and variance equal to the noise variance.
- For a codeword of length n, the n-dimensional noise variance is nN
- Thus, with high Prob., the received vector is contained in a sphere of radius $\sqrt{nN}$ around the true codeword.



- If we receive any vector inside the sphere, we assign it to the given codeword.
- An error will happen if the received vector falls outside the sphere, which has low Prob. of error.

- Similarly, other codewords will be assigned spheres.
- The Total Energy of the Received vectors is $n(P+N)$. So, they lie in a sphere of radius $\sqrt{n(P+N)}$
- The question is: How many non-overlapping spheres you can fit inside the big sphere?



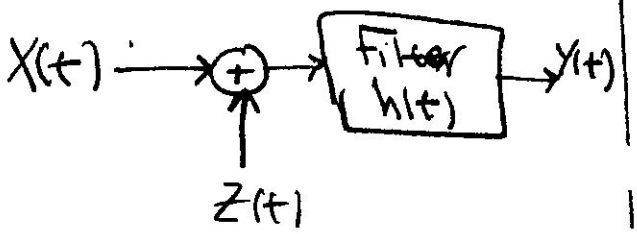
- The Volume of an-n-dimensional sphere is $A_n r^n$

∴ The maximum number of non-intersating spheres is

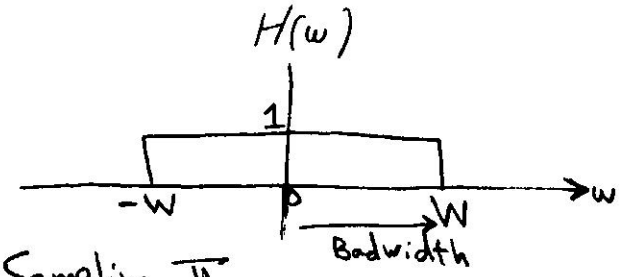
$$\frac{A_n [n(P+N)]^{\frac{n}{2}}}{A_n (nN)^{\frac{n}{2}}} = 2^{\frac{n}{2}\log(1+\frac{P}{N})}$$

compare to $2^{nC} \Rightarrow C = \frac{1}{2}\log(1+\frac{P}{N})$

10.3 Band-Limited Channels



$$Y(t) = (X(t) + Z(t)) * h(t)$$



Sampling Theorem

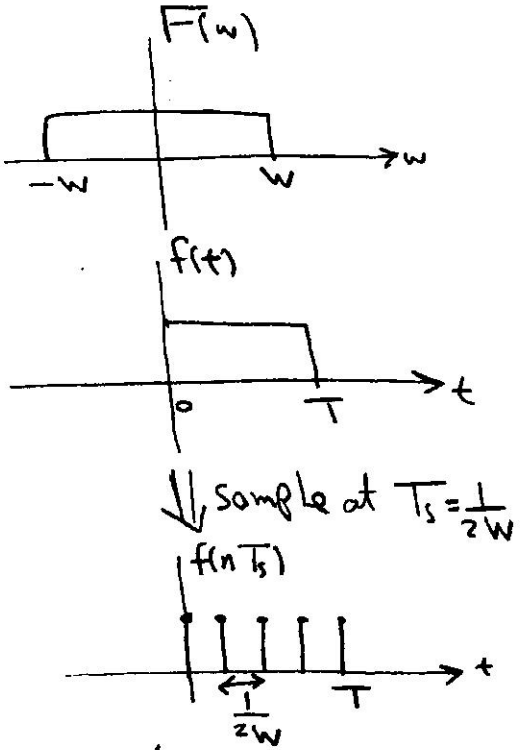
Suppose a function $f(t)$ is band-limited to W, then the function is completely determined by samples of the function spaced $\frac{1}{2W}$ seconds apart.

$\Rightarrow$ sampling frequency $f_s = 2W$ Hz.

$T_s = \frac{1}{2W}$ Second

Almost time-limited Almost band-limited functions

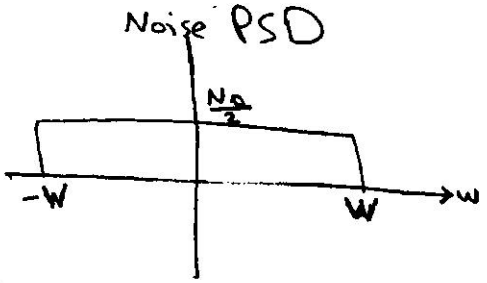
— Most of the energy in bandwidth W and most of the energy in a finite time interval $(0, T)$.



$\Rightarrow$ number of sample in $f(t)$ in the period $(0, T)$ is equal to $2WT$

$\Rightarrow$ The sampled function is a vector in a vector space of $2TW$ dimensions

Band-Limited Noise



- If the noise has power spectral density $\frac{N_0}{2}$ and bandwidth W

$\Rightarrow$ The noise power $= \frac{N_0}{2}(2W) = N_0 W$

- The noise samples in one interval T are i.i.d Gaussian R.V with variance equal to $\frac{N_0 WT}{2WT} = \frac{N_0}{2}$

Capacity of Band-Limited Channels

Recal that the capacity of Gaussian channels is

$$C = \frac{1}{2}\log_2\left(1+\frac{P}{N}\right) \text{ bits per transmission}$$

- Let the channel be used over the time interval $[0,T]$. In this case, the power per sample is $\frac{PT}{2WT} = \frac{P}{2W}$

- The noise variance per sample is $\frac{N_0}{2}$

$\Rightarrow$ The capacity per sample is

$$C = \frac{1}{2}\log\left(1+\frac{\frac{P}{2W}}{\frac{N_0}{2}}\right)$$

$$C = \frac{1}{2}\log\left(1+\frac{P}{N_0 W}\right) \text{ bits per sample.}$$

(P: Signal Power; N_0W: Noise Power)

Since there are 2W samples each second

$\Rightarrow$ the capacity can be rewritten as

$$C = W\log\left(1+\frac{P}{N_0 W}\right) \text{ bits per second}$$

For Reliable Comm.

$$R < C$$

$$R < W \log_2\left(1 + \frac{P}{N_0 W}\right)$$

Recall that the bandwidth efficiency is:

$$\frac{R}{W} = r < \log_2\left(1 + \frac{P}{N_0 W}\right)$$

Since $E_b = \frac{E_s}{\log_2 M} = \frac{P T_s}{\log_2 M} = \frac{P}{R}$

$$\therefore r < \log_2\left(1 + \frac{E_b R}{N_0 W}\right)$$

$$r < \log_2\left(1 + r\frac{E_b}{N_0}\right)$$

$$\therefore 2^r < 1 + r\frac{E_b}{N_0}$$

$$\Rightarrow \frac{E_b}{N_0} > \frac{2^r - 1}{r}$$

The minimum value

$\frac{E_b}{N_0}$ for which reliable communication is possible is obtained by letting $r \to 0$

$$\Rightarrow \frac{E_b}{N_0} > \ln 2 \approx 0.693 \approx -1.6 \text{ dB}$$

* This is the minimum value for $\frac{E_b}{N_0}$, No system can transmit reliably below this limit.

* In order to achieve this limit, we need to let $r \to 0$ or $W \to \infty$

Practice Problems

Problems 6.4, 6.5
6.10, 6.11
6.15, 6.42
6.43, 6.46
6.58, 6.69, 6.72