ECE 8700: Communication Systems Engineering

Module 5
An Introduction to Information Theory

Mojtaba Vaezi
http://www.ece.villanova.edu/ mvaezi

Villanova University

April 10, 2025

Contents

1 Information Theory Basics **4**

1.1 Information 8

1.2 Entropy 9

1.2.1 Joint Entropy 14

1.2.2 Conditional Entropy 15

1.3 Mutual Information 17

2 Data Compression **21**

2.1 Lossless Compression 23

2.2 Variable-Length Source Coding 25

2.2.1 Huffman Coding 32

2.3 Differential Entropy of Continuous Random Variables 39

2.3.1 Mutual Information for Continuous Random Variables 40

3 Channel Models and Channel Capacity **41**

3.1 Channel Models 41

3.1.1 Binary Symmetric Channel (BSC) 41

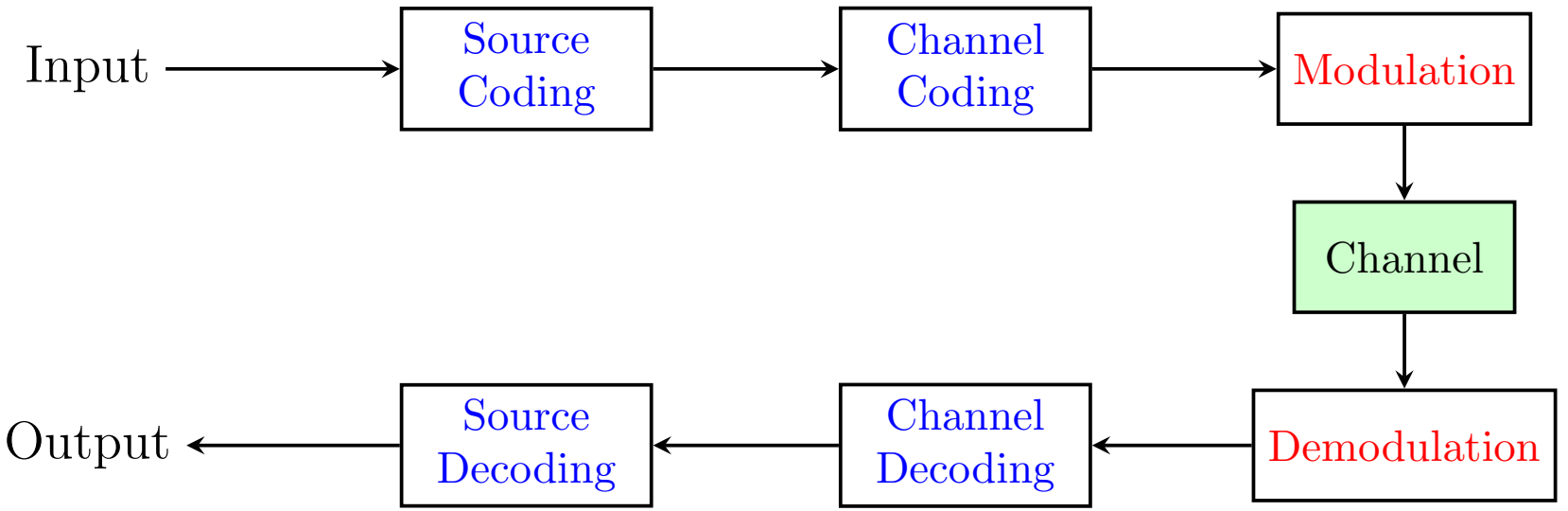
3.1.2 Discrete Memoryless Channel (DMC) 43

3.1.3 The Discrete-Input Continuous-Output Channel 44

3.1.4 The Discrete-Time AWGN Channel 45
3.2 Channel Capacity 46
3.2.1 The Capacity of Discrete-Time AWGN Channel 49
3.2.2 The Capacity of Discrete-Time Binary-Input AWGN Channel 51
3.2.3 The Capacity of Band-Limited AWGN Channel 53

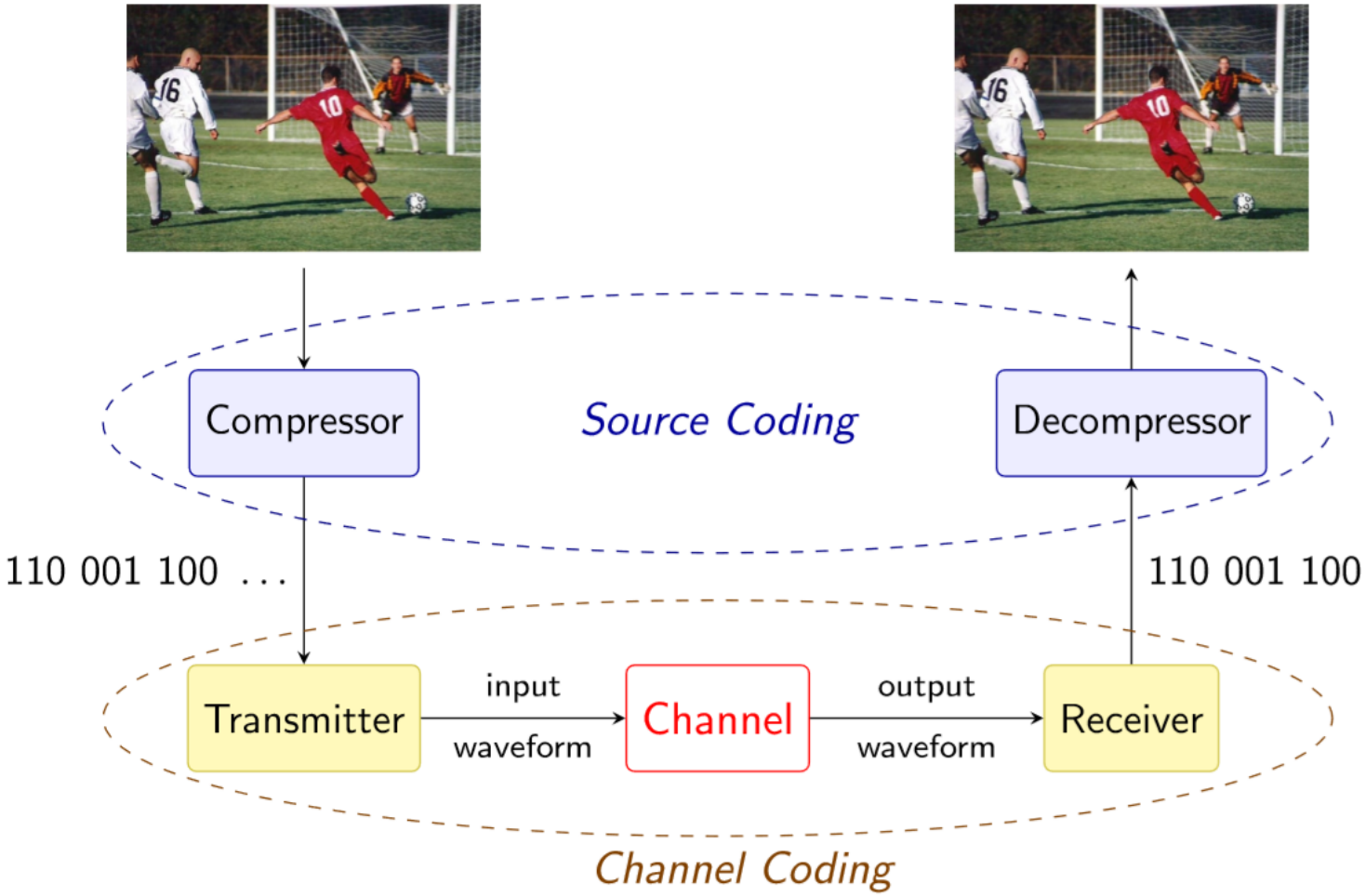
1 Information Theory Basics

- Recall that a digital transmission scheme generally involves



- Input/output are considered digital (if analogue sampled/quantized)
- Your cell-phone, for example, contains a pair of transmitter and receiver, consisting of a **CODEC** and **MODEM**

Entropy & Capacity

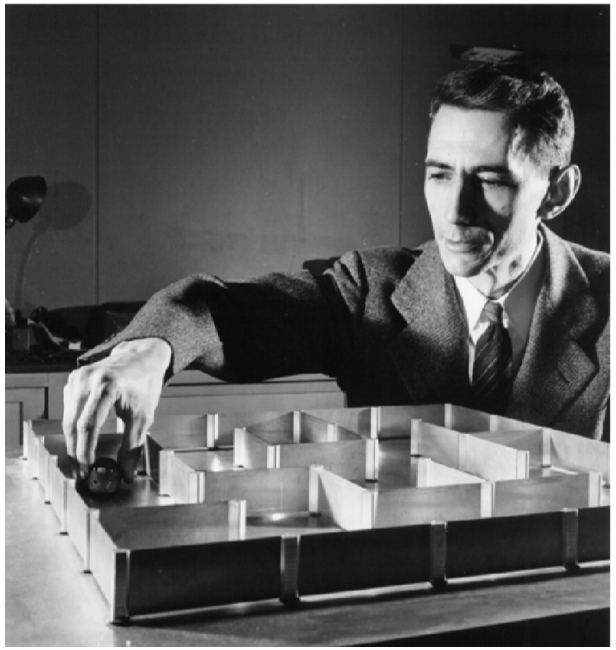


Shannon's 1948 Paper (A Mathematical Theory of Communication)

Shannon posed and answered two fundamental questions:

- Given a source of data, how much can we compress it? (source coding)
- Given a channel, at what rate can we transmit data? (channel coding)

- Claude Elwood Shannon (1916-2001) was a mathematician, electrical engineer, computer scientist, cryptographer and inventor, known as the "father of information theory." [source: Wiki]



Entropy & Capacity

Information theory provides answers to two fundamental questions (among others)

1. What is the irreducible complexity, below which a signal cannot be compressed? (source coding)
1. What is the ultimate transmission rate for reliable communication over a noisy channel? (channel coding)

The answers to these two questions lie in the entropy of a source and the capacity of a channel, respectively:

1. **Entropy** is defined in terms of the probabilistic behavior of a source of information; it is so named in deference to the parallel use of this concept in thermodynamics (gas molecules spread randomly).
2. **Capacity** is defined as the intrinsic ability of a channel to convey information; it is naturally related to the noise characteristics of the channel.

1.1 Information

Q: What is **information**? How to measure it (**unit**)?

Q: Which of the following statements carries more information? Why?

1. There will be a snowstorm in Florida tomorrow.
2. There will be a snowstorm in Alaska tomorrow.

A good measure of information must quantify **uncertainty**, surprise, or knowledge gain.

- Information quantifies how surprising an event is!
- The information content of an event X with probability $p(x)$ may be quantify as

$$I(x) = \log_2 \frac{1}{p(x)} = -\log_2 p(x) \quad (\textbf{bits})$$

1.2 Entropy

The entropy of a discrete random variable X with alphabet $\mathcal{X}$ and probability mass function (pmf) $p(x) = Pr(X = x)$ is defined by

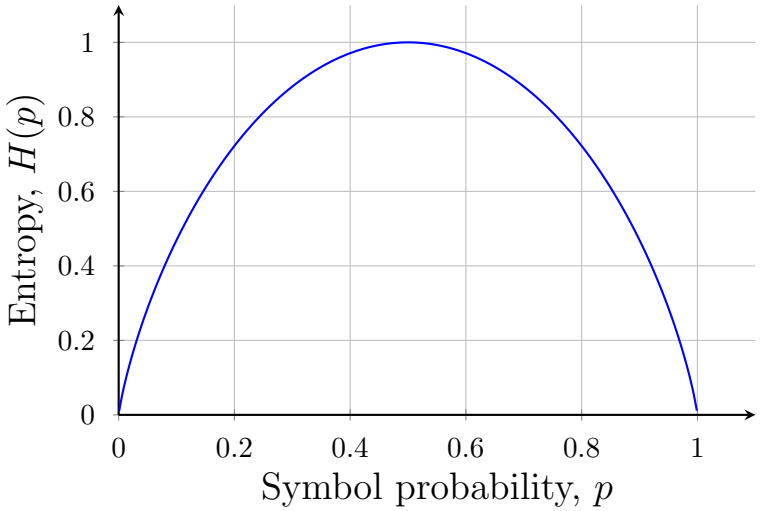
$$H(X) = \sum_{x \in \mathcal{X}} -p(x) \log_2 p(x) \quad \text{bits or bits/symbol}$$

- $H(X)$ is the average information (or **uncertainty**) of all possible outcomes
- Can be written as $\mathbb{E}\left[\log_2 p(x)\right]$
- $H(X)$ is the **uncertainty** associated with the random variable X
- For x such that $p(x) = 0$, "$0 \log \frac{1}{0} = -0 \log 0 = 0$"

- **Binary entropy function** is defined as the entropy of a Bernoulli process (i.i.d. binary variable) with probability p of one of two values, and is given by

$$X = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1 - p. \end{cases}$$

$$H(p) = -p \log_2 p - (1 - p) \log_2 (1 - p) \qquad \text{bits/symbol}$$



Example 1. *Which event has greater entropy?*

- *Let rv X represent the event of US winning the World Cup.*
  $X = 1$ with probability 0.1, and $X = 0$ with probability 0.9
- *Let the rv Y represent the event of rain tomorrow.*
  $Y = 1$ with probability 0.4, and $Y = 0$ with probability 0.6

Properties of Entropy

For a discrete random variable X taking values in $\mathcal{X}$:

1. $H(X) \geq 0$ (because $\frac{1}{p(x)} \geq 1$ implies $\log \frac{1}{p(x)} \geq 0$)
2. If we denote the alphabet size by $|\mathcal{X}|$, then $H(X) \leq \log |\mathcal{X}|$
3. Among all random variables taking values in $\mathcal{X}$, the equiprobable distribution $\left(\frac{1}{|\mathcal{X}|}, \ldots, \frac{1}{|\mathcal{X}|}\right)$ has the maximum entropy, equal to $\log |\mathcal{X}|$.

Example 2. *Let X be a random variable uniformly distributed over 32 outcomes. To represent each outcome, 5-bit strings suffice.*
The entropy is

$$H(X) = -\sum_{i=1}^{32} p(i) \log p(i) = -\sum_{i=1}^{32} \frac{1}{32} \log \frac{1}{32} = \log 32 = 5 \text{ bits/symbol},$$

which matches the number of bits needed to describe X when all outcomes are equally likely.

Example 3. *Suppose a horse race has 8 horses with winning probabilities*

$$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right).$$

Suppose that we wish to send a message indicating which horse won the race. What is the minimum number of bits we need for this?

Solution:

- One alternative is to send the index of the winning horse. This description requires 3 bits for any of the horses.

  However, the entropy of this distribution is

$$\begin{aligned} H(X) &= -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{16}\log_2\frac{1}{16} + 4\times\frac{1}{64}\log_2\frac{1}{64}\right) \\ &= 2 \text{ bits/symbol.} \end{aligned}$$

  So, there must be a more efficient approach.

- To efficiently encode the outcome of the race, we can assign shorter binary codes to more probable outcomes

Horse 1:	0
Horse 2:	10
Horse 3:	110
Horse 4:	1110
Horse 5:	111100
Horse 6:	111101
Horse 7:	111110
Horse 8:	111111

  Then, the average description length is

$$\text{average length} = \frac{1}{2}\cdot 1 \text{ bit} + \frac{1}{4}\cdot 2 \text{ bits} + \frac{1}{8}\cdot 3 \text{ bits} + \frac{1}{16}\cdot 4 \text{ bits} + 4\cdot\frac{1}{64}\cdot 6 \text{ bits} = 2 \text{ bits}$$

- With the above encoding, the average description length matches the entropy, achieving efficient representation of the outcomes.

1.2.1 Joint Entropy

- The **joint entropy** of two random variables X and Y is defined as

$$H(X,Y) = -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)\log_2 p(x,y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2 p(x,y)$$

Example 4. *Let X be the event that tomorrow is cloudy, and Y be the event that it will rain tomorrow. Joint PMF P_{XY} is given by*

Cloudy \ *Rain*	*Yes*	*No*
Yes	$\frac{3}{8}$	$\frac{3}{8}$
No	0	$\frac{1}{4}$

Find $H(X,Y)$ and $H(X)$.

Solution:

$$H(X,Y) = -(\frac{3}{8}\log_2\frac{3}{8} + \frac{3}{8}\log_2\frac{3}{8} + 0\log_2 0 + \frac{1}{4}\log_2\frac{1}{4}) = 1.56 \text{ bits}$$

$$P(X = \text{cloudy}) = \frac{3}{8} + \frac{3}{8} = \frac{3}{4}, \quad P(X = \text{not cloudy}) = \frac{1}{4} \implies H(X) = 0.81 \text{ bits}$$

1.2.2 Conditional Entropy

- *Entropy* is the uncertainty of a single random variable. We can define conditional entropy $H(X|Y)$, which is the entropy of a random variable conditional on the knowledge of another random variable.
- The **conditional entropy** $H(Y \mid X)$, is defined as

$$H(Y \mid X) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(y \mid x)$$

- We can show that $H(Y \mid X) = H(X, Y) - H(X)$

Example 5. *In Example 4, find $H(Y \mid X)$ using both the identity $H(Y \mid X) = H(X,Y) - H(X)$ and the definition of $H(Y \mid X)$.*

Solution:

1. Using the identity we get

$$H(Y \mid X) = H(X,Y) - H(X) = 1.56 - 0.81 = 0.75 \text{ bits}$$

2. To use the definition we need to evaluate $p(y|x)$ first.

$$P(Y = \text{rainy} \mid X = \text{cloudy}) = \frac{\frac{3}{8}}{\frac{3}{8} + \frac{3}{8}} = \frac{1}{2} \qquad P(Y = \text{not rainy} \mid X = \text{cloudy}) = \frac{1}{2}$$

$$P(Y = \text{rainy} \mid X = \text{not cloudy}) = \frac{0}{0 + \frac{1}{4}} = 0 \qquad P(Y = \text{not rainy} \mid X = \text{not cloudy}) = 1$$

$$H(Y \mid X) = -\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P(X = x, Y = y) \log P(Y = y \mid X = x)$$

$$= -\left(\frac{3}{8} \log_2 \frac{1}{2} + \frac{3}{8} \log_2 \frac{1}{2} + 0 \log_2 0 + \frac{1}{4} \log_2 1 \right) = \frac{3}{4} = 0.75 \text{ bits}$$

1.3 Mutual Information

For two random variables X and Y, the mutual information is defined by

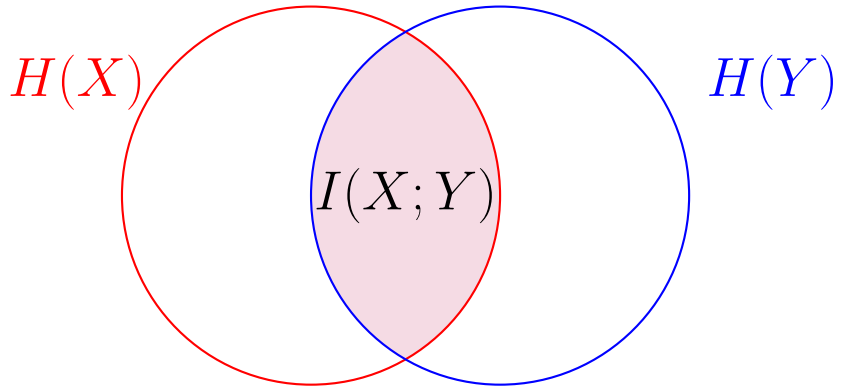
$$I(X;Y) = \sum_{x \in \mathcal{X},\, y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- The mutual information $I(X;Y)$ is a measure of the dependence between the two random variables.
- It says how much information X carries about Y.
- It has the following **properties**
  1. It is symmetric in X and Y, i.e., $I(X;Y) = I(Y;X)$
  2. $I(X;Y) \geq 0$ with equality if and only if X and Y are independent.
  3. $I(X;Y) \leq \min\{H(X), H(Y)\}$
  4. $I(X;X) = H(X)$

- Using the definitions, we can show that

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

- Thus, the reduction in uncertainty due to another random variable is called the *mutual information*.



The union represents $H(X,Y)$

Entropy & Capacity

- Both entropy and mutual information depend on the probabilities of the random variables, not on the specific values they take.
- If an information source is deterministic (i.e., one value of X has probability 1, and all others have probability 0), then the entropy of the source is zero.
- In this case, there is no ambiguity in the source, and it conveys no information.

Example 6. *Evaluate $I(Y;X)$ for Example 4.*

Cloudy \ *Rain*	*Yes*	*No*
Yes	$\frac{3}{8}$	$\frac{3}{8}$
No	0	$\frac{1}{4}$

Solution: From Examples 4 and 5, we have

$$H(X) = 0.81, \qquad H(X,Y) = 1.56, \qquad H(Y \mid X) = 0.75$$

To compute the mutual information $I(X;Y)$, we first calculate $H(Y)$ or $H(Y|X)$.

$$P(Y = \text{rainy}) = \frac{3}{8} + 0 = \frac{3}{8}, \qquad P(Y = \text{not rainy}) = \frac{3}{8} + \frac{1}{4} = \frac{5}{8}$$

$$H(Y) = -\left(\frac{3}{8}\log_2\frac{3}{8} + \frac{5}{8}\log_2\frac{5}{8}\right) \approx 0.96 \text{ bits},$$

$$I(X;Y) = H(Y) - H(Y|X) \approx 0.96 - 0.75 = 0.21 \text{ bits}.$$

2 Data Compression

Data compression is the process of reducing the number of bits needed to represent information.

Shannon's First Theorem (Lossless Source Coding Theorem)

Theorem 1. *Let X denote a discrete memoryless source (DMS) with entropy $H(X)$. Then,*

1. *For any rate $R \geq H(X)$, there exists a lossless source code for X.*
2. *No lossless code exists for X if $R < H(X)$.*

An immediate implication of the above theorem is

The Limit of Lossless Compression

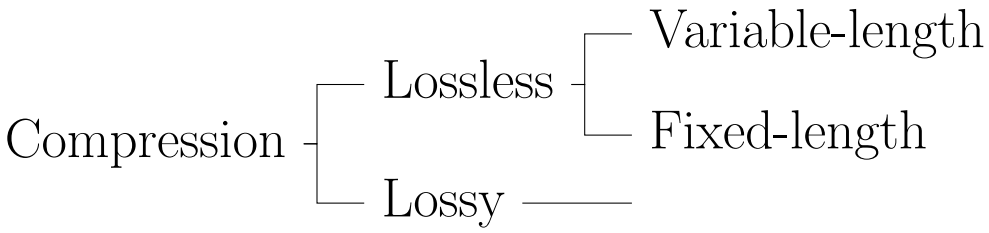
- $H(X)$ is the *fundamental limit* for lossless compression of X. That is, entropy gives the *minimum* average number of bits needed per symbol.

Why compression works?

Most real-world data contains **redundancy**. Compression removes or encodes this redundancy.

- **Example:** In English, letter `e` occurs more frequently than `z`.
- Assigning shorter codewords to frequent symbols saves bits!

In general, compression techniques could be categorized into **lossless** and **lossy**.



2.1 Lossless Compression

- In lossless compression no information is lost (e.g., text, code).
  - Used when exact reproduction is critical.
  - If $H(X) = 2$, no **lossless method** can compress below 2 bits per symbol on average.

    Revisit the horse race example (Example 3)!
- Main lossless compression techniques
  - **Huffman Coding**
    - A **lossless** method that builds a binary tree based on symbol frequencies.
    - Produces optimal **prefix codes** (no code is a prefix of another).
    - Is a variable-length coding algorithm.
  - **Lempel-Ziv Coding:**
    - Dictionary-based method used in ZIP, PNG, and GIF formats.
    - Replaces repeated strings with references to earlier occurrences.
    - Is a fixed-length coding algorithm.

2. Lossy Compression

- In lossy compression some information is discarded (e.g., images, audio).
  - Used when exact reproduction is not necessary.
  - Common in multimedia: JPEG (images), MP3 (audio), MPEG (video).
  - Offers much higher compression at the cost of some quality.
  - Lossy methods dominate in multimedia applications for efficiency.

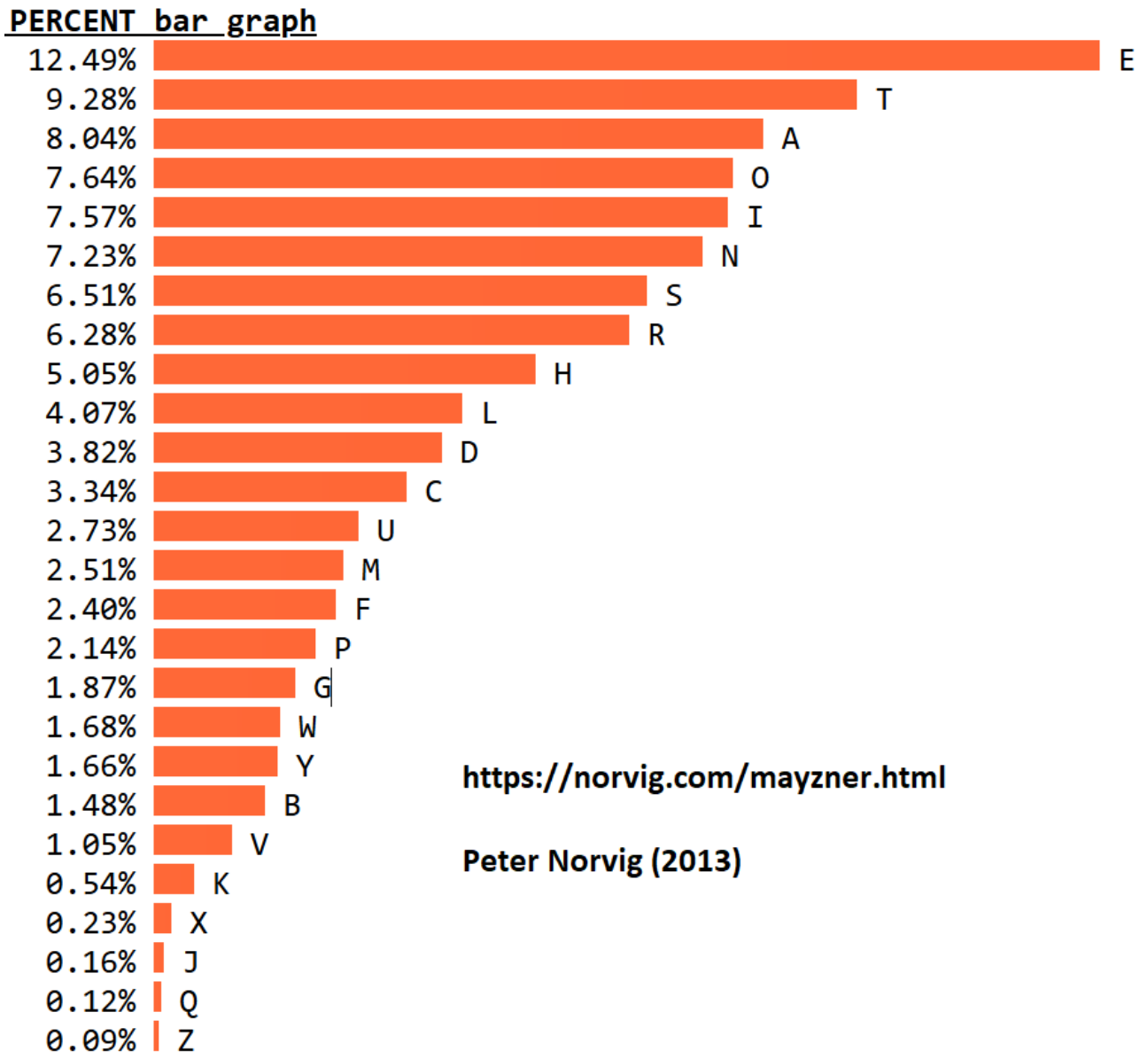
2.2 Variable-Length Source Coding

- When source symbols are not equally probable, an efficient encoding method is to use variable-length codewords.
- An **example** of such encoding is the Morse code (19th century).
  - assigns short codewords to letters that occur more frequently
  - infrequent letters are assigned longer codewords.

A: .-
B: -...
C: -.-.
E: .
H: ....
T: -
O: ---

- Histogram of letter occurrence in English text was evaluated in 2013! (next page)

Entropy & Capacity



- Following this philosophy, we can use the probabilities of occurrence of source letters to guide the selection of codewords.
- The problem is one of how to optimally select and assign codewords to symbols.
- We explain this using the following example.

Example 7. *Suppose that a discrete memoryless source (DMS) with output letters* a_1, a_2, a_3, a_4 *and corresponding probabilities*

- $P(a_1) = \frac{1}{2}$,
- $P(a_2) = \frac{1}{4}$,
- $P(a_3) = P(a_4) = \frac{1}{8}$,

is encoded as shown in Table below. Find the average number of bits per source letter and compare Codes I, II, and III in terms of decodability.

Letter	$P[a_k]$	**Code I**	**Code II**	**Code III**
a_1	$\frac{1}{2}$	1	0	0
a_2	$\frac{1}{4}$	00	10	01
a_3	$\frac{1}{8}$	01	110	011
a_4	$\frac{1}{8}$	10	111	111
$\bar{R}$		1.5	1.75	1.75

Solution:

- All three codes are variable-length codes.
- The average number of bits per source letter $\bar{R}$ (**average codeword length**), is given by

$$\bar{R} = \sum_{k=1}^{L} n_k \, P(a_k)$$

where

  - $\bar{R}$: Average codeword length
  - n_k: Number of bits (length of the codeword) for symbol a_k
  - $P(a_k)$: Probability of symbol a_k

With this we have

  - $\bar{R}_{\text{Code I}} = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 2 + \frac{1}{8} \cdot 2 = 1.5$
  - $\bar{R}_{\text{Code II}} = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = 1.75$
  - $\bar{R}_{\text{Code III}} = \bar{R}_{\text{Code II}}$

- Thus, in terms of codword length, Code I is better than Code II and Code III.

- However, Code I a variable-length code that has a basic flaw.
  - To see the flaw, suppose we are presented with the bit sequence `001001`....
  - The first symbol, corresponding to `00`, is clearly a_2.
  - However, the next 4 bits, `1001`, are ambiguous and not uniquely decodable.
  - They could represent either the sequence a_4a_3 or $a_1a_2a_1$.
  - While the ambiguity might be resolved by waiting for more bits, such decoding delay is undesirable in practice.

> We prefer codes that are decodable instantaneously — i.e., without delay. Such codes are called **instantaneous codes**.

- Code II is both **uniquely decodable** and **instantaneous**.
  - The digit `0` indicates the end of a codeword for the first three codewords.
  - This characteristic, along with the fact that no codeword is longer than three binary digits, ensures that the code is instantaneously decodable.

- Code III is **uniquely decodable** but **not instantaneously decodable**.
  - This means that while each encoded sequence can be decoded unambiguously, decoding cannot begin until the entire sequence is received.
  - Clearly, this code does not satisfy the **prefix condition** — at least one codeword is a prefix of another.
  - As a result, immediate (real-time) decoding is not possible with Code III.
- Finally, Code II is the best as it is the only uniquely and instantaneously decodable code.

2.2.1 Huffman Coding

Huffman coding is a popular algorithm for **lossless** data compression.

1. Arrange all symbols in descending order of probabilities.
2. Build a binary tree by repeatedly merging the two least probable symbols into a new node.
3. Assign '0'/'1' to the higher/less probable branches respectively, in the node.
4. If there is more than one symbol (node) left, return to step 2.
5. Extract the Huffman codewords from the different branches (bottom-up).

Properties of Huffman Coding

- No codeword is a prefix of another, enabling unique and instantaneous decodability.
- Minimizes the average codeword length among all prefix codes.
- Used in formats like JPEG, MP3, and some ZIP implementations.

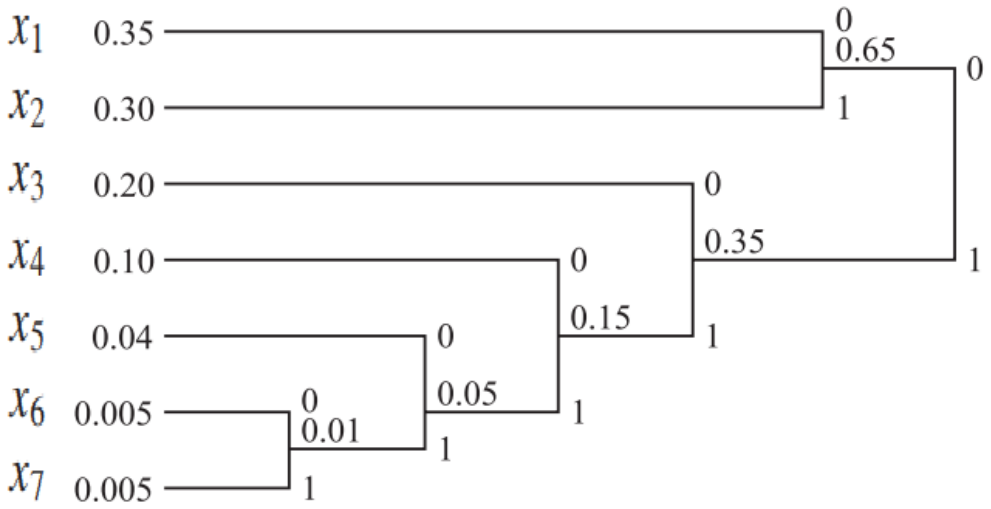
Example 8. *Consider a DMS with seven possible symbols $x_1, \ldots, x_7$ having the probabilities of occurrence listed in the table below.*

Letter	Probability
x_1	0.35
x_2	0.30
x_3	0.20
x_4	0.10
x_5	0.04
x_6	0.005
x_7	0.005

1. *Find the entropy $H(X)$ of the source and the average number of binary digits per symbol for this code, denoted by $\bar{R}$.*
2. *Illustrate your solution using a binary tree diagram.*

Solution:

$$H(X) = -(0.35\log_2 0.35 + 0.3\log_2 0.3 + 0.2\log_2 0.2 + 0.1\log_2 0.1 + 0.04\log_2 0.04 + 2\times\log_2 0.005\log_2 0.005) = 2.11 \quad \text{bits/symbol.}$$



An example of variable-length source encoding for a DMS.

Letter	Probability	Self-information	Code
x_1	0.35	1.5146	00
x_2	0.30	1.7370	01
x_3	0.20	2.3219	10
x_4	0.10	3.3219	110
x_5	0.04	4.6439	1110
x_6	0.005	7.6439	11110
x_7	0.005	7.6439	11111
	$H(X) = 2.11$	$\bar{R} = 2.21$	

- We note that the generated code is **not unique**. There are two reasons:

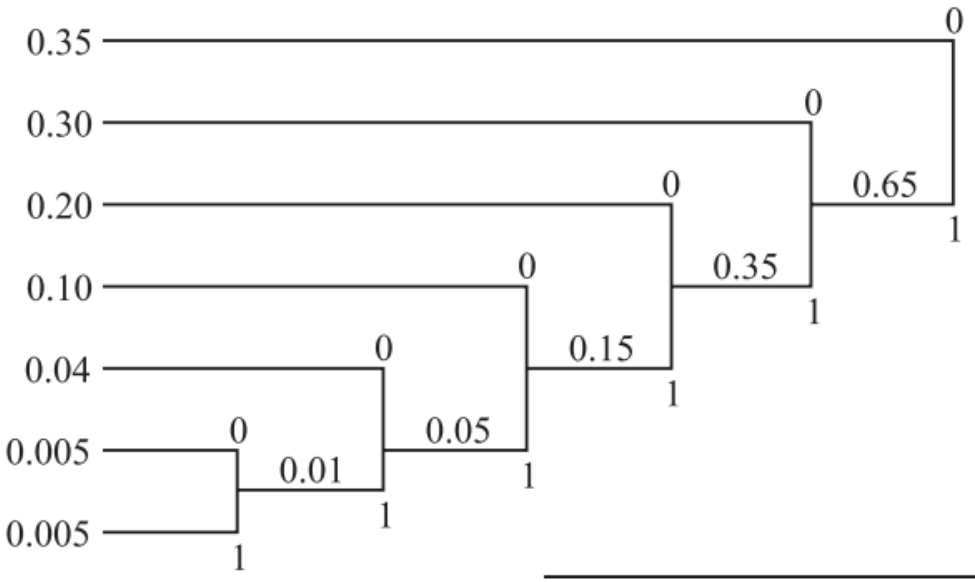
1. Ties in symbol probabilities can lead to different valid trees.

- For example, in the next-to-last step of encoding, x_1 and x_3' have the same probability, i.e., 0.35.
  - In the previous construction, we chose to pair x_1 with x_2
  - An alternative would be to pair x_2 with x_3'.
  - If we choose this alternative pairing, the resulting code (shown in next page) still has the same average number of bits per source symbol $\bar{R} = 2.21$.

- Code efficiency remains unchanged, even with different pairings.

2. We can simply reverse the assignment of 0 and 1 on the branches and still obtain an efficient code.

- The resulting code will continue to satisfy the **prefix condition** and maintain the same average length.



Letter	Code
x_1	0
x_2	10
x_3	110
x_4	1110
x_5	11110
x_6	111110
x_7	111111

$\bar{R} = 2.21$

Example 9. *Consider a DMS with symbol probabilities*

$$0.36, 0.14, 0.13, 0.12, 0.10, 0.09, 0.04, 0.02.$$

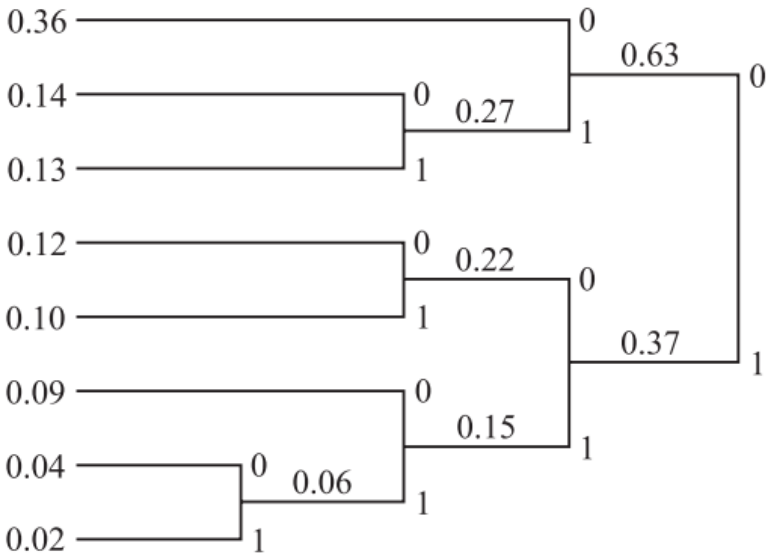
1. *Find the Hufman code for the output of the above source.*
2. *Illustrate your solution using a binary tree diagram.*
3. *What is the average length and the code efficiency?*

Solution:

$$\begin{aligned} H(X) = -(&0.36 \log_2 0.36 + 0.14 \log_2 0.14 + 0.13 \log_2 0.13 + 0.12 \log_2 0.12 + 0.1 \log_2 0.1 \\ &+ 0.09 \log_2 0.09 + \log_2 0.04 \log_2 0.04 + \log_2 0.02 \log_2 0.02) = 2.62 \text{ bits/symbol} \end{aligned}$$

$$\begin{aligned} \bar{R} = &0.36 \times 2 + 0.14 \times 3 + 0.13 \times 3 + 0.12 \times 3 + 0.1 \times 3 + 0.09 \times 3 \\ &+ 0.04 \times 4 + 0.02 \times 4 = 2.7 \end{aligned}$$

$$\text{Code efficiency} = \frac{2.63}{2.7} = 97\%$$



Letter	Code
x_1	00
x_2	010
x_3	011
x_4	100
x_5	101
x_6	110
x_7	1110
x_8	1111
$H(X) = 2.63$	$\bar{R} = 2.70$

2.3 Differential Entropy of Continuous Random Variables

- A continuous random variable requires an infinite number of binary digits to represent it exactly, as a result, its self-information is infinite.
- Therefore, the entropy of a continuous random variable is also infinite.
- To address this, we define the **differential entropy**

$$h(X) = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx$$

- Unlike discrete entropy, differential entropy does not have a direct physical meaning as self-information.

Example 10. *Show that the differential entropy of* $\mathcal{N}(0, \sigma^2)$ *is* $\frac{1}{2}\log_2 2\pi e\sigma^2$.

2.3.1 Mutual Information for Continuous Random Variables

- The definition of mutual information for discrete random variables extends naturally to continuous random variables.
- Let X and Y be **continuous random variables** with joint PDF $p(x,y)$ and marginal PDFs $p(x)$ and $p(y)$. The mutual information between X and Y is defined as

$$I(X;Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)dx\,dy$$

Conditional Entropy and Mutual Information

- The average conditional entropy of X given Y is defined as:

$$h(X|Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\log p(x|y)\,dx\,dy$$

- Using this definition, the average mutual information can be expressed as:

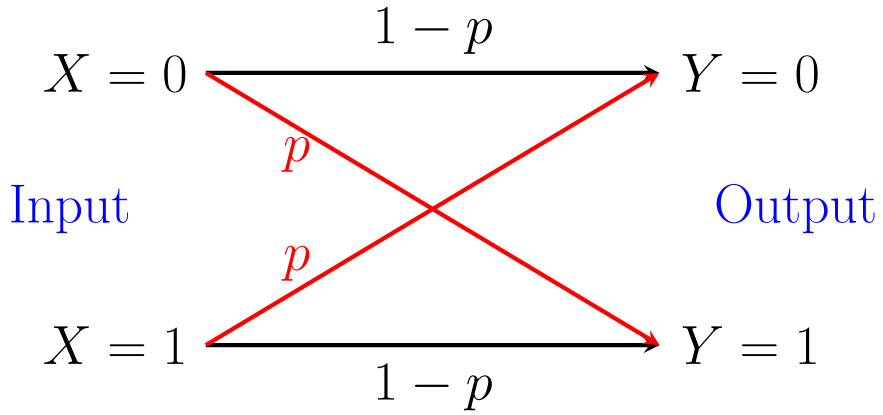
$$I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$$

- This highlights the symmetry of mutual information: $I(X;Y) = I(Y;X)$

3 Channel Models and Channel Capacity

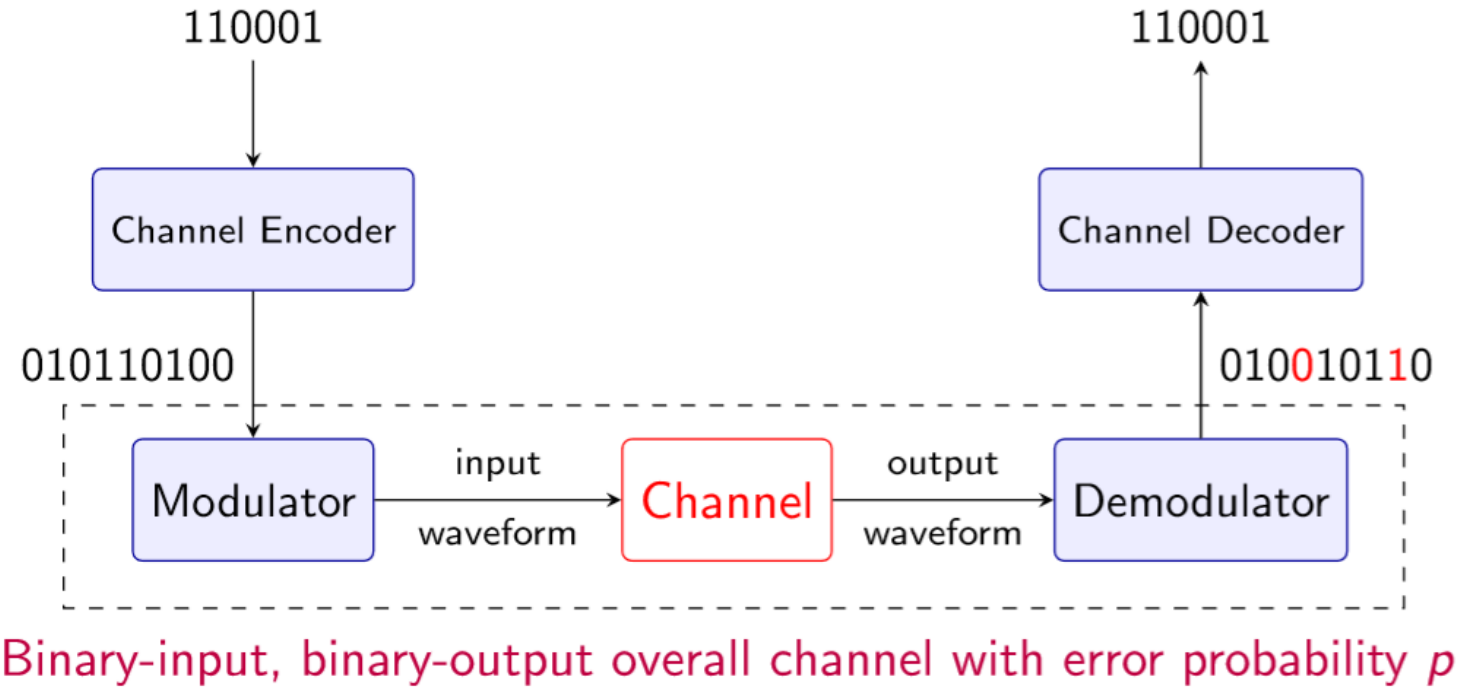
3.1 Channel Models

3.1.1 Binary Symmetric Channel (BSC)



$$P(Y=0 \mid X=0) = 1-p, \qquad P(Y=1 \mid X=0) = p,$$
$$P(Y=1 \mid X=1) = 1-p, \qquad P(Y=0 \mid X=1) = p.$$

The BSC is a simple, yet highly practical and important channel model.


110001
Channel Encoder
010110100
Modulator
input
waveform
Channel
output
waveform
Demodulator
010010110
Channel Decoder
110001
Binary-input, binary-output overall channel with error probability p

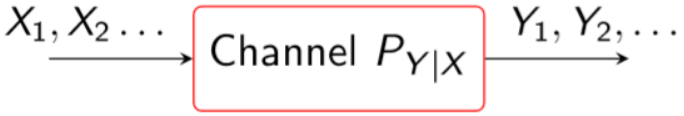
3.1.2 Discrete Memoryless Channel (DMC)

- A **discrete memoryless channel (DMC)** is a channel where
  - The input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ are discrete sets.
  - The channel is **memoryless**, meaning the current output depends only on the current input and not on previous transmissions.



A discrete memoryless channel (DMC) is a system consisting of an input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and a set of *transition* probabilities

$$P_{Y|X}(b|a) = Pr(Y = b|X = a) \quad \textit{for all } a \in \mathcal{X} \textit{ and } b \in \mathcal{Y}.$$



- The BSC is a special case of the DMC in which the input and output are binary.

3.1.3 The Discrete-Input Continuous-Output Channel

- Channel has a **discrete input alphabet** $\mathcal{X}$, with $|\mathcal{X}| = M$, and a **continuous output** $Y \in \mathbb{R}$.
- Characterized by a set of conditional probability density functions

$$p(y \mid x), \quad x \in \mathcal{X},\ y \in \mathbb{R}$$

- A **key example** is the AWGN channel

$$Y = X + N$$

  where $N \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise.
- For a given input $X = x$, the output $Y \sim \mathcal{N}(x, \sigma^2)$, i.e.,

$$p(y \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right)$$

- For input sequence X_i, the output sequence is

$$Y_i = X_i + N_i, \quad i = 1, 2, \ldots, n$$

3.1.4 The Discrete-Time AWGN Channel

- In the discrete-time AWGN channel, both input and output are **real-valued**, i.e., $X, Y \in \mathbb{R}$.
- At each time i, the received symbol is

$$y_i = x_i + n_i$$

  where $n_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. Gaussian noise variables.
- The channel input is subject to a **power constraint**

$$\mathbb{E}[X^2] \leq P$$

- For an input sequence $x = (x_1, x_2, \ldots, x_n)$, with high probability

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n}\|x\|^2 \leq P$$

- **Geometric interpretation**: Valid input sequences lie within an n-dimensional sphere of radius $\sqrt{nP}$ centered at the origin.

3.2 Channel Capacity

Shannon's Second Theorem (Noisy Channel Coding Theorem)

Theorem 2. ***Reliable communication*** *over a discrete memoryless channel (DMC) is possible if the communication rate R satisfies $R < C$, where C is the channel capacity. At rates higher than capacity, reliable communication is impossible.*

- *For an arbitrary DMC, the capacity is given by*

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

where the maximization is over all probability mass functions of the form

$$p = \left(p_1, p_2, \ldots, p_{|\mathcal{X}|}\right)$$

on the input alphabet $\mathcal{X}$.

- *The units of channel capacity C is bits per transmission or bits per channel use.*

In simpler words, Shannon says

1. For any rate $R < C$, there exists a way for reliable communication.
2. Reliable communication is impossible for $R > C$.

- The **units of channel capacity** C depend on the base of the logarithm:
  - *Bits* per transmission if $\log_2$ is used.
  - *Nats* per transmission if the natural logarithm ln (base e) is used.

The following **three units are equivalent** and may be used alternatively.

- Bits per transmission
- Bits per channel use
- Bits per second per Hertz (bps/Hz)

- If a symbol enters the channel every τ_s seconds, then

$$\text{Channel capacity rate} = \frac{C}{\tau_s} \quad (\textbf{bits/s} \text{ or nats/s})$$

Entropy & Capacity

- Shannon's theorem is fundamental to modern communication theory.
- Shannon's proof of the noisy channel coding theorem is **nonconstructive**.
- It uses a powerful technique known as **random coding**.
- Instead of searching for the best coding scheme (a hard task), Shannon
  - Considers *all possible coding schemes*.
  - Analyzes the *average performance* over all such codes.
- He proves that if the communication rate $R < C$, then:

$$\text{Average error probability} \rightarrow 0$$

- This implies that there must exist **at least one** code for which the error probability also tends to zero.
- Thus, reliable communication is theoretically possible even if a specific code is not explicitly given.

3.2.1 The Capacity of Discrete-Time AWGN Channel

- **Discrete-time AWGN channel:** At each time i, the received symbol is

$$y_i = x_i + n_i$$

  where $n_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. Gaussian noise variables.

- The capacity C is given by

$$\boxed{C = \frac{1}{2}\log_2\left(1 + \frac{P}{\sigma^2}\right) \quad \textbf{bits/transmission}}$$

  - This expression is obtained by maximizing the mutual information $I(X;Y)$ over all input distributions $p(x)$ that satisfy the average power constraint

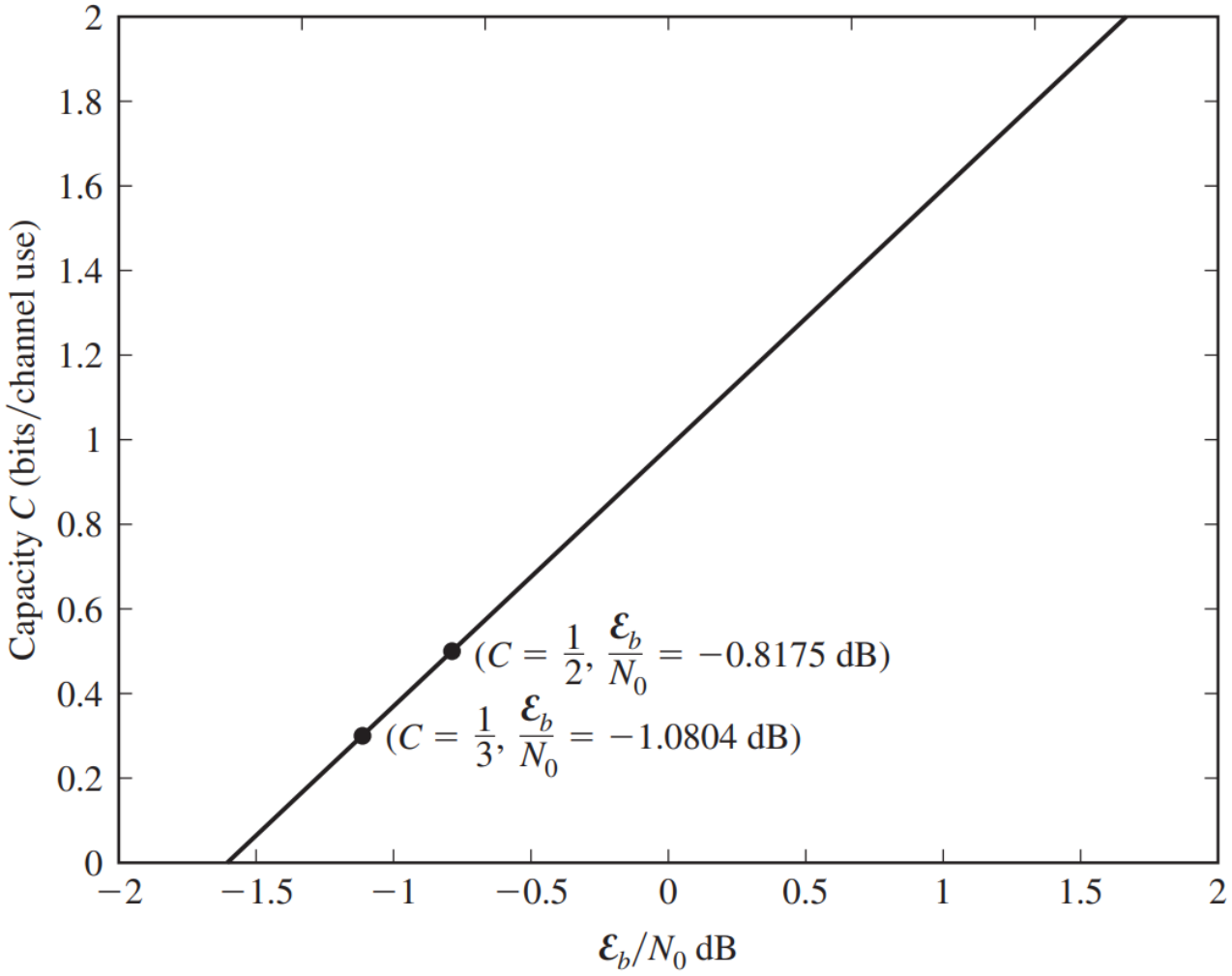
$$\mathbb{E}[X^2] \leq P$$

  - The input distribution that achieves capacity is $X \sim \mathcal{N}(0, P)$ — a **zero-mean Gaussian** distribution with variance P.

- The above formula is equivalent to

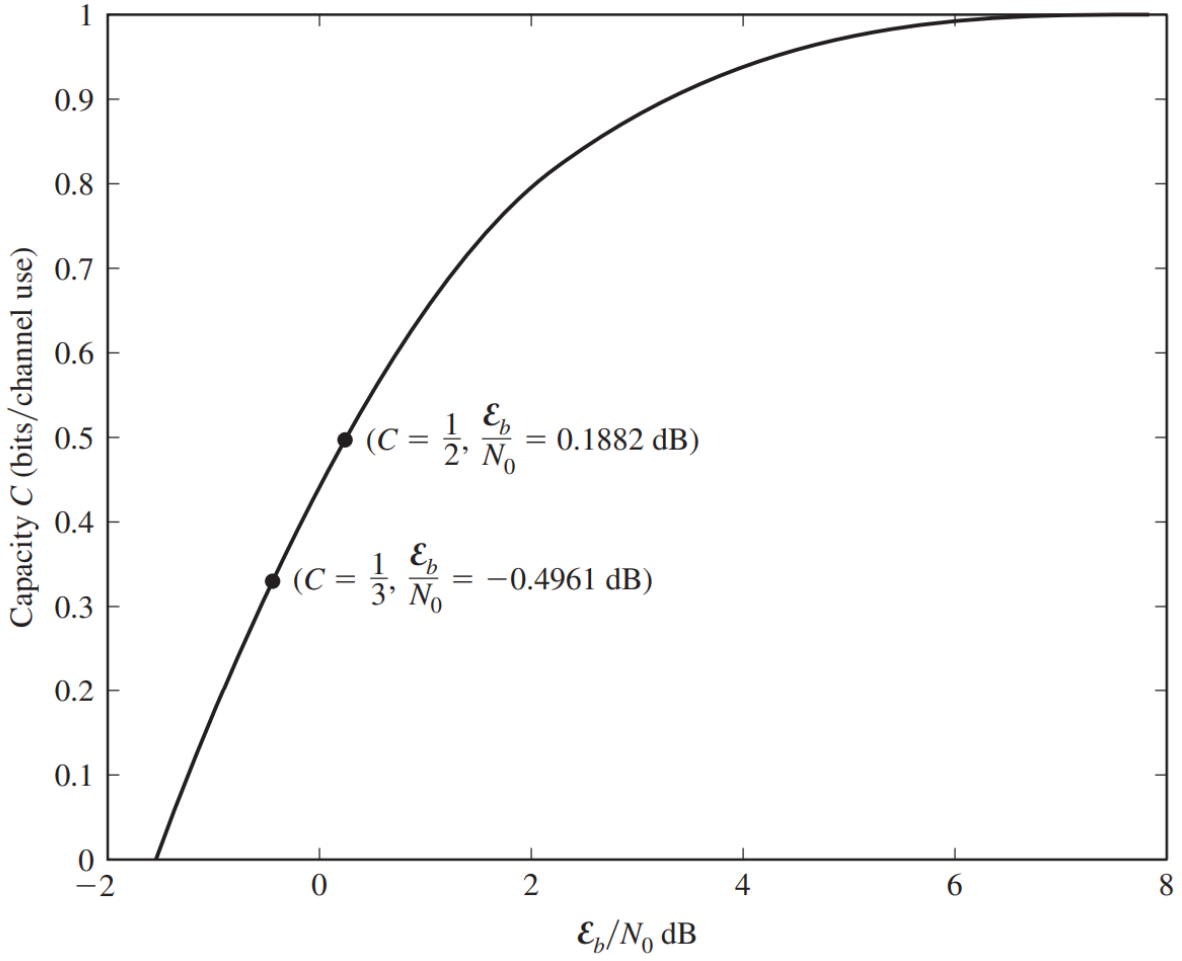
$$\boxed{C = \frac{1}{2}\log_2\left(1 + \text{SNR}\right) \quad \textbf{bits/transmission}}$$

- A plot of channel capacity versus SNR per bit

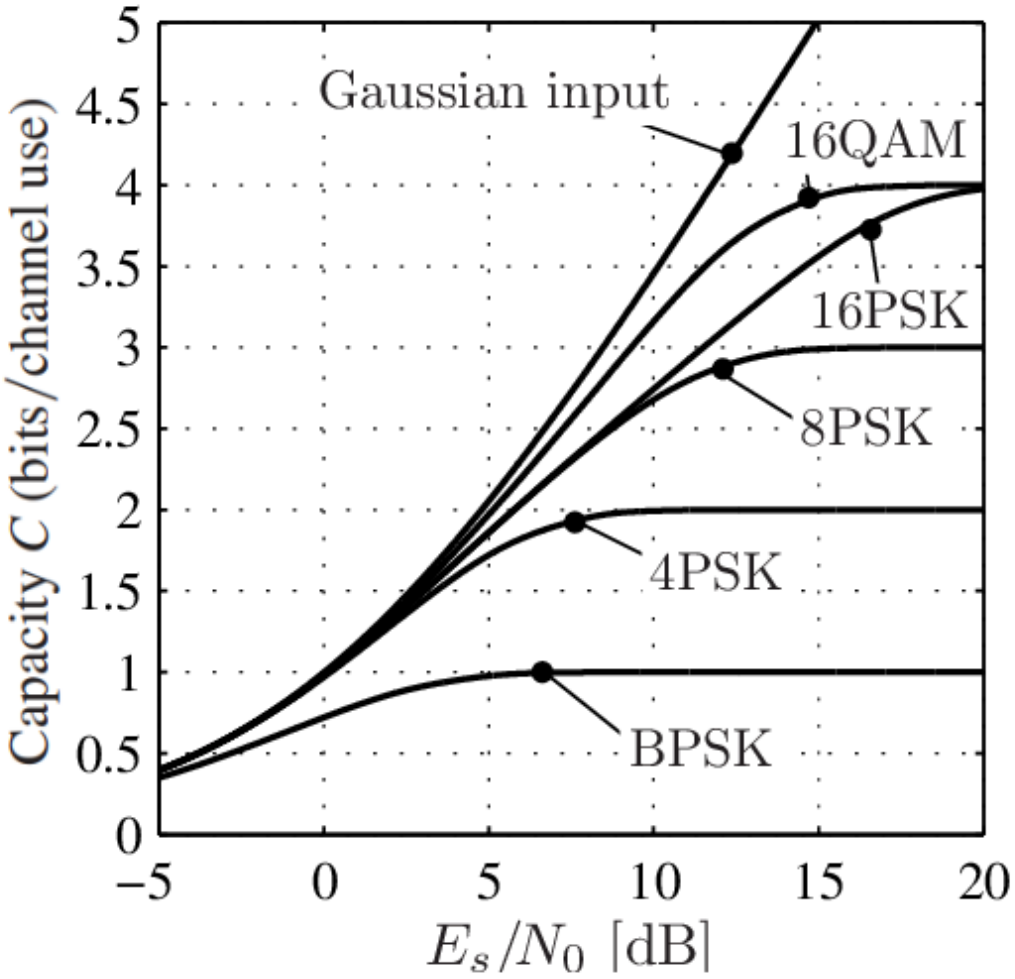


3.2.2 The Capacity of Discrete-Time Binary-Input AWGN Channel

- **Binary-AWGN channel capacity** versus SNR per bit.



- M**-ary AWGN channel capacity** versus SNR per bit



3.2.3 The Capacity of Band-Limited AWGN Channel

- A band-limited AWGN channel with bandwidth W and power constraint P is equivalent to a discrete-time AWGN channel used $2W$ times per second.
- The discrete-time channel SNR is
  - Input power/channel use $= \frac{P}{2W}$, Noise variance $= \frac{N_0}{2} \implies \text{SNR} = \frac{P}{N_0 W}$
  - **Or:** Input power $= P$, Noise power $= \frac{N_0}{2}.2W = N_0 W \implies \text{SNR} = \frac{P}{N_0 W}$
- Its capacity per use is

$$\boxed{C = \frac{1}{2}\log_2\left(1 + \frac{P}{N_0 W}\right) \quad \textbf{bits/channel use}}$$

Hence, the capacity of the continuous-time channel is

$$\boxed{C = W\log_2\left(1 + \frac{P}{N_0 W}\right) \quad \textbf{bits/second}}$$

This is the famous **Shannon capacity** formula for a band-limited AWGN channel.

Asymptotic Behaviors

- As $P \to \infty$, capacity $C \to \infty$, but at a logarithmic rate. Thus,

  Infinite power yields infinite capacity.

- Increasing W increases capacity (more transmissions), but also increases noise power, reducing SNR $= \frac{P}{N_0 W}$.
- As $W \to \infty$, use $\ln(1+x) \approx x$ for small x, yielding:

$$C_\infty = \lim_{W \to \infty} W \log_2 \left(1 + \frac{P}{N_0 W}\right) = \log_2 e \cdot \frac{P}{N_0} \approx 1.44 \cdot \frac{P}{N_0} \quad \text{bits/s}$$

- Thus,

  Infinite bandwidth does not yield infinite capacity—it's limited by power.

- For reliable communication over a band-limited AWGN channel we need

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

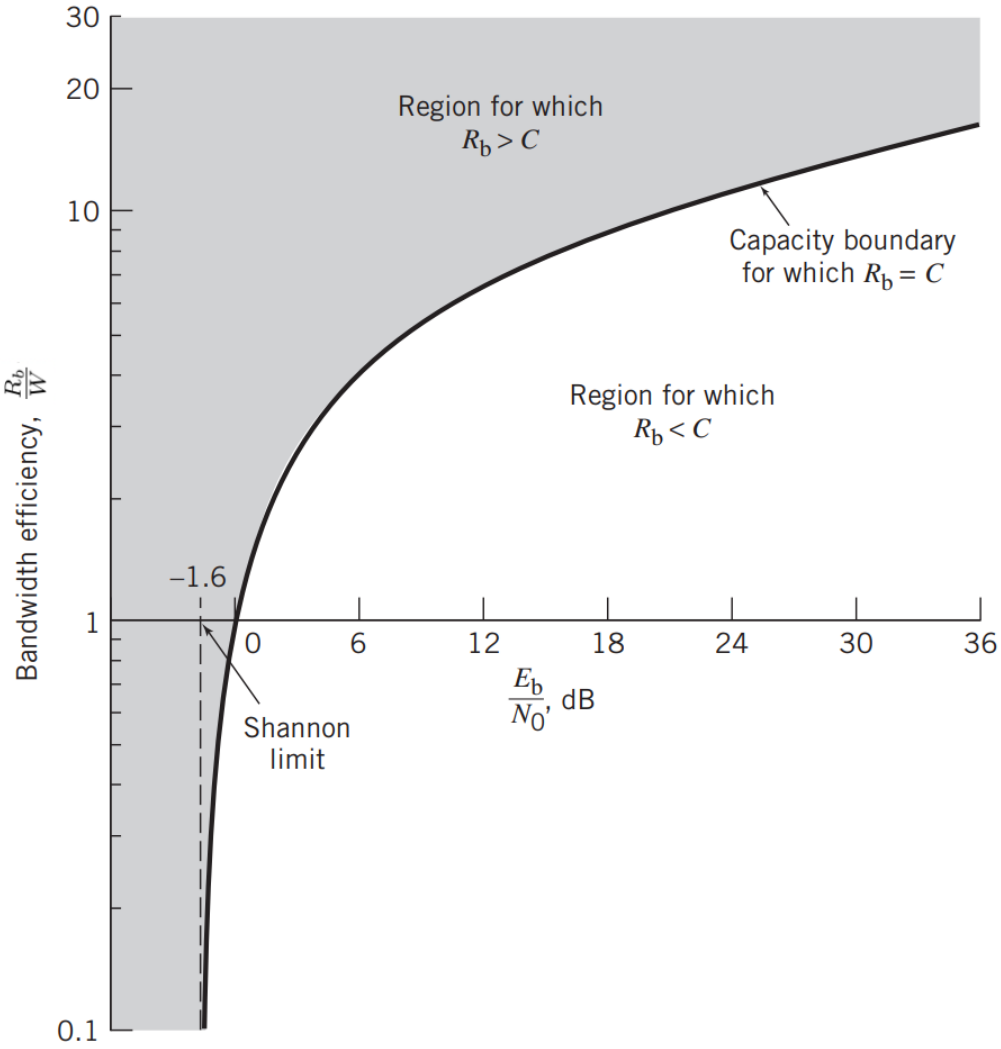
- Since, $P = R_b \mathcal{E}_b$

$$\frac{R_b}{W} < \log_2 \left(1 + \frac{\mathcal{E}_b R_b}{N_0 W} \right) \implies \frac{\mathcal{E}_b}{N_0} > \frac{2^{\frac{R_b}{W}} - 1}{\frac{R_b}{W}}$$

- Let $r = \frac{R_b}{W}$. The minimum value of $\frac{\mathcal{E}_b}{N_0}$ is

$$\lim_{r \to 0} \frac{2^r - 1}{r} = \ln 2 \approx 0.693 \sim -1.6 \text{ dB}$$

No communication system can operate reliably if $\frac{\mathcal{E}_b}{N_0} < -1.6\text{dB}$.



Capacity Region

- This diagram relates the following three parameters
  - bandwidth efficiency (r)
  - capacity (probability of error)
  - $\frac{\mathcal{E}_b}{N_0}$

References

[1] John G. Proakis and Masoud Salehi, *Digital Communications*, McGraw-Hill, 2008.

[2] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley & Sons, 2006.

[3] Simon Haykin, *Digital Communication Systems*, John Wiley & Sons, 2014.

Elements of information theory.