EE4152
Digital Communications

A/Prof Li Kwok Hung
EEE-S1-B1b-62
Ext-5028
Ekhli@ntu.edu.sg
EEE, NTU

Topics for Information Theory

- Measure of Information
- Entropy of an Information Source
- Source Encoding
- Channel Representation
- Joint and Conditional Entropy
- Channel Capacity
- Shannon’s Theorems

Textbook & References

- **B P Lathi and Z Ding**, *Modern Digital and Analog Communication Systems*, 4/Ed, Oxford University Press, 2010
- **S Haykin and M Moher**, *Communication Systems*, 5/Ed, John Wiley, 2010.
- **J G Proakis and M Salehi**, *Communication Systems Engineering*, 2/Ed, Prentice-Hall, 2002

Measure of Information

Introduction

Consider the following three statements:

- I shall continue my lecture next time
- Another professor will lecture next time
- There will be no more EE4152 lectures

Which one conveys more information?

Probability & Information

It appears that the lower the **probability** of a statement, the greater is the **information** conveyed by that statement. Accordingly, the third statement provides more information to you.

Measure of Information

Let x_j be an event which occurs with probability $p(x_j)$. Then the information associated with x_j is

$$I(x_j) = \log_a \frac{1}{p(x_j)} = -\log_a p(x_j)$$

The base of the logarithm is arbitrary and determines the units. If $a = 2$, then the unit of information is *bit*.

Example

Consider a random experiment with 8 equally likely outcomes. The information associated with each outcome is

$$I(x_j) = -\log_2 \frac{1}{8} = \log_2 8 = 3\,\text{bits}$$

where j ranges from 1 to 8. Note that the information is greater than one bit.

Entropy of an Information Source

Entropy

The average information associated with a discrete source X is

$$H(X) = \sum_{j=1}^{n} p(x_j)\log_2 \frac{1}{p(x_j)}$$

$$= -\sum_{j=1}^{n} p(x_j)\log_2 p(x_j)$$

where n is the total number of possible outcomes.

Entropy …

- Entropy can be regarded as average **uncertainty** associated with the information source.
- The uncertainty is **maximum** when all outcomes are **equally likely** (see the following example).

Example

For a binary source X with $p(1) = \alpha$ and $p(0) = 1 - \alpha$. Derive the entropy of the source as a function of α and sketch $H(X)$ as α varies from 0 to 1.

$$H(X) = \Omega(\alpha) \equiv -\alpha \log_2 \alpha - (1-\alpha)\log_2(1-\alpha)$$

where $\Omega(\alpha)$ is the **entropy function**. The plot is shown in Fig. 1. Note that $\alpha = 0.5$ yields the maximum value of $H(X)$.

Example …

If α is not equal to 0.5, then one symbol is more likely to occur than the other. We thus have less uncertainty about the source output.

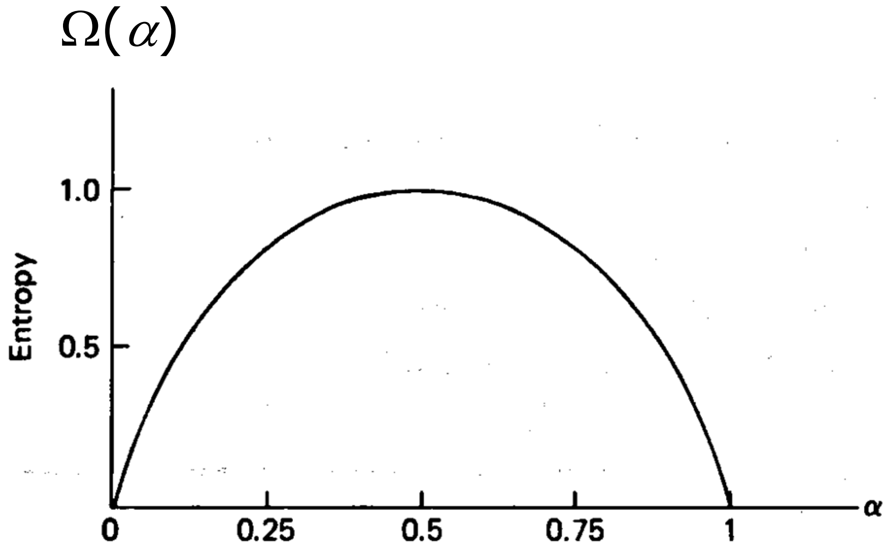


Fig.1: Entropy of a binary source.

Example

For a source X emitting 2000 symbols/sec from an alphabet of size 4 with probabilities ½, ¼, 1/8 and 1/8, respectively. The entropy is

$$H(X) = \frac{1}{2} \times \log_2(2) + \frac{1}{4} \times \log_2(4) + \frac{1}{8} \times \log_2(8) + \frac{1}{8} \times \log_2(8)$$

$$= 1.75 \quad \text{bits/symbol}$$

The information rate is thus

$$R_b = 2000 \times 1.75 = 3500 \text{ bits/sec}$$

Exercise

A message source generates one of four messages randomly every microsecond. The probabilities of these messages are 0.4, 0.3, 0.2 and 0.1. Each emitted message is independent of the other messages in the sequence. Compute the source entropy and the information rate.

Answers: $H(X) = 1.8464$ bits/symbol

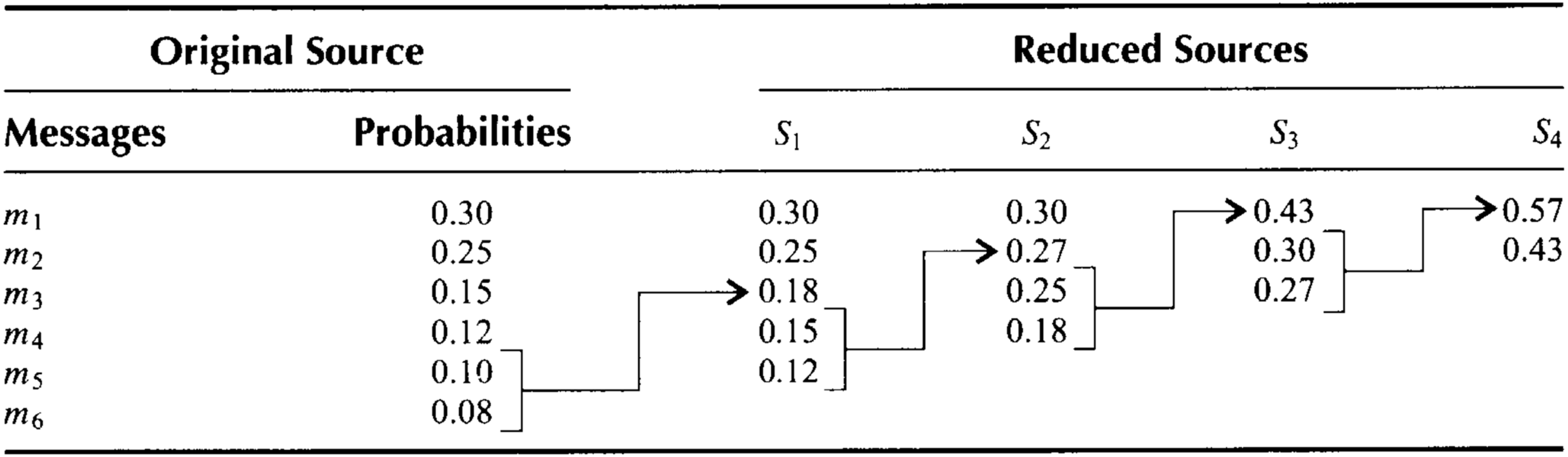
$R_b = 1.8464 \times 10^6$ bits/s

Source Encoding

Introduction

Encoding a source X with entropy $H(X)$ needs, on the average, a minimum of $H(X)$ bits per message. Thus the average word length of an optimum code is $H(X)$. Unfortunately, to attend this limit, we have to encode a sequence of N messages ($N \rightarrow \infty$) at a time.

Optimum (Huffman) Code



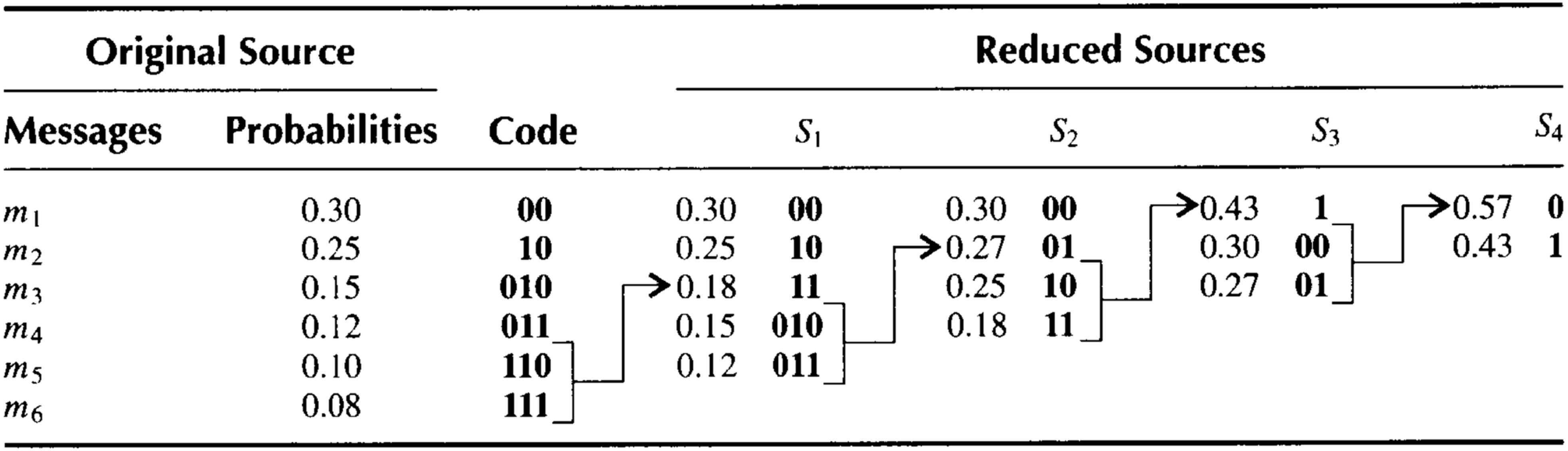
Original Source		Reduced Sources			
Messages	Probabilities	S_1	S_2	S_3	S_4
m_1	0.30	0.30	0.30	0.43	0.57
m_2	0.25	0.25	0.27	0.30	0.43
m_3	0.15	0.18	0.25	0.27	
m_4	0.12	0.15	0.18		
m_5	0.10	0.12			
m_6	0.08				

Fig. 2: Example of Huffman encoding.

Coding Procedure

- The last two messages are now assigned 0 and 1 as their first digits.
- We now go back and assign the numbers 0 and 1 to the second digit for the two messages that were combined in the previous step.
- We keep regressing this way until the first column is reached.

Optimum Encoding



Original Source			Reduced Sources							
Messages	Probabilities	Code	S_1		S_2		S_3		S_4	
m_1	0.30	**00**	0.30	**00**	0.30	**00**	0.43	**1**	0.57	**0**
m_2	0.25	**10**	0.25	**10**	0.27	**01**	0.30	**00**	0.43	**1**
m_3	0.15	**010**	0.18	**11**	0.25	**10**	0.27	**01**		
m_4	0.12	**011**	0.15	**010**	0.18	**11**				
m_5	0.10	**110**	0.12	**011**						
m_6	0.08	**111**								

Fig. 3: Example of resulting codewords.

Compact Code

The optimum code obtained in this way is called a **compact code**. The average length of the compact code is

$$L = \sum_{i=1}^{6} P_i L_i$$

$$= 0.3(2) + 0.25(2) + 0.15(3) + 0.12(3) + 0.10(3) + 0.08(3)$$

$$= 2.45 \text{ bits}$$

Compact Code …

The entropy $H(X)$ of the source is

$$H(X)=\sum_{i=1}^{6}P_i\log_2\frac{1}{P_i}=2.422\text{ bits}$$

Code efficiency

$$\eta=H(X)/L=2.422/2.45=0.9887$$

Redundancy

$$\gamma=1-\eta=0.011$$

Uniquely Decodable

Huffman code is uniquely decodable. If we receive a sequence of Huffman-coded messages, it can be decoded only one way.

Exercise

Suppose a discrete memoryless source M emits messages m_1, m_2 and m_3 with probabilities 0.1, 0.2 and 0.7, respectively.

(a) Find the entropy of the information source M.

(b) Design a Huffman code for the source M and determine its code efficiency.

Answer:

(a) 1.1568 bits/message

(b) η = 1.1568/1.3 = 0.8898

How to Increase Efficiency?

To achieve code efficiency close to 1, we need to encode a sequence of N messages ($N \rightarrow \infty$) at a time. Normally, the use of $N = 2$ or 3, called **second- or third-order extension**, can yield an efficiency close to 1.

Example

Message	probability	Code
m_1	0.8	0
m_2	0.2	1

A zero-memory source emits messages m_1 and m_2 with probabilities 0.8 and 0.2, respectively. Find the Huffman binary code for this codes for N = 1, 2 and 3.

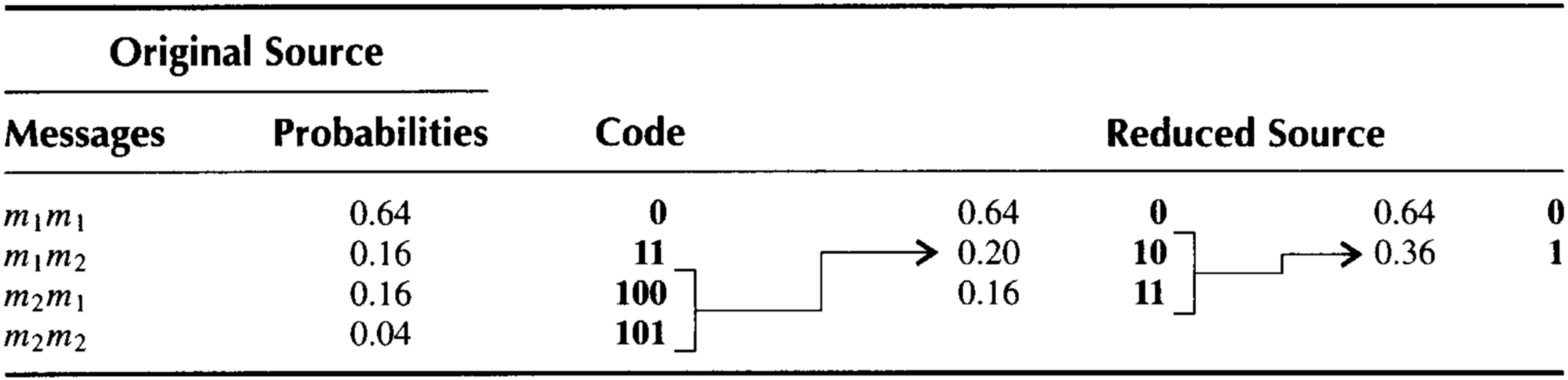
For N = 1, the Huffman code for the source is simply 0 and 1. Hence,

$$H(X) = -(0.8\log_2 0.8 + 0.2\log_2 0.2) = 0.72 \text{ bit}$$

$$\eta = H(X)/L = 0.72/1 = 0.72$$

Example …

For $N = 2$, there are 4 composite messages.



Original Source			Reduced Source			
Messages	Probabilities	Code				
m_1m_1	0.64	0	0.64	0	0.64	0
m_1m_2	0.16	11	0.20	10	0.36	1
m_2m_1	0.16	100	0.16	11		
m_2m_2	0.04	101				

Fig. 4: Second-order extension.

$$\boxed{p(m_1m_1) = (0.8)(0.8) = 0.64, \quad p(m_1m_2) = (0.8)(0.2) = 0.16}$$

Example …

In this case, the average word length is

$$L' = 0.64(1) + 0.16(2) + 0.16(3) + 0.04(3)$$

$$= 1.56 \text{ bits}$$

For each message, the average word length is

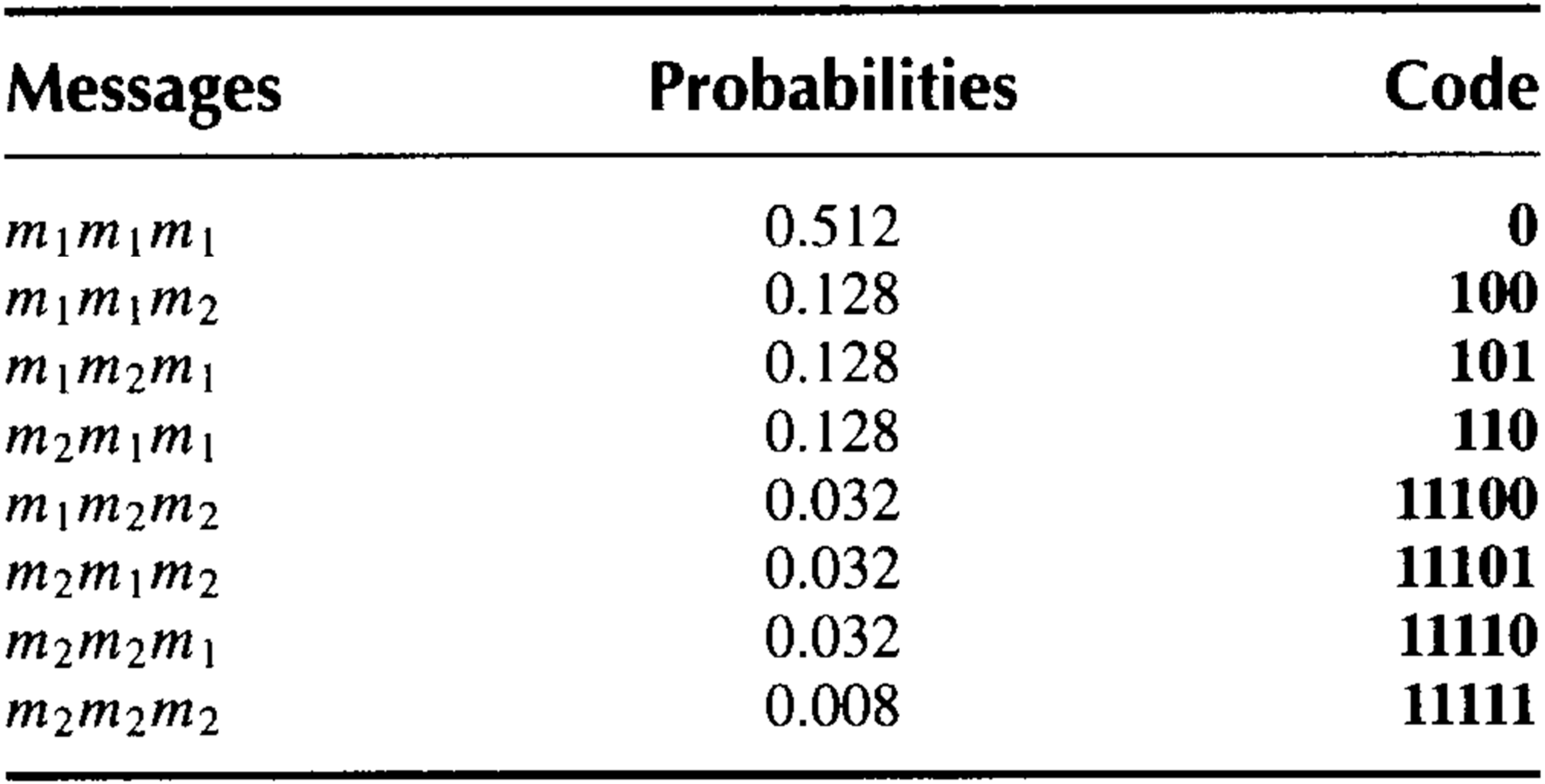
$$L = L'/2 = 1.56/2 = 0.78 \text{ bit}$$

and

$$\eta = \frac{H(X)}{L} = \frac{0.72}{0.78} = 0.923$$

$p(m_1m_1m_1)=(0.8)(0.8)(0.8)=0.512$, $p(m_1m_1m_2)=(0.8)(0.8)(0.2)=0.128$

Example …



Messages	Probabilities	Code
$m_1m_1m_1$	0.512	0
$m_1m_1m_2$	0.128	100
$m_1m_2m_1$	0.128	101
$m_2m_1m_1$	0.128	110
$m_1m_2m_2$	0.032	11100
$m_2m_1m_2$	0.032	11101
$m_2m_2m_1$	0.032	11110
$m_2m_2m_2$	0.008	11111

Fig. 5: Third-order extension.

Example …

In this case the average word length is

$$L'' = 0.512(1) + 0.128(3)(3) + 0.032(5)(3) + 0.08(5)$$

$$= 2.184 \text{ bits}$$

For each message, the average word length is

$$L = L''/3 = 2.184/3 = 0.728 \text{ bit}$$

and

$$\eta = \frac{H(X)}{L} = \frac{0.72}{0.728} = 0.989 \quad (\text{close to } 1)$$

Exercise

Continue the exercise on p. 24. Design a Huffman code for the second-order extension of the information source *M* (i.e., take two messages at a time). Determine the code efficiency in this case.

Answer:

$$\eta = \frac{H(X)}{L} = \frac{1.1568}{2.33/2} = 0.9930$$

Channel Representations

Memoryless Discrete Channels

The channel output at a given time is a function of the channel input **at that time** and is not a function of previous channel inputs. Memoryless discrete channels are completely described by **channel transition probabilities** and **input probabilities**.

Channel Diagram

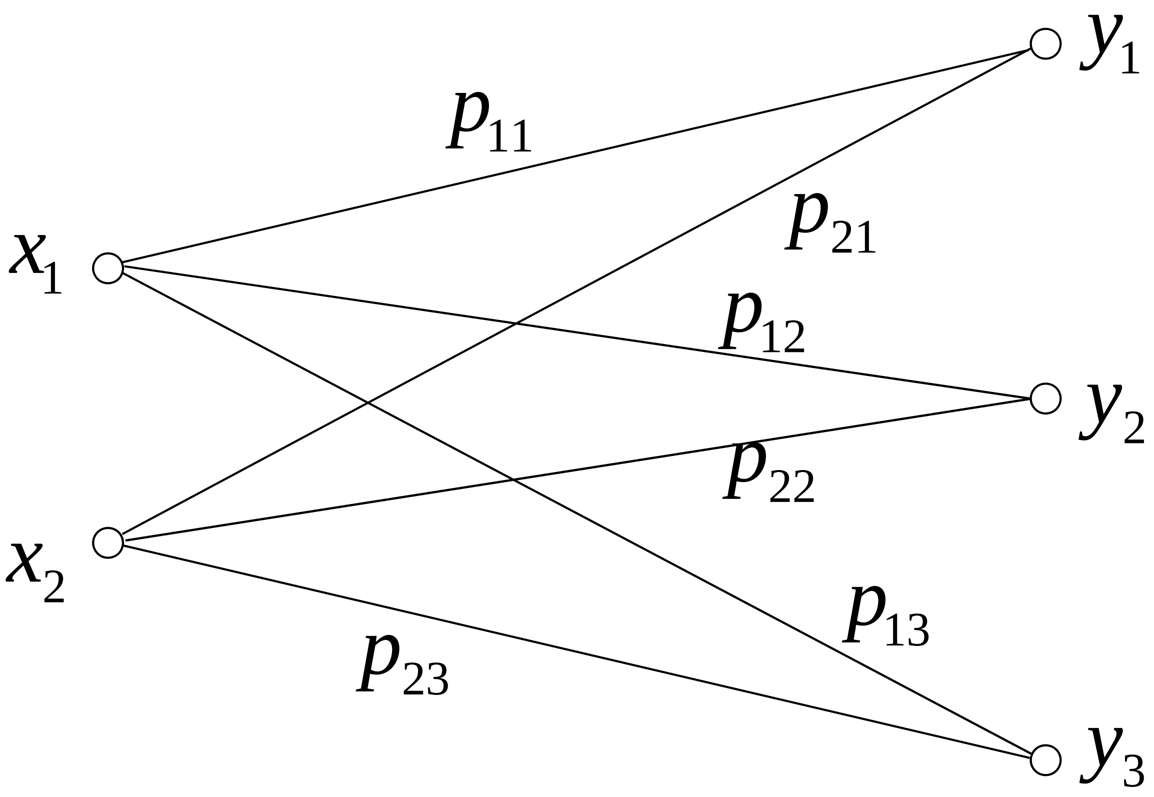


Fig. 6: A channel with 2 inputs and 3 outputs

Transition Probability

- Each input-to-output path is indicated along with a transition probability

$$p_{ij} = p(y_j \mid x_i)$$

- It specifies the conditional probability of obtaining output y_j given that the input is x_i.

Matrix of Transition Probabilities

- For the channel shown in Fig. 6, the matrix of transition probabilities is

$$\mathbf{P}(\mathbf{Y}\,|\,\mathbf{X}) = \begin{bmatrix} p(y_1\,|\,x_1) & p(y_2\,|\,x_1) & p(y_3\,|\,x_1) \\ p(y_1\,|\,x_2) & p(y_2\,|\,x_2) & p(y_3\,|\,x_2) \end{bmatrix}$$

- Since each input to the channel results in some output, each row of the channel matrix must sum to unity.

$$\sum_{j=1}^{3} p(y_j\,|\,x_i) = 1, \quad i = 1, 2$$

Input & Output Probabilities

For the channel shown in Fig. 6, we can write input probabilities compactly as

$$\mathbf{P(X)} = \left[p(x_1) \quad p(x_2) \right]$$

and

$$\mathbf{P(Y)} = \left[p(y_1) \quad p(y_2) \quad p(y_3) \right]$$

It is easy to see that

$$\mathbf{P(Y)} = \mathbf{P(X)P(Y \mid X)}$$

$$\text{vector} \Rightarrow \mathbf{P(X)} = \begin{bmatrix} p(x_1) & p(x_2) \end{bmatrix}, \quad \text{diagonal matrix} \Rightarrow \mathbf{P(X)} = \begin{bmatrix} p(x_1) & 0 \\ 0 & p(x_2) \end{bmatrix}$$

Joint Probability Matrix

If **P(X)** is written as a diagonal matrix, then **P(X)P(Y|X)** yields a matrix **P(X,Y)**. This matrix is called the **joint probability matrix**. The term $p(x_i, y_j)$ is the joint probability of transmitting x_i and receiving y_j. Note that

$$\sum_i \sum_j p(x_i, y_j) = 1$$

Example

$$\mathbf{P(Y \mid X)} = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

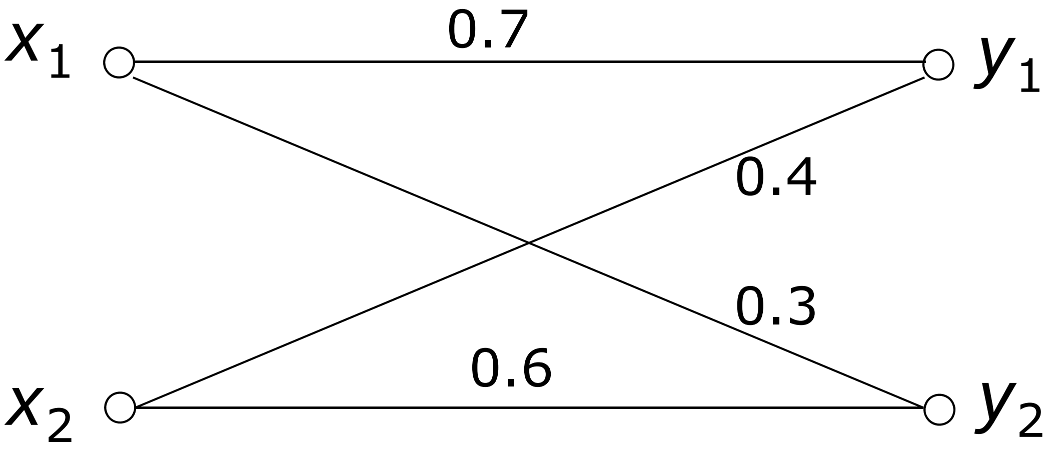


Fig. 7: Binary channel.

Example …

- If the input probabilities are $p(x_1) = p(x_2) = 0.5$, the output probabilities are

$$\mathbf{P(Y)} = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.55 & 0.45 \end{bmatrix}$$

- The joint probability matrix is

$$\mathbf{P(X, Y)} = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.35 & 0.15 \\ 0.2 & 0.3 \end{bmatrix}$$

Joint & Conditional Entropy

$H(X)$ & $H(Y)$

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

$H(X)$ is the average uncertainty of the source.

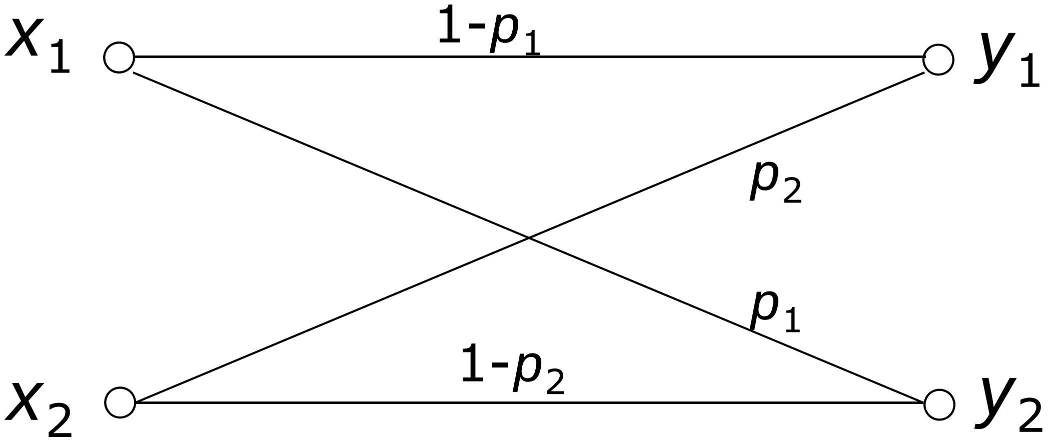
$$H(Y) = -\sum_{j=1}^{m} p(y_j)\log_2 p(y_j)$$

$H(Y)$ is the average uncertainty of the received symbol.

Exercise

$$p(x_1) = \alpha$$

$$p(x_2) = 1 - \alpha$$



Binary asymmetric channel.

Determine the output probability vector and its entropy $H(Y) = \Omega(p_2+(1\text{-}p_1\text{-}p_2)\alpha)$.

$H(Y|X)$

$$H(Y \mid X) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 p(y_j \mid x_i)$$

The function $H(Y|X)$ is the average uncertainty of the received symbol given that X was transmitted.

H(*X*|*Y*)

$$H(X|Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 p(x_i|y_j)$$

The function $H(X|Y)$ is the average uncertainty of the transmitted symbol after we have received a symbol.

H(*X*,*Y*)

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 p(x_i, y_j)$$

The joint entropy $H(X,Y)$ is the average uncertainty of the communication system as a whole. Two important relationships are

$$\begin{aligned} H(X,Y) &= H(X\,|\,Y) + H(Y) \\ &= H(Y\,|\,X) + H(X) \end{aligned}$$

Exercise

Continue the exercise on p. 43. Compute **P(X, Y)** and $H(Y|X)$.

Answer:

$$\mathbf{P(X,Y)} = \begin{bmatrix} \alpha(1-p_1) & \alpha p_1 \\ p_2(1-\alpha) & (1-\alpha)(1-p_2) \end{bmatrix}$$

$$H(Y \mid X) = \alpha\,\Omega(p_1) + (1-\alpha)\Omega(p_2)$$

Channel Capacity

Introduction

Consider an observer at the channel output. The observer's average uncertainty concerning the channel input is $H(X)$ before the reception of an output. His average uncertainty of the input will usually decrease when the output is received. In other words,

$$H(X \mid Y) \leq H(X)$$

Mutual Information

The decrease in the observer's uncertainty of the transmitted signal when the output is received as a measure of the average transmitted information. It is called the **mutual information**

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

and can be expressed as

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

Channel Capacity

The *channel capacity* C_s is defined as the maximum average information per symbol that can be transmitted through the channel, i.e.,

$$C_s = \max\left[I(X;Y)\right]$$

The maximization is with respect to the **source probabilities** since the transition probabilities are fixed by the channel.

Example: Noiseless Discrete channel

- $I(X;Y) = H(X) - H(X|Y)$
- It is obvious that we know the input when the output is given.
- The uncertainty $H(X|Y)$ is zero.
- $I(X;Y) = H(X)$
- $H(X)$ is maximum if all source symbols are equally likely.

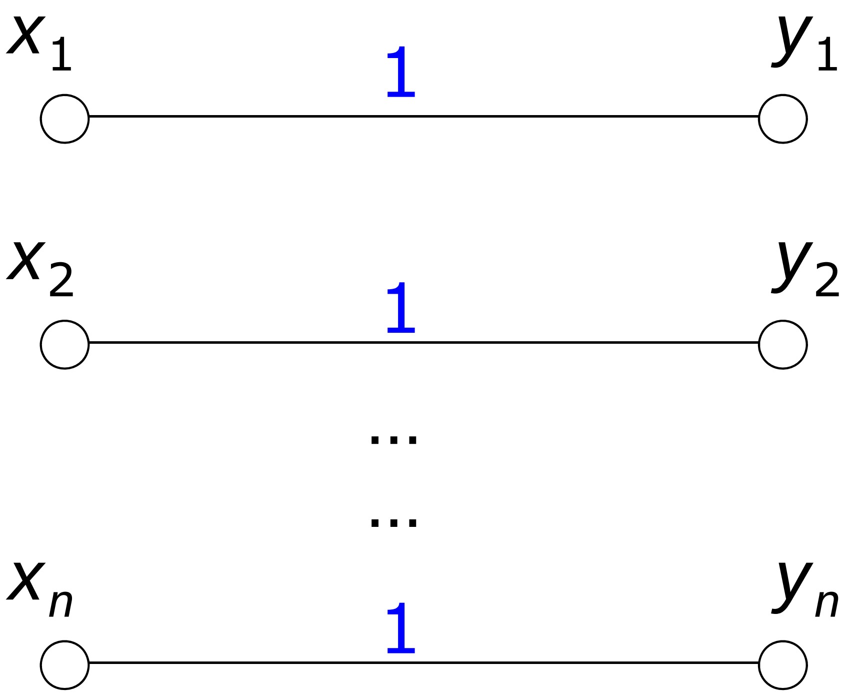


Fig. 8: Noiseless channel.

Example: Noiseless Discrete channel …

- $I(X;Y)$ is maximum if all input symbols are equally likely.
- Hence, the channel capacity for the noiseless discrete channel is

$$C_s = \sum_{i=1}^{n} \frac{1}{n} \log_2 n = \log_2 n$$

Example: Binary Symmetric Channel

$p(x_1) = \alpha$

$p(x_2) = 1 - \alpha$

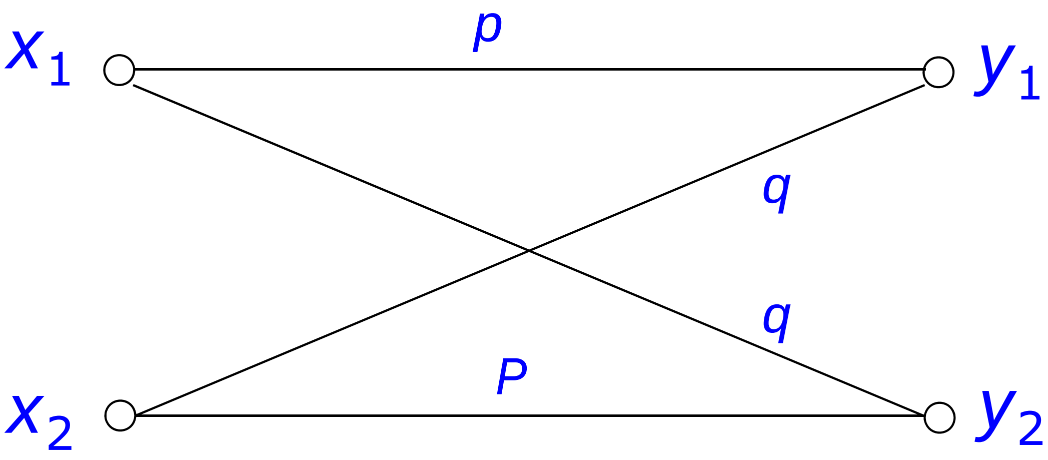


Fig. 9: Binary symmetric channel

$$\boxed{\mathbf{P(Y\,|\,X)} = \begin{bmatrix} p & q \\ q & p \end{bmatrix}, \quad p + q = 1}$$

Example: BSC …

$$\boxed{\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \end{aligned}}$$

- We will determine C_s by maximizing

$$I(X;Y) = H(Y) - H(Y \mid X)$$

- Using the probabilities defined in Fig. 9, we have

$$H(Y \mid X) = -\sum_{i=1}^{2}\sum_{j=1}^{2} p(x_i, y_j)\log_2 p(y_j \mid x_i)$$

$$= -p\log_2 p - q\log_2 q \equiv \Omega(p)$$

Example: BSC …

- Thus

$$I(X;Y) = H(Y) + p\log_2 p + q\log_2 q$$

- Since p and q are fixed by the channel, $I(X;Y)$ is maximized when $H(Y)$ is maximum. We should choose the outputs to be equally likely and $H(Y) = 1$. The channel capacity is

$$C_s = 1 + p\log_2 p + q\log_2 q = 1 - \Omega(p)$$

Equally likely $p(y_1) = p(y_2) = 0.5$ ➔ $H(Y) = 1$

Example: BSC …

- If $p = 0$ or 1, the channel output is completely determined by the channel input, and the capacity is 1 bit per symbol.
- If $p = ½$, the capacity is zero.

$C_s = 1 - \Omega(p)$

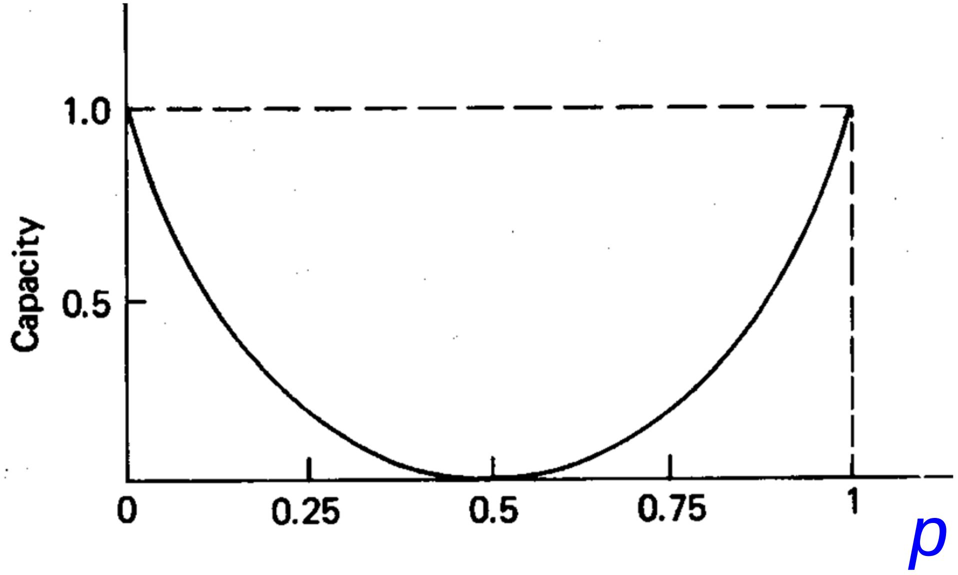


Fig. 10: Capacity of BSC.

Shannon's Theorems

Noiseless Coding Theorem

Given a channel and a source that generates information at a rate (i.e., $H(X)$ bits/symbol) less than the channel capacity (i.e., C_s bits/symbol), it is possible to encode the source output in such a manner that it can be transmitted through the channel.

[Hint: using extension]

Fundamental Theorem of Information Theory

Given a discrete memoryless channel (each symbol is perturbed by noise independent of all other symbols) with capacity C_s and a source with positive rate R, where $R < C_s$, there exists a code such that the output of the source can be transmitted over the channel with an arbitrarily small probability of error.

Fundamental Theorem of Information Theory …

Shannon’s theorem predicts essentially error-free transmission in the presence of noise. The main effect on our communication system is the transmission rate, not error rate.

Unfortunately, the theorem tells us only the existence of codes and tells nothing of how to construct these codes.

Cc [bits/sec]= *Cs* [bits/symbol] x *Rs* [symbols/sec]

The Continuous Channel

The capacity, in **bits per second**, of a continuous channel with additive white Gaussian noise (AWGN) is given by

$$C_c = B \log_2\left(1 + \frac{S}{N}\right)$$

where *B* is the channel bandwidth in Hz and *S/N* is the **signal-to-noise power ratio**. This is known as the **Shannon-Hartley law**.

Tradeoff between *B* and *S*/*N*

For infinite signal-to-noise ratio, which is the noiseless case, the capacity is infinite for any nonzero bandwidth. However, the capacity cannot be made arbitrarily large by increasing bandwidth B if noise is present.

Large-Bandwidth Case

In order to understand the behavior of the Shannon-Hartley law for the large-bandwidth case, the equation can be rewritten as

$$C_c = B\log_2\left(1+\frac{S}{N_0 B}\right) = \frac{S}{N_0}\log_2\left[\left(1+\frac{S}{N_0 B}\right)^{N_0 B/S}\right]$$

where S is the signal power and N_0 is the one-sided power spectral density of noise.

$$\boxed{N = N_0 B, \qquad \log_a x^y = y\log_a x}$$

Large-Bandwidth Case …

Since

$$\lim_{x\to 0}(1+x)^{1/x} = e$$

we have

$$\lim_{B\to\infty} C_c = \frac{S}{N_0}\log_2 e$$

Note that $S = E_b R_b$, where E_b is the energy per bit and R_b is the bit rate. This yields

$$\lim_{B\to\infty} C_c = \frac{E_b R_b}{N_0}\log_2 e$$

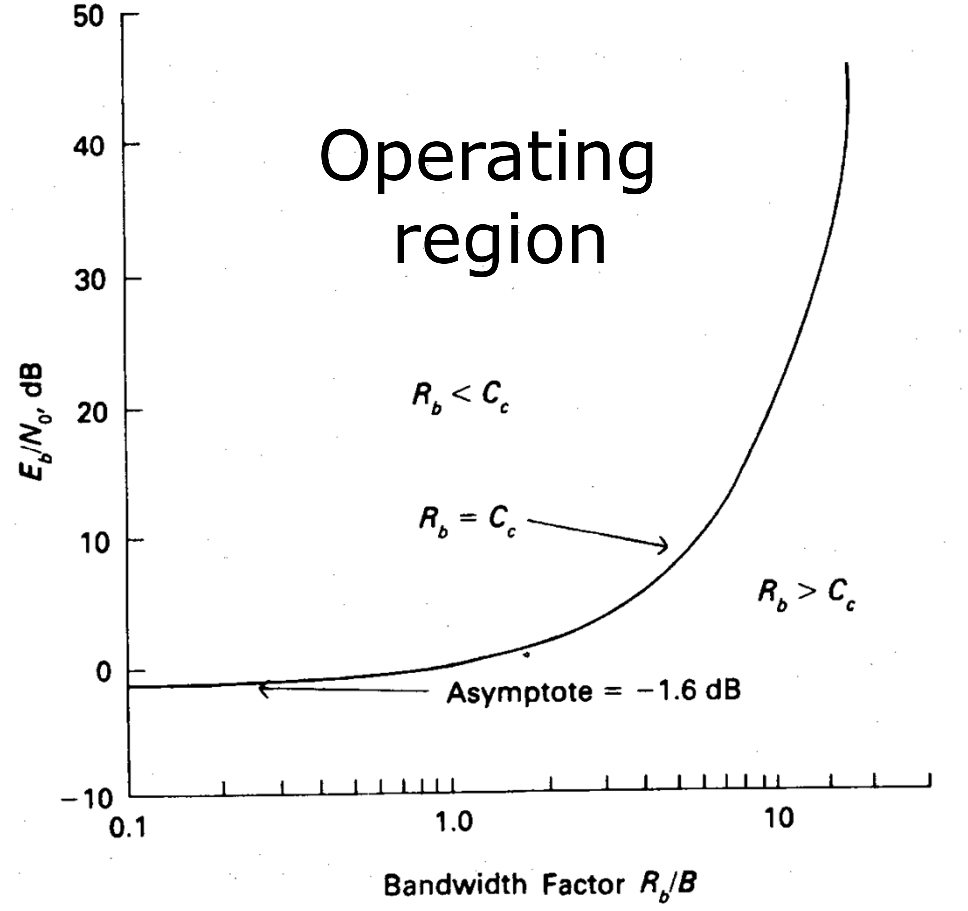
Large-Bandwidth Case …

For R_b approaches C_c, we obtain

$$\frac{E_b}{N_0} = \frac{1}{\log_2 e} = \ln 2 = -1.6\,\text{dB} \qquad \text{as } B \to \infty$$

When $E_b/N_0 > -1.6\,\text{dB}$, we can communicate with zero error. However, reliable communication is not generally possible at lower signal-to-noise ratios.

Tradeoff Plot



$$C_c = B\log_2\left(1+\frac{S}{N_0 B}\right)$$

As R_b approaches C_c,

$$\frac{R_b}{B} = \log_2\left(1+\frac{E_b}{N_0}\frac{R_b}{B}\right)$$

Fig. 11: $R_b = C_c$ relationship for AWGN channel

Power-Limited Operation

If the available bandwidth B is large so that $R_b << B$, operation in the $R_b < C_c$ region requires only that E_b/N_0 is slightly greater than -1.6 dB. Therefore, operation in this region is desired for **power-limited operation**.

Bandwidth-Limited Operation

If the available bandwidth B is limited so that $R_b >> B$, operation in the $R_b < C_c$ region requires a much larger value of E_b/N_0. This is referred to as **bandwidth-limited operation**.

Tradeoff

In an AWGN channel where the Shannon-Hartley law applies, a tradeoff exists between power and bandwidth. This tradeoff is of fundamental important in the design of communication systems.