ELG5375: Digital Communications

Introduction to Information Theory

Dr. Sergey Loyka

EECS, University of Ottawa

April 10, 2026

Big Picture

- Information theory: quantifies info. generation, transmission, storage
- Determines fundamental limits to reliable communication
- Two fundamental parts:
  - **Source coding**: compression
  - **Channel coding**: reliable communication
- Key quantities:
  - entropy
  - mutual information
  - channel capacity

A Simple Guessing Game

Suppose someone chooses one message and you try to guess it.

- If there are only 2 equally likely messages, one yes/no question is enough
- If there are 4 (or 8) equally likely messages, you need more questions
- If one possibility is much more likely than the others, guessing becomes easier

Key lesson:

more uncertainty $\Longrightarrow$ **more information needed to describe the outcome**

What is Information?

- A highly predictable event carries little information
- A surprising event carries more information
- Information can be viewed as **reduction in uncertainty**

Examples:

- fair coin flip $\rightarrow$ uncertain
- deterministic output $\rightarrow$ no surprise
- rare event $\rightarrow$ more informative

Self-Information

For an event $X = x$ with probability $p(x)$, define

$$I(x) = \log \frac{1}{p(x)} = -\log p(x).$$

- rare events have larger self-information
- if $p(x) = 1$, then $I(x) = 0$
- independent events add:

$$I(x, y) = I(x) + I(y)$$

- base-2 logarithm gives units of **bits**

Self-Information

Summary of important properties:

$$I(x_i) = 0 \text{ if } p(x_i) = 1 \tag{1}$$

$$I(x_i) \geq 0 \tag{2}$$

$$I(x_i) > I(x_j) \text{ iff } p(x_i) < p(x_j) \tag{3}$$

$$I(x_i, x_j) = I(x_i) + I(x_j) \text{ iff } x_i \text{ \& } x_j \text{ are independent} \tag{4}$$

Entropy

- entropy is the **average self-information**

$$H(X) = -\sum_{m=1}^{M} p(a_m) \log p(a_m), \; X \in \; [a_1, a_2, ...a_M]$$

- it measures the uncertainty in X "on average"
- it is also the fundamental limit of data compression
- it depends on probabilities $p(a_m) = p_m$ only, not on a_m:

$$H(X) = -\sum_{m} p_m \log p_m$$

Properties of Entropy I

- **Nonnegativity:**

$$H(X) \geq 0$$

- **Deterministic case:**

$$H(X) = 0 \iff X \text{ is deterministic}$$

- **Upper bound:** if $X \in [a_1, a_2, ...a_M]$, then

$$0 \leq H(X) \leq \log M$$

- Equality holds when ???

An Example: A Binary Source

Let $X \sim \text{Bernoulli}(p)$, so

$$P(X = a_1) = p, \qquad P(X = a_2) = 1 - p.$$

$$H(X) = H_2(p) = -p \log_2 p - (1-p) \log_2 (1-p).$$

If $p = 0.5$, then $H(X) = 1$ bit.
If $p = 0.9$, then

$$H(X) = -0.9 \log_2 0.9 - 0.1 \log_2 0.1 \approx 0.47 \text{ bits.}$$

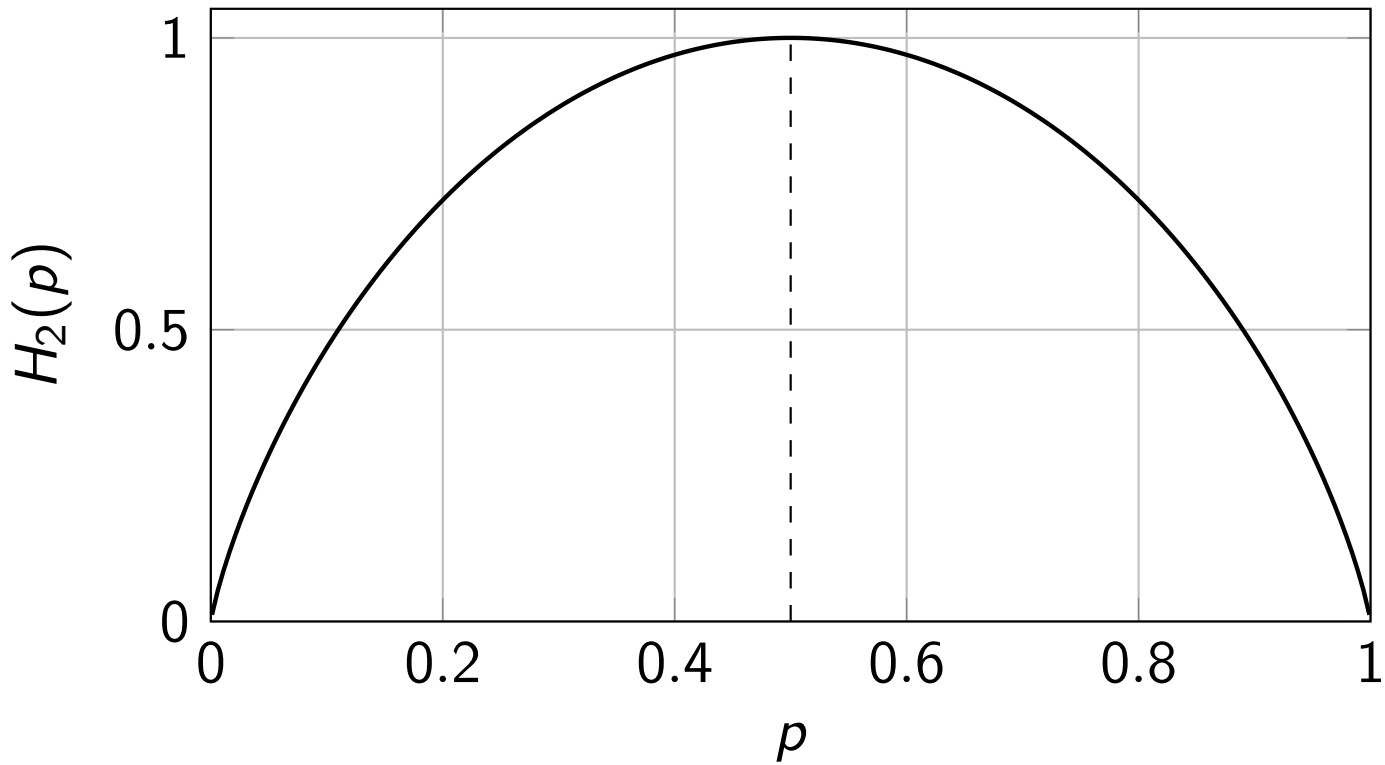
Interpretation: a heavily biased source is easier to guess, so it has lower entropy.

Q: what is the best guessing strategy? Probability of error?

Binary Entropy Function

For $X \sim \text{Bernoulli}(p)$,

$$H(X) = H_2(p) = -p \log p - (1-p) \log(1-p).$$



- maximum at $p = 0.5$, symmetric about $p = 0.5$

Joint and Conditional Entropy

- $H(X, Y)$: uncertainty in the pair (X, Y)

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y)$$

- $H(X|Y)$: uncertainty in X after observing Y

$$H(X|Y) = \sum_{y} p(y)\, H(X|Y = y).$$

- **chain rule**:

$$H(X, Y) = H(Y) + H(X|Y)$$

- **Q:** prove it

Properties of Entropy II

- **Chain rule:**

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

- **Conditioning reduces entropy:**

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

- **Independence implies additivity:** if X and Y are independent, then

$$H(X, Y) = H(X) + H(Y)$$

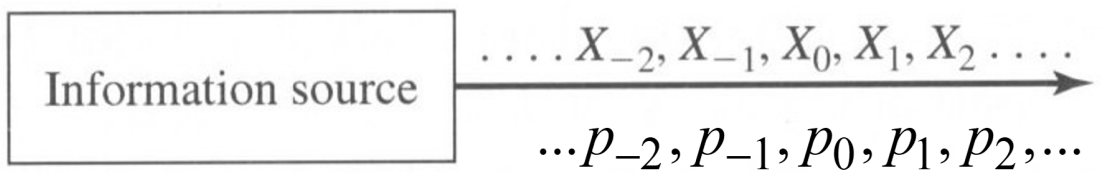
- **Q:** prove all of the above

Data Compression (Source Coding)

- the most important (and useful) application of entropy
- how to compress file/data stream losslessly?
- what is the fundamental limit of compression?
- simple yet fundamental model: **discrete memoryless source** (DMS):

$$X^n = [X_1, X_2, .., X_n], \; X_i \sim \text{i.i.d.} \tag{5}$$
$$X_i \in \{a_1, a_2, .., a_M\}, \; \Pr\{X_i = a_m\} = p_m$$

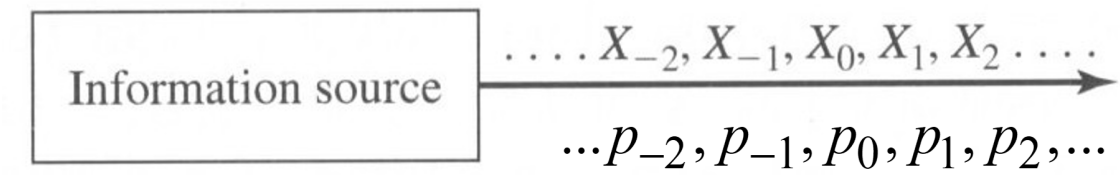
where $\{a_1, a_2, .., a_M\}$ is the source alphabet



- e.g. bandlimited analog (continuous) source $\Rightarrow$ sampling theorem $\Rightarrow$ discrete-time source

Data Compression (Source Coding)

- the most important (and useful) application of entropy



- **Q1**: what is the minimum data rate needed to transmit this source losslessly?
- **Q2:** how many bits are needed to save X^n in a file?

Data Compression (Source Coding)

- the most important (and useful) application of entropy
- **Source Coding Theorem**[1]: any DMS can be transmitted losslessly with rate R [bit/sym.] if $R > H(X)$. Lossless transmission is not possible if $R < H(X)$.
- equivalently, for large n, sequence X^n can compressed losslessly to any N_b [bits] if $N_b > nH(X)$. Lossless compression is not possible if $N_b < nH(X)$.
- i.e. $H(X)$ is a **sharp boundary** of what is possible and what is not in lossless data compression
- $H(X) \approx$ **absolutely minimum number of bits/symbol** required to transmit/save the source losslessly

[1] this is a "rough" formulation that gives a big picture; consult any IT book for precise details and definitions of source codes.

Data Compression (Source Coding)

- **Source Coding Theorem**[2]: any DMS can be transmitted/stored losslessly with rate R if $R > H(X)$. Lossless transmission/storage is not possible if $R < H(X)$.

- proof: via LLN
- can be extended to sources with memory (less bits are needed)
- can be applied to continuous (analog) sources using quantization
- used extensively in practice (.zip, .jpeg, .mp3, .mp4, etc.)

[2] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, 2006.

An Example: A Binary Source

Consider DMS with $X_i \sim \text{Bernoulli}(p)$ and $p = 0.9$,

$$P(X = a_1) = p, \qquad P(X = a_2) = 1 - p.$$

and $n = 10^6$.

- **Q1:** minimum R to transmit the source losslessly?
- **Q2:** minimum number of bits to store it losslessly?
- **Q3:** how do these change if (a) $p = 0.5$? (b) $p = 1$? (c) $p = 0$?
- **Q4:** how do the answers to Q1 and Q2 change if $X_1 = X_2 = ... = X_n$?

Example: Rate of a Bandlimited Source

- Baseband, bandlimited source, $F_{\max} = 4$ kHz, sampled at Nyquist rate
- Samples are quantized to $[-2, -1, 0, 1, 2]$, and the corresponding probabilities are $[1/16,\ 1/8,\ 1/2,\ 1/4,\ 1/16]$.
- Find the rate [bit/s] of the source:

$$H(X) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{8}\log 8 + \frac{2}{16}\log 16 = \frac{15}{8}\ \text{bit/sample}$$

$$R = H(X) \cdot f_s = H(X) \cdot 2F_{\max} = 15\ \text{kbit/s}$$

- **Q1:** how the answer would change for a uniformly-distributed source?
- **Q2:** if only 2 symbols have non-zero probability? only one?

Mutual Information $I(X;Y)$

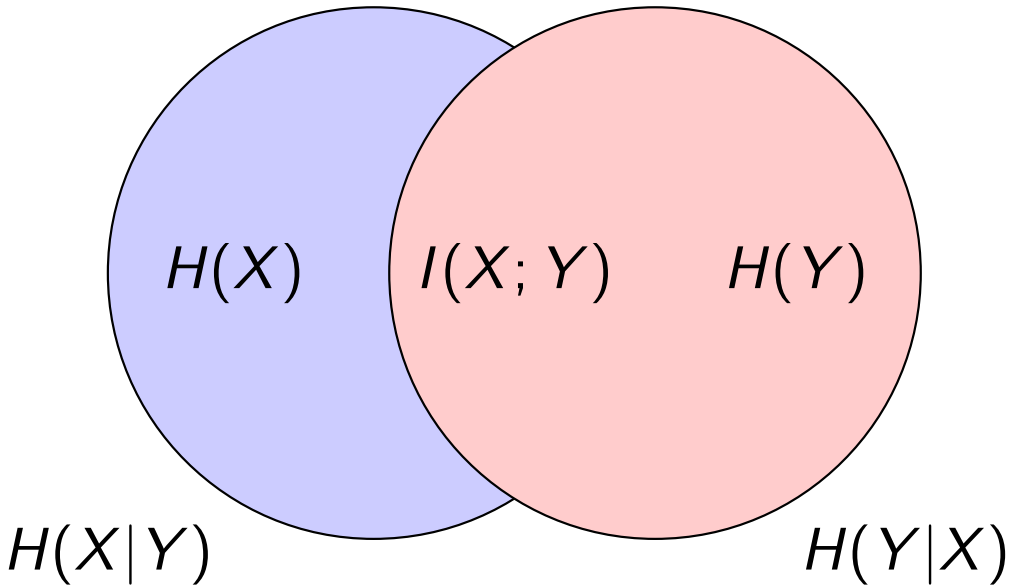
- measures how much knowing Y tells us about X

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

- also measures shared information between X and Y
- Equivalent forms:

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

Mutual Information (MI): Picture



- overlap corresponds to shared information
- left-only part corresponds to $H(X|Y)$
- right-only part corresponds to $H(Y|X)$

Properties of Mutual Information (MI)

- **Nonnegativity:**

$$I(X;Y) \geq 0$$

- **Symmetry:**

$$I(X;Y) = I(Y;X)$$

- **Zero iff independence:**

$$I(X;Y) = 0 \iff P_{XY} = P_X P_Y$$

- **Upper bounds:**

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

Mutual Information (MI)

- a measure of dependance
- appears everywhere in information/communication theory, statistics, and machine learning
- most important application of MI: a measure of reliable data rate
- maximum MI = **channel capacity**

Channel as a model

- "Channel" as a model of communication system

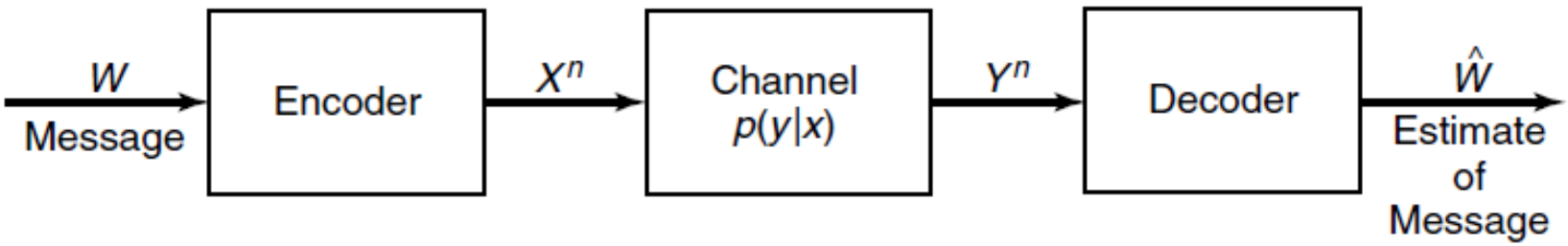


FIGURE 7.1. Communication system.

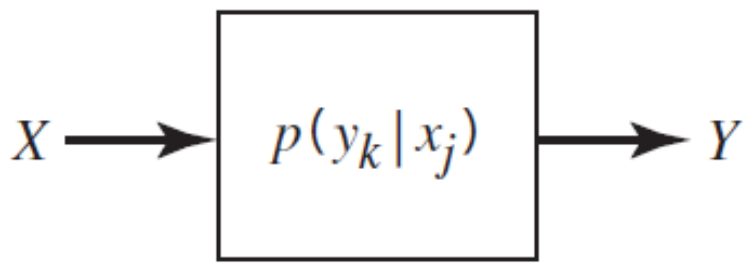
T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, 2006.

- Channel as a stochastic mapping $X^n \to Y^n$,
- or conditional distribution $p(y^n|x^n)$; reminder: $x^n = [x_1, x_2, ..., x_n]$.

Discrete Memoryless Channel (DMC)

DMC: most simple (yet fundamental) model, no memory:

$$X_i \to Y_i \text{ via } p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i)$$



Example: additive noise channel (with memoryless noise),

$$y_i = x_i + z_i, \;\; p(y_i|x_i) = p_z(y_i - x_i)$$

z_i = i.i.d. noise

Discrete Memoryless Channel (DMC)

Channel capacity: maximum rate of reliable data transmission.

DMC capacity is maximum mutual information (between X and Y):

$$C = \max_{p(x)} I(X;Y) \text{ [bit/ch. use]}$$

Also applies to AWGN channel.

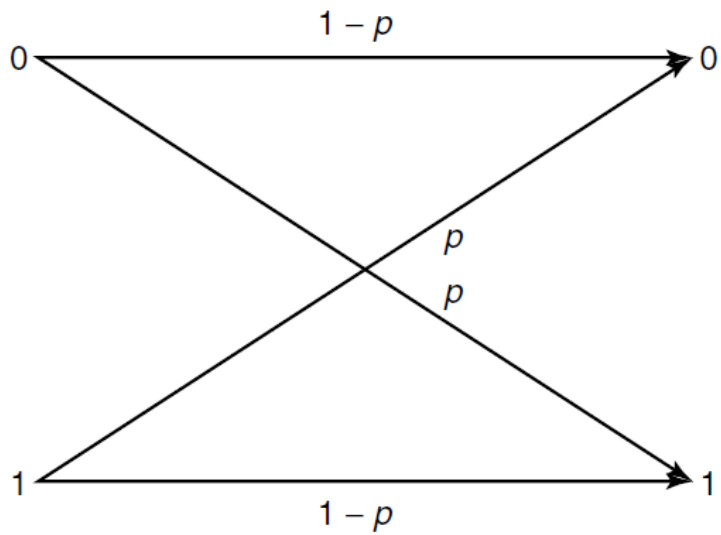
Can be extended to many other channels, e.g. with memory.

Example: Binary Symmetric Channel (BSC)

Channel:

$$y = x + z, \ \Pr\{Y \neq X\} = p = P_e$$

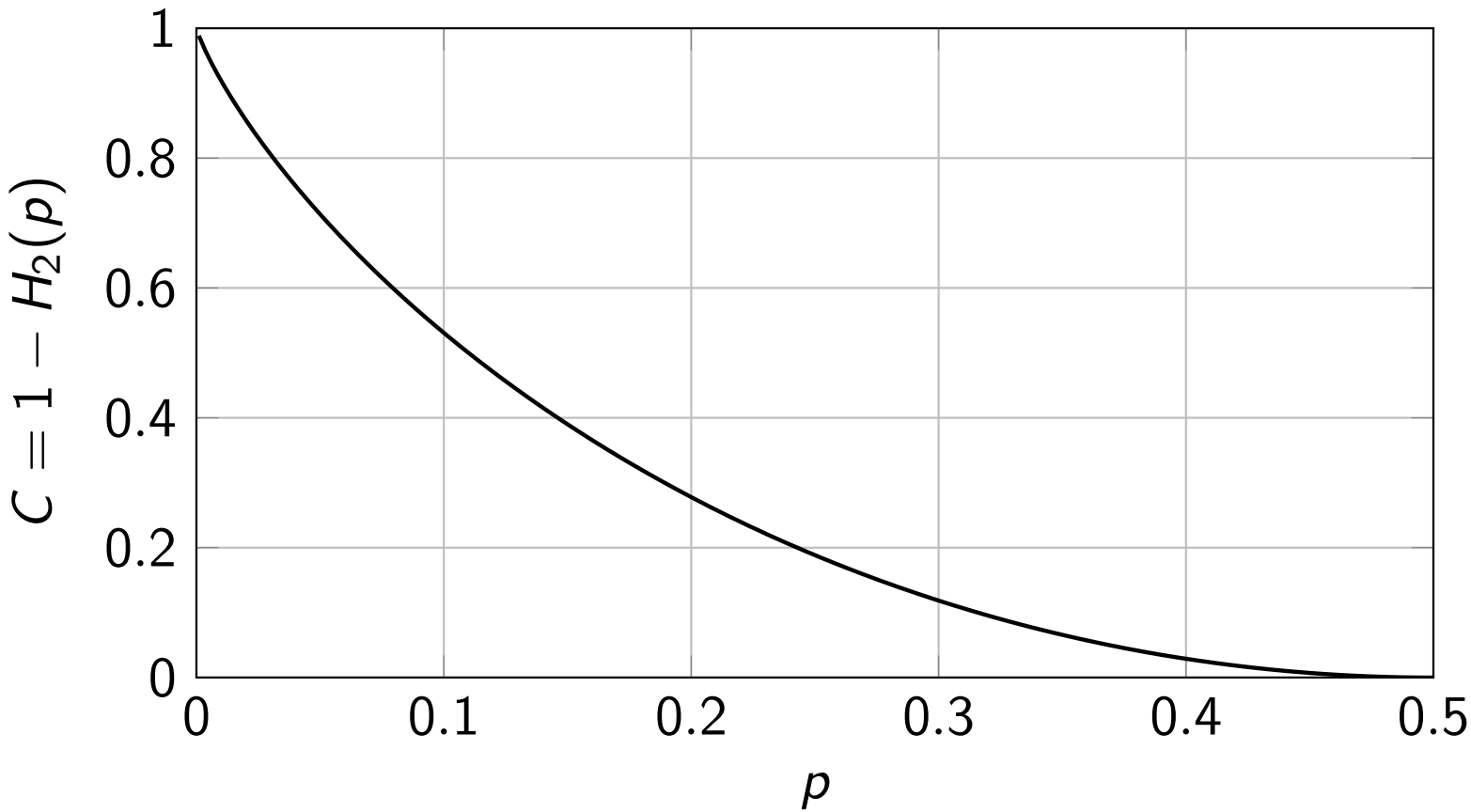
$x, y, z =$ binary, $0/1$;
$p = \Pr\{Z = 1\} =$ "cross-over" (error or flip) probability



T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, 2006.

Its **capacity** is $C = 1 - H_2(p)$

BSC Capacity vs Crossover Probability



Summary

- Introduction to information theory.
- Entropy: measure of information (and uncertainty).
- Properties. Joint/conditional entropies.
- Data compression (source coding). Fundamental limit.
- Discrete memoryless source.
- Mutual information.
- Channel capacity. DMC and BSC.

Reading

- S. Haykin, Digital Communication Systems, Wiley, 2014. Ch. 5.
- R.E. Ziemer, W.H. Tranter, Principles of Communications, Wiley, New York, 2009. Ch. 11.
- B.P.Lathi, Z. Ding, Modern Digital and Analog Communication Systems, Oxford University Press, 2009.
- T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley, 2006.

Note: Do not forget to do end-of-chapter problems. Remember the learning efficiency pyramid!