Information Theory for Single-User Systems

Part I

by

Fady Alajaji† and Po-Ning Chen‡

†Department of Mathematics & Statistics,
Queen’s University, Kingston, ON K7L 3N6, Canada
Email: fady@polya.mast.queensu.ca

‡Department of Electrical and Computer Engineering
Institute of Communications Engineering
National Chiao Tung University
1001, Ta Hsueh Road
Hsin Chu, Taiwan 30010
Republic of China
Email: poning@faculty.nctu.edu.tw

August 28, 2014
Preface

The reliable transmission of information bearing signals over a noisy communication channel is at the heart of what we call communication. Information theory—founded by Claude E. Shannon in 1948—provides a mathematical framework for the theory of communication; it describes the fundamental limits to how efficiently one can encode information and still be able to recover it with negligible loss. These lecture notes will examine basic and advanced concepts of this theory with a focus on single-user (point-to-point) systems. What follows is a tentative list of topics to be covered.

1. Part I:

   (a) Information measures for discrete systems: self-information, entropy, mutual information and divergence, data processing theorem, Fano’s inequality, Pinsker’s inequality, simple hypothesis testing and the Neyman-Pearson lemma.

   (b) Fundamentals of lossless source coding (data compression): discrete memoryless sources, fixed-length (block) codes for asymptotically lossless compression, Asymptotic Equipartition Property (AEP) for discrete memoryless sources, coding for stationary ergodic sources, entropy rate and redundancy, variable-length codes for lossless compression, prefix codes, Kraft inequality, Huffman codes, Shannon-Fano-Elias codes and Lempel-Ziv codes.

   (c) Fundamentals of channel coding: discrete memoryless channels, block codes for data transmission, channel capacity, coding theorem for discrete memoryless channels, calculation of channel capacity, channels with symmetry structures.

   (d) Information measures for continuous alphabet systems and Gaussian channels: differential entropy, mutual information and divergence, AEP for continuous memoryless sources, capacity and channel coding theorem of discrete-time memoryless Gaussian channels, capacity of uncorrelated parallel Gaussian channels and the water-filling principle, capacity of correlated Gaussian channels, non-Gaussian discrete-
time memoryless channels, capacity of band-limited (continuous-time) white Gaussian channels.

(e) Fundamentals of lossy source coding and joint source-channel coding: distortion measures, rate-distortion theorem for memoryless sources, rate-distortion function and its properties, rate-distortion function for memoryless Gaussian sources, lossy joint source-channel coding theorem.

(f) Overview on suprema and limits (Appendix A).

(g) Overview in probability and random processes (Appendix B): random variable and random process, statistical properties of random processes, Markov chains, convergence of sequences of random variables, ergodicity and laws of large numbers, central limit theorem, concavity and convexity, Jensen’s inequality, Lagrange multipliers and the Karush-Kuhn-Tucker condition for the optimization of convex functions.

2. Part II:

(a) General information measures: information spectra and quantiles and their properties, Rényi’s information measures.

(b) Lossless data compression for arbitrary sources with memory: fixed-length lossless data compression theorem for arbitrary sources, variable-length lossless data compression theorem for arbitrary sources, entropy of English, Lempel-Ziv codes.

(c) Randomness and resolvability: resolvability and source coding, approximation of output statistics for arbitrary systems with memory.

(d) Coding for arbitrary channels with memory: channel capacity for arbitrary single-user channels, optimistic Shannon coding theorem, strong converse, $\varepsilon$-capacity.

(e) Lossy data compression for arbitrary sources with memory.

(f) Hypothesis testing: error exponent and divergence, large deviations theory, Berry-Essen theorem.

(g) Channel reliability: random coding exponent, expurgated exponent, partitioning exponent, sphere-packing exponent, the asymptotic largest minimum distance of block codes, Elias bound, Varshamov-Gilbert bound, Bhattacharyya distance.

(h) Information theory of networks: distributed detection, data compression for distributed sources, capacity of multiple access channels, degraded broadcast channel, Gaussian multiple terminal channels.
As shown from the list, the lecture notes are divided into two parts. The first part is suitable for a 12-week introductory course such as the one taught at the Department of Mathematics and Statistics, Queen’s University at Kingston, Ontario, Canada. It also meets the need of a fundamental course for senior undergraduates as the one given at the Department of Computer Science and Information Engineering, National Chi Nan University, Taiwan. For an 18-week graduate course such as the one given in the Department of Electrical and Computer Engineering, National Chiao-Tung University, Taiwan, the lecturer can selectively add advanced topics covered in the second part to enrich the lecture content, and provide a more complete and advanced view on information theory to students.

The authors are very much indebted to all people who provided insightful comments on these lecture notes. Special thanks are devoted to Prof. Yunghsiang S. Han with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, for his enthusiasm in testing these lecture notes at his previous school (National Chi-Nan University), and providing the authors with valuable feedback.

Remarks to the reader. In these notes, all assumptions, claims, conjectures, corollaries, definitions, examples, exercises, lemmas, observations, properties, and theorems are numbered under the same counter. For example, the lemma that immediately follows Theorem 2.1 is numbered as Lemma 2.2, instead of Lemma 2.1.

In addition, one may obtain the latest version of the lecture notes from http://shannon.cm.nctu.edu.tw. The interested reader is welcome to submit comments to poning@faculty.nctu.edu.tw.
Acknowledgements

Thanks are given to our families for their full support during the period of writing these lecture notes.
# Table of Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>x</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Overview</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Communication system model</td>
<td>2</td>
</tr>
<tr>
<td>2 Information Measures for Discrete Systems</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Entropy, joint entropy and conditional entropy</td>
<td>5</td>
</tr>
<tr>
<td>2.1.1 Self-information</td>
<td>5</td>
</tr>
<tr>
<td>2.1.2 Entropy</td>
<td>8</td>
</tr>
<tr>
<td>2.1.3 Properties of entropy</td>
<td>10</td>
</tr>
<tr>
<td>2.1.4 Joint entropy and conditional entropy</td>
<td>12</td>
</tr>
<tr>
<td>2.1.5 Properties of joint entropy and conditional entropy</td>
<td>14</td>
</tr>
<tr>
<td>2.2 Mutual information</td>
<td>16</td>
</tr>
<tr>
<td>2.2.1 Properties of mutual information</td>
<td>16</td>
</tr>
<tr>
<td>2.2.2 Conditional mutual information</td>
<td>17</td>
</tr>
<tr>
<td>2.3 Properties of entropy and mutual information for multiple random variables</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Data processing inequality</td>
<td>20</td>
</tr>
<tr>
<td>2.5 Fano’s inequality</td>
<td>22</td>
</tr>
<tr>
<td>2.6 Divergence and variational distance</td>
<td>26</td>
</tr>
<tr>
<td>2.7 Convexity/concavity of information measures</td>
<td>36</td>
</tr>
<tr>
<td>2.8 Fundamentals of hypothesis testing</td>
<td>39</td>
</tr>
<tr>
<td>3 Lossless Data Compression</td>
<td>47</td>
</tr>
<tr>
<td>3.1 Principles of data compression</td>
<td>47</td>
</tr>
<tr>
<td>3.2 Block codes for asymptotically lossless compression</td>
<td>49</td>
</tr>
<tr>
<td>3.2.1 Block codes for discrete memoryless sources</td>
<td>49</td>
</tr>
</tbody>
</table>
3.2.2 Block codes for stationary ergodic sources .......................... 58
3.2.3 Redundancy for lossless block data compression .................. 63
3.3 Variable-length codes for lossless data compression ............... 64
  3.3.1 Non-singular codes and uniquely decodable codes .............. 64
  3.3.2 Prefix or instantaneous codes .................................. 68
  3.3.3 Examples of binary prefix codes .................................. 74
    A) Huffman codes: optimal variable-length codes ................. 74
    B) Shannon-Fano-Elias code ......................................... 78
3.3.4 Examples of universal lossless variable-length codes .......... 80
  A) Adaptive Huffman code ............................................. 80
  B) Lempel-Ziv codes .................................................... 82

4 Data Transmission and Channel Capacity ............................... 90
  4.1 Principles of data transmission .................................... 90
  4.2 Discrete memoryless channels ...................................... 92
  4.3 Block codes for data transmission over DMCs ..................... 99
  4.4 Calculating channel capacity ...................................... 111
    4.4.1 Symmetric, weakly-symmetric and quasi-symmetric channels 112
    4.4.2 Karuch-Kuhn-Tucker condition for channel capacity ....... 118

5 Differential Entropy and Gaussian Channels ............................. 128
  5.1 Differential entropy .................................................. 129
  5.2 Joint and conditional differential entropies, divergence and mutual
       information ............................................................. 135
  5.3 AEP for continuous memoryless sources ............................. 146
  5.4 Capacity and channel coding theorem for the discrete-time mem-
       oryless Gaussian channel ............................................ 147
  5.5 Capacity of uncorrelated parallel Gaussian channels: The water-
       filling principle ....................................................... 158
  5.6 Capacity of correlated parallel Gaussian channels ............... 161
  5.7 Non-Gaussian discrete-time memoryless channels ..................... 163
  5.8 Capacity of the band-limited white Gaussian channel ............. 165

6 Lossy Data Compression and Transmission .............................. 170
  6.1 Fundamental concept on lossy data compression .................. 170
    6.1.1 Motivations ....................................................... 170
    6.1.2 Distortion measures .............................................. 171
    6.1.3 Frequently used distortion measures .......................... 173
  6.2 Fixed-length lossy data compression codes ....................... 175
  6.3 Rate-distortion function for discrete memoryless
       sources ................................................................. 176
  6.4 Property and calculation of rate-distortion functions ........... 185
6.4.1 Rate-distortion function for binary sources and Hamming
additive distortion measure .................................. 185
6.4.2 Rate-distortion function for Gaussian source and squared
error distortion measure ........................................ 187
6.5 Joint source-channel information transmission .................. 189

A Overview on Suprema and Limits................................ 194
A.1 Supremum and maximum ...................................... 194
A.2 Infimum and minimum ........................................ 196
A.3 Boundedness and suprema operations ......................... 197
A.4 Sequences and their limits ................................... 199
A.5 Equivalence ..................................................... 204

B Overview in Probability and Random Processes ................. 205
B.1 Probability space .............................................. 205
B.2 Random variable and random process ........................ 206
B.3 Distribution functions versus probability space ............. 207
B.4 Relation between a source and a random process ............ 207
B.5 Statistical properties of random sources ..................... 208
B.6 Convergence of sequences of random variables ............. 213
B.7 Ergodicity and laws of large numbers ........................ 216
  B.7.1 Laws of large numbers .................................... 216
  B.7.2 Ergodicity and law of large numbers .................... 219
B.8 Central limit theorem .......................................... 222
B.9 Convexity, concavity and Jensen’s inequality ................ 222
B.10 Lagrange multipliers technique and Karush-Kuhn-Tucker (KKT)
    condition ...................................................... 224
# List of Tables

<table>
<thead>
<tr>
<th>Number</th>
<th>Table Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>An example of the $\delta$-typical set with $n = 2$ and $\delta = 0.4$, where $\mathcal{F}_2(0.4) = { \text{AB, AC, BA, BB, BC, CA, CB} }$. The codeword set is ${ \text{001(AB), 010(AC), 011(BA), 100(BB), 101(BC), 110(CA), 111(CB), 000(AA, AD, BD, CC, CD, DA, DB, DC, DD)} }$, where the parenthesis following each binary codeword indicates those sourcewords that are encoded to this codeword. The source distribution is $P_X(A) = 0.4$, $P_X(B) = 0.3$, $P_X(C) = 0.2$ and $P_X(D) = 0.1$.</td>
<td>54</td>
</tr>
<tr>
<td>5.1</td>
<td>Quantized random variable $q_n(X)$ under an $n$-bit accuracy: $H(q_n(X))$ and $H(q_n(X)) - n$ versus $n$.</td>
<td>133</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Number</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>3</td>
</tr>
<tr>
<td>2.1</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>17</td>
</tr>
<tr>
<td>2.3</td>
<td>21</td>
</tr>
<tr>
<td>2.4</td>
<td>24</td>
</tr>
<tr>
<td>3.1</td>
<td>49</td>
</tr>
<tr>
<td>3.2</td>
<td>57</td>
</tr>
<tr>
<td>3.3</td>
<td>58</td>
</tr>
<tr>
<td>3.4</td>
<td>69</td>
</tr>
<tr>
<td>3.5</td>
<td>70</td>
</tr>
<tr>
<td>3.6</td>
<td>77</td>
</tr>
<tr>
<td>3.7</td>
<td>82</td>
</tr>
<tr>
<td>3.8</td>
<td>83</td>
</tr>
<tr>
<td>3.9</td>
<td>84</td>
</tr>
</tbody>
</table>
4.1 A data transmission system, where $W$ represents the message for transmission, $X^n$ denotes the codeword corresponding to message $W$, $Y^n$ represents the received word due to channel input $X^n$, and $\hat{W}$ denotes the reconstructed message from $Y^n$.

4.2 Binary symmetric channel.

4.3 Binary erasure channel.

4.4 Binary symmetric erasure channel.

4.5 Ultimate channel coding rate $R$ versus channel capacity $C$ and behavior of the probability of error as blocklength $n$ goes to infinity for a DMC.

5.1 The water-pouring scheme for uncorrelated parallel Gaussian channels. The horizontal dashed line, which indicates the level where the water rises to, indicates the value of $\theta$ for which $\sum_{i=1}^{k} P_i = P$.

5.2 Band-limited waveform channel with additive white Gaussian noise.

5.3 Water-pouring for the band-limited colored Gaussian channel.

6.1 Example for applications of lossy data compression codes.

6.2 “Grouping” as one kind of lossy data compression.

6.3 The Shannon limits for $(2,1)$ and $(3,1)$ codes under antipodal binary-input AWGN channel.

6.4 The Shannon limits for $(2,1)$ and $(3,1)$ codes under continuous-input AWGN channels.

A.1 Illustration of Lemma A.17.

B.1 General relations of random processes.

B.2 Relation of ergodic random processes respectively defined through time-shift invariance and ergodic theorem.

B.3 The support line $y = ax + b$ of the convex function $f(x)$. 
Chapter 1

Introduction

1.1 Overview

Since its inception, the main role of *Information Theory* has been to provide the engineering and scientific communities with a mathematical framework for the theory of communication by establishing the *fundamental limits* on the performance of various communication systems. The birth of Information Theory was initiated with the publication of the groundbreaking works [43, 45] of Claude Elwood Shannon (1916-2001) who asserted that it is possible to send information-bearing signals at a *fixed positive rate* through a noisy communication channel with an arbitrarily small probability of error as long as the transmission rate is below a certain fixed quantity that depends on the channel statistical characteristics; he “baptized” this quantity with the name of *channel capacity*. He further proclaimed that random (stochastic) sources, representing data, speech or image signals, can be compressed distortion-free at a minimal rate given by the source’s intrinsic amount of information, which he called *source entropy* and defined in terms of the source statistics. He went on proving that if a source has an entropy that is less than the capacity of a communication channel, then the source can be reliably transmitted (with asymptotically vanishing probability of error) over the channel. He further generalized these “coding theorems” from the lossless (distortionless) to the lossy context where the source can be compressed and reproduced (possibly after channel transmission) within a tolerable distortion threshold [44].

Inspired and guided by the pioneering ideas of Shannon,\(^1\) information theorists gradually expanded their interests beyond communication theory, and investigated fundamental questions in several other related fields. Among them we cite:

\(^1\)See [47] for accessing most of Shannon’s works, including his yet untapped doctoral dissertation on an algebraic framework for population genetics.
statistical physics (thermodynamics, quantum information theory);
computer science (algorithmic complexity, resolvability);
probability theory (large deviations, limit theorems);
statistics (hypothesis testing, multi-user detection, Fisher information, estimation);
economics (gambling theory, investment theory);
biology (biological information theory);
cryptography (data security, watermarking);
data networks (self-similarity, traffic regulation theory).

In this textbook, we focus our attention on the study of the basic theory of communication for single-user (point-to-point) systems for which Information Theory was originally conceived.

1.2 Communication system model

A simple block diagram of a general communication system is depicted in Figure 1.1.

Let us briefly describe the role of each block in the figure.

- **Source**: The source, which usually represents data or multimedia signals, is modelled as a random process (the necessary background regarding random processes is introduced in Appendix B). It can be discrete (finite or countable alphabet) or continuous (uncountable alphabet) in value and in time.

- **Source Encoder**: Its role is to represent the source in a compact fashion by removing its unnecessary or redundant content (i.e., by compressing it).

- **Channel Encoder**: Its role is to enable the reliable reproduction of the source encoder output after its transmission through a noisy communication channel. This is achieved by adding redundancy (using usually an algebraic structure) to the source encoder output.

- **Modulator**: It transforms the channel encoder output into a waveform suitable for transmission over the physical channel. This is typically accomplished by varying the parameters of a sinusoidal signal in proportion with the data provided by the channel encoder output.
• **Physical Channel**: It consists of the noisy (or unreliable) medium that the transmitted waveform traverses. It is usually modelled via a sequence of conditional (or transition) probability distributions of receiving an output given that a specific input was sent.

• **Receiver Part**: It consists of the demodulator, the channel decoder and the source decoder where the reverse operations are performed. The destination represents the sink where the source estimate provided by the source decoder is reproduced.

In this text, we will model the concatenation of the modulator, physical channel and demodulator via a discrete-time channel with a given sequence of conditional probability distributions. Given a source and a discrete channel, our objectives will include determining the fundamental limits of how well we can construct a (source/channel) coding scheme so that:

- the smallest number of source encoder symbols can represent each source symbol distortion-free or within a prescribed distortion level $D$, where $D > 0$ and the channel is noiseless;

---

2Except for a brief interlude with the continuous-time (waveform) Gaussian channel in Chapter 5, we will consider discrete-time communication systems throughout the text.
• the largest rate of information can be transmitted over a noisy channel between the channel encoder input and the channel decoder output with an arbitrarily small probability of decoding error;

• we can guarantee that the source is transmitted over a noisy channel and reproduced at the destination within distortion $D$, where $D > 0$. 
Chapter 2

Information Measures for Discrete Systems

In this chapter, we define information measures for discrete-time discrete-alphabet\(^1\) systems from a probabilistic standpoint and develop their properties. Elucidating the operational significance of probabilistically defined information measures vis-a-vis the fundamental limits of coding constitutes a main objective of this book; this will be seen in the subsequent chapters.

2.1 Entropy, joint entropy and conditional entropy

2.1.1 Self-information

Let \( E \) be an event belonging to a given event space and having probability \( \Pr(E) \triangleq p_E \), where \( 0 \leq p_E \leq 1 \). Let \( I(E) \) – called the self-information of \( E \) – represent the amount of information one gains when learning that \( E \) has occurred (or equivalently, the amount of uncertainty one had about \( E \) prior to learning that it has happened). A natural question to ask is “what properties should \( I(E) \) have?” Although the answer to this question may vary from person to person, here are some common properties that \( I(E) \) is reasonably expected to have.

1. \( I(E) \) should be a decreasing function of \( p_E \).

In other words, this property first states that \( I(E) = I(p_E) \), where \( I(\cdot) \) is a real-valued function defined over \([0, 1]\). Furthermore, one would expect that the less likely event \( E \) is, the more information is gained when one

\(^1\)By discrete alphabets, one usually means finite or countably infinite alphabets. We however mostly focus on finite alphabet systems, although the presented information measures allow for countable alphabets (when they exist).
learns it has occurred. In other words, \( I(p_E) \) is a decreasing function of \( p_E \).

2. \( I(p_E) \) should be continuous in \( p_E \).

   Intuitively, one should expect that a small change in \( p_E \) corresponds to a small change in the amount of information carried by \( E \).

3. If \( E_1 \) and \( E_2 \) are independent events, then \( I(E_1 \cap E_2) = I(E_1) + I(E_2) \), or equivalently, \( I(p_{E_1} \times p_{E_2}) = I(p_{E_1}) + I(p_{E_2}) \).

   This property declares that when events \( E_1 \) and \( E_2 \) are independent from each other (i.e., when they do not affect each other probabilistically), the amount of information one gains by learning that both events have jointly occurred should be equal to the sum of the amounts of information of each individual event.

Next, we show that the only function that satisfies properties 1-3 above is the logarithmic function.

**Theorem 2.1** The only function defined over \( p \in [0, 1] \) and satisfying

1. \( I(p) \) is monotonically decreasing in \( p \);
2. \( I(p) \) is a continuous function of \( p \) for \( 0 \leq p \leq 1 \);
3. \( I(p_1 \times p_2) = I(p_1) + I(p_2) \);

is \( I(p) = -c \cdot \log_b(p) \), where \( c \) is a positive constant and the base \( b \) of the logarithm is any number larger than one.

**Proof:**

**Step 1: Claim.** For \( n = 1, 2, 3, \cdots \),

\[
I \left( \frac{1}{n} \right) = -c \cdot \log_b \left( \frac{1}{n} \right),
\]

where \( c > 0 \) is a constant.

**Proof:** First note that for \( n = 1 \), condition 3 directly shows the claim, since it yields that \( I(1) = I(1) + I(1) \). Thus \( I(1) = 0 = -c \log_b(1) \).

Now let \( n \) be a fixed positive integer greater than 1. Conditions 1 and 3 respectively imply

\[
n < m \Rightarrow I \left( \frac{1}{n} \right) < I \left( \frac{1}{m} \right)
\]  

(2.1.1)
and
\[ I \left( \frac{1}{mn} \right) = I \left( \frac{1}{m} \right) + I \left( \frac{1}{n} \right) \] (2.1.2)
where \( n, m = 1, 2, 3, \ldots \). Now using (2.1.2), we can show by induction (on \( k \)) that
\[ I \left( \frac{1}{n^k} \right) = k \cdot I \left( \frac{1}{n} \right) \] (2.1.3)
for all non-negative integers \( k \).
Now for any positive integer \( r \), there exists a non-negative integer \( k \) such that
\[ n^k \leq 2^r < n^{k+1}. \]
By (2.1.1), we obtain
\[ I \left( \frac{1}{n^k} \right) \leq I \left( \frac{1}{2^r} \right) < I \left( \frac{1}{n^{k+1}} \right), \]
which together with (2.1.3), yields
\[ k \cdot I \left( \frac{1}{n} \right) \leq r \cdot I \left( \frac{1}{2} \right) < (k + 1) \cdot I \left( \frac{1}{n} \right). \]
Hence, since \( I(1/n) > I(1) = 0, \)
\[ \frac{k}{r} \leq \frac{I(1/2)}{I(1/n)} \leq \frac{k + 1}{r}. \]
On the other hand, by the monotonicity of the logarithm, we obtain
\[ \log_b n^k \leq \log_b 2^r \leq \log_b n^{k+1} \Leftrightarrow \frac{k}{r} \leq \frac{\log_b (2)}{\log_b (n)} \leq \frac{k + 1}{r}. \]
Therefore,
\[ \left| \frac{\log_b (2)}{\log_b (n)} - \frac{I(1/2)}{I(1/n)} \right| < \frac{1}{r}. \]
Since \( n \) is fixed, and \( r \) can be made arbitrarily large, we can let \( r \to \infty \) to get:
\[ I \left( \frac{1}{n} \right) = c \cdot \log_b (n). \]
where \( c = I(1/2)/\log_b (2) > 0 \). This completes the proof of the claim.
Step 2: Claim. $I(p) = -c \cdot \log_b(p)$ for positive rational number $p$, where $c > 0$ is a constant.

Proof: A positive rational number $p$ can be represented by a ratio of two integers, i.e., $p = r/s$, where $r$ and $s$ are both positive integers. Then condition 3 yields that

$$I\left(\frac{1}{s}\right) = I\left(\frac{r}{s \cdot r}\right) = I\left(\frac{r}{s}\right) + I\left(\frac{1}{r}\right),$$

which, from Step 1, implies that

$$I(p) = I\left(\frac{r}{s}\right) = I\left(\frac{1}{s}\right) - I\left(\frac{1}{r}\right) = c \cdot \log_b s - c \cdot \log_b r = -c \cdot \log_b p.$$

Step 3: For any $p \in [0, 1]$, it follows by continuity and the density of the rationals in the reals that

$$I(p) = \lim_{a \uparrow p, \text{ } a \text{ rational}} I(a) = \lim_{b \downarrow p, \text{ } b \text{ rational}} I(b) = -c \cdot \log_b(p).$$

The constant $c$ above is by convention normalized to $c = 1$. Furthermore, the base $b$ of the logarithm determines the type of units used in measuring information. When $b = 2$, the amount of information is expressed in bits (i.e., binary digits). When $b = e$ – i.e., the natural logarithm (ln) is used – information is measured in nats (i.e., natural units or digits). For example, if the event $E$ concerns a Heads outcome from the toss of a fair coin, then its self-information is $I(E) = -\log_2(1/2) = 1$ bit or $-\ln(1/2) = 0.693$ nats.

More generally, under base $b > 1$, information is in $b$-ary units or digits. For the sake of simplicity, we will throughout use the base-2 logarithm unless otherwise specified. Note that one can easily convert information units from bits to $b$-ary units by dividing the former by $\log_2(b)$.

2.1.2 Entropy

Let $X$ be a discrete random variable taking values in a finite alphabet $\mathcal{X}$ under a probability distribution or probability mass function (pmf) $P_X(x) \triangleq P[X = x]$ for all $x \in \mathcal{X}$. Note that $X$ generically represents a memoryless source, which is a random process $\{X_n\}_{n=1}^{\infty}$ with independent and identically distributed (i.i.d.) random variables (cf. Appendix B).
Definition 2.2 (Entropy) The entropy of a discrete random variable $X$ with pmf $P_X(\cdot)$ is denoted by $H(X)$ or $H(P_X)$ and defined by

$$H(X) \triangleq -\sum_{x \in \mathcal{X}} P_X(x) \cdot \log_2 P_X(x) \quad \text{(bits)}.$$ 

Thus $H(X)$ represents the statistical average (mean) amount of information one gains when learning that one of its $|\mathcal{X}|$ outcomes has occurred, where $|\mathcal{X}|$ denotes the size of alphabet $\mathcal{X}$. Indeed, we directly note from the definition that

$$H(X) = E[-\log_2 P_X(X)] = E[I(X)]$$

where $I(x) \triangleq -\log_2 P_X(x)$ is the self-information of the elementary event $[X = x]$.

When computing the entropy, we adopt the convention

$$0 \cdot \log_2 0 = 0,$$

which can be justified by a continuity argument since $x \log_2 x \to 0$ as $x \to 0$. Also note that $H(X)$ only depends on the probability distribution of $X$ and is not affected by the symbols that represent the outcomes. For example when tossing a fair coin, we can denote Heads by 2 (instead of 1) and Tail by 100 (instead of 0), and the entropy of the random variable representing the outcome would remain equal to $\log_2(2) = 1$ bit.

Example 2.3 Let $X$ be a binary (valued) random variable with alphabet $\mathcal{X} = \{0, 1\}$ and pmf given by $P_X(1) = p$ and $P_X(0) = 1 - p$, where $0 \leq p \leq 1$ is fixed. Then $H(X) = -p \cdot \log_2 p - (1 - p) \cdot \log_2 (1 - p)$. This entropy is conveniently called the binary entropy function and is usually denoted by $h_b(p)$: it is illustrated in Figure 2.1. As shown in the figure, $h_b(p)$ is maximized for a uniform distribution (i.e., $p = 1/2$).

The units for $H(X)$ above are in bits as base-2 logarithm is used. Setting

$$H_D(X) \triangleq -\sum_{x \in \mathcal{X}} P_X(x) \cdot \log_D P_X(x)$$

yields the entropy in $D$-ary units, where $D > 1$. Note that we abbreviate $H_2(X)$ as $H(X)$ throughout the book since bits are common measure units for a coding system, and hence

$$H_D(X) = \frac{H(X)}{\log_2 D}.$$ 

Thus

$$H_e(X) = \frac{H(X)}{\log_2(e)} = (\ln 2) \cdot H(X)$$

gives the entropy in nats, where $e$ is the base of the natural logarithm.
2.1.3 Properties of entropy

When developing or proving the basic properties of entropy (and other information measures), we will often use the following fundamental inequality on the logarithm (its proof is left as an exercise).

Lemma 2.4 (Fundamental inequality (FI)) For any $x > 0$ and $D > 1$, we have that

$$\log_D(x) \leq \log_D(e) \cdot (x - 1)$$

with equality if and only if (iff) $x = 1$.

Setting $y = 1/x$ and using FI above directly yield that for any $y > 0$, we also have that

$$\log_D(y) \geq \log_D(e) \left(1 - \frac{1}{y}\right),$$

also with equality iff $y = 1$. In the above the base-$D$ logarithm was used. Specifically, for a logarithm with base-2, the above inequalities become

$$\log_2(e) \left(1 - \frac{1}{x}\right) \leq \log_2(x) \leq \log_2(e) \cdot (x - 1)$$

with equality iff $x = 1$.

Lemma 2.5 (Non-negativity) $H(X) \geq 0$. Equality holds iff $X$ is deterministic (when $X$ is deterministic, the uncertainty of $X$ is obviously zero).
Proof: $0 \leq P_X(x) \leq 1$ implies that $\log_2[1/P_X(x)] \geq 0$ for every $x \in \mathcal{X}$. Hence,

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{1}{P_X(x)} \geq 0,$$

with equality holding iff $P_X(x) = 1$ for some $x \in \mathcal{X}$. $\square$

**Lemma 2.6 (Upper bound on entropy)** If a random variable $X$ takes values from a finite set $\mathcal{X}$, then

$$H(X) \leq \log_2 |\mathcal{X}|,$$

where $|\mathcal{X}|$ denotes the size of the set $\mathcal{X}$. Equality holds iff $X$ is equiprobable or uniformly distributed over $\mathcal{X}$ (i.e., $P_X(x) = \frac{1}{|\mathcal{X}|}$ for all $x \in \mathcal{X}$).

**Proof:**

$$\log_2 |\mathcal{X}| - H(X) = \log_2 |\mathcal{X}| \times \left[ \sum_{x \in \mathcal{X}} P_X(x) \right] - \left[ - \sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x) \right]$$

$$= \sum_{x \in \mathcal{X}} P_X(x) \times \log_2 |\mathcal{X}| + \sum_{x \in \mathcal{X}} P_X(x) \log_2 P_X(x)$$

$$\geq \sum_{x \in \mathcal{X}} P_X(x) \cdot \log_2(e) \left( 1 - \frac{1}{|\mathcal{X}| \times P_X(x)} \right)$$

$$= \log_2(e) \sum_{x \in \mathcal{X}} \left( P_X(x) - \frac{1}{|\mathcal{X}|} \right)$$

$$= \log_2(e) \cdot (1 - 1) = 0$$

where the inequality follows from the FI Lemma, with equality iff $(\forall x \in \mathcal{X}), |\mathcal{X}| \times P_X(x) = 1$, which means $P_X(\cdot)$ is a uniform distribution on $\mathcal{X}$. $\square$

Intuitively, $H(X)$ tells us how random $X$ is. Indeed, $X$ is deterministic (not random at all) iff $H(X) = 0$. If $X$ is uniform (equiprobable), $H(X)$ is maximized, and is equal to $\log_2 |\mathcal{X}|$.

**Lemma 2.7 (Log-sum inequality)** For non-negative numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} \left( a_i \log_D \frac{a_i}{b_i} \right) \geq \left( \sum_{i=1}^{n} a_i \right) \log_D \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \quad (2.1.4)$$

with equality holding iff, $(\forall 1 \leq i \leq n) (a_i/b_i) = (a_1/b_1)$, a constant independent of $i$. (By convention, $0 \cdot \log_D(0) = 0$, $0 \cdot \log_D(0/0) = 0$ and $a \cdot \log_D(a/0) = \infty$ if $a > 0$. Again, this can be justified by “continuity.”)
Proof: Let \( a \triangleq \sum_{i=1}^{n} a_i \) and \( b \triangleq \sum_{i=1}^{n} b_i \). Then

\[
\sum_{i=1}^{n} a_i \log_D \frac{a_i}{b_i} - a \log_D \frac{a}{b} = a \left[ \sum_{i=1}^{n} \frac{a_i}{a} \log_D \frac{a_i}{b_i} - \left( \sum_{i=1}^{n} \frac{a_i}{a} \right) \log_D \frac{a}{b} \right]
\]

\[
\geq a \log_D(e) \sum_{i=1}^{n} \frac{a_i}{a} \left[ 1 - \frac{b_i}{a_i b} \right]
\]

\[
= a \log_D(e) \left( \sum_{i=1}^{n} \frac{a_i}{a} - \sum_{i=1}^{n} \frac{b_i}{b} \right)
\]

\[
= a \log_D(e) (1 - 1) = 0
\]

where the inequality follows from the FI Lemma, with equality holding iff \( \frac{a_i b_i}{b_i a} = 1 \) for all \( i \); i.e., \( \frac{a_i}{b_i} = \frac{a}{b} \) \( \forall i \).

We also provide another proof using Jensen’s inequality (cf. Theorem B.17 in Appendix B). Without loss of generality, assume that \( a_i > 0 \) and \( b_i > 0 \) for every \( i \). Jensen’s inequality states that

\[
\sum_{i=1}^{n} \alpha_i f(t_i) \geq f \left( \sum_{i=1}^{n} \alpha_i t_i \right)
\]

for any strictly convex function \( f(\cdot) \), \( \alpha_i \geq 0 \), and \( \sum_{i=1}^{n} \alpha_i = 1 \); equality holds iff \( t_i \) is a constant for all \( i \). Hence by setting \( \alpha_i = \frac{b_i}{\sum_{j=1}^{n} b_j} \), \( t_i = a_i/b_i \), and \( f(t) = t \cdot \log_D(t) \), we obtain the desired result. \( \square \)

### 2.1.4 Joint Entropy and Conditional Entropy

Given a pair of random variables \((X,Y)\) with a joint pmf \( P_{X,Y}(\cdot, \cdot) \) defined on \( X \times Y \), the self-information of the (two-dimensional) elementary event \([X = x, Y = y]\) is defined by

\[
I(x, y) \triangleq -\log_2 P_{X,Y}(x, y).
\]

This leads us to the definition of joint entropy.

**Definition 2.8 (Joint entropy)** The joint entropy \( H(X, Y) \) of random variables \((X,Y)\) is defined by

\[
H(X, Y) \triangleq -\sum_{(x,y) \in X \times Y} P_{X,Y}(x, y) \cdot \log_2 P_{X,Y}(x, y)
\]

12
The conditional entropy can also be similarly defined as follows.

**Definition 2.9 (Conditional entropy)** Given two jointly distributed random variables $X$ and $Y$, the conditional entropy $H(Y|X)$ of $Y$ given $X$ is defined by

$$H(Y|X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) \left( - \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \cdot \log_2 P_{Y|X}(y|x) \right)$$  \hspace{1cm} (2.1.5)

where $P_{Y|X}(\cdot|\cdot)$ is the conditional pmf of $Y$ given $X$.

Equation (2.1.5) can be written into three different but equivalent forms:

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

$$= E[- \log_2 P_{Y|X}(Y|X)]$$

$$= \sum_{x \in \mathcal{X}} P_X(x) \cdot H(Y|X = x)$$

where $H(Y|X = x) \triangleq - \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \log_2 P_{Y|X}(y|x)$.

The relationship between joint entropy and conditional entropy is exhibited by the fact that the entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other.

**Theorem 2.10 (Chain rule for entropy)**

$$H(X,Y) = H(X) + H(Y|X).$$ \hspace{1cm} (2.1.6)

**Proof:** Since

$$P_{X,Y}(x,y) = P_X(x)P_{Y|X}(y|x),$$

we directly obtain that

$$H(X,Y) = E[- \log P_{X,Y}(X,Y)]$$

$$= E[- \log_2 P_{X}(X)] + E[- \log_2 P_{Y|X}(Y|X)]$$

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

By its definition, joint entropy is commutative; i.e., $H(X,Y) = H(Y,X)$. Hence,

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

13
which implies that
\[ H(X) - H(X|Y) = H(Y) - H(Y|X). \] (2.1.7)

The above quantity is exactly equal to the mutual information which will be introduced in the next section.

The conditional entropy can be thought of in terms of a channel whose input is the random variable \( X \) and whose output is the random variable \( Y \). \( H(X|Y) \) is then called the \textit{equivocation}\(^2\) and corresponds to the uncertainty in the channel input from the \textit{receiver’s point-of-view}. For example, suppose that the set of possible outcomes of random vector \( (X, Y) \) is \( \{(0, 0), (0, 1), (1, 0), (1, 1)\} \), where none of the elements has zero probability mass. When the receiver \( Y \) receives 1, he still cannot determine exactly what the sender \( X \) observes (it could be either 1 or 0); therefore, the uncertainty, from the receiver’s point of view, depends on the probabilities \( P_{X|Y}(0|1) \) and \( P_{X|Y}(1|1) \).

Similarly, \( H(Y|X) \), which is called \textit{prevarication},\(^3\) is the uncertainty in the channel output from the \textit{transmitter’s point-of-view}. In other words, the sender knows exactly what he sends, but is uncertain on what the receiver will finally obtain.

A case that is of specific interest is when \( H(X|Y) = 0 \). By its definition, \( H(X|Y) = 0 \) if \( X \) becomes deterministic after observing \( Y \). In such case, the uncertainty of \( X \) after giving \( Y \) is completely zero.

The next corollary can be proved similarly to Theorem 2.10.

**Corollary 2.11 (Chain rule for conditional entropy)**

\[ H(X, Y|Z) = H(X|Z) + H(Y|X, Z). \]

### 2.1.5 Properties of joint entropy and conditional entropy

**Lemma 2.12 (Conditioning never increases entropy)** Side information \( Y \) decreases the uncertainty about \( X \):

\[ H(X|Y) \leq H(X) \]

with equality holding iff \( X \) and \( Y \) are independent. In other words, “conditioning” reduces entropy.

\(^2\)Equivocation is an ambiguous statement one uses deliberately in order to deceive or avoid speaking the truth.

\(^3\)Prevarication is the deliberate act of deviating from the truth (it is a synonym of “equivocation”).
Proof:

\[ H(X) - H(X|Y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_{X,Y}(x,y) \cdot \log_2 \frac{P_{X|Y}(x|y)}{P_X(x)} \]

\[ = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_{X,Y}(x,y) \cdot \log_2 \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} \]

\[ \geq \left( \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_{X,Y}(x,y) \right) \log_2 \frac{\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_{X,Y}(x,y)}{\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_X(x)P_Y(y)} \]

\[ = 0 \]

where the inequality follows from the log-sum inequality, with equality holding if

\[ \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} = \text{constant} \quad \forall \; (x,y) \in \mathcal{X} \times \mathcal{Y}. \]

Since probability must sum to 1, the above constant equals 1, which is exactly the case of \( X \) being independent of \( Y \).

\[ \square \]

**Lemma 2.13** Entropy is additive for independent random variables; i.e.,

\[ H(X,Y) = H(X) + H(Y) \quad \text{for independent } X \text{ and } Y. \]

**Proof:** By the previous lemma, independence of \( X \) and \( Y \) implies \( H(Y|X) = H(Y) \). Hence

\[ H(X,Y) = H(X) + H(Y|X) = H(X) + H(Y). \]

\[ \square \]

Since conditioning never increases entropy, it follows that

\[ H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y). \quad (2.1.8) \]

The above lemma tells us that equality holds for (2.1.8) only when \( X \) is independent of \( Y \).

A result similar to (2.1.8) also applies to conditional entropy.

**Lemma 2.14** Conditional entropy is lower additive; i.e.,

\[ H(X_1, X_2|Y_1, Y_2) \leq H(X_1|Y_1) + H(X_2|Y_2). \]
Equality holds iff
\[ P_{X_1,X_2|Y_1,Y_2}(x_1, x_2|y_1, y_2) = P_{X_1|Y_1}(x_1|y_1) P_{X_2|Y_2}(x_2|y_2) \]
for all \( x_1, x_2, y_1 \) and \( y_2 \).

**Proof:** Using the chain rule for conditional entropy and the fact that conditioning reduces entropy, we can write
\[
H(X_1, X_2|Y_1, Y_2) = H(X_1|Y_1, Y_2) + H(X_2|X_1, Y_1, Y_2) \\
\leq H(X_1|Y_1) + H(X_2|Y_2). \tag{2.1.9}
\]
For (2.1.9), equality holds iff \( X_1 \) and \( X_2 \) are conditionally independent given \((Y_1, Y_2)\):
\[
P_{X_1,X_2|Y_1,Y_2}(x_1, x_2|y_1, y_2) = P_{X_1|Y_1}(x_1|y_1) P_{X_2|Y_2}(x_2|y_2). \tag{2.1.9}
\]
For (2.1.10), equality holds iff \( X_1 \) is conditionally independent of \( Y_2 \) given \( Y_1 \) (i.e., \( P_{X_1|Y_1,Y_2}(x_1|y_1, y_2) = P_{X_1|Y_1}(x_1|y_1) \)) and \( X_2 \) is conditionally independent of \( Y_1 \) given \( Y_2 \) (i.e., \( P_{X_2|Y_1,Y_2}(x_2|y_1, y_2) = P_{X_2|Y_2}(x_2|y_2) \)). Hence, the desired equality condition of the lemma is obtained. \( \square \)

### 2.2 Mutual information

For two random variables \( X \) and \( Y \), the **mutual information** between \( X \) and \( Y \) is the reduction in the *uncertainty* of \( Y \) due to the knowledge of \( X \) (or vice versa).

A dual definition of mutual information states that it is the average amount of information that \( Y \) has (or contains) about \( X \) or \( X \) has (or contains) about \( Y \).

We can think of the mutual information between \( X \) and \( Y \) in terms of a channel whose input is \( X \) and whose output is \( Y \). Thereby the reduction of the uncertainty is by definition the total uncertainty of \( X \) (i.e., \( H(X) \)) minus the uncertainty of \( X \) after observing \( Y \) (i.e., \( H(X|Y) \)). Mathematically, it is
\[
\text{mutual information} = I(X; Y) \triangleq H(X) - H(X|Y). \tag{2.2.1}
\]
It can be easily verified from (2.1.7) that mutual information is symmetric; i.e., \( I(X; Y) = I(Y; X) \).

#### 2.2.1 Properties of mutual information

**Lemma 2.15**

1. \( I(X; Y) = \sum_{x \in X} \sum_{y \in Y} P_{X,Y}(x, y) \log_2 \frac{P_{X,Y}(x, y)}{P_X(x)P_Y(y)} \).
2. \( I(X; Y) = I(Y; X) \).
3. \( I(X; Y) = H(X) + H(Y) - H(X, Y) \).
4. \( I(X; Y) \leq H(X) \) with equality holding iff \( X \) is a function of \( Y \) (i.e., \( X = f(Y) \) for some function \( f(\cdot) \)).
5. \( I(X; Y) \geq 0 \) with equality holding iff \( X \) and \( Y \) are independent.
6. \( I(X; Y) \leq \min\{\log_2 |X|, \log_2 |Y|\} \).

**Proof:** Properties 1, 2, 3, and 4 follow immediately from the definition. Property 5 is a direct consequence of Lemma 2.12. Property 6 holds iff \( I(X; Y) \leq \log_2 |X| \) and \( I(X; Y) \leq \log_2 |Y| \). To show the first inequality, we write \( I(X; Y) = H(X) - H(X|Y) \), use the fact that \( H(X|Y) \) is non-negative and apply Lemma 2.6. A similar proof can be used to show that \( I(X; Y) \leq \log_2 |Y| \). □

The relationships between \( H(X) \), \( H(Y) \), \( H(X, Y) \), \( H(X|Y) \), \( H(Y|X) \) and \( I(X; Y) \) can be illustrated by the Venn diagram in Figure 2.2.

### 2.2.2 Conditional mutual information

The conditional mutual information, denoted by \( I(X; Y|Z) \), is defined as the common uncertainty between \( X \) and \( Y \) under the knowledge of \( Z \). It is mathematically defined by

\[
I(X; Y|Z) \triangleq H(X|Z) - H(X|Y, Z). \tag{2.2.2}
\]
Lemma 2.16 (Chain rule for mutual information)

\[ I(X; Y, Z) = I(X; Y) + I(X; Z|Y) = I(X; Z) + I(X; Y|Z). \]

**Proof:** Without loss of generality, we only prove the first equality:

\[
I(X; Y, Z) = H(X) - H(X|Y, Z) = H(X) - H(X|Y) + H(X|Y) - H(X|Y, Z) = I(X; Y) + I(X; Z|Y).
\]

The above lemma can be read as: the information that \((Y, Z)\) has about \(X\) is equal to the information that \(Y\) has about \(X\) plus the information that \(Z\) has about \(X\) when \(Y\) is already known.

### 2.3 Properties of entropy and mutual information for multiple random variables

**Theorem 2.17 (Chain rule for entropy)** Let \(X_1, X_2, \ldots, X_n\) be drawn according to \(P_{X^n}(x^n) \triangleq P_{X_1,...,X_n}(x_1, \ldots, x_n)\), where we use the common superscript notation to denote an \(n\)-tuple: \(X^n \triangleq (X_1, \ldots, X_n)\) and \(x^n \triangleq (x_1, \ldots, x_n)\). Then

\[
H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1),
\]

where \(H(X_i|X_{i-1}, \ldots, X_1) \triangleq H(X_1)\) for \(i = 1\). (The above chain rule can also be written as:

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

where \(X^i \triangleq (X_1, \ldots, X_i)\).)

**Proof:** From (2.1.6),

\[
H(X_1, X_2, \ldots, X_n) = H(X_1, X_2, \ldots, X_{n-1}) + H(X_n|X_{n-1}, \ldots, X_1). \quad (2.3.1)
\]

Once again, applying (2.1.6) to the first term of the right-hand-side of (2.3.1), we have

\[
H(X_1, X_2, \ldots, X_{n-1}) = H(X_1, X_2, \ldots, X_{n-2}) + H(X_{n-1}|X_{n-2}, \ldots, X_1).
\]

The desired result can then be obtained by repeatedly applying (2.1.6). \(\Box\)
Theorem 2.18 (Chain rule for conditional entropy)

\[ H(X_1, X_2, \ldots, X_n | Y) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1, Y). \]

**Proof:** The theorem can be proved similarly to Theorem 2.17. \(\square\)

Theorem 2.19 (Chain rule for mutual information)

\[ I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1), \]

where \(I(X_i; Y | X_{i-1}, \ldots, X_1) \triangleq I(X_1; Y)\) for \(i = 1\).

**Proof:** This can be proved by first expressing mutual information in terms of entropy and conditional entropy, and then applying the chain rules for entropy and conditional entropy. \(\square\)

Theorem 2.20 (Independence bound on entropy)

\[ H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i). \]

Equality holds iff all the \(X_i\)'s are independent from each other.\(^4\)

**Proof:** By applying the chain rule for entropy,

\[ H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \leq \sum_{i=1}^{n} H(X_i). \]

Equality holds iff each conditional entropy is equal to its associated entropy, that is, \(X_i\) is independent of \((X_{i-1}, \ldots, X_1)\) for all \(i\). \(\square\)

\(^4\)This condition is equivalent to requiring that \(X_i\) be independent of \((X_{i-1}, \ldots, X_1)\) for all \(i\). The equivalence can be directly proved using the chain rule for joint probabilities, i.e., \(P_{X^n}(x^n) = \prod_{i=1}^{n} P_{X_i | X_{i-1}}(x_i | x_{i-1})\); it is left as an exercise.
Theorem 2.21 (Bound on mutual information) If \( \{(X_i, Y_i)\}_{i=1}^n \) is a process satisfying the conditional independence assumption \( P_{Y^n|X^n} = \prod_{i=1}^n P_{Y_i|X_i} \), then

\[
I(X_1, \ldots, X_n; Y_1, \ldots, Y_n) \leq \sum_{i=1}^n I(X_i; Y_i)
\]

with equality holding iff \( \{X_i\}_{i=1}^n \) are independent.

**Proof:** From the independence bound on entropy, we have

\[
H(Y_1, \ldots, Y_n) \leq \sum_{i=1}^n H(Y_i).
\]

By the conditional independence assumption, we have

\[
H(Y_1, \ldots, Y_n|X_1, \ldots, X_n) = E \left[ -\log_2 P_{Y^n|X^n}(Y^n|X^n) \right]
= E \left[ -\sum_{i=1}^n \log_2 P_{Y_i|X_i}(Y_i|X_i) \right]
= \sum_{i=1}^n H(Y_i|X_i).
\]

Hence

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

with equality holding iff \( \{Y_i\}_{i=1}^n \) are independent, which holds iff \( \{X_i\}_{i=1}^n \) are independent.

\[
\square
\]

2.4 Data processing inequality

Lemma 2.22 (Data processing inequality) (This is also called the data processing lemma.) If \( X \to Y \to Z \), then \( I(X; Y) \geq I(X; Z) \).

**Proof:** The Markov chain relationship \( X \to Y \to Z \) means that \( X \) and \( Z \) are conditional independent given \( Y \) (cf. Appendix B); we directly have that \( I(X; Z|Y) = 0 \). By the chain rule for mutual information,

\[
I(X; Z) + I(X; Y|Z) = I(X; Y, Z) \tag{2.4.1}
\]
“By processing, we can only reduce (mutual) information, but the processed information may be in a more *useful* form!”

Figure 2.3: Communication context of the data processing lemma.

\[
I(U; V) \leq I(X; Y)
\]

Since \(I(X; Y|Z) \geq 0\), we obtain that \(I(X; Y) \geq I(X; Z)\) with equality holding iff \(I(X; Y|Z) = 0\).

The data processing inequality means that the mutual information will not increase after processing. This result is somewhat counter-intuitive since given two random variables \(X\) and \(Y\), we might believe that applying a well-designed processing scheme to \(Y\), which can be generally represented by a mapping \(g(Y)\), could possibly increase the mutual information. However, for any \(g(\cdot)\), \(X \rightarrow Y \rightarrow g(Y)\) forms a Markov chain which implies that data processing cannot increase mutual information. A communication context for the data processing lemma is depicted in Figure 2.3, and summarized in the next corollary.

**Corollary 2.23** For jointly distributed random variables \(X\) and \(Y\) and any function \(g(\cdot)\), we have \(X \rightarrow Y \rightarrow g(Y)\) and

\[
I(X; Y) \geq I(X; g(Y)).
\]

We also note that if \(Z\) obtains all the information about \(X\) through \(Y\), then knowing \(Z\) will not help increase the mutual information between \(X\) and \(Y\); this is formalized in the following.

**Corollary 2.24** If \(X \rightarrow Y \rightarrow Z\), then

\[
I(X; Y|Z) \leq I(X; Y).
\]

**Proof:** The proof directly follows from (2.4.1) and (2.4.2).

It is worth pointing out that it is possible that \(I(X; Y|Z) > I(X; Y)\) when \(X, Y\) and \(Z\) do *not* form a Markov chain. For example, let \(X\) and \(Y\) be independent equiprobable binary zero-one random variables, and let \(Z = X + Y\). Then,

\[
I(X; Y|Z) = H(X|Z) - H(X|Y, Z)
\]
\[ H(X|Z) = P_Z(0)H(X|z = 0) + P_Z(1)H(X|z = 1) + P_Z(2)H(X|z = 2) = 0 + 0.5 + 0 = 0.5 \text{ bits}, \]

which is clearly larger than \( I(X; Y) = 0 \).

Finally, we observe that we can extend the data processing inequality for a sequence of random variables forming a Markov chain:

**Corollary 2.25** If \( X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n \), then for any \( i, j, k, l \) such that \( 1 \leq i \leq j \leq k \leq l \leq n \), we have that

\[ I(X_i; X_l) \leq I(X_j; X_k). \]

### 2.5 Fano’s inequality

Fano’s inequality is a quite useful tool widely employed in Information Theory to prove converse results for coding theorems (as we will see in the following chapters).

**Lemma 2.26 (Fano’s inequality)** Let \( X \) and \( Y \) be two random variables, correlated in general, with alphabets \( \mathcal{X} \) and \( \mathcal{Y} \), respectively, where \( \mathcal{X} \) is finite but \( \mathcal{Y} \) can be countably infinite. Let \( \hat{X} \triangleq g(Y) \) be an estimate of \( X \) from observing \( Y \), where \( g : \mathcal{Y} \rightarrow \mathcal{X} \) is a given estimation function. Define the probability of error as

\[ P_e \triangleq \Pr[\hat{X} \neq X]. \]

Then the following inequality holds

\[ H(X|Y) \leq h_b(P_e) + P_e \cdot \log_2(|\mathcal{X}| - 1), \tag{2.5.1} \]

where \( h_b(x) \triangleq -x \log_2 x - (1 - x) \log_2 (1 - x) \) for \( 0 \leq x \leq 1 \) is the binary entropy function.

**Observation 2.27**

- Note that when \( P_e = 0 \), we obtain that \( H(X|Y) = 0 \) (see (2.5.1)) as intuition suggests, since if \( P_e = 0 \), then \( \hat{X} = g(Y) = X \) (with probability 1) and thus \( H(X|Y) = H(g(Y)|Y) = 0 \).
• Fano’s inequality yields upper and lower bounds on $P_e$ in terms of $H(X|Y)$. This is illustrated in Figure 2.4, where we plot the region for the pairs $(P_e, H(X|Y))$ that are permissible under Fano’s inequality. In the figure, the boundary of the permissible (dashed) region is given by the function

$$f(P_e) \triangleq h_b(P_e) + P_e \cdot \log_2(|\mathcal{X}| - 1),$$

the right-hand side of (2.5.1). We obtain that when

$$\log_2(|\mathcal{X}| - 1) < H(X|Y) \leq \log_2(|\mathcal{X}|),$$

$P_e$ can be upper and lower bounded as follows:

$$0 < \inf\{a : f(a) \geq H(X|Y)\} \leq P_e \leq \sup\{a : f(a) \geq H(X|Y)\} < 1.$$ 

Furthermore, when

$$0 < H(X|Y) \leq \log_2(|\mathcal{X}| - 1),$$

only the lower bound holds:

$$P_e \geq \inf\{a : f(a) \geq H(X|Y)\} > 0.$$ 

Thus for all non-zero values of $H(X|Y)$, we obtain a lower bound (of the same form above) on $P_e$: the bound implies that if $H(X|Y)$ is bounded away from zero, $P_e$ is also bounded away from zero.

• A weaker but simpler version of Fano’s inequality can be directly obtained from (2.5.1) by noting that $h_b(P_e) \leq 1$:

$$H(X|Y) \leq 1 + P_e \log_2(|\mathcal{X}| - 1), \quad (2.5.2)$$

which in turn yields that

$$P_e \geq \frac{H(X|Y) - 1}{\log_2(|\mathcal{X}| - 1)} \quad (\text{for } |\mathcal{X}| > 2)$$

which is weaker than the above lower bound on $P_e$.

**Proof of Lemma 2.26:**

Define a new random variable,

$$E \triangleq \begin{cases} 1, & \text{if } g(Y) \neq X \\ 0, & \text{if } g(Y) = X \end{cases}.$$
Then using the chain rule for conditional entropy, we obtain

\[
\]

Observe that \( E \) is a function of \( X \) and \( Y \); hence, \( H(E|X, Y) = 0 \). Since conditioning never increases entropy, \( H(E|Y) \leq H(E) = h_b(P_e) \). The remaining term, \( H(X|E, Y) \), can be bounded as follows:

\[
H(X|E, Y) = \Pr[E = 0]H(X|Y, E = 0) + \Pr[E = 1]H(X|Y, E = 1) \\
\leq (1 - P_e) \cdot 0 + P_e \cdot \log_2(|X| - 1),
\]

since \( X = g(Y) \) for \( E = 0 \), and given \( E = 1 \), we can upper bound the conditional entropy by the logarithm of the number of remaining outcomes, i.e., \((|X| - 1)\).

Combining these results completes the proof. \( \Box \)

Fano’s inequality cannot be improved in the sense that the lower bound, \( H(X|Y) \), can be achieved for some specific cases. Any bound that can be achieved in some cases is often referred to as sharp.\(^5\) From the proof of the above lemma, we can observe that equality holds in Fano’s inequality, if \( H(E|Y) = H(E) \) and \( H(X|Y, E = 1) = \log_2(|X| - 1) \). The former is equivalent to \( E \) being independent of \( Y \), and the latter holds iff \( P_{X|Y}(\cdot|y) \) is uniformly distributed over

\(^5\) **Definition.** A bound is said to be **sharp** if the bound is achievable for some specific cases. A bound is said to be **tight** if the bound is achievable for all cases.
the set $X - \{g(y)\}$. We can therefore create an example in which equality holds in Fano’s inequality.

**Example 2.28** Suppose that $X$ and $Y$ are two independent random variables which are both uniformly distributed on the alphabet $\{0, 1, 2\}$. Let the estimating function be given by $g(y) = y$. Then

$$P_e = \Pr[g(Y) \neq X] = \Pr[Y \neq X] = 1 - \sum_{x=0}^{2} P_X(x)P_Y(x) = \frac{2}{3}.$$  

In this case, equality is achieved in Fano’s inequality, i.e.,

$$h_b \left( \frac{2}{3} \right) + \frac{2}{3} \cdot \log_2(3 - 1) = H(X|Y) = H(X) = \log_2 3.$$

To conclude this section, we present an alternative proof for Fano’s inequality to illustrate the use of the data processing inequality and the FI Lemma.

**Alternative Proof of Fano’s inequality:** Noting that $X \rightarrow Y \rightarrow \hat{X}$ form a Markov chain, we directly obtain via the data processing inequality that

$$I(X; Y) \geq I(X; \hat{X}),$$

which implies that

$$H(X|Y) \leq H(X|\hat{X}).$$

Thus, if we show that $H(X|\hat{X})$ is no larger than the right-hand side of (2.5.1), the proof of (2.5.1) is complete.

Noting that

$$P_e = \sum_{x \in X} \sum_{\hat{x} \in X: \hat{x} \neq x} P_{X,\hat{X}}(x, \hat{x})$$

and

$$1 - P_e = \sum_{x \in X} \sum_{\hat{x} \in X: \hat{x} = x} P_{X,\hat{X}}(x, \hat{x}) = \sum_{x \in X} P_{X,\hat{X}}(x, x),$$

we obtain that

$$H(X|\hat{X}) - h_b(P_e) - P_e \log_2(|X| - 1)$$

$$= \sum_{x \in X} \sum_{\hat{x} \in X: \hat{x} \neq x} P_{X,\hat{X}}(x, \hat{x}) \log_2 \frac{1}{P_{X|\hat{X}}(x|\hat{x})} + \sum_{x \in X} P_{X,\hat{X}}(x, x) \log_2 \frac{1}{P_{X|\hat{X}}(x|x)}$$

$$- \left[ \sum_{x \in X} \sum_{\hat{x} \in X: \hat{x} \neq x} P_{X,\hat{X}}(x, \hat{x}) \right] \log_2 \left( \frac{|X| - 1}{P_e} \right) + \left[ \sum_{x \in X} P_{X,\hat{X}}(x, x) \right] \log_2 (1 - P_e)$$

25
\[
= \sum_{x \in X} \sum_{\hat{x} \in X; \hat{x} \neq x} P_{X,\hat{X}}(x, \hat{x}) \log_2 \frac{P_e}{P_{X|\hat{X}}(x|\hat{x})(|X| - 1)} \\
+ \sum_{x \in X} P_{X,\hat{X}}(x, x) \log_2 \frac{1 - P_e}{P_{X|\hat{X}}(x|x)}
\]
\[
\leq \log_2(e) \sum_{x \in X} \sum_{\hat{x} \in X; \hat{x} \neq x} P_{X,\hat{X}}(x, \hat{x}) \left[ \frac{P_e}{P_{X|\hat{X}}(x|\hat{x})(|X| - 1)} - 1 \right] \\
+ \log_2(e) \sum_{x \in X} P_{X,\hat{X}}(x, x) \left[ \frac{1 - P_e}{P_{X|\hat{X}}(x|x)} - 1 \right]
\]
\[
= \log_2(e) \left[ \frac{P_e}{(|X| - 1)} \sum_{x \in X} \sum_{\hat{x} \in X; \hat{x} \neq x} P_{X}(\hat{x}) - \sum_{x \in X} \sum_{\hat{x} \in X; \hat{x} \neq x} P_{X,\hat{X}}(x, \hat{x}) \right] \\
+ \log_2(e) \left[ (1 - P_e) \sum_{x \in X} P_{X}(x) - \sum_{x \in X} P_{X,\hat{X}}(x, x) \right]
\]
\[
= \log_2(e) \left[ \frac{P_e}{(|X| - 1)} (|X| - 1) - P_e \right] + \log_2(e) [(1 - P_e) - (1 - P_e)]
\]
\[
= 0
\]

where the inequality follows by applying the FI Lemma to each logarithm term in (2.5.3).

2.6 Divergence and variational distance

In addition to the probabilistically defined entropy and mutual information, another measure that is frequently considered in information theory is divergence or relative entropy. In this section, we define this measure and study its statistical properties.

**Definition 2.29 (Divergence)** Given two discrete random variables \( X \) and \( \hat{X} \) defined over a common alphabet \( \mathcal{X} \), the divergence (other names are Kullback-Leibler divergence or distance, relative entropy and discrimination) is denoted by \( D(X||\hat{X}) \) or \( D(P_X||P_{\hat{X}}) \) and defined by\(^6\)

\[
D(X||\hat{X}) = D(P_X||P_{\hat{X}}) \triangleq E_X \left[ \log_2 \frac{P_X(X)}{P_{\hat{X}}(X)} \right] = \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{P_X(x)}{P_{\hat{X}}(X)}.
\]

\(^6\)In order to be consistent with the units (in bits) adopted for entropy and mutual information, we will also use the base-2 logarithm for divergence unless otherwise specified.
In other words, the divergence $D(P_X\|P_{\hat{X}})$ is the expectation (with respect to $P_X$) of the log-likelihood ratio $\log_2[P_X/P_{\hat{X}}]$ of distribution $P_X$ against distribution $P_{\hat{X}}$. $D(X\|\hat{X})$ can be viewed as a measure of “distance” or “dissimilarity” between distributions $P_X$ and $P_{\hat{X}}$. $D(X\|\hat{X})$ is also called relative entropy since it can be regarded as a measure of the inefficiency of mistakenly assuming that the distribution of a source is $P_{\hat{X}}$ when the true distribution is $P_X$. For example, if we know the true distribution $P_X$ of a source, then we can construct a lossless data compression code with average codeword length achieving entropy $H(X)$ (this will be studied in the next chapter). If, however, we mistakenly thought that the “true” distribution is $P_{\hat{X}}$ and employ the “best” code corresponding to $P_{\hat{X}}$, then the resultant average codeword length becomes

$$\sum_{x \in \mathcal{X}} [-P_X(x) \cdot \log_2 P_{\hat{X}}(x)].$$

As a result, the relative difference between the resultant average codeword length and $H(X)$ is the relative entropy $D(X\|\hat{X})$. Hence, divergence is a measure of the system cost (e.g., storage consumed) paid due to mis-classifying the system statistics.

Note that when computing divergence, we follow the convention that

$$0 \cdot \log_2 \frac{0}{p} = 0 \quad \text{and} \quad p \cdot \log_2 \frac{p}{0} = \infty \quad \text{for} \ p > 0.$$

We next present some properties of the divergence and discuss its relation with entropy and mutual information.

**Lemma 2.30 (Non-negativity of divergence)**

$$D(X\|\hat{X}) \geq 0$$

with equality iff $P_X(x) = P_{\hat{X}}(x)$ for all $x \in \mathcal{X}$ (i.e., the two distributions are equal).

**Proof:**

$$D(X\|\hat{X}) = \sum_{x \in \mathcal{X}} P_X(x) \cdot \log_2 \frac{P_X(x)}{P_{\hat{X}}(x)}$$

$$\geq \left( \sum_{x \in \mathcal{X}} P_X(x) \right) \cdot \log_2 \frac{\sum_{x \in \mathcal{X}} P_X(x)}{\sum_{x \in \mathcal{X}} P_{\hat{X}}(x)}$$

$$= 0$$
where the second step follows from the log-sum inequality with equality holding iff for every \( x \in \mathcal{X} \),

\[
\frac{P_X(x)}{P_{\hat{X}}(x)} = \frac{\sum_{a \in \mathcal{X}} P_X(a)}{\sum_{b \in \mathcal{X}} P_{\hat{X}}(b)},
\]

or equivalently \( P_X(x) = P_{\hat{X}}(x) \) for all \( x \in \mathcal{X} \).

\[\square\]

**Lemma 2.31 (Mutual information and divergence)**

\[I(X;Y) = D(P_{X,Y} \parallel P_X \times P_Y),\]

where \( P_{X,Y}(\cdot,\cdot) \) is the joint distribution of the random variables \( X \) and \( Y \) and \( P_X(\cdot) \) and \( P_Y(\cdot) \) are the respective marginals.

**Proof:** The observation follows directly from the definitions of divergence and mutual information. \[\square\]

**Definition 2.32 (Refinement of distribution)** Given distribution \( P_X \) on \( \mathcal{X} \), divide \( \mathcal{X} \) into \( k \) mutually disjoint sets, \( \mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_k \), satisfying

\[\mathcal{X} = \bigcup_{i=1}^{k} \mathcal{U}_i.\]

Define a new distribution \( P_U \) on \( \mathcal{U} = \{1, 2, \cdots, k\} \) as

\[P_U(i) = \sum_{x \in \mathcal{U}_i} P_X(x).\]

Then \( P_X \) is called a refinement (or more specifically, a \( k \)-refinement) of \( P_U \).

Let us briefly discuss the relation between the processing of information and its refinement. Processing of information can be modeled as a (many-to-one) mapping, and refinement is actually the reverse operation. Recall that the data processing lemma shows that mutual information can never increase due to processing. Hence, if one wishes to increase mutual information, he should simultaneously “anti-process” (or refine) the involved statistics.

From Lemma 2.31, the mutual information can be viewed as the divergence of a joint distribution against the product distribution of the marginals. It is therefore reasonable to expect that a similar effect due to processing (or a reverse effect due to refinement) should also apply to divergence. This is shown in the next lemma.
Lemma 2.33 (Refinement cannot decrease divergence) Let \( P_X \) and \( \hat{P}_X \) be the refinements (\( k \)-refinements) of \( P_U \) and \( \hat{P}_U \) respectively. Then
\[
D(P_X \| P_X) \geq D(P_U \| \hat{P}_U).
\]

Proof: By the log-sum inequality, we obtain that for any \( i \in \{1, 2, \cdots, k\} \)
\[
\sum_{x \in U_i} P_X(x) \log_2 \frac{P_X(x)}{\hat{P}_X(x)} \geq \left( \sum_{x \in U_i} P_X(x) \right) \log_2 \frac{\sum_{x \in U_i} P_X(x)}{\sum_{x \in U_i} \hat{P}_X(x)} = P_U(i) \log_2 \frac{P_U(i)}{\hat{P}_U(i)},
\]
with equality iff
\[
\frac{P_X(x)}{\hat{P}_X(x)} = \frac{P_U(i)}{\hat{P}_U(i)},
\]
for all \( x \in U \). Hence,
\[
D(P_X \| P_X) = \sum_{i=1}^{k} \sum_{x \in U_i} P_X(x) \log_2 \frac{P_X(x)}{\hat{P}_X(x)} \geq \sum_{i=1}^{k} P_U(i) \log_2 \frac{P_U(i)}{\hat{P}_U(i)} = D(P_U \| \hat{P}_U),
\]
with equality iff
\[
(\forall i) (\forall x \in U_i) \frac{P_X(x)}{\hat{P}_X(x)} = \frac{P_U(i)}{\hat{P}_U(i)}.
\]

Observation 2.34 One drawback of adopting the divergence as a measure between two distributions is that it does not meet the symmetry requirement of a true distance,\(^7\) since interchanging its two arguments may yield different quantities. In other words, \( D(P_X \| P_X) \neq D(P_X \| P_X) \) in general. (It also does not satisfy the triangular inequality.) Thus divergence is not a true distance or metric. Another measure which is a true distance, called \textit{variational distance}, is sometimes used instead.

\(^7\)Given a non-empty set \( A \), the function \( d : A \times A \rightarrow [0, \infty) \) is called a distance or metric if it satisfies the following properties.

1. Non-negativity: \( d(a, b) \geq 0 \) for every \( a, b \in A \) with equality holding iff \( a = b \).
2. Symmetry: \( d(a, b) = d(b, a) \) for every \( a, b \in A \).
3. Triangular inequality: \( d(a, b) + d(b, c) \geq d(a, c) \) for every \( a, b, c \in A \).
Lemma 2.36 The variational distance satisfies

\[ \|P_X - P_{\hat{X}}\| = 2 \cdot \sup_{E \subseteq \mathcal{X}} |P_X(E) - P_{\hat{X}}(E)| = 2 \cdot \sum_{x \in \mathcal{X} : P_X(x) > P_{\hat{X}}(x)} [P_X(x) - P_{\hat{X}}(x)]. \]

Proof: We first show that \( \|P_X - P_{\hat{X}}\| = 2 \cdot \sum_{x \in \mathcal{X} : P_X(x) > P_{\hat{X}}(x)} [P_X(x) - P_{\hat{X}}(x)] \). Setting \( \mathcal{A} \triangleq \{ x \in \mathcal{X} : P_X(x) > P_{\hat{X}}(x) \} \), we have

\[
\|P_X - P_{\hat{X}}\| = \sum_{x \in \mathcal{X}} |P_X(x) - P_{\hat{X}}(x)|
= \sum_{x \in \mathcal{A}} |P_X(x) - P_{\hat{X}}(x)| + \sum_{x \in A^c} |P_X(x) - P_{\hat{X}}(x)|
= \sum_{x \in \mathcal{A}} [P_X(x) - P_{\hat{X}}(x)] + \sum_{x \in A^c} [P_X(x) - P_{\hat{X}}(x)]
= \sum_{x \in \mathcal{A}} [P_X(x) - P_{\hat{X}}(x)] + P_X(A^c) - P_{\hat{X}}(A^c)
= \sum_{x \in \mathcal{A}} [P_X(x) - P_{\hat{X}}(x)] + P_X(A) - P_{\hat{X}}(A)
= \sum_{x \in \mathcal{A}} [P_X(x) - P_{\hat{X}}(x)] + \sum_{x \in A^c} [P_X(x) - P_{\hat{X}}(x)]
= 2 \cdot \sum_{x \in A} [P_X(x) - P_{\hat{X}}(x)]
\]

where \( A^c \) denotes the complement set of \( A \).

We next prove that \( \|P_X - P_{\hat{X}}\| = 2 \cdot \sup_{E \subseteq \mathcal{X}} |P_X(E) - P_{\hat{X}}(E)| \) by showing that each quantity is greater than or equal to the other. For any set \( E \subseteq \mathcal{X} \), we can write

\[
\|P_X - P_{\hat{X}}\| = \sum_{x \in \mathcal{X}} |P_X(x) - P_{\hat{X}}(x)|
= \sum_{x \in E} |P_X(x) - P_{\hat{X}}(x)| + \sum_{x \in E^c} |P_X(x) - P_{\hat{X}}(x)|
\geq \left| \sum_{x \in E} [P_X(x) - P_{\hat{X}}(x)] \right| + \left| \sum_{x \in E^c} [P_X(x) - P_{\hat{X}}(x)] \right|
\]
\[ = |P_X(E) - \hat{P}_X(E)| + |P_X(E^c) - \hat{P}_X(E^c)| \\
= |P_X(E) - \hat{P}_X(E)| + |P_X(E) - P_X(E)| \\
= 2 \cdot |P_X(E) - \hat{P}_X(E)|. \]

Thus \( \|P_X - \hat{P}_X\| \geq 2 \cdot \sup_{E \subset X} |P_X(E) - \hat{P}_X(E)|. \) Conversely, we have that
\[
2 \cdot \sup_{E \subset X} |P_X(E) - \hat{P}_X(E)| \geq 2 \cdot |P_X(A) - \hat{P}_X(A)|
= \left| \sum_{x \in A} [P_X(x) - \hat{P}_X(x)] \right| + \left| \sum_{x \in A^c} [\hat{P}_X(x) - P_X(x)] \right|
= \sum_{x \in A} |P_X(x) - \hat{P}_X(x)| + \sum_{x \in A^c} |P_X(x) - \hat{P}_X(x)|
= \|P_X - \hat{P}_X\|.
\]

Therefore, \( \|P_X - \hat{P}_X\| = 2 \cdot \sup_{E \subset X} |P_X(E) - \hat{P}_X(E)|. \) \( \square \)

**Lemma 2.37 (Variational distance vs divergence: Pinsker’s inequality)**

\[
D(X\|\hat{X}) \geq \frac{\log_2(e)}{2} \cdot \|P_X - \hat{P}_X\|^2.
\]

This result is referred to as Pinsker’s inequality.

**Proof:**

1. With \( A \triangleq \{ x \in X : P_X(x) > \hat{P}_X(x) \} \), we have from the previous lemma that
\[
\|P_X - \hat{P}_X\| = 2|P_X(A) - \hat{P}_X(A)|.
\]

2. Define two random variables \( U \) and \( \hat{U} \) as:
\[
U = \begin{cases} 
1, & \text{if } X \in A; \\
0, & \text{if } X \in A^c,
\end{cases}
\]
and
\[
\hat{U} = \begin{cases} 
1, & \text{if } \hat{X} \in A; \\
0, & \text{if } \hat{X} \in A^c.
\end{cases}
\]

Then \( P_X \) and \( P_\hat{X} \) are refinements (2-refinements) of \( P_U \) and \( P_\hat{U} \), respectively. From Lemma 2.33, we obtain that
\[
D(P_X\|P_\hat{X}) \geq D(P_U\|P_\hat{U}).
\]

31
3. The proof is complete if we show that

\[
D(P_U \| P_{\hat{U}}) \geq 2 \log_2(e) [P_X(A) - P_{\hat{X}}(A)]^2
= 2 \log_2(e) [P_U(1) - P_{\hat{U}}(1)]^2.
\]

For ease of notations, let \( p = P_U(1) \) and \( q = P_{\hat{U}}(1) \). Then to prove the above inequality is equivalent to show that

\[
p \cdot \ln \frac{p}{q} + (1 - p) \cdot \ln \frac{1 - p}{1 - q} \geq 2(p - q)^2.
\]

Define

\[
f(p, q) \triangleq p \cdot \ln \frac{p}{q} + (1 - p) \cdot \ln \frac{1 - p}{1 - q} - 2(p - q)^2,
\]

and observe that

\[
\frac{df(p, q)}{dq} = (p - q) \left( 4 - \frac{1}{q(1 - q)} \right) \leq 0 \quad \text{for } q \leq p.
\]

Thus, \( f(p, q) \) is non-increasing in \( q \) for \( q \leq p \). Also note that \( f(p, q) = 0 \) for \( q = p \). Therefore,

\[
f(p, q) \geq 0 \quad \text{for } q \leq p.
\]

The proof is completed by noting that

\[
f(p, q) \geq 0 \quad \text{for } q \geq p,
\]

since \( f(1 - p, 1 - q) = f(p, q) \).

\[\square\]

**Observation 2.38** The above lemma tells us that for a sequence of distributions \( \{(P_{X_n}, P_{\hat{X}_n})\}_{n \geq 1} \), when \( D(P_{X_n} \| P_{\hat{X}_n}) \) goes to zero as \( n \) goes to infinity, \( \| P_{X_n} - P_{\hat{X}_n} \| \) goes to zero as well. But the converse does not necessarily hold. For a quick counterexample, let

\[
P_{X_n}(0) = 1 - P_{X_n}(1) = 1/n > 0
\]

and

\[
P_{\hat{X}_n}(0) = 1 - P_{\hat{X}_n}(1) = 0.
\]

In this case,

\[
D(P_{X_n} \| P_{\hat{X}_n}) \to \infty
\]

since by convention, \( (1/n) \cdot \log_2((1/n)/0) \to \infty \). However,

\[
\| P_X - P_{\hat{X}} \|
= 2 \left[ P_X(\{x : P_X(x) > P_{\hat{X}}(x)\}) - P_{\hat{X}}(\{x : P_X(x) > P_{\hat{X}}(x)\}) \right]
\]

32
We however can upper bound $D(P_X \| P_\hat{X})$ by the variational distance between $P_X$ and $P_\hat{X}$ when $D(P_X \| P_\hat{X}) < \infty$.

**Lemma 2.39** If $D(P_X \| P_\hat{X}) < \infty$, then

$$D(P_X \| P_\hat{X}) \leq \frac{\log_2(e)}{\min\{x : P_X(x) > 0\} \cdot \min\{P_X(x), P_\hat{X}(x)\}} \cdot \|P_X - P_\hat{X}\|.$$ 

**Proof:** Without loss of generality, we assume that $P_X(x) > 0$ for all $x \in \mathcal{X}$. Since $D(P_X \| P_\hat{X}) < \infty$, we have that for any $x \in \mathcal{X}$, $P_X(x) > 0$ implies that $P_\hat{X}(x) > 0$. Let

$$t \triangleq \min\{x \in \mathcal{X} : P_X(x) > 0\} \cdot \min\{P_X(x), P_\hat{X}(x)\}.$$ 

Then for all $x \in \mathcal{X}$,

$$\ln \frac{P_X(x)}{P_\hat{X}(x)} \leq \ln \frac{P_X(x)}{P_\hat{X}(x)} \leq \max_{\min\{P_X(x), P_\hat{X}(x)\} \leq s \leq \max\{P_X(x), P_\hat{X}(x)\}} \frac{d \ln(s)}{ds} \cdot |P_X(x) - P_\hat{X}(x)| = \frac{1}{\min\{P_X(x), P_\hat{X}(x)\}} \cdot |P_X(x) - P_\hat{X}(x)| \leq \frac{1}{t} \cdot |P_X(x) - P_\hat{X}(x)|.$$

Hence,

$$D(P_X \| P_\hat{X}) = \log_2(e) \sum_{x \in \mathcal{X}} P_X(x) \cdot \ln \frac{P_X(x)}{P_\hat{X}(x)} \leq \frac{\log_2(e)}{t} \sum_{x \in \mathcal{X}} P_X(x) \cdot |P_X(x) - P_\hat{X}(x)| \leq \frac{\log_2(e)}{t} \sum_{x \in \mathcal{X}} |P_X(x) - P_\hat{X}(x)|$$

$$= \frac{\log_2(e)}{t} \cdot \|P_X - P_\hat{X}\|.$$ 

The next lemma discusses the effect of side information on divergence. As stated in Lemma 2.12, side information usually reduces entropy; it, however,
increases divergence. One interpretation of these results is that side information is useful. Regarding entropy, side information provides us more information, so uncertainty decreases. As for divergence, it is the measure or index of how easy one can differentiate the source from two candidate distributions. The larger the divergence, the easier one can tell apart between these two distributions and make the right guess. At an extreme case, when divergence is zero, one can never tell which distribution is the right one, since both produce the same source. So, when we obtain more information (side information), we should be able to make a better decision on the source statistics, which implies that the divergence should be larger.

**Definition 2.40 (Conditional divergence)** Given three discrete random variables, $X$, $\hat{X}$ and $Z$, where $X$ and $\hat{X}$ have a common alphabet $\mathcal{X}$, we define the conditional divergence between $X$ and $\hat{X}$ given $Z$ by

$$D(X\|\hat{X}|Z) = D(P_{X,Z}\|P_{\hat{X}|Z}) = \sum_{z \in Z} \sum_{x \in \mathcal{X}} P_{X,Z}(x,z) \log \frac{P_{X,Z}(x|z)}{P_{\hat{X}|Z}(x|z)}.$$  

In other words, it is the expected value with respect to $P_{X,Z}$ of the log-likelihood ratio $\log \frac{P_{X,Z}}{P_{\hat{X}|Z}}$.

**Lemma 2.41 (Conditional mutual information and conditional divergence)** Given three discrete random variables $X$, $Y$ and $Z$ with alphabets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, respectively, and joint distribution $P_{X,Y,Z}$, then

$$I(X;Y|Z) = D(P_{X,Y|Z}\|P_{X|Z}P_{Y|Z})$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} P_{X,Y,Z}(x,y,z) \log_2 \frac{P_{X,Y,Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}.$$  

where $P_{X,Y|Z}$ is conditional joint distribution of $X$ and $Y$ given $Z$, and $P_{X|Z}$ and $P_{Y|Z}$ are the conditional distributions of $X$ and $Y$, respectively, given $Z$.

**Proof:** The proof follows directly from the definition of conditional mutual information (2.2.2) and the above definition of conditional divergence.

**Lemma 2.42 (Chain rule for divergence)** For three discrete random variables, $X$, $\hat{X}$ and $Z$, where $X$ and $\hat{X}$ have a common alphabet $\mathcal{X}$, we have that

$$D(P_{X,Z}\|P_{\hat{X},Z}) = D(P_X\|P_{\hat{X}}) + D(P_{Z|X}\|P_{Z|\hat{X}}).$$  

**Proof:** The proof readily follows from the divergence definitions.
Lemma 2.43 (Conditioning never decreases divergence) For three discrete random variables, $X$, $\hat{X}$ and $Z$, where $X$ and $\hat{X}$ have a common alphabet $\mathcal{X}$, we have that

$$D(P_{X|Z}||P_{\hat{X}|Z}) \geq D(P_X||P_{\hat{X}}).$$

Proof:

$$D(P_{X|Z}||P_{\hat{X}|Z}) - D(P_X||P_{\hat{X}})$$

$$= \sum_{z \in Z} \sum_{x \in \mathcal{X}} P_{X,Z}(x, z) \cdot \log_2 \frac{P_{X|Z}(x|z)}{P_{\hat{X}|Z}(x|z)} - \sum_{x \in \mathcal{X}} P_X(x) \cdot \log_2 \frac{P_X(x)}{P_{\hat{X}}(x)}$$

$$= \sum_{z \in Z} \sum_{x \in \mathcal{X}} P_{X,Z}(x, z) \cdot \log_2 \frac{P_{X|Z}(x|z)}{P_{\hat{X}|Z}(x|z)} - \sum_{x \in \mathcal{X}} \left( \sum_{z \in Z} P_{X,Z}(x, z) \right) \cdot \log_2 \frac{P_X(x)}{P_{\hat{X}}(x)}$$

$$= \sum_{z \in Z} \sum_{x \in \mathcal{X}} P_{X,Z}(x, z) \cdot \log_2(e) \left( 1 - \frac{P_{X|Z}(x|z)P_X(x)}{P_{\hat{X}|Z}(x|z)P_{\hat{X}}(x)} \right) \quad \text{(by the FI Lemma)}$$

$$= \log_2(e) \left( 1 - \sum_{x \in \mathcal{X}} \frac{P_X(x)}{P_{\hat{X}}(x)} \sum_{z \in Z} P_Z(z)P_{X|Z}(x|z) \right)$$

$$= \log_2(e) \left( 1 - \sum_{x \in \mathcal{X}} \frac{P_X(x)}{P_{\hat{X}}(x)}P_X(x) \right)$$

$$= \log_2(e) \left( 1 - \sum_{x \in \mathcal{X}} P_X(x) \right) = 0$$

with equality holding iff for all $x$ and $z$,

$$\frac{P_X(x)}{P_{\hat{X}}(x)} = \frac{P_{X|Z}(x|z)}{P_{\hat{X}|Z}(x|z)}.$$

\[\square\]

Note that it is not necessary that

$$D(P_{X|Z}||P_{\hat{X}|Z}) \geq D(P_X||P_{\hat{X}}).$$

In other words, the side information is helpful for divergence only when it provides information on the similarity or difference of the two distributions. For the above case, $Z$ only provides information about $X$, and $\hat{Z}$ provides information about $\hat{X}$; so the divergence certainly cannot be expected to increase. The next lemma shows that if $(Z, \hat{Z})$ are independent of $(X, \hat{X})$, then the side information of $(Z, \hat{Z})$ does not help in improving the divergence of $X$ against $\hat{X}$.
Lemma 2.44 (Independent side information does not change divergence) If \((X, \hat{X})\) is independent of \((Z, \hat{Z})\), then
\[
D(P_{X|Z} \parallel P_{\hat{X}|\hat{Z}}) = D(P_X \parallel P_{\hat{X}}),
\]
where
\[
D(P_{X|Z} \parallel P_{\hat{X}|\hat{Z}}) \triangleq \sum_{x \in X} \sum_{z \in Z} P_{X,Z}(x, z) \log_2 \frac{P_{X|Z}(x|z)}{P_{\hat{X}|\hat{Z}}(x|z)}.
\]

Proof: This can be easily justified by the definition of divergence. \(\square\)

Lemma 2.45 (Additivity of divergence under independence)
\[
D(P_{X,Z} \parallel P_{\hat{X},\hat{Z}}) = D(P_X \parallel P_{\hat{X}}) + D(P_Z \parallel P_{\hat{Z}}),
\]
provided that \((X, \hat{X})\) is independent of \((Z, \hat{Z})\).

Proof: This can be easily proved from the definition. \(\square\)

2.7 Convexity/concavity of information measures

We next address the convexity/concavity properties of information measures with respect to the distributions on which they are defined. Such properties will be useful when optimizing the information measures over distribution spaces.

Lemma 2.46

1. \(H(P_X)\) is a concave function of \(P_X\), namely
\[
H(\lambda P_X + (1 - \lambda)P_{\hat{X}}) \geq \lambda H(P_X) + (1 - \lambda)H(P_{\hat{X}}).
\]

2. Noting that \(I(X; Y)\) can be re-written as \(I(P_X, P_{Y|X})\), where
\[
I(P_X, P_{Y|X}) \triangleq \sum_{x \in X} \sum_{y \in Y} P_{Y|X}(y|x)P_X(x) \log_2 \frac{P_{Y|X}(y|x)}{\sum_{a \in X} P_{Y|X}(y|a)P_X(a)},
\]
then \(I(X; Y)\) is a concave function of \(P_X\) (for fixed \(P_{Y|X}\)), and a convex function of \(P_{Y|X}\) (for fixed \(P_X\)).
3. \( D(P_X \| P_X) \) is convex with respect to both the first argument \( P_X \) and the second argument \( P_X \). It is also convex in the pair \( (P_X, P_X) \); i.e., if \( (P_X, P_X) \) and \( (Q_X, Q_X) \) are two pairs of probability mass functions, then

\[
D(\lambda P_X + (1 - \lambda) Q_X \| \lambda P_X + (1 - \lambda) Q_X) \\
\leq \lambda \cdot D(P_X \| P_X) + (1 - \lambda) \cdot D(Q_X \| Q_X),
\]

for all \( \lambda \in [0,1] \).

Proof:

1. The proof uses the log-sum inequality:

\[
H(\lambda P_X + (1 - \lambda) P_X) - \left[ \lambda H(P_X) + (1 - \lambda) H(P_X) \right] \\
= \lambda \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{P_X(x)}{\lambda P_X(x) + (1 - \lambda) P_X(x)} \\
+ (1 - \lambda) \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{P_X(x)}{\lambda P_X(x) + (1 - \lambda) P_X(x)} \\
\geq \lambda \left( \sum_{x \in \mathcal{X}} P_X(x) \right) \log_2 \frac{\sum_{x \in \mathcal{X}} P_X(x)}{\lambda P_X(x) + (1 - \lambda) P_X(x)} \\
+ (1 - \lambda) \left( \sum_{x \in \mathcal{X}} P_X(x) \right) \log_2 \frac{\sum_{x \in \mathcal{X}} P_X(x)}{\lambda P_X(x) + (1 - \lambda) P_X(x)} \\
= 0,
\]

with equality holding iff \( P_X(x) = P_X(x) \) for all \( x \).

2. We first show the concavity of \( I(P_X, P_{Y|X}) \) with respect to \( P_X \). Let \( \bar{\lambda} = 1 - \lambda \).

\[
I(\lambda P_X + \bar{\lambda} P_X, P_{Y|X}) - \lambda I(P_X, P_{Y|X}) - \bar{\lambda} I(P_X, P_{Y|X}) \\
= \lambda \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log_2 \frac{P_X(x) P_{Y|X}(y|x)}{\sum_{x \in \mathcal{X}} [\lambda P_X(x) + \bar{\lambda} P_X(x)] P_{Y|X}(y|x)} \\
+ \bar{\lambda} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log_2 \frac{P_X(x) P_{Y|X}(y|x)}{\sum_{x \in \mathcal{X}} [\lambda P_X(x) + \bar{\lambda} P_X(x)] P_{Y|X}(y|x)} \\
\geq 0 \quad \text{(by the log-sum inequality)}
\]
with equality holding iff
\[
\sum_{x \in \mathcal{X}} P_X(x) P_{Y|X}(y|x) = \sum_{x \in \mathcal{X}} P_{\tilde{X}}(x) P_{Y|X}(y|x)
\]
for all \( y \in \mathcal{Y} \). We now turn to the convexity of \( I(P_X, P_{Y|X}) \) with respect to \( P_{Y|X} \). For ease of notation, let \( P_{Y,\lambda}(y) \triangleq \lambda P_Y(y) + \bar{\lambda} P_{\tilde{Y}}(y) \), and \( P_{Y,\lambda|X}(y|x) \triangleq \lambda P_{Y|X}(y|x) + \bar{\lambda} P_{\tilde{Y}|X}(y|x) \). Then
\[
\lambda I(P_X, P_{Y|X}) + \bar{\lambda} I(P_X, P_{\tilde{Y}|X}) - I(P_X, \lambda P_{Y|X} + \bar{\lambda} P_{\tilde{Y}|X})
\]
\[
= \lambda \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log_2 \frac{P_{Y|X}(y|x)}{P_Y(y)}
\]
\[
+ \bar{\lambda} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{\tilde{Y}|X}(y|x) \log_2 \frac{P_{\tilde{Y}|X}(y|x)}{P_{\tilde{Y}}(y)}
\]
\[
- \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y,\lambda|X}(y|x) \log_2 \frac{P_{Y,\lambda|X}(y|x)}{P_{Y,\lambda}(y)}
\]
\[
= \lambda \log_2(e) \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \left( 1 - \frac{P_Y(y) P_{Y,\lambda|X}(y|x)}{P_{Y|X}(y|x) P_{Y,\lambda}(y)} \right)
\]
\[
+ \bar{\lambda} \log_2(e) \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{\tilde{Y}|X}(y|x) \left( 1 - \frac{P_{\tilde{Y}}(y) P_{Y,\lambda|X}(y|x)}{P_{\tilde{Y}|X}(y|x) P_{Y,\lambda}(y)} \right)
\]
\[
= 0,
\]
where the inequality follows from the FI Lemma, with equality holding iff
\[
(\forall x \in \mathcal{X}, y \in \mathcal{Y}) \quad \frac{P_Y(y)}{P_{Y|X}(y|x)} = \frac{P_{\tilde{Y}}(y)}{P_{\tilde{Y}|X}(y|x)}.
\]

3. For ease of notation, let \( P_{X,\lambda}(x) \triangleq \lambda P_X(x) + (1 - \lambda) P_{\tilde{X}}(x) \).
\[
\lambda D(P_X \| P_X) + (1 - \lambda) D(P_{\tilde{X}} \| P_X) - D(P_{X,\lambda} \| P_X)
\]
\[
= \lambda \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{P_X(x)}{P_{X,\lambda}(x)} + (1 - \lambda) \sum_{x \in \mathcal{X}} P_{\tilde{X}}(x) \log_2 \frac{P_{\tilde{X}}(x)}{P_{X,\lambda}(x)}
\]
\[
= \lambda D(P_X \| P_{X,\lambda}) + (1 - \lambda) D(P_{\tilde{X}} \| P_{X,\lambda})
\]

38
\[ \geq 0 \]

by the non-negativity of the divergence, with equality holding iff \( P_X(x) = P_\tilde{X}(x) \) for all \( x \).

Similarly, by letting \( P_{\hat{X}}(x) \equiv \lambda P_X(x) + (1 - \lambda) P_\tilde{X}(x) \), we obtain:

\[
\lambda D(P_X \| P_{\hat{X}}) + (1 - \lambda) D(P_X \| P_\tilde{X}) - D(P_X \| P_{\hat{X}}) \\
= \lambda \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{P_{\hat{X}}(x)}{P_X(x)} + (1 - \lambda) \sum_{x \in \mathcal{X}} P_X(x) \log_2 \frac{P_{\hat{X}}(x)}{P_\tilde{X}(x)} \\
\geq \frac{\lambda}{\ln 2} \sum_{x \in \mathcal{X}} P_X(x) \left( 1 - \frac{P_X(x)}{P_{\hat{X}}(x)} \right) + \frac{(1 - \lambda)}{\ln 2} \sum_{x \in \mathcal{X}} P_X(x) \left( 1 - \frac{P_\tilde{X}(x)}{P_{\hat{X}}(x)} \right) \\
= \log_2(e) \left( 1 - \sum_{x \in \mathcal{X}} P_X(x) \frac{\lambda P_X(x) + (1 - \lambda) P_\tilde{X}(x)}{P_{\hat{X}}(x)} \right) \\
= 0,
\]

where the inequality follows from the FI Lemma, with equality holding iff \( P_\tilde{X}(x) = P_{\hat{X}}(x) \) for all \( x \).

Finally, by the log-sum inequality, for each \( x \in \mathcal{X} \), we have

\[
(\lambda P_X(x) + (1 - \lambda) Q_X(x)) \log_2 \frac{\lambda P_X(x) + (1 - \lambda) Q_X(x)}{\lambda P_X(x) + (1 - \lambda) Q_\tilde{X}(x)} \\
\leq \lambda P_X(x) \log_2 \frac{\lambda P_X(x)}{\lambda P_X(x)} + (1 - \lambda) Q_X(x) \log_2 \frac{(1 - \lambda) Q_X(x)}{(1 - \lambda) Q_\tilde{X}(x)}.
\]

Summing over \( x \), we yield (2.7.1).

Note that the last result (convexity of \( D(P_X \| P_\tilde{X}) \) in the pair \( (P_X, P_\tilde{X}) \)) actually implies the first two results: just set \( P_\tilde{X} = Q_X \) to show convexity in the first argument \( P_X \), and set \( P_X = Q_X \) to show convexity in the second argument \( P_\tilde{X} \).

\[ \square \]

### 2.8 Fundamentals of hypothesis testing

One of the fundamental problems in statistics is to decide between two alternative explanations for the observed data. For example, when gambling, one may wish to test whether it is a fair game or not. Similarly, a sequence of observations on the market may reveal the information that whether a new product is successful
or not. This is the simplest form of the hypothesis testing problem, which is usually named \textit{simple hypothesis testing}.

It has quite a few applications in information theory. One of the frequently cited examples is the alternative interpretation of the law of large numbers. Another example is the computation of the true coding error (for universal codes) by testing the empirical distribution against the true distribution. All of these cases will be discussed subsequently.

The simple hypothesis testing problem can be formulated as follows:

\textbf{Problem:} Let $X_1, \ldots, X_n$ be a sequence of observations which is possibly drawn according to either a “null hypothesis” distribution $P_{X^n}$ or an “alternative hypothesis” distribution $P_{\hat{X}^n}$. The hypotheses are usually denoted by:

- $H_0 : P_{X^n}$
- $H_1 : P_{\hat{X}^n}$

Based on one sequence of observations $x^n$, one has to decide which of the hypotheses is true. This is denoted by a decision mapping $\phi(\cdot)$, where

$$
\phi(x^n) = \begin{cases} 
0, & \text{if distribution of } X^n \text{ is classified to be } P_{X^n}; \\
1, & \text{if distribution of } X^n \text{ is classified to be } P_{\hat{X}^n}.
\end{cases}
$$

Accordingly, the possible observed sequences are divided into two groups:

- Acceptance region for $H_0$ : $\{x^n \in \mathcal{X}^n : \phi(x^n) = 0\}$
- Acceptance region for $H_1$ : $\{x^n \in \mathcal{X}^n : \phi(x^n) = 1\}$.

Hence, depending on the true distribution, there are possibly two types of probability of errors:

- Type I error : $\alpha_n = \alpha_n(\phi) \triangleq P_{X^n}(\{x^n \in \mathcal{X}^n : \phi(x^n) = 1\})$
- Type II error : $\beta_n = \beta_n(\phi) \triangleq P_{\hat{X}^n}(\{x^n \in \mathcal{X}^n : \phi(x^n) = 0\})$.

The choice of the decision mapping is dependent on the optimization criterion. Two of the most frequently used ones in information theory are:

1. \textbf{Bayesian hypothesis testing.}

   Here, $\phi(\cdot)$ is chosen so that the Bayesian cost

   $$
   \pi_0 \alpha_n + \pi_1 \beta_n
   $$

   is minimized, where $\pi_0$ and $\pi_1$ are the prior probabilities for the null and alternative hypotheses, respectively. The mathematical expression for Bayesian testing is:

   $$
   \min_{\phi} \left[ \pi_0 \alpha_n(\phi) + \pi_1 \beta_n(\phi) \right].
   $$
2. Neyman Pearson hypothesis testing subject to a fixed test level.

Here, $\phi(\cdot)$ is chosen so that the type II error $\beta_n$ is minimized subject to a constant bound on the type I error; i.e.,

$$\alpha_n \leq \varepsilon$$

where $\varepsilon > 0$ is fixed. The mathematical expression for Neyman-Pearson testing is:

$$\min_{\{\phi : \alpha_n(\phi) \leq \varepsilon\}} \beta_n(\phi).$$

The set $\{\phi\}$ considered in the minimization operation could have two different ranges: range over deterministic rules, and range over randomization rules. The main difference between a randomization rule and a deterministic rule is that the former allows the mapping $\phi(x^n)$ to be random on $\{0, 1\}$ for some $x^n$, while the latter only accept deterministic assignments to $\{0, 1\}$ for all $x^n$. For example, a randomization rule for specific observations $\tilde{x}^n$ can be

$$\phi(\tilde{x}^n) = 0, \text{ with probability } 0.2;$$

$$\phi(\tilde{x}^n) = 1, \text{ with probability } 0.8.$$  

The Neyman-Pearson lemma shows the well-known fact that the likelihood ratio test is always the optimal test.

**Lemma 2.47 (Neyman-Pearson Lemma)** For a simple hypothesis testing problem, define an acceptance region for the null hypothesis through the likelihood ratio as

$$A_n(\tau) \triangleq \left\{ x^n \in X^n : \frac{P_{X^n}(x^n)}{P_{\hat{X}_n}(x^n)} > \tau \right\},$$

and let

$$\alpha^*_{n} \triangleq P_{X^n} \left\{ A^c_n(\tau) \right\} \quad \text{and} \quad \beta^*_{n} \triangleq P_{\hat{X}_n} \left\{ A_n(\tau) \right\}.$$  

Then for type I error $\alpha_n$ and type II error $\beta_n$ associated with another choice of acceptance region for the null hypothesis, we have

$$\alpha_n \leq \alpha^*_{n} \Rightarrow \beta_n \geq \beta^*_{n}.$$  

**Proof:** Let $B$ be a choice of acceptance region for the null hypothesis. Then

$$\alpha_n + \tau \beta_n = \sum_{x^n \in B^c} P_{X^n}(x^n) + \tau \sum_{x^n \in B} P_{\hat{X}_n}(x^n)$$

$$= \sum_{x^n \in B^c} P_{X^n}(x^n) + \tau \left[ 1 - \sum_{x^n \in B^c} P_{\hat{X}_n}(x^n) \right]$$

41
\[ = \tau + \sum_{x^n \in B} \left[ P_{X^n}(x^n) - \tau P_{\hat{X}^n}(x^n) \right]. \tag{2.8.1} \]

Observe that (2.8.1) is minimized by choosing \( B = \mathcal{A}_n(\tau) \). Hence,
\[ \alpha_n + \tau \beta_n \geq \alpha^*_n + \tau \beta^*_n, \]
which immediately implies the desired result. \( \square \)

The Neyman-Pearson lemma indicates that no other choices of acceptance regions can simultaneously improve both type I and type II errors of the likelihood ratio test. Indeed, from (2.8.1), it is clear that for any \( \alpha_n \) and \( \beta_n \), one can always find a likelihood ratio test that performs as good. Therefore, the likelihood ratio test is an optimal test. The statistical properties of the likelihood ratio thus become essential in hypothesis testing. Note that, when the observations are i.i.d. under both hypotheses, the divergence, which is the statistical expectation of the log-likelihood ratio, plays an important role in hypothesis testing (for non-memoryless observations, one is then concerned with the divergence rate, an extended notion of divergence for systems with memory which will be defined in a following chapter).

Problems

1. Prove the FI Lemma.

2. Show that the two conditions in Footnote 4 are equivalent.

3. For a finite-alphabet random variable \( X \), show that \( H(X) \leq \log_2 |\mathcal{X}| \) using the log-sum inequality.

4. Given a pair of random variables \((X, Y)\), is \( H(X|Y) = H(Y|X) \)? When do we have equality?

5. Given a random variable \( X \), what is the relationship between \( H(X) \) and \( H(Y) \) when \( Y \) is defined as follows.
   (a) \( Y = \log_2(X) \)?
   (b) \( Y = \sin(X) \)?

6. Show that the entropy of a real-valued function \( f \) of \( X \) is less than or equal to the entropy of \( X \)?
   \( \text{Hint:} \) By the chain rule for entropy, \[ H(X, f(X)) = H(X) + H(f(X)|X) = H(f(X)) + H(X|f(X)). \]
7. Show that $H(Y|X) = 0$ iff $Y$ is a function of $X$.

8. Give examples of:
   (a) $I(X;Y|Z) < I(X;Y)$.
   (b) $I(X;Y|Z) > I(X;Y)$.

   *Hint:* For (a), create example for $I(X;Y|Z) = 0$ and $I(X;Y) > 0$. For (b), create example for $I(X;Y) = 0$ and $I(X;Y|Z) > 0$.

9. Let the joint distribution of $X$ and $Y$ be:

   \[
   \begin{array}{c|cc}
   Y & 0 & 1 \\
   \hline
   0 & \frac{1}{4} & 0 \\
   1 & \frac{1}{2} & \frac{1}{4} \\
   \end{array}
   \]

   Draw the Venn diagram for $H(X)$, $H(Y)$, $H(X|Y)$, $H(Y|X)$, $H(X,Y)$ and $I(X;Y)$, and indicate the quantities (in bits) for each area of the Venn diagram.

10. *Inequalities:* Which of the following inequalities are $\geq$, $=$, $\leq$? Label each with $\geq$, $=$, or $\leq$ and justify your answer.

   (a) $H(X|Z) - H(X|Y)$ versus $I(X;Y|Z)$.
   (b) $H(X|g(Y))$ versus $H(X|Y)$ for a real function $g$.
   (c) $H(X_{1000}|X_{999})$ versus $\lim_{n \to \infty} (1/n)H(X_n, X_{n-1}, \cdots, X_1)$ for a stationary Markov source $\{X_n\}_{n=1}^{\infty}$.

11. Given random variables $X$ and $Y$ with finite alphabets, define a new random variable $Z$ as $Z = X + Y$.

   (a) Prove that $H(Z|X) = H(Y|X)$.
   (b) Show that $H(Z) \geq H(X)$ and $H(Z) \geq H(Y)$.
   (c) Give an example such that $H(X) > H(Z)$ and $H(Y) > H(Z)$.

   *Hint:* You may create an example with $H(Z) = 0$, but $H(X) > 0$ and $H(Y) > 0$.
   (d) Under what condition does $H(Z) = H(X) + H(Y)$?
12. Let \( X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \cdots \rightarrow X_n \) form a Markov chain. Show that:

(a) \( I(X_1; X_2, \ldots, X_n) = I(X_1; X_2) \).

(b) For any \( n \), \( H(X_1|X_{n-1}) \leq H(X_1|X_n) \).

13. Prove that refinement cannot decrease \( H(X) \). Namely, given a random variable \( X \) with distribution \( P_X \) and a partition \((U_1, \ldots, U_m)\) on its alphabet \( \mathcal{X} \) such that \( P_X(U_i) = P_X(U_i) \) for \( 1 \leq i \leq m \), show that \( H(X) \geq H(U) \).

14. Provide examples for the following inequalities.

(a) \( D(P_X|Z||P_X) > D(P_X^\hat{X}|Z) \).

(b) \( D(P_X|Z||P_X^\hat{X}) < D(P_X||P_X^\hat{X}) \).

15. Prove that the binary divergence defined by

\[
D(p||q) \triangleq p \log_2 \frac{p}{q} + (1 - p) \log_2 \frac{1 - p}{1 - q}
\]

satisfies

\[
D(p||q) \leq \log_2(e) \frac{(p - q)^2}{q(1 - q)}
\]

for \( 0 < p < 1 \) and \( 0 < q < 1 \).

Hint: Use the FI Lemma.

16. Fano’s inequality for list decoding: Let \( X \) and \( Y \) be two random variables with alphabets \( \mathcal{X} \) and \( \mathcal{Y} \), respectively, where \( \mathcal{X} \) is finite and \( \mathcal{Y} \) can be countably many. Given a fixed integer \( m \geq 1 \), define

\[
\hat{X}^m \triangleq (g_1(Y), g_2(Y), \ldots, g_m(Y))
\]

as the list of estimates of \( X \) obtained by observing \( Y \), where \( g_i: \mathcal{Y} \rightarrow \mathcal{X} \) is a given estimation function for \( i = 1, 2, \ldots, m \). Define the probability of list decoding error as

\[
P_e^{(m)} \triangleq \Pr \left[ \hat{X}_1 \neq X, \hat{X}_2 \neq X, \ldots, \hat{X}_m \neq X \right].
\]

Show that

\[
H(X|Y) \leq \log_2(\sum_{x \in \mathcal{X}} \sum_{\hat{x}^m \in \hat{X}^m : x = \hat{x} \text{ for some } i} P_{\hat{x}^m}(\hat{x}^m)) + (1 - P_e^{(m)}) \log_2 (|\mathcal{X}|-u) + (1 - P_e^{(m)}) \log_2 (u),
\]

where

\[
u \triangleq \sum_{x \in \mathcal{X}} \sum_{\hat{x}^m \in \hat{X}^m : x = \hat{x} \text{ for some } i} P_{\hat{x}^m}(\hat{x}^m).
\]
Note: When \( m = 1 \), we obtain that \( u = 1 \) and the right-hand side of (2.8.2) reduces to the original Fano inequality (cf. the right-hand side of (2.5.1)).

Hint: Show that \( H(X|Y) \leq H(X|\hat{X}^m) \) and that \( H(X|\hat{X}^m) \) is less than the right-hand-side of the above inequality.

17. Fano’s inequality for ternary partitioning of the observation space: In Problem 16, \( P_e^{(m)} \) and \( u \) can actually be expressed as

\[
P_e^{(m)} = \sum_{x \in \mathcal{X}} \sum_{\hat{x}^m \in \mathcal{U}_x} P_{X,\hat{X}^m}(x, \hat{x}^m) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}_x} P_{X,Y}(x, y)
\]

and

\[
u = \sum_{x \in \mathcal{X}} \sum_{\hat{x}^m \in \mathcal{U}_x} P_{\hat{X}^m}(\hat{x}^m) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}_x} P_Y(y),
\]

respectively, where

\[
\mathcal{U}_x \triangleq \{ \hat{x}^m \in \mathcal{X}^m : \hat{x}_i = x \text{ for some } i \}
\]

and

\[
\mathcal{Y}_x \triangleq \{ y \in \mathcal{Y} : g_i(y) = x \text{ for some } i \}.
\]

Thus, given \( x \in \mathcal{X}, \mathcal{Y}_x \) and \( \mathcal{Y}_x^c \) form a binary partition on the observation space \( \mathcal{Y} \).

Now consider again random variables \( X \) and \( Y \) with alphabets \( \mathcal{X} \) and \( \mathcal{Y} \), respectively, where \( \mathcal{X} \) is finite and \( \mathcal{Y} \) can be countably many, and assume that for each \( x \in \mathcal{X} \), we are given a ternary partition \( \{ \mathcal{S}_x, \mathcal{T}_x, \mathcal{V}_x \} \) on the observation space \( \mathcal{Y} \), where the sets \( \mathcal{S}_x, \mathcal{T}_x \) and \( \mathcal{V}_x \) are mutually disjoint and their union equals \( \mathcal{Y} \).

Define

\[
p \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{S}_x} P_{X,Y}(x, y), \quad q \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{T}_x} P_{X,Y}(x, y), \quad r \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{V}_x} P_{X,Y}(x, y)
\]

and

\[
s \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{S}_x} P_Y(y), \quad t \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{T}_x} P_Y(y), \quad v \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{V}_x} P_Y(y).
\]

Note that \( p + q + r = 1 \) and \( s + t + v = |\mathcal{X}| \). Show that

\[
H(X|Y) \leq H(p, q, r) + p \log_2(s) + q \log_2(t) + r \log_2(v), \tag{2.8.3}
\]

where

\[
H(p, q, r) = p \log_2 \frac{1}{p} + q \log_2 \frac{1}{q} + r \log_2 \frac{1}{r}.
\]

Note: When \( \mathcal{Y}_x = \emptyset \) for all \( x \in \mathcal{X} \), we obtain that \( \mathcal{S}_x = \mathcal{T}_x^c \) for all \( x \), \( r = v = 0 \) and \( p = 1 - q \); as a result, inequality (2.8.3) acquires a similar expression as (2.8.2) with \( p \) standing for the probability of error.

45
18. $\epsilon$-Independence: Let $X$ and $Y$ be two jointly distributed random variables with finite respective alphabets $\mathcal{X}$ and $\mathcal{Y}$ and joint pmf $P_{X,Y}$ defined on the $\mathcal{X} \times \mathcal{Y}$. Given a fixed $\epsilon > 0$, random variable $Y$ is said to be $\epsilon$-independent from random variable $X$ if

$$\sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} |P_{Y|X}(y|x) - P_Y(y)| < \epsilon$$

where $P_X$ and $P_Y$ are the marginal pmf’s of $X$ and $Y$, respectively, and $P_{Y|X}$ is the conditional pmf of $Y$ given $X$. Show that

$$I(X;Y) < \frac{\log_2(e)}{2} \epsilon^2$$

is a sufficient condition for $Y$ to be $\epsilon$-independent from $X$, where $I(X;Y)$ is the mutual information (in bits) between $X$ and $Y$. 

46
Chapter 3

Lossless Data Compression

3.1 Principles of data compression

As mentioned in Chapter 1, data compression describes methods of representing a source by a code whose average codeword length (or code rate) is acceptably small. The representation can be: lossless (or asymptotically lossless) where the reconstructed source is identical (or asymptotically identical) to the original source; or lossy where the reconstructed source is allowed to deviate from the original source, usually within an acceptable threshold. We herein focus on lossless data compression.

Since a memoryless source is modelled as a random variable, the averaged codeword length of a codebook is calculated based on the probability distribution of that random variable. For example, a ternary memoryless source $X$ exhibits three possible outcomes with

$$
P_X(x = \text{outcome}_A) = 0.5;
$$
$$
P_X(x = \text{outcome}_B) = 0.25;
$$
$$
P_X(x = \text{outcome}_C) = 0.25.
$$

Suppose that a binary code book is designed for this source, in which outcome$_A$, outcome$_B$ and outcome$_C$ are respectively encoded as 0, 10, and 11. Then the average codeword length (in bits/source outcome) is

$$
\text{length}(0) \times P_X(\text{outcome}_A) + \text{length}(10) \times P_X(\text{outcome}_B) \\
+\text{length}(11) \times P_X(\text{outcome}_C)
$$

$$
= 1 \times 0.5 + 2 \times 0.25 + 2 \times 0.25
$$

$$
= 1.5 \text{ bits}.
$$

There are usually no constraints on the basic structure of a code. In the case where the codeword length for each source outcome can be different, the
code is called a variable-length code. When the codeword lengths of all source outcomes are equal, the code is referred to as a fixed-length code. It is obvious that the minimum average codeword length among all variable-length codes is no greater than that among all fixed-length codes, since the latter is a subclass of the former. We will see in this chapter that the smallest achievable average code rate for variable-length and fixed-length codes coincide for sources with good probabilistic characteristics, such as stationarity and ergodicity. But for more general sources with memory, the two quantities are different (cf. Part II of the book).

For fixed-length codes, the sequence of adjacent codewords are concatenated together for storage or transmission purposes, and some punctuation mechanism—such as marking the beginning of each codeword or delineating internal sub-blocks for synchronization between encoder and decoder—is normally considered an implicit part of the codewords. Due to constraints on space or processing capability, the sequence of source symbols may be too long for the encoder to deal with all at once; therefore, segmentation before encoding is often necessary. For example, suppose that we need to encode using a binary code the grades of a class with 100 students. There are three grade levels: A, B and C. By observing that there are $3^{100}$ possible grade combinations for 100 students, a straightforward code design requires $\lceil \log_2(3^{100}) \rceil = 159$ bits to encode these combinations. Now suppose that the encoder facility can only process 16 bits at a time. Then the above code design becomes infeasible and segmentation is unavoidable. Under such constraint, we may encode grades of 10 students at a time, which requires $\lceil \log_2(3^{10}) \rceil = 16$ bits. As a consequence, for a class of 100 students, the code requires 160 bits in total.

In the above example, the letters in the grade set \{A, B, C\} and the letters from the code alphabet \{0, 1\} are often called source symbols and code symbols, respectively. When the code alphabet is binary (as in the previous two examples), the code symbols are referred to as code bits or simply bits (as already used). A tuple (or grouped sequence) of source symbols is called a sourceword and the resulting encoded tuple consisting of code symbols is called a codeword. (In the above example, each sourceword consists of 10 source symbols (students) and each codeword consists of 16 bits.)

Note that, during the encoding process, the sourceword lengths do not have to be equal. In this text, we however only consider the case where the sourcewords have a fixed length throughout the encoding process (except for the Lempel-Ziv code briefly discussed at the end of this chapter), but we will allow the codewords to have fixed or variable lengths as defined earlier. The block diagram of a source

\[\text{In other words, our fixed-length codes are actually “fixed-to-fixed length codes” and our variable-length codes are “fixed-to-variable length codes” since, in both cases, a fixed number}\]
coding system is depicted in Figure 3.1.

When adding segmentation mechanisms to fixed-length codes, the codes can be loosely divided into two groups. The first consists of block codes in which the encoding (or decoding) of the next segment of source symbols is independent of the previous segments. If the encoding/decoding of the next segment, somehow, retains and uses some knowledge of earlier segments, the code is called a fixed-length tree code. As will not investigate such codes in this text, we can use “block codes” and “fixed-length codes” as synonyms.

In this chapter, we first consider data compression for block codes in Section 3.2. Data compression for variable-length codes is then addressed in Section 3.3.

3.2 Block codes for asymptotically lossless compression

3.2.1 Block codes for discrete memoryless sources

We first focus on the study of asymptotically lossless data compression of discrete memoryless sources via block (fixed-length) codes. Such sources were already defined in Appendix B and the previous chapter; but we nevertheless recall their definition.

**Definition 3.1 (Discrete memoryless source)** A discrete memoryless source (DMS) \( \{X_n\}_{n=1}^{\infty} \) consists of a sequence of independent and identically distributed (i.i.d.) random variables, \( X_1, X_2, X_3, \ldots \), all taking values in a common finite alphabet \( \mathcal{X} \). In particular, if \( P_X(\cdot) \) is the common distribution or probability mass function (pmf) of the \( X_i \)'s, then

\[
P_{X^n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P_X(x_i).
\]

of source symbols is mapped onto codewords with fixed and variable lengths, respectively.
Definition 3.2 An \((n, M)\) block code of blocklength \(n\) and size \(M\) (which can be a function of \(n\) in general,\(^2\) i.e., \(M = M_n\)) for a discrete source \(\{X_n\}_{n=1}^\infty\) is a set \(\{c_1, c_2, \ldots, c_M\} \subseteq X^n\) consisting of \(M\) reproduction (or reconstruction) words, where each reproduction word is a sourceword (an \(n\)-tuple of source symbols).\(^3\)

The block code’s operation can be symbolically represented as

\[
(x_1, x_2, \ldots, x_n) \rightarrow c_m \in \{c_1, c_2, \ldots, c_M\}.
\]

This procedure will be repeated for each consecutive block of length \(n\), i.e.,

\[
\cdots (x_{3n}, \ldots, x_{31})(x_{2n}, \ldots, x_{21})(x_{1n}, \ldots, x_{11}) \rightarrow \cdots |c_{m_3}|c_{m_2}|c_{m_1},
\]

where "\(|\)" reflects the necessity of “punctuation mechanism” or ”synchronization mechanism” for consecutive source block coders.

The next theorem provides a key tool for proving Shannon’s source coding theorem.

\(^2\)In the literature, both \((n, M)\) and \((M, n)\) have been used to denote a block code with blocklength \(n\) and size \(M\). For example, [53, p. 149] adopts the former one, while [13, p. 193] uses the latter. We use the \((n, M)\) notation since \(M = M_n\) is a function of \(n\) in general.

\(^3\)One can binary-index the reproduction words in \(\{c_1, c_2, \ldots, c_M\}\) using \(k \equiv \lceil \log_2 M \rceil\) bits. As such \(k\)-bit words in \(\{0, 1\}^k\) are usually stored for retrieval at a later date, the \((n, M)\) block code can be represented by an encoder-decoder pair of functions \((f, g)\), where the encoding function \(f : X^n \rightarrow \{0, 1\}^k\) maps each sourceword \(x^n\) to a \(k\)-bit word \(f(x^n)\) which we call a codeword. Then the decoding function \(g : \{0, 1\}^k \rightarrow \{c_1, c_2, \ldots, c_M\}\) is a retrieving operation that produces the reproduction words. Since the codewords are binary-valued, such a block code is called a binary code. More generally, a \(D\)-ary block code (where \(D > 1\) is an integer) would use an encoding function \(f : X^n \rightarrow \{0, 1, \ldots, D - 1\}^k\) where each codeword \(f(x^n)\) contains \(k\) \(D\)-ary code symbols.

Furthermore, since the behavior of block codes is investigated for sufficiently large \(n\) and \(M\) (tending to infinity), it is legitimate to replace \(\lceil \log_2 M \rceil\) by \(\log_2 M\) for the case of binary codes. With this convention, the data compression rate or code rate is

\[
\text{bits required per source symbol} = \frac{k}{n} = \frac{1}{n} \log_2 M.
\]

Similarly, for \(D\)-ary codes, the rate is

\[
\text{\(D\)-ary code symbols required per source symbol} = \frac{k}{n} = \frac{1}{n} \log_D M.
\]

For computational convenience, \textit{nats} (under the natural logarithm) can be used instead of \textit{bits} or \(D\)-ary code symbols; in this case, the code rate becomes:

\[
\text{nats required per source symbol} = \frac{1}{n} \log M.
\]

\(^4\)When one uses an encoder-decoder pair \((f, g)\) to describe the block code, the code’s operation can be expressed as: \(c_m = g(f(x^n))\).
Theorem 3.3 (Shannon-McMillan) (Asymptotic equipartition property or AEP\(^5\)) If \(\{X_n\}_{n=1}^{\infty}\) is a DMS with entropy \(H(X)\), then
\[
-\frac{1}{n} \log_2 P_{X^n}(X_1, \ldots, X_n) \to H(X) \text{ in probability.}
\]
In other words, for any \(\delta > 0\),
\[
\lim_{n \to \infty} \Pr \left\{ \left| -\frac{1}{n} \log_2 P_{X^n}(X_1, \ldots, X_n) - H(X) \right| > \delta \right\} = 0.
\]
Proof: This theorem follows by first observing that for an i.i.d. sequence \(\{X_n\}_{n=1}^{\infty}\),
\[
-\frac{1}{n} \log_2 P_{X^n}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} [-\log_2 P_X(X_i)]
\]
and that the sequence \(\{-\log_2 P_X(X_i)\}_{i=1}^{\infty}\) is i.i.d., and then applying the weak law of large numbers (WLLN) on the later sequence.

The AEP indeed constitutes an “information theoretic” analog of WLLN as it states that if \(\{-\log_2 P_X(X_i)\}_{i=1}^{\infty}\) is an i.i.d. sequence, then for any \(\delta > 0\),
\[
\Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} [-\log_2 P_X(X_i)] - H(X) \right| \leq \delta \right\} \to 1 \text{ as } n \to \infty.
\]
As a consequence of the AEP, all the probability mass will be ultimately placed on the weakly \(\delta\)-typical set, which is defined as
\[
\mathcal{F}_n(\delta) \triangleq \left\{ x^n \in \mathcal{X}^n : \left| -\frac{1}{n} \log_2 P_{X^n}(x^n) - H(X) \right| \leq \delta \right\} = \left\{ x^n \in \mathcal{X}^n : \left| -\frac{1}{n} \sum_{i=1}^{n} \log_2 P_X(x_i) - H(X) \right| \leq \delta \right\}.
\]
Note that since the source is memoryless, for any \(x^n \in \mathcal{F}_n(\delta)\), \(-1/n \log_2 P_{X^n}(x^n)\), the normalized self-information of \(x^n\), is equal to \((1/n) \sum_{i=1}^{n} [-\log_2 P_X(x_i)]\), which is the empirical (arithmetic) average self-information or “apparent” entropy of the source. Thus, a source-word \(x^n\) is \(\delta\)-typical if it yields an apparent source entropy within \(\delta\) of the “true” source entropy \(H(X)\). Note that the source-words in \(\mathcal{F}_n(\delta)\) are nearly equiprobable or equally surprising (cf. Property 1 of Theorem 3.4); this justifies naming Theorem 3.3 by AEP.

Theorem 3.4 (Consequence of the AEP) Given a DMS \(\{X_n\}_{n=1}^{\infty}\) with entropy \(H(X)\) and any \(\delta\) greater than zero, then the weakly \(\delta\)-typical set \(\mathcal{F}_n(\delta)\) satisfies the following.

\(^5\)This is also called the entropy stability property.
1. If \( x^n \in \mathcal{F}_n(\delta) \), then
\[
2^{-n(H(X) + \delta)} \leq P_{X^n}(x^n) \leq 2^{-n(H(X) - \delta)}.
\]

2. \( P_{X^n}(\mathcal{F}_n^c(\delta)) < \delta \) for sufficiently large \( n \), where the superscript “\( c \)” denotes the complementary set operation.

3. \(|\mathcal{F}_n(\delta)| > (1 - \delta)2^{n(H(X) - \delta)}\) for sufficiently large \( n \), and \(|\mathcal{F}_n(\delta)| \leq 2^{n(H(X) + \delta)}\) for every \( n \), where \(|\mathcal{F}_n(\delta)|\) denotes the number of elements in \( \mathcal{F}_n(\delta) \).

**Note:** The above theorem also holds if we define the typical set using the base-\( D \) logarithm \( \log_D \) for any \( D > 1 \) instead of the base-2 logarithm; in this case, one just needs to appropriately change the base of the exponential terms in the above theorem (by replacing \( 2^x \) terms with \( D^x \) terms) and also substitute \( H(X) \) with \( H_D(X) \).

**Proof:** Property 1 is an immediate consequence of the definition of \( \mathcal{F}_n(\delta) \).

Property 2 is a direct consequence of the AEP, since the AEP states that for a fixed \( \delta > 0 \), \( \lim_{n \to \infty} P_{X^n}(\mathcal{F}_n(\delta)) = 1 \); i.e., \( \forall \varepsilon > 0 \), there exists \( n_0 = n_0(\varepsilon) \) such that for all \( n \geq n_0 \),
\[
P_{X^n}(\mathcal{F}_n(\delta)) > 1 - \varepsilon.
\]
In particular, setting \( \varepsilon = \delta \) yields the result. We nevertheless provide a direct proof of Property 2 as we give an explicit expression for \( n_0 \): observe that by Chebyshev’s inequality, \(^6\)
\[
P_{X^n}(\mathcal{F}_n^c(\delta)) = P_{X^n}\left\{ x^n \in \mathcal{X}^n : \left| -\frac{1}{n} \log_2 P_{X^n}(x^n) - H(X) \right| > \delta \right\}
\leq \frac{\sigma_X^2}{n \delta^2} < \delta,
\]
for \( n > \sigma_X^2/\delta^3 \), where the variance
\[
\sigma_X^2 \triangleq \text{Var}[- \log_2 P_X(X)] = \sum_{x \in \mathcal{X}} P_X(x) [\log_2 P_X(x)]^2 - (H(X))^2
\]
is a constant\(^7\) independent of \( n \).

\(^6\)The detail for Chebyshev’s inequality as well as its proof can be found on p. 217 in Appendix B.

\(^7\)In the proof, we assume that the variance \( \sigma_X^2 = \text{Var}[\log_2 P_X(X)] < \infty \). This holds since the source alphabet is finite:
\[
\text{Var}[\log_2 P_X(X)] \leq E[(\log_2 P_X(X))^2] = \sum_{x \in \mathcal{X}} P_X(x)(\log_2 P_X(x))^2
\leq \sum_{x \in \mathcal{X}} \frac{4}{e^2} \left( \log_2(e) \right)^2 = \frac{4}{e^2} \left( \log_2(e) \right)^2 \times |\mathcal{X}| < \infty.
\]
To prove Property 3, we have from Property 1 that
\[ 1 \geq \sum_{x^n \in \mathcal{F}_n(\delta)} P_{X^n}(x^n) \geq \sum_{x^n \in \mathcal{F}_n(\delta)} 2^{-n(H(X)+\delta)} = |\mathcal{F}_n(\delta)|2^{-n(H(X)+\delta)}, \]
and, using Properties 2 and 1, we have that
\[ 1 - \delta < 1 - \frac{\sigma_X^2}{n\delta^2} \leq \sum_{x^n \in \mathcal{F}_n(\delta)} P_{X^n}(x^n) \leq \sum_{x^n \in \mathcal{F}_n(\delta)} 2^{-n(H(X)-\delta)} = |\mathcal{F}_n(\delta)|2^{-n(H(X)-\delta)}, \]
for \( n \geq \sigma_X^2/\delta^3 \).

Note that for any \( n > 0 \), a block code \( \mathcal{C}_n = (n, M) \) is said to be uniquely decodable or completely lossless if its set of reproduction words is trivially equal to the set of all source \( n \)-tuples: \( \{c_1, c_2, \ldots, c_M\} = \mathcal{X}^n \). In this case, if we are binary-indexing the reproduction words using an encoding-decoding pair \((f, g)\), every sourceword \( x^n \) will be assigned to a distinct binary codeword \( f(x^n) \) of length \( k = \log_2 M \) and all the binary \( k \)-tuples are the image under \( f \) of some sourceword. In other words, \( f \) is a bijective (injective and surjective) map and hence invertible with the decoding map \( g = f^{-1} \) and \( M = |\mathcal{X}|^n = 2^k \). Thus the code rate is \( (1/n) \log_2 M = \log_2 |\mathcal{X}| \) bits/source symbol.

Now the question becomes: can we achieve a better (i.e., smaller) compression rate? The answer is affirmative: we can achieve a compression rate equal to the source entropy \( H(X) \) (in bits), which can be significantly smaller than \( \log_2 M \) when this source is strongly non-uniformly distributed, if we give up unique decodability (for every \( n \)) and allow \( n \) to be sufficiently large to asymptotically achieve lossless reconstruction by having an arbitrarily small (but positive) probability of decoding error \( P_e(\mathcal{C}_n) \triangleq P_{X^n}\{x^n \in \mathcal{X}^n : g(f(x^n)) \neq x^n\} \).

Thus, block codes herein can perform data compression that is asymptotically lossless with respect to blocklength; this contrasts with variable-length codes which can be completely lossless (uniquely decodable) for every finite blocklength.

We now can formally state and prove Shannon’s asymptotically lossless source coding theorem for block codes. The theorem will be stated for general \( D \)-ary block codes, representing the source entropy \( H_D(X) \) in \( D \)-ary code symbol/source symbol as the smallest (infimum) possible compression rate for asymptotically lossless \( D \)-ary block codes. Without loss of generality, the theorem will be proved for the case of \( D = 2 \). The idea behind the proof of the forward (achievability) part is basically to binary-index the source sequence in the weakly \( \delta \)-typical set \( \mathcal{F}_n(\delta) \) to a binary codeword (starting from index one with corresponding \( k \)-tuple codeword 0 \( \cdots \) 01); and to encode all sourcewords outside \( \mathcal{F}_n(\delta) \) to a default all-zero binary codeword, which certainly cannot be reproduced distortionless due to its many-to-one-mapping property. The resultant
code rate is \( \left( \frac{1}{n} \right) \left\lceil \log_2(|\mathcal{F}_n(\delta)| + 1) \right\rceil \) bits per source symbol. As revealed in the Shannon-McMillan AEP theorem and its Consequence, almost all the probability mass will be on \( \mathcal{F}_n(\delta) \) as \( n \) sufficiently large, and hence, the probability of non-reconstructable source sequences can be made arbitrarily small. A simple example for the above coding scheme is illustrated in Table 3.1. The converse part of the proof will establish (by expressing the probability of correct decoding in terms of the \( \delta \)-typical set and also using the Consequence of the AEP) that for any sequence of \( D \)-ary codes with rate strictly below the source entropy, their probability of error cannot asymptotically vanish (is bounded away from zero). Actually a stronger result is proven: it is shown that their probability of error not only does not asymptotically vanish, it actually ultimately grows to 1 (this is why we call this part a “strong” converse).

Table 3.1: An example of the \( \delta \)-typical set with \( n = 2 \) and \( \delta = 0.4 \), where \( \mathcal{F}_2(0.4) = \{ AB, AC, BA, BB, BC, CA, CB \} \). The codeword set is \{ 001(AB), 010(AC), 011(BA), 100(BB), 101(BC), 110(CA), 111(CB), 000(AA, AD, BD, CC, CD, DA, DB, DC, DD) \}, where the parenthesis following each binary codeword indicates those source-words that are encoded to this codeword. The source distribution is \( P_X(A) = 0.4 \), \( P_X(B) = 0.3 \), \( P_X(C) = 0.2 \) and \( P_X(D) = 0.1 \).
Theorem 3.5 (Shannon’s source coding theorem)  Given integer \( D > 1 \), consider a discrete memoryless source \( \{X_n\}_{n=1}^\infty \) with entropy \( H_D(X) \). Then the following hold.

- **Forward part (achievability):** For any \( 0 < \varepsilon < 1 \), there exists \( 0 < \delta < \varepsilon \) and a sequence of \( D \)-ary block codes \( \{\mathcal{C}_n = (n, M_n)\}_{n=1}^\infty \) with
  \[
  \limsup_{n \to \infty} \frac{1}{n} \log_D M_n \leq H_D(X) + \delta
  \]  
  satisfying
  \[
  P_e(\mathcal{C}_n) < \varepsilon
  \]  
  for all sufficiently large \( n \), where \( P_e(\mathcal{C}_n) \) denotes the probability of decoding error for block code \( \mathcal{C}_n \).

- **Strong converse part:** For any \( 0 < \varepsilon < 1 \), any sequence of \( D \)-ary block codes \( \{\mathcal{C}_n = (n, M_n)\}_{n=1}^\infty \) with
  \[
  \limsup_{n \to \infty} \frac{1}{n} \log_D M_n < H_D(X)
  \]  
  satisfies
  \[
  P_e(\mathcal{C}_n) > 1 - \varepsilon
  \]  
  for all \( n \) sufficiently large.

**Proof:**

*Forward Part:* Without loss of generality, we will prove the result for the case of binary codes (i.e., \( D = 2 \)). Also recall that subscript \( D \) in \( H_D(X) \) will be dropped (i.e., omitted) specifically when \( D = 2 \).

Given \( 0 < \varepsilon < 1 \), fix \( \delta \) such that \( 0 < \delta < \varepsilon \) and choose \( n > 2/\delta \). Now construct a binary \( \mathcal{C}_n \) block code by simply mapping the \( \delta/2 \)-typical source words \( x^n \) onto distinct not all-zero binary codewords of length \( k \triangleq \lceil \log_2 M_n \rceil \) bits. In other words, binary-index (cf. the footnote in Definition 3.2) the source words in \( F_n(\delta/2) \) with the following encoding map:

\[
\begin{cases}
  x^n \to \text{binary index of } x^n, & \text{if } x^n \in F_n(\delta/2); \\
  x^n \to \text{all-zero codeword}, & \text{if } x^n \not\in F_n(\delta/2).
\end{cases}
\]  

\( \text{(3.2.2)} \) is equivalent to \( \limsup_{n \to \infty} P_e(\mathcal{C}_n) \leq \varepsilon \). Since \( \varepsilon \) can be made arbitrarily small, the forward part actually indicates the existence of a sequence of \( D \)-ary block codes \( \{\mathcal{C}_n\}_{n=1}^\infty \) satisfying (3.2.1) such that \( \limsup_{n \to \infty} P_e(\mathcal{C}_n) = 0 \).

Based on this, the converse should be that any sequence of \( D \)-ary block codes satisfying (3.2.3) satisfies \( \limsup_{n \to \infty} P_e(\mathcal{C}_n) > 0 \). However, the so-called strong converse actually gives a stronger consequence: \( \limsup_{n \to \infty} P_e(\mathcal{C}_n) = 1 \) (as \( \varepsilon \) can be made arbitrarily small).
Then by the Shannon-McMillan AEP theorem, we obtain that

\[ M_n = |\mathcal{F}_n(\delta/2)| + 1 \leq 2^{n(H(X)+\delta/2)} + 1 < 2 \cdot 2^{n(H(X)+\delta/2)} < 2^{n(H(X)+\delta)}, \]

for \( n > 2/\delta \). Hence, a sequence of \( \mathcal{C}_n = (n, M_n) \) block code satisfying (3.2.1) is established. It remains to show that the error probability for this sequence of \( (n, M_n) \) block code can be made smaller than \( \varepsilon \) for all sufficiently large \( n \).

By the Shannon-McMillan AEP theorem,

\[ P_{X^n}(\mathcal{F}_n^c(\delta/2)) < \frac{\delta}{2} \quad \text{for all sufficiently large } n. \]

Consequently, for those \( n \) satisfying the above inequality, and being bigger than \( 2/\delta \),

\[ P_e(\mathcal{C}_n) \leq P_{X^n}(\mathcal{F}_n^c(\delta/2)) < \delta \leq \varepsilon. \]

(For the last step, the readers can refer to Table 3.1 to confirm that only the “ambiguous” sequences outside the typical set contribute to the probability of error.)

**Strong Converse Part:** Fix any sequence of block codes \( \{\mathcal{C}_n\}_{n=1}^\infty \) with

\[ \limsup_{n \to \infty} \frac{1}{n} \log_2 |\mathcal{C}_n| < H(X). \]

Let \( \mathcal{S}_n \) be the set of source symbols that can be correctly decoded through \( \mathcal{C}_n \)-coding system. (A quick example is depicted in Figure 3.2.) Then \( |\mathcal{S}_n| = |\mathcal{C}_n| \).

By choosing \( \delta \) small enough with \( \varepsilon/2 > \delta > 0 \), and also by definition of limsup operation, we have

\[ (\exists N_0)(\forall n > N_0) \quad \frac{1}{n} \log_2 |\mathcal{S}_n| = \frac{1}{n} \log_2 |\mathcal{C}_n| < H(X) - 2\delta, \]

which implies

\[ |\mathcal{S}_n| < 2^{n(H(X)-2\delta)}. \]

Furthermore, from Property 2 of the Consequence of the AEP, we obtain that

\[ (\exists N_1)(\forall n > N_1) \quad P_{X^n}(\mathcal{F}_n^c(\delta)) < \delta. \]

Consequently, for \( n > N \triangleq \max\{N_0, N_1, \log_2(2/\varepsilon)/\delta\} \), the probability of correctly block decoding satisfies

\[ 1 - P_e(\mathcal{C}_n) = \sum_{x^n \in \mathcal{S}_n} P_{X^n}(x^n) = \sum_{x^n \in \mathcal{S}_n \cap \mathcal{F}_n^c} P_{X^n}(x^n) + \sum_{x^n \in \mathcal{S}_n \cap \mathcal{F}_n} P_{X^n}(x^n) \]

56
Source Symbols
\[ S_n \]

Codewords

Figure 3.2: Possible codebook \( \mathcal{C}_n \) and its corresponding \( S_n \). The solid box indicates the decoding mapping from \( \mathcal{C}_n \) back to \( S_n \).

\[
\begin{align*}
\Pr[X^n(F_n^c(\delta))] & \leq \Pr[X^n(F_n^c(\delta))] + |S_n \cap F_n(\delta)| \cdot \max_{x^n \in F_n(\delta)} \Pr[X^n(x^n)] \\
& < \delta + |S_n| \cdot \max_{x^n \in F_n(\delta)} \Pr[X^n(x^n)] \\
& < \frac{\varepsilon}{2} + 2^{n(H(X)-2\delta)} \cdot 2^{-n(H(X)-\delta)} \\
& < \frac{\varepsilon}{2} + 2^{-n\delta} \\
& < \varepsilon,
\end{align*}
\]

which is equivalent to \( P_e(\mathcal{C}_n) > 1 - \varepsilon \) for \( n > N \).

**Observation 3.6** The results of the above theorem is illustrated in Figure 3.3, where \( R = \limsup_{n \to \infty} (1/n) \log D M_n \) is usually called the ultimate (or asymptotic) code rate of block codes for compressing the source. It is clear from the figure that the (ultimate) rate of any block code with arbitrarily small decoding error probability must be greater than or equal to the source entropy.\(^9\) Conversely, the probability of decoding error for any block code of rate smaller than entropy ultimately approaches 1 (and hence is bounded away from zero). Thus for a DMS, the source entropy \( H_D(X) \) is the infimum of all “achievable” source (block) coding rates; i.e., it is the infimum of all rates for which there exists a sequence of \( D \)-ary block codes with asymptotically vanishing (as the blocklength goes to infinity) probability of decoding error.

For a source with (statistical) memory, Shannon-McMillan’s theorem cannot be directly applied in its original form, and thereby Shannon’s source coding

\(^9\)Note that it is clear from the statement and proof of the forward part of Theorem 3.5 that the source entropy can be achieved as an ultimate compression rate as long as \( (1/n) \log D M_n \) approaches it from above with increasing \( n \).
Figure 3.3: (Ultimate) Compression rate $R$ versus source entropy $H_D(X)$ and behavior of the probability of block decoding error as block length $n$ goes to infinity for a discrete memoryless source.

Theorem appears restricted to only memoryless sources. However, by exploring the concept behind these theorems, one can find that the key for the validity of Shannon’s source coding theorem is actually the existence of a set $A_n = \{x_1^n, x_2^n, \ldots, x_M^n\}$ with $M \approx D^n H_D(X)$ and $P_{X^n}(A_n^c) \to 0$, namely, the existence of a “typical-like” set $A_n$ whose size is prohibitively small and whose probability mass is asymptotically large. Thus, if we can find such typical-like set for a source with memory, the source coding theorem for block codes can be extended for this source. Indeed, with appropriate modifications, the Shannon-McMillan theorem can be generalized for the class of stationary ergodic sources and hence a block source coding theorem for this class can be established; this is considered in the next subsection. The block source coding theorem for general (e.g., non-stationary non-ergodic) sources in terms of a “generalized entropy” measure (see the end of the next subsection for a brief description) will be studied in detail in Part II of the book.

3.2.2 Block codes for stationary ergodic sources

In practice, a stochastic source used to model data often exhibits memory or statistical dependence among its random variables; its joint distribution is hence not a product of its marginal distributions. In this subsection, we consider the asymptotic lossless data compression theorem for the class of stationary ergodic sources.\textsuperscript{10}

Before proceeding to generalize the block source coding theorem, we need to first generalize the “entropy” measure for a sequence of dependent random variables $X_n$ (which certainly should be backward compatible to the discrete memoryless cases). A straightforward generalization is to examine the limit of the normalized block entropy of a source sequence, resulting in the concept of entropy rate.

Definition 3.7 (Entropy rate) The entropy rate for a source $\{X_n\}_{n=1}^\infty$ is de-

\textsuperscript{10}The definitions of stationarity and ergodicity can be found on p. 210 in Appendix B.
noted by $H(\mathcal{X})$ and defined by

$$H(\mathcal{X}) \triangleq \lim_{n \to \infty} \frac{1}{n} H(X^n)$$

provided the limit exists, where $X^n = (X_1, \cdots, X_n)$.

Next we will show that the entropy rate exists for stationary sources (here, we do not need ergodicity for the existence of entropy rate).

**Lemma 3.8** For a stationary source $\{X_n\}_{n=1}^\infty$, the conditional entropy

$$H(X_n|X_{n-1}, \ldots, X_1)$$

is non-increasing in $n$ and also bounded from below by zero. Hence by Lemma A.20, the limit

$$\lim_{n \to \infty} H(X_n|X_{n-1}, \ldots, X_1)$$

exists.

**Proof:** We have

$$H(X_n|X_{n-1}, \ldots, X_1) \leq H(X_n|X_{n-1}, \ldots, X_2)$$  \hspace{1cm} (3.2.4)

$$= H(X_n, \ldots, X_2) - H(X_{n-1}, \ldots, X_2)$$

$$= H(X_{n-1}, \ldots, X_1) - H(X_{n-2}, \ldots, X_1)$$  \hspace{1cm} (3.2.5)

where (3.2.4) follows since conditioning never increases entropy, and (3.2.5) holds because of the stationarity assumption. Finally, recall that each conditional entropy $H(X_n|X_{n-1}, \ldots, X_1)$ is non-negative. \hfill \Box

**Lemma 3.9 (Cesaro-mean theorem)** If $a_n \to a$ as $n \to \infty$ and $b_n = (1/n) \sum_{i=1}^n a_i$, then $b_n \to a$ as $n \to \infty$.

**Proof:** $a_n \to a$ implies that for any $\varepsilon > 0$, there exists $N$ such that for all $n > N$, $|a_n - a| < \varepsilon$. Then

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

$$\leq \frac{1}{n} \sum_{i=1}^n |a_i - a|$$
\[\frac{1}{n} \sum_{i=1}^{N} |a_i - a| + \frac{1}{n} \sum_{i=N+1}^{n} |a_i - a| \]

\[\leq \frac{1}{n} \sum_{i=1}^{N} |a_i - a| + \frac{n - N}{n} \varepsilon.\]

Hence, \(\lim_{n \to \infty} |b_n - a| \leq \varepsilon.\) Since \(\varepsilon\) can be made arbitrarily small, the lemma holds. \(\Box\)

**Theorem 3.10** For a stationary source \(\{X_n\}_{n=1}^{\infty}\), its entropy rate always exists and is equal to

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

**Proof:** The result directly follows by writing

\[\frac{1}{n} H(X^n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \text{ (chain-rule for entropy)}\]

and applying the Cesaro-mean theorem. \(\Box\)

**Observation 3.11** It can also be shown that for a stationary source, \((1/n)H(X^n)\) is non-increasing in \(n\) and \((1/n)H(X^n) \geq H(X_n|X_{n-1}, \ldots, X_1)\) for all \(n \geq 1\). (The proof is left as an exercise. See Problem 3.)

It is obvious that when \(\{X_n\}_{n=1}^{\infty}\) is a discrete memoryless source, \(H(X^n) = n \times H(X)\) for every \(n\). Hence,

\[H(X) = \lim_{n \to \infty} \frac{1}{n} H(X^n) = H(X).\]

For a first-order stationary Markov source,

\[H(X) = \lim_{n \to \infty} \frac{1}{n} H(X^n) = \lim_{n \to \infty} H(X_n|X_{n-1}, \ldots, X_1) = H(X_2|X_1),\]

where

\[H(X_2|X_1) \triangleq -\sum_{x_1 \in \mathcal{X}} \sum_{x_2 \in \mathcal{X}} \pi(x_1) P_{X_2|X_1}(x_2|x_1) \cdot \log P_{X_2|X_1}(x_2|x_1),\]

and \(\pi(\cdot)\) is the stationary distribution for the Markov source. Furthermore, if the Markov source is binary with \(P_{X_2|X_1}(0|1) = \alpha\) and \(P_{X_2|X_1}(1|0) = \beta\), then

\[H(X) = \frac{\beta}{\alpha + \beta} h_b(\alpha) + \frac{\alpha}{\alpha + \beta} h_b(\beta),\]

where \(h_b(\alpha) \triangleq -\alpha \log_2 \alpha - (1 - \alpha) \log_2 (1 - \alpha)\) is the binary entropy function.
Theorem 3.12 (Generalized AEP or Shannon-McMillan-Breiman Theorem [13]) If \( \{X_n\}_{n=1}^{\infty} \) is a stationary ergodic source, then
\[
-\frac{1}{n} \log_2 P_{X^n}(X_1, \ldots, X_n) \xrightarrow{a.s.} H(X).
\]

Since the AEP theorem (law of large numbers) is valid for stationary ergodic sources, all consequences of AEP will follow, including Shannon’s lossless source coding theorem.

Theorem 3.13 (Shannon’s source coding theorem for stationary ergodic sources) Given integer \( D > 1 \), let \( \{X_n\}_{n=1}^{\infty} \) be a stationary ergodic source with entropy rate (in base \( D \))
\[
H_D(X) \triangleq \lim_{n \to \infty} \frac{1}{n} H_D(X^n).
\]
Then the following hold.

- **Forward part ( achievability):** For any \( 0 < \varepsilon < 1 \), there exists \( \delta \) with \( 0 < \delta < \varepsilon \) and a sequence of \( D \)-ary block codes \( \{C_n = (n, M_n)\}_{n=1}^{\infty} \) with
\[
\limsup_{n \to \infty} \frac{1}{n} \log_D M_n < H_D(X) + \delta,
\]
and probability of decoding error satisfied
\[
P_e(C_n) < \varepsilon
\]
for all sufficiently large \( n \).

- **Strong converse part:** For any \( 0 < \varepsilon < 1 \), any sequence of \( D \)-ary block codes \( \{C_n = (n, M_n)\}_{n=1}^{\infty} \) with
\[
\limsup_{n \to \infty} \frac{1}{n} \log_D M_n < H_D(X)
\]
satisfies
\[
P_e(C_n) > 1 - \varepsilon
\]
for all \( n \) sufficiently large.

A discrete memoryless (i.i.d.) source is stationary and ergodic (so Theorem 3.5 is clearly a special case of Theorem 3.13). In general, it is hard to check whether a stationary process is ergodic or not. It is known though that if a stationary process is a mixture of two or more stationary ergodic processes,
i.e., its \( n \)-fold distribution can be written as the mean (with respect to some distribution) of the \( n \)-fold distributions of stationary ergodic processes, then it is not ergodic.\(^{11}\)

For example, let \( P \) and \( Q \) be two distributions on a finite alphabet \( \mathcal{X} \) such that the process \( \{X_n\}_{n=1}^{\infty} \) is i.i.d. with distribution \( P \) and the process \( \{Y_n\}_{n=1}^{\infty} \) is i.i.d. with distribution \( Q \). Flip a biased coin (with Heads probability equal to \( \theta \), \( 0 < \theta < 1 \)) once and let

\[
Z_i = \begin{cases} 
X_i & \text{if Heads} \\
Y_i & \text{if Tails}
\end{cases}
\]

for \( i = 1, 2, \cdots \). Then the resulting process \( \{Z_i\}_{i=1}^{\infty} \) has its \( n \)-fold distribution as a mixture of the \( n \)-fold distributions of \( \{X_n\}_{n=1}^{\infty} \) and \( \{Y_n\}_{n=1}^{\infty} \):

\[
P_{Z^n}(a^n) = \theta P_{X^n}(a^n) + (1-\theta)P_{Y^n}(a^n)
\]

for all \( a^n \in \mathcal{X}^n \), \( n = 1, 2, \cdots \). Then the process \( \{Z_i\}_{i=1}^{\infty} \) is stationary but not ergodic.

A specific case for which ergodicity can be easily verified (other than the case of i.i.d. sources) is the case of stationary Markov sources. Specifically, if a (finite-alphabet) stationary Markov source is irreducible,\(^{12}\) then it is ergodic and hence the Generalized AEP holds for this source. Note that irreducibility can be verified in terms of the source’s transition probability matrix.

In more complicated situations such as when the source is non-stationary (with time-varying statistics) and/or non-ergodic, the source entropy rate \( H(X) \) (if the limit exists; otherwise one can look at the \( \lim\inf/\lim\sup \) of \( \frac{1}{n}H(X^n) \)) has no longer an operational meaning as the smallest possible compression rate. This causes the need to establish new entropy measures which appropriately characterize the operational limits of an arbitrary stochastic system with memory. This is achieved in [24] where Han and Verdú introduce the notions of \( \inf/\sup \)-entropy rates and illustrate the key role these entropy measures play in proving a general lossless block source coding theorem. More specifically, they demonstrate that for an arbitrary finite-alphabet source \( X = \{X^n = (X_1, X_2, \ldots, X_n)\}_{n=1}^{\infty} \) (not necessarily stationary and ergodic), the expression for the minimum achievable (block) source coding rate is given by the \( \sup \)-entropy rate \( \bar{H}(X) \), defined by

\[
\bar{H}(X) \triangleq \inf_{\beta \in \mathbb{R}} \left\{ \beta : \limsup_{n \to \infty} \Pr \left[ \frac{1}{n} \log P_{X^n}(X^n) > \beta \right] = 0 \right\}.
\]

More details will be provided in Part II of the book.

\(^{11}\)The converse is also true; i.e., if a stationary process cannot be represented as a mixture of stationary ergodic processes, then it is ergodic.

\(^{12}\)See p. 212 in Appendix B for the definition of irreducibility of Markov sources.
3.2.3 Redundancy for lossless block data compression

Shannon’s block source coding theorem establishes that the smallest data compression rate for achieving arbitrarily small error probability for stationary ergodic sources is given by the entropy rate. Thus one can define the source redundancy as the reduction in coding rate one can achieve via asymptotically lossless block source coding versus just using uniquely decodable (completely lossless for any value of the sourceword blocklength \( n \)) block source coding. In light of the fact that the former approach yields a source coding rate equal to the entropy rate while the later approach provides a rate of \( \log_2 |\mathcal{X}| \), we therefore define the total block source-coding redundancy \( \rho_t \) (in bits/source symbol) for a stationary ergodic source \( \{X_n\}_{n=1}^\infty \) as

\[
\rho_t \triangleq \log_2 |\mathcal{X}| - H(\mathcal{X}).
\]

Hence \( \rho_t \) represents the amount of “useless” (or superfluous) statistical source information one can eliminate via binary\(^{13} \) block source coding.

If the source is i.i.d. and uniformly distributed, then its entropy rate is equal to \( \log_2 |\mathcal{X}| \) and as a result its redundancy is \( \rho_t = 0 \). This means that the source is incompressible, as expected, since in this case every sourceword \( x^n \) will belong to the \( \delta \)-typical set \( \mathcal{F}_n(\delta) \) for every \( n > 0 \) and \( \delta > 0 \) (i.e., \( \mathcal{F}_n(\delta) = \mathcal{X}^n \)) and hence there are no superfluous sourcewords that can be dispensed of via source coding. If the source has memory or has a non-uniform marginal distribution, then its redundancy is strictly positive and can be classified into two parts:

- Source redundancy due to the non-uniformity of the source marginal distribution \( \rho_d \):
  \[
  \rho_d \triangleq \log_2 |\mathcal{X}| - H(X_1).
  \]

- Source redundancy due to the source memory \( \rho_m \):
  \[
  \rho_m \triangleq H(X_1) - H(\mathcal{X}).
  \]

As a result, the source total redundancy \( \rho_t \) can be decomposed in two parts:

\[
\rho_t = \rho_d + \rho_m.
\]

We can summarize the redundancy of some typical stationary ergodic sources in the following table.

\(^{13}\)Since we are measuring \( \rho_t \) in code bits/source symbol, all logarithms in its expression are in base 2 and hence this redundancy can be eliminated via asymptotically lossless binary block codes (one can also change the units to \( D \)-ary code symbol/source symbol by using base-\( D \) logarithms for the case of \( D \)-ary block codes).
### 3.3 Variable-length codes for lossless data compression

#### 3.3.1 Non-singular codes and uniquely decodable codes

We next study variable-length (completely) lossless data compression codes.

**Definition 3.14** Consider a discrete source $\{X_n\}_{n=1}^{\infty}$ with finite alphabet $\mathcal{X}$ along with a $D$-ary code alphabet $\mathcal{B} = \{0, 1, \ldots, D-1\}$, where $D > 1$ is an integer. Fix integer $n \geq 1$, then a $D$-ary $n$-th order variable-length code (VLC) is a map $f : \mathcal{X}^n \to \mathcal{B}^*$ mapping (fixed-length) sourcewords of length $n$ to $D$-ary codewords in $\mathcal{B}^*$ of variable lengths, where $\mathcal{B}^*$ denotes the set of all finite-length strings from $\mathcal{B}$ (i.e., $c \in \mathcal{B}^* \iff \exists$ integer $l \geq 1$ such that $c \in \mathcal{B}^l$).

The codebook $\mathcal{C}$ of a VLC is the set of all codewords:

$$\mathcal{C} = f(\mathcal{X}^n) = \{f(x^n) \in \mathcal{B}^* : x^n \in \mathcal{X}^n\}.$$  

A variable-length lossless data compression code is a code in which the source symbols can be completely reconstructed without distortion. In order to achieve this goal, the source symbols have to be encoded unambiguously in the sense that any two different source symbols (with positive probabilities) are represented by different codewords. Codes satisfying this property are called *non-singular codes*. In practice however, the encoder often needs to encode a sequence of source symbols, which results in a concatenated sequence of codewords. If any concatenation of codewords can also be unambiguously reconstructed without punctuation, then the code is said to be *uniquely decodable*. In

<table>
<thead>
<tr>
<th>Source</th>
<th>$\rho_d$</th>
<th>$\rho_m$</th>
<th>$\rho_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>i.i.d. uniform</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>i.i.d. non-uniform</td>
<td>$\log_2</td>
<td>\mathcal{X}</td>
<td>- H(X_1)$</td>
</tr>
<tr>
<td>1st-order symmetric Markov $^{14}$</td>
<td>0</td>
<td>$H(X_1) - H(X_2</td>
<td>X_1)$</td>
</tr>
<tr>
<td>1st-order non-symmetric Markov</td>
<td>$\log_2</td>
<td>\mathcal{X}</td>
<td>- H(X_1)$</td>
</tr>
</tbody>
</table>

$^{14}$A first-order Markov process is symmetric if for any $x_1$ and $\hat{x}_1$,

$$\{a : a = P_{X_2|X_1}(y|x_1) \text{ for some } y\} = \{a : a = P_{X_2|X_1}(y|\hat{x}_1) \text{ for some } y\}.$$
other words, a VLC is uniquely decodable if all finite sequences of sourcewords 
\((x^n \in X^n)\) are mapped onto distinct strings of codewords; i.e., for any \(m\) and \(m'\), 
\((x^n_1, x^n_2, \cdots, x^n_m) \neq (y^n_1, y^n_2, \cdots, y^n_{m'})\) implies that
\[
(f(x^n_1), f(x^n_2), \cdots, f(x^n_m)) \neq (f(y^n_1), f(y^n_2), \cdots, f(y^n_{m'})).
\]

Note that a non-singular VLC is not necessarily uniquely decodable. For example, consider a binary (first-order) code for the source with alphabet
\[X = \{A, B, C, D, E, F\}\]
given by
\[
\begin{align*}
\text{code of } A &= 0, \\
\text{code of } B &= 1, \\
\text{code of } C &= 00, \\
\text{code of } D &= 01, \\
\text{code of } E &= 10, \\
\text{code of } F &= 11.
\end{align*}
\]
The above code is clearly non-singular; it is however not uniquely decodable because the codeword sequence, 010, can be reconstructed as \(ABA, DA\) or \(AE\) (i.e., \((f(A), f(B), f(A)) = (f(D), f(A)) = (f(A), f(E))\) even if \((A, B, A), (D, A)\) and \((A, E)\) are all non-equal).

One important objective is to find out how “efficiently” we can represent a given discrete source via a uniquely decodable \(n\)-th order VLC and provide a construction technique that (at least asymptotically, as \(n \to \infty\)) attains the optimal “efficiency.” In other words, we want to determine what is the smallest possible average code rate (or equivalently, average codeword length) can an \(n\)-th order uniquely decodable VLC have when (losslessly) representing a given source, and we want to give an explicit code construction that can attain this smallest possible rate (at least asymptotically in the sourceword length \(n\)).

**Definition 3.15** Let \(C\) be a \(D\)-ary \(n\)-th order VLC code
\[f : X^n \to \{0, 1, \cdots, D - 1\}^*\]
for a discrete source \(\{X_n\}_{n=1}^\infty\) with alphabet \(X\) and distribution \(P_{X^n}(x^n), x^n \in X^n\). Setting \(\ell(c_{x^n})\) as the length of the codeword \(c_{x^n} = f(x^n)\) associated with sourceword \(x^n\), then the average codeword length for \(C\) is given by
\[\bar{\ell} \triangleq \sum_{x^n \in X^n} P_{X^n}(x^n) \ell(c_{x^n})\]
and its average code rate (in $D$-ary code symbols/source symbol) is given by

$$R_n \triangleq \frac{\ell}{n} = \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \ell(e_{x^n}).$$

The following theorem provides a strong condition which a uniquely decodable code must satisfy.

**Theorem 3.16 (Kraft inequality for uniquely decodable codes)** Let $C$ be a uniquely decodable $D$-ary $n$-th order VLC for a discrete source $\{X_n\}_{n=1}^{\infty}$ with alphabet $\mathcal{X}$. Let the $M = |\mathcal{X}|^n$ codewords of $C$ have lengths $\ell_1, \ell_2, \ldots, \ell_M$, respectively. Then the following inequality must hold

$$\sum_{m=1}^{M} D^{-\ell_m} \leq 1.$$

**Proof:** Suppose that we use the codebook $C$ to encode $N$ sourcewords $(x^n_i \in \mathcal{X}^n, i = 1, \cdots, N)$ arriving in a sequence; this yields a concatenated codeword sequence $e_1e_2e_3\cdots e_N$.

Let the lengths of the codewords be respectively denoted by

$$\ell(e_1), \ell(e_2), \ldots, \ell(e_N).$$

Consider

$$\left(\sum_{e_1 \in C} \sum_{e_2 \in C} \cdots \sum_{e_N \in C} D^{-[\ell(e_1)+\ell(e_2)+\cdots+\ell(e_N)]}\right).$$

It is obvious that the above expression is equal to

$$\left(\sum_{e \in C} D^{-\ell(e)}\right)^N = \left(\sum_{m=1}^{M} D^{-\ell_m}\right)^N.$$

(Note that $|C| = M$.) On the other hand, all the code sequences with length

$$i = \ell(e_1) + \ell(e_2) + \cdots + \ell(e_N)$$

contribute equally to the sum of the identity, which is $D^{-i}$. Let $A_i$ denote the number of $N$-codeword sequences that have length $i$. Then the above identity can be re-written as

$$\left(\sum_{m=1}^{M} D^{-\ell_m}\right)^N = \sum_{i=1}^{LN} A_i D^{-i},$$
where 

\[ L \triangleq \max_{c \in C} \ell(c). \]

Since \( C \) is by assumption a uniquely decodable code, the codeword sequence must be unambiguously decodable. Observe that a code sequence with length \( i \) has at most \( D^i \) unambiguous combinations. Therefore, \( A_i \leq D^i \), and

\[
\left( \sum_{m=1}^{M} D^{-\ell_m} \right)^N = \sum_{i=1}^{LN} A_i D^{-i} \leq \sum_{i=1}^{LN} D^i D^{-i} = LN,
\]

which implies that

\[
\sum_{m=1}^{M} D^{-\ell_m} \leq (LN)^{1/N}.
\]

The proof is completed by noting that the above inequality holds for every \( N \), and the upper bound \( (LN)^{1/N} \) goes to 1 as \( N \) goes to infinity. \(\square\)

The Kraft inequality is a very useful tool, especially for showing that the fundamental lower bound of the average rate of uniquely decodable VLCs for discrete memoryless sources is given by the source entropy.

**Theorem 3.17** The average rate of every uniquely decodable \( D \)-ary \( n \)-th order VLC for a discrete memoryless source \( \{X_n\}_{n=1}^{\infty} \) is lower-bounded by the source entropy \( H_D(X) \) (measured in \( D \)-ary code symbols/source symbol).

**Proof:** Consider a uniquely decodable \( D \)-ary \( n \)-th order VLC code for the source \( \{X_n\}_{n=1}^{\infty} \)

\[ f : \mathcal{X}^n \to \{0, 1, \cdots, D - 1\}^* \]

and let \( \ell(c_{x^n}) \) denote the length of the codeword \( c_{x^n} = f(x^n) \) for sourceword \( x^n \). Hence,

\[
\bar{R}_n - H_D(X) = \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \ell(c_{x^n}) - \frac{1}{n} H_D(X^n)
\]

\[
= \frac{1}{n} \left[ \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \ell(c_{x^n}) - \sum_{x^n \in \mathcal{X}^n} (-P_{X^n}(x^n) \log_D P_{X^n}(x^n)) \right]
\]

\[
= \frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \log_D \frac{P_{X^n}(x^n)}{D^{-\ell(c_{x^n})}}
\]

\[
\geq \frac{1}{n} \left[ \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \right] \log_D \left[ \frac{\sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n)}{\sum_{x^n \in \mathcal{X}^n} D^{-\ell(c_{x^n})}} \right]
\]

(\text{log-sum inequality})

67
\[ n \log \left( \sum_{x^n \in \mathcal{X}^n} D_{-\ell(c, x^n)} \right) \geq 0 \]

where the last inequality follows from the Kraft inequality for uniquely decodable codes and the fact that the logarithm is a strictly increasing function. \( \square \)

From the above theorem, we know that the average code rate is no smaller than the source entropy. Indeed a lossless data compression code, whose average code rate achieves entropy, should be optimal (since if its average code rate is below entropy, the Kraft inequality is violated and the code is no longer uniquely decodable). We summarize

1. Uniquely decodability \( \Rightarrow \) the Kraft inequality holds.
2. Uniquely decodability \( \Rightarrow \) average code rate of VLCs for memoryless sources is lower bounded by the source entropy.

**Exercise 3.18**

1. Find a non-singular and also non-uniquely decodable code that violates the Kraft inequality. (Hint: The answer is already provided in this subsection.)
2. Find a non-singular and also non-uniquely decodable code that beats the entropy lower bound.

### 3.3.2 Prefix or instantaneous codes

A *prefix code* is a VLC which is self-punctuated in the sense that there is no need to append extra symbols for differentiating adjacent codewords. A more precise definition follows:

**Definition 3.19 (Prefix code)** A VLC is called a *prefix code* or an *instantaneous code* if no codeword is a prefix of any other codeword.

A prefix code is also named an *instantaneous code* because the codeword sequence can be decoded *instantaneously* (it is immediately recognizable) without the reference to future codewords in the same sequence. Note that a uniquely decodable code is not necessarily prefix-free and may not be decoded instantaneously. The relationship between different codes encountered thus far is depicted in Figure 3.4.

A \( D \)-ary prefix code can be represented graphically as an initial segment of a \( D \)-ary tree. An example of a tree representation for a binary (\( D = 2 \)) prefix code is shown in Figure 3.5.
Theorem 3.20 (Kraft inequality for prefix codes) There exists a $D$-ary $n$th-order prefix code for a discrete source $\{X_n\}_{n=1}^{\infty}$ with alphabet $\mathcal{X}$ iff the codewords of length $\ell_m$, $m = 1, \ldots, M$, satisfy the Kraft inequality, where $M = |\mathcal{X}|^n$.

Proof: Without loss of generality, we provide the proof for the case of $D = 2$ (binary codes).

1. [The forward part] Prefix codes satisfy the Kraft inequality.

The codewords of a prefix code can always be put on a tree. Pick up a length $\ell_{\text{max}} \triangleq \max_{1 \leq m \leq M} \ell_m$.

A tree has originally $2^{\ell_{\text{max}}}$ nodes on level $\ell_{\text{max}}$. Each codeword of length $\ell_m$ obstructs $2^{\ell_{\text{max}}-\ell_m}$ nodes on level $\ell_{\text{max}}$. In other words, when any node is chosen as a codeword, all its children will be excluded from being codewords (as for a prefix code, no codeword can be a prefix of any other code). There are exactly $2^{\ell_{\text{max}}-\ell_m}$ excluded nodes on level $\ell_{\text{max}}$ of the tree. We therefore say that each codeword of length $\ell_m$ obstructs $2^{\ell_{\text{max}}-\ell_m}$ nodes on level $\ell_{\text{max}}$. Note that no two codewords obstruct the same nodes on level $\ell_{\text{max}}$. Hence the number of totally obstructed codewords on level $\ell_{\text{max}}$ should be less than $2^{\ell_{\text{max}}}$, i.e.,

$$\sum_{m=1}^{M} 2^{\ell_{\text{max}}-\ell_m} \leq 2^{\ell_{\text{max}}},$$

which immediately implies the Kraft inequality:

$$\sum_{m=1}^{M} 2^{-\ell_m} \leq 1.$$
Figure 3.5: Tree structure of a binary prefix code. The codewords are those residing on the leaves, which in this case are 00, 01, 10, 110, 1110 and 1111.

(This part can also be proven by stating the fact that a prefix code is a uniquely decodable code. The objective of adding this proof is to illustrate the characteristics of a tree-like prefix code.)

2. [The converse part] Kraft inequality implies the existence of a prefix code.

Suppose that $\ell_1, \ell_2, \ldots, \ell_M$ satisfy the Kraft inequality. We will show that there exists a binary tree with $M$ selected nodes where the $i^{th}$ node resides on level $\ell_i$.

Let $n_i$ be the number of nodes (among the $M$ nodes) residing on level $i$ (namely, $n_i$ is the number of codewords with length $i$ or $n_i = |\{m : \ell_m = i\}|$), and let

$$\ell_{\text{max}} \triangleq \max_{1 \leq m \leq M} \ell_m.$$ 

Then from the Kraft inequality, we have

$$n_12^{-1} + n_22^{-2} + \cdots + n_{\ell_{\text{max}}}2^{-\ell_{\text{max}}} \leq 1.$$ 

The above inequality can be re-written in a form that is more suitable for this proof as:

$$n_i2^{-1} \leq 1$$
\[ n_1 2^{-1} + n_2 2^{-2} \leq 1 \]
\[ \ldots \]
\[ n_1 2^{-1} + n_2 2^{-2} + \ldots + n_{\ell_{\text{max}}} 2^{-\ell_{\text{max}}} \leq 1. \]

Hence,
\[ n_1 \leq 2 \]
\[ n_2 \leq 2^2 - n_1 2^1 \]
\[ \ldots \]
\[ n_{\ell_{\text{max}}} \leq 2^{\ell_{\text{max}}} - n_1 2^{\ell_{\text{max}}-1} - \ldots - n_{\ell_{\text{max}}-1} 2^1, \]

which can be interpreted in terms of a tree model as: the 1st inequality says that the number of codewords of length 1 is less than the available number of nodes on the 1st level, which is 2. The 2nd inequality says that the number of codewords of length 2 is less than the total number of nodes on the 2nd level, which is 2\(^2\), minus the number of nodes obstructed by the 1st level nodes already occupied by codewords. The succeeding inequalities demonstrate the availability of a sufficient number of nodes at each level after the nodes blocked by shorter length codewords have been removed. Because this is true at every codeword length up to the maximum codeword length, the assertion of the theorem is proved.

Theorems 3.16 and 3.20 unveil the following relation between a variable-length uniquely decodable code and a prefix code.

**Corollary 3.21** A uniquely decodable \( D \)-ary \( n \)-th order code can always be replaced by a \( D \)-ary \( n \)-th order prefix code with the same average codeword length (and hence the same average code rate).

The following theorem interprets the relationship between the average code rate of a prefix code and the source entropy.

**Theorem 3.22** Consider a discrete memoryless source \( \{X_n\}_{n=1}^{\infty} \).

1. For any \( D \)-ary \( n \)-th order prefix code for the source, the average code rate is no less than the source entropy \( H_D(X) \).

2. There must exist a \( D \)-ary \( n \)-th order prefix code for the source whose average code rate is no greater than \( H_D(X) + \frac{1}{n} \), namely,

\[ \overline{R}_n \triangleq \frac{1}{n} \sum_{x^n \in X^n} P_{X^n}(x^n) \ell(c_{x^n}) \leq H_D(X) + \frac{1}{n}, \]

where \( c_{x^n} \) is the codeword for sourceword \( x^n \), and \( \ell(c_{x^n}) \) is the length of codeword \( c_{x^n} \).
Proof: A prefix code is uniquely decodable, and hence it directly follows from Theorem 3.17 that its average code rate is no less than the source entropy.

To prove the second part, we can design a prefix code satisfying both (3.3.1) and the Kraft inequality, which immediately implies the existence of the desired code by Theorem 3.20. Choose the codeword length for sourceword $x^n$ as

$$\ell(c_{x^n}) = \left\lfloor -\log_D P_{X^n}(x^n) \right\rfloor + 1.$$  \hspace{1cm} (3.3.2)

Then

$$D^{-\ell(c_{x^n})} \leq P_{X^n}(x^n).$$

Summing both sides over all source symbols, we obtain

$$\sum_{x^n \in X^n} D^{-\ell(c_{x^n})} \leq 1,$$

which is exactly the Kraft inequality. On the other hand, (3.3.2) implies

$$\ell(c_{x^n}) \leq -\log_D P_{X^n}(x^n) + 1,$$

which in turn implies

$$\sum_{x^n \in X^n} P_{X^n}(x^n) \ell(c_{x^n}) \leq \sum_{x^n \in X^n} \left[ - P_{X^n}(x^n) \log_D P_{X^n}(x^n) \right] + \sum_{x^n \in X^n} P_{X^n}(x^n)$$

$$= H_D(X^n) + 1 = nH_D(X) + 1,$$

where the last equality holds since the source is memoryless. \hfill \Box

We note that $n$-th order prefix codes (which encode sourcewords of length $n$) for memoryless sources can yield an average code rate arbitrarily close to the source entropy when allowing $n$ to grow without bound. For example, a memoryless source with alphabet

$$\{A, B, C\}$$

and probability distribution

$$P_X(A) = 0.8, \quad P_X(B) = P_X(C) = 0.1$$

has entropy being equal to

$$-0.8 \cdot \log_2 0.8 - 0.1 \cdot \log_2 0.1 - 0.1 \cdot \log_2 0.1 = 0.92 \text{ bits.}$$

One of the best binary first-order or single-letter encoding (with $n = 1$) prefix codes for this source is given by $c(A) = 0$, $c(B) = 10$ and $c(C) = 11$, where $c(\cdot)$ is the encoding function. Then the resultant average code rate for this code is

$$0.8 \times 1 + 0.2 \times 2 = 1.2 \text{ bits} \geq 0.92 \text{ bits.}$$

72
Now if we consider a second-order (with \( n = 2 \)) prefix code by encoding two consecutive source symbols at a time, the new source alphabet becomes

\[ \{AA, AB, AC, BA, BB, BC, CA, CB, CC\}, \]

and the resultant probability distribution is calculated by

\[
(\forall x_1, x_2 \in \{A, B, C\}) \; \; P_{X^2}(x_1, x_2) = P_X(x_1)P_X(x_2)
\]
as the source is memoryless. Then one of the best binary prefix codes for the source is given by

\[
\begin{align*}
    c(AA) &= 0 \\
    c(AB) &= 100 \\
    c(AC) &= 101 \\
    c(BA) &= 110 \\
    c(BB) &= 11100 \\
    c(BC) &= 11101 \\
    c(CA) &= 1110 \\
    c(CB) &= 11110 \\
    c(CC) &= 11111.
\end{align*}
\]

The average code rate of this code now becomes

\[
\frac{0.64(1 \times 1) + 0.08(3 \times 3 + 4 \times 1) + 0.01(6 \times 4)}{2} = 0.96 \text{ bits},
\]

which is closer to the source entropy of 0.92 bits. As \( n \) increases, the average code rate will be brought closer to the source entropy.

From Theorems 3.17 and 3.22, we obtain the lossless variable-length source coding theorem for discrete memoryless sources.

**Theorem 3.23 (Lossless variable-length source coding theorem)** Fix integer \( D > 1 \) and consider a discrete memoryless source \( \{X_n\}_{n=1}^\infty \) with distribution \( P_X \) and entropy \( H_D(X) \) (measured in \( D \)-ary units). Then the following hold.

- **Forward part (achievability):** For any \( \varepsilon > 0 \), there exists a \( D \)-ary \( n \)-th order prefix (hence uniquely decodable) code
  \[
  f : \mathcal{X}^n \rightarrow \{0, 1, \cdots, D-1\}^*
  \]
  for the source with an average code rate \( \overline{R}_n \) satisfying
  \[
  \overline{R}_n \leq H_D(X) + \varepsilon
  \]
  for \( n \) sufficiently large.
• **Converse part:** Every uniquely decodable code

\[ f : \mathcal{X}^n \to \{0, 1, \ldots, D - 1\}^* \]

for the source has an average code rate \( \overline{R}_n \geq H_D(X) \).

Thus, for a discrete memoryless source, its entropy \( H_D(X) \) (measured in \( D \)-ary units) represents the smallest variable-length lossless compression rate for \( n \) sufficiently large.

**Proof:** The forward part follows directly from Theorem 3.22 by choosing \( n \) large enough such that \( 1/n < \varepsilon \), and the converse part is already given by Theorem 3.17.

**Observation 3.24** Theorem 3.23 actually also holds for the class of stationary sources by replacing the source entropy \( H_D(X) \) with the source entropy rate

\[ H_D(\mathcal{X}) \triangleq \lim_{n \to \infty} \frac{1}{n} H_D(\mathcal{X}^n), \]

measured in \( D \)-ary units. The proof is very similar to the proofs of Theorems 3.17 and 3.22 with slight modifications (such as using the fact that \( \frac{1}{n} H_D(\mathcal{X}^n) \) is non-increasing with \( n \) for stationary sources).

### 3.3.3 Examples of binary prefix codes

A) **Huffman codes:** optimal variable-length codes

Given a discrete source with alphabet \( \mathcal{X} \), we next construct an optimal binary first-order (single-letter) uniquely decodable variable-length code

\[ f : \mathcal{X} \to \{0, 1\}^*, \]

where optimality is in the sense that the code’s average codeword length (or equivalently, its average code rate) is minimized over the class of all binary uniquely decodable codes for the source. Note that finding optimal \( n \)-th order codes with \( n > 1 \) follows directly by considering \( \mathcal{X}^n \) as a new source with expanded alphabet (i.e., by mapping \( n \) source symbols at a time).

By Corollary 3.21, we remark that in our search for optimal uniquely decodable codes, we can restrict our attention to the (smaller) class of optimal prefix codes. We thus proceed by observing the following necessary conditions of optimality for binary prefix codes.

74
Lemma 3.25 Let $C$ be an optimal binary prefix code with codeword lengths $\ell_i$, $i = 1, \ldots, M$, for a source with alphabet $X = \{a_1, \ldots, a_M\}$ and symbol probabilities $p_1, \ldots, p_M$. We assume, without loss of generality, that

$$p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_M,$$

and that any group of source symbols with identical probability is listed in order of increasing codeword length (i.e., if $p_i = p_{i+1} = \cdots = p_{i+s}$, then $\ell_i \leq \ell_{i+1} \leq \cdots \leq \ell_{i+s}$). Then the following properties hold.

1. Higher probability source symbols have shorter codewords: $p_i > p_j$ implies $\ell_i \leq \ell_j$, for $i, j = 1, \ldots, M$.

2. The two least probable source symbols have codewords of equal length:

$$\ell_{M-1} = \ell_M.$$

3. Among the codewords of length $\ell_M$, two of the codewords are identical except in the last digit.

Proof:

1) If $p_i > p_j$ and $\ell_i > \ell_j$, then it is possible to construct a better code $C'$ by interchanging (“swapping”) codewords $i$ and $j$ of $C$, since

$$\bar{\ell}(C') - \bar{\ell}(C) = p_i \ell_j + p_j \ell_i - (p_i \ell_i + p_j \ell_j) = (p_i - p_j)(\ell_j - \ell_i) < 0.$$

Hence code $C'$ is better than code $C$, contradicting the fact that $C$ is optimal.

2) We first know that $\ell_{M-1} \leq \ell_M$, since:

- If $p_{M-1} > p_M$, then $\ell_{M-1} \leq \ell_M$ by result 1) above.
- If $p_{M-1} = p_M$, then $\ell_{M-1} \leq \ell_M$ by our assumption about the ordering of codewords for source symbols with identical probability.

Now, if $\ell_{M-1} < \ell_M$, we may delete the last digit of codeword $M$, and the deletion cannot result in another codeword since $C$ is a prefix code. Thus the deletion forms a new prefix code with a better average codeword length than $C$, contradicting the fact that $C$ is optimal. Hence, we must have that $\ell_{M-1} = \ell_M$.

3) Among the codewords of length $\ell_M$, if no two codewords agree in all digits except the last, then we may delete the last digit in all such codewords to obtain a better codeword. \qed
The above observation suggests that if we can construct an optimal code for the entire source except for its two least likely symbols, then we can construct an optimal overall code. Indeed, the following lemma due to Huffman follows from Lemma 3.25.

**Lemma 3.26 (Huffman)** Consider a source with alphabet $\mathcal{X} = \{a_1, \ldots, a_M\}$ and symbol probabilities $p_1, \ldots, p_M$ such that $p_1 \geq p_2 \geq \cdots \geq p_M$. Consider the reduced source alphabet $\mathcal{Y}$ obtained from $\mathcal{X}$ by combining the two least likely source symbols $a_{M-1}$ and $a_M$ into an equivalent symbol $a_{M-1,M}$ with probability $p_{M-1} + p_M$. Suppose that $C'$, given by $f' : \mathcal{Y} \to \{0,1\}^*$, is an optimal code for the reduced source $\mathcal{Y}$. We now construct a code $C, f : \mathcal{X} \to \{0,1\}^*$, for the original source $\mathcal{X}$ as follows:

- The codewords for symbols $a_1, a_2, \ldots, a_{M-2}$ are exactly the same as the corresponding codewords in $C'$: 
  
  $$f(a_1) = f'(a_1), f(a_2) = f'(a_2), \ldots, f(a_{M-2}) = f'(a_{M-2}).$$

- The codewords associated with symbols $a_{M-1}$ and $a_M$ are formed by appending a “0” and a “1”, respectively, to the codeword $f'(a_{M-1,M})$ associated with the letter $a_{M-1,M}$ in $C'$:
  
  $$f(a_{M-1}) = [f'(a_{M-1,M})0] \quad \text{and} \quad f(a_M) = [f'(a_{M-1,M})1].$$

Then code $C$ is optimal for the original source $\mathcal{X}$.

Hence the problem of finding the optimal code for a source of alphabet size $M$ is reduced to the problem of finding an optimal code for the reduced source of alphabet size $M - 1$. In turn we can reduce the problem to that of size $M - 2$ and so on. Indeed the above lemma yields a recursive algorithm for constructing optimal binary prefix codes.

**Huffman encoding algorithm:** Repeatedly apply the above lemma until one is left with a reduced source with two symbols. An optimal binary prefix code for this source consists of the codewords 0 and 1. Then proceed backwards, constructing (as outlined in the above lemma) optimal codes for each reduced source until one arrives at the original source.

**Example 3.27** Consider a source with alphabet $\{1, 2, 3, 4, 5, 6\}$ and symbol probabilities 0.25, 0.25, 0.25, 0.1, 0.1 and 0.05, respectively. By following the Huffman encoding procedure as shown in Figure 3.6, we obtain the Huffman code as

$$00, 01, 10, 110, 1110, 1111.$$
Observation 3.28

• Huffman codes are not unique for a given source distribution; e.g., by inverting all the code bits of a Huffman code, one gets another Huffman code, or by resolving ties in different ways in the Huffman algorithm, one also obtains different Huffman codes (but all of these codes have the same minimal $R_n$).

• One can obtain optimal codes that are not Huffman codes; e.g., by interchanging two codewords of the same length of a Huffman code, one can get another non-Huffman (but optimal) code. Furthermore, one can construct an optimal suffix code (i.e., a code in which no codeword can be a suffix of another codeword) from a Huffman code (which is a prefix code) by reversing the Huffman codewords.

• Binary Huffman codes always satisfy the Kraft inequality with equality (their code tree is “saturated”); e.g., see [14, p. 72].

• Any $n$-th order binary Huffman code $f : \mathcal{X}^n \rightarrow \{0, 1\}^*$ for a stationary source $\{X_n\}_{n=1}^\infty$ with finite alphabet $\mathcal{X}$ satisfies:

$$H(\mathcal{X}) \leq \frac{1}{n} H(X^n) \leq \overline{R_n} < \frac{1}{n} H(X^n) + \frac{1}{n}.$$
Thus, as \( n \) increases to infinity, \( R_n \to H(\mathcal{X}) \) but the complexity as well as encoding-decoding delay grows exponentially with \( n \).

- Finally, note that non-binary (i.e., for \( D > 2 \)) Huffman codes can also be constructed in a mostly similar way as for the case of binary Huffman codes by designing a \( D \)-ary tree and iteratively applying Lemma 3.26, where now the \( D \) least likely source symbols are combined at each stage. The only difference from the case of binary Huffman codes is that we have to ensure that we are ultimately left with \( D \) symbols at the last stage of the algorithm to guarantee the code’s optimality. This is remedied by expanding the original source alphabet \( \mathcal{X} \) by adding “dummy” symbols (each with zero probability) so that the alphabet size of the expanded source \( |\mathcal{X}'| \) is the smallest positive integer greater than or equal to \( |\mathcal{X}| \) with

\[ |\mathcal{X}'| = 1 \pmod{D-1}. \]

For example, if \( |\mathcal{X}| = 6 \) and \( D = 3 \) (ternary codes), we obtain that \( |\mathcal{X}'| = 7 \), meaning that we need to enlarge the original source \( \mathcal{X} \) by adding one dummy (zero-probability) source symbol.

We thus obtain that the necessary conditions for optimality of Lemma 3.25 also hold for \( D \)-ary prefix codes when replacing \( \mathcal{X} \) with the expanded source \( \mathcal{X}' \) and replacing “two” with “\( D \)” in the statement of the lemma. The resulting \( D \)-ary Huffman code will be an optimal code for the original source \( \mathcal{X} \) (e.g., see [19, Chap. 3] and [37, Chap. 11]).

**B) Shannon-Fano-Elias code**

Assume \( \mathcal{X} = \{1, \ldots, M\} \) and \( P_X(x) > 0 \) for all \( x \in \mathcal{X} \). Define

\[ F(x) \triangleq \sum_{a \leq x} P_X(a), \]

and

\[ \bar{F}(x) \triangleq \sum_{a < x} P_X(a) + \frac{1}{2} P_X(x). \]

**Encoder:** For any \( x \in \mathcal{X} \), express \( \bar{F}(x) \) in decimal binary form, say

\[ \bar{F}(x) = .c_1 c_2 \ldots c_k \ldots, \]

and take the first \( k \) (fractional) bits as the codeword of source symbol \( x \), i.e.,

\( (c_1, c_2, \ldots, c_k) \),
where \( k \triangleq \lceil \log_2(1/P_X(x)) \rceil + 1 \).

**Decoder:** Given codeword \((c_1, \ldots, c_k)\), compute the cumulative sum of \( F(\cdot) \) starting from the smallest element in \( \{1, 2, \ldots, M\} \) until the first \( x \) satisfying
\[
F(x) \geq .c_1 \ldots c_k.
\]

Then \( x \) should be the original source symbol.

**Proof of decodability:** For any number \( a \in [0, 1] \), let \([a]_k\) denote the operation that chops the binary representation of \( a \) after \( k \) bits (i.e., removing the \((k+1)\)th bit, the \((k+2)\)th bit, etc). Then
\[
\bar{F}(x) - [\bar{F}(x)]_k < \frac{1}{2^k}.
\]

Since \( k = \lceil \log_2(1/P_X(x)) \rceil + 1 \),
\[
\frac{1}{2^k} \leq \frac{1}{2} P_X(x)
= \left[ \sum_{a<x} P_X(a) + \frac{P_X(x)}{2} \right] - \sum_{a \leq x-1} P_X(a)
= \bar{F}(x) - F(x-1).
\]

Hence,
\[
F(x-1) = \left[ F(x-1) + \frac{1}{2^k} \right] - \frac{1}{2^k} \leq \bar{F}(x) - \frac{1}{2^k} < [\bar{F}(x)]_k.
\]

In addition,
\[
F(x) > \bar{F}(x) \geq [\bar{F}(x)]_k.
\]

Consequently, \( x \) is the first element satisfying
\[
F(x) \geq .c_1c_2 \ldots c_k.
\]

**Average codeword length:**
\[
\bar{\ell} = \sum_{x \in X} P_X(x) \left[ \log_2 \frac{1}{P_X(x)} \right] + 1
< \sum_{x \in X} P_X(x) \log_2 \frac{1}{P_X(x)} + 2
= (H(X) + 2) \text{ bits}.
\]

**Observation 3.29** The Shannon-Fano-Elias code is a prefix code.
3.3.4 Examples of universal lossless variable-length codes

In Section 3.3.3, we assume that the source distribution is known. Thus we can use either Huffman codes or Shannon-Fano-Elias codes to compress the source. What if the source distribution is not a known priori? Is it still possible to establish a completely lossless data compression code which is universally good (or asymptotically optimal) for all interested sources? The answer is affirmative. Two of the examples are the adaptive Huffman codes and the Lempel-Ziv codes (which unlike Huffman and Shannon-Fano-Elias codes map variable-length source words onto codewords).

A) Adaptive Huffman code

A straightforward universal coding scheme is to use the empirical distribution (or relative frequencies) as the true distribution, and then apply the optimal Huffman code according to the empirical distribution. If the source is i.i.d., the relative frequencies will converge to its true marginal probability. Therefore, such universal codes should be good for all i.i.d. sources. However, in order to get an accurate estimation of the true distribution, one must observe a sufficiently long source sequence under which the coder will suffer a long delay. This can be improved by using the adaptive universal Huffman code [20].

The working procedure of the adaptive Huffman code is as follows. Start with an initial guess of the source distribution (based on the assumption that the source is DMS). As a new source symbol arrives, encode the data in terms of the Huffman coding scheme according to the current estimated distribution, and then update the estimated distribution and the Huffman codebook according to the newly arrived source symbol.

To be specific, let the source alphabet be \( X = \{a_1, \ldots, a_M\} \). Define

\[
N(a_i|x^n) \triangleq \text{number of } a_i \text{ occurrence in } x_1, x_2, \ldots, x_n.
\]

Then the (current) relative frequency of \( a_i \) is \( N(a_i|x^n)/n \). Let \( c_n(a_i) \) denote the Huffman codeword of source symbol \( a_i \) with respect to the distribution

\[
\left[ \frac{N(a_1|x^n)}{n}, \frac{N(a_2|x^n)}{n}, \ldots, \frac{N(a_M|x^n)}{n} \right].
\]

Now suppose that \( x_{n+1} = a_j \). The codeword \( c_n(a_j) \) is set as output, and the relative frequency for each source outcome becomes:

\[
\frac{N(a_j|x^{n+1})}{n+1} = n \times \left( \frac{N(a_j|x^n)}{n} \right) + 1
\]

80
and
\[ \frac{N(a_i|x^{n+1})}{n+1} = \frac{n \times (N(a_i|x^n)/n)}{n+1} \quad \text{for } i \neq j. \]

This observation results in the following distribution updated policy.

\[ P^{(n+1)}_X(a_j) = \frac{n P^{(n)}_X(a_j) + 1}{n+1} \]

and
\[ P^{(n+1)}_X(a_i) = \frac{n}{n+1} P^{(n)}_X(a_i) \quad \text{for } i \neq j, \]

where \( P^{(n+1)}_X \) represents the estimate of the true distribution \( P_X \) at time \((n+1)\).

Note that in the Adaptive Huffman coding scheme, the encoder and decoder need not be re-designed at every time, but only when a sufficient change in the estimated distribution occurs such that the so-called \textit{sibling property} is violated.

\textbf{Definition 3.30 (Sibling property)} A prefix code is said to have the \textit{sibling property} if its codetree satisfies:

1. every node in the code-tree (except for the root node) has a sibling (i.e., the code-tree is saturated), and
2. the node can be listed in non-decreasing order of probabilities with each node being adjacent to its sibling.

The next observation indicates the fact that the Huffman code is the only prefix code satisfying the sibling property.

\textbf{Observation 3.31} A prefix code is a Huffman code iff it satisfies the sibling property.

An example for a code tree satisfying the sibling property is shown in Figure 3.7. The first requirement is satisfied since the tree is saturated. The second requirement can be checked by the node list in Figure 3.7.

If the next observation (say at time \( n = 17 \)) is \( a_3 \), then its codeword 100 is set as output (using the Huffman code corresponding to \( P^{(16)}_X \)). The estimated distribution is updated as

\[
\begin{align*}
P^{(17)}_X(a_1) &= \frac{16 \times (3/8)}{17} = \frac{6}{17}, \quad P^{(17)}_X(a_2) = \frac{16 \times (1/4)}{17} = \frac{4}{17}, \\
P^{(17)}_X(a_3) &= \frac{16 \times (1/8) + 1}{17} = \frac{3}{17}, \quad P^{(17)}_X(a_4) = \frac{16 \times (1/8)}{17} = \frac{2}{17}
\end{align*}
\]
Figure 3.7: Example of the sibling property based on the code tree from $P_X^{(16)}$. The arguments inside the parenthesis following $a_j$ respectively indicate the codeword and the probability associated with $a_j$. “b” is used to denote the internal nodes of the tree with the assigned (partial) code as its subscript. The number in the parenthesis following $b$ is the probability sum of all its children.

$$P_X^{(17)}(a_5) = \frac{16 \times [1/(16)]}{17} = \frac{1}{17}, \quad P_X^{(17)}(a_6) = \frac{16 \times [1/(16)]}{17} = \frac{1}{17}.$$ 

The sibling property is then violated (cf. Figure 3.8). Hence, codebook needs to be updated according to the new estimated distribution, and the observation at $n = 18$ shall be encoded using the new codebook in Figure 3.9. Details about Adaptive Huffman codes can be found in [20].

**B) Lempel-Ziv codes**

We now introduce a well-known and feasible universal coding scheme, which is named after its inventors, Lempel and Ziv (e.g., cf. [13]). These codes, unlike Huffman and Shannon-Fano-Elias codes, map variable-length sourcewords (as
opposed to fixed-length codewords) onto codewords.

Suppose the source alphabet is binary. Then the Lempel-Ziv encoder can be described as follows.

**Encoder:**

1. Parse the input sequence into strings that have never appeared before. For example, if the input sequence is 1011010100010..., the algorithm first grabs the first letter 1 and finds that it has never appeared before. So 1 is the first string. Then the algorithm scoops the second letter 0 and also determines that it has not appeared before, and hence, put it to be the next string. The algorithm moves on to the next letter 1, and finds that this string has appeared. Hence, it hits another letter 1 and yields a new string 11, and so on. Under this procedure, the source sequence is parsed into the strings

$$1, 0, 11, 01, 010, 00, 10.$$
2. Let $L$ be the number of distinct strings of the parsed source. Then we need $\log_2 L$ bits to index these strings (starting from one). In the above example, the indices are:

<table>
<thead>
<tr>
<th>parsed source</th>
<th>1</th>
<th>0</th>
<th>11</th>
<th>01</th>
<th>010</th>
<th>00</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>index</td>
<td>001</td>
<td>010</td>
<td>011</td>
<td>100</td>
<td>101</td>
<td>110</td>
<td>111</td>
</tr>
</tbody>
</table>

The codeword of each string is then the index of its prefix concatenated with the last bit in its source string. For example, the codeword of source string 010 will be the index of 01, i.e., 100, concatenated with the last bit of the source string, i.e., 0. Through this procedure, encoding the above parsed strings with $L = 3$ yields the codeword sequence

$$(000, 1)(000, 0)(001, 1)(010, 1)(100, 0)(010, 0)(001, 0)$$

or equivalently,

0001000000110101100001000010.

Figure 3.9: (Continuation of Figure 3.8) Updated Huffman code. The sibling property holds now for the new code.
Note that the conventional Lempel-Ziv encoder requires two passes: the first pass to decide \( L \), and the second pass to generate the codewords. The algorithm, however, can be modified so that it requires only one pass over the entire source string. Also note that the above algorithm uses an \textit{equal} number of bits—\( \log_2 L \)—to all the location index, which can also be relaxed by proper modification.

\textit{Decoder:} The decoding is straightforward from the encoding procedure.

\textbf{Theorem 3.32} The above algorithm asymptotically achieves the entropy rate of any stationary ergodic source (with unknown statistics).

\textbf{Proof:} Refer to [13, Sec. 13.5]. \( \square \)

\textbf{Problems}

1. A binary discrete memoryless source \( \{X_n\}_{n=1}^\infty \) has distribution \( P_X(1) = 0.005 \). A binary codeword is provided for every sequence of 100 source digits containing three or fewer ones. In other words, the set of sourcewords of length 100 that are encoded to distinct block codewords is

\[ \mathcal{A} \triangleq \{ x^{100} \in \{0, 1\}^{100} : \text{number of 1's in } x^{100} \leq 3 \} . \]

(a) Show that \( \mathcal{A} \) is indeed a typical set \( \mathcal{F}_{100}(0.2) \) defined using the natural logarithm.

(b) Find the minimum codeword blocklength for the block coding scheme.

(c) Find the probability for sourcewords not in \( \mathcal{A} \).

(d) Use Chebyshev’s inequality to bound the probability of observing a sourceword outside \( \mathcal{A} \). Compare this bound with the actual probability computed in part (c).

\textit{Hint:} Let \( X_i \) represent the binary random digit at instance \( i \), and let \( S_n = X_1 + \cdots + X_n \). Note that \( \Pr[S_{100} \geq 4] \) is equal to

\[ \Pr \left[ \left| \frac{1}{100} S_{100} - 0.005 \right| \geq 0.35 \right] . \]

2. \textit{Weak converse to the Fixed-Length Source Coding Theorem:} Recall (see Footnote 3) that an \((n, M)\) fixed-length source code for a discrete memoryless source (DMS) \( \{X_n\}_{n=1}^\infty \) with finite alphabet \( \mathcal{X} \) consists of an encoder \( f : \mathcal{X}^n \to \{1, 2, \cdots, M\} \), and a decoder \( g : \{1, 2, \cdots, M\} \to \mathcal{X}^n \). The rate of the code is

\[ R_n \triangleq \frac{1}{n} \log_2 M \text{ bits/source symbol}, \]
and its probability of decoding error is

\[ P_e = \Pr[X^n \neq \hat{X}^n], \]

where \( \hat{X}^n = g(f(X^n)) \).

(a) Show that any fixed-length source code \((n, M)\) for a DMS satisfies

\[ P_e \geq \frac{H(X) - R_n}{\log_2 |X|} - \frac{1}{n \log_2 |X|}, \]

where \( H(X) \) is the source entropy.

Hint: Show that \( \log_2 M \geq I(X^n; \hat{X}^n) \), and use Fano’s inequality.

(b) Deduce the (weak) converse to the fixed-length source coding theorem for DMS’s by proving that for any \((n, M)\) source code with

\[ \limsup_{n \to \infty} R_n < H(X), \]

its \( P_e \) is bounded away from zero for \( n \) sufficiently large.

3. For a stationary source \( \{X_n\}_{n=1}^{\infty} \), show that for any integer \( n > 1 \),

(a) \[ \frac{1}{n} H(X^n) \leq \frac{1}{n-1} H(X^{n-1}) \]

(b) \[ \frac{1}{n} H(X^n) \geq H(X_n|X^{n-1}). \]

Hint: Use the chain rule for entropy and the fact that

\[ H(X_i|X_{i-1}, \ldots, X_1) = H(X_n|X_{n-1}, \ldots, X_{n-i+1}) \]

for every \( i \).

4. Random walk: A person walks on a line of integers. Each time, he may walk forward with probability 0.9, or he may walk backwards with probability 0.1. Let \( X_i \) be the number he stands on at time instance \( i \), and let \( X_0 = 0 \) (with probability one).

(a) Find \( H(X_1, X_2, \ldots, X_n) \).

(b) Find the entropy rate of the process \( \{X_n\}_{n=1}^{\infty} \), i.e.,

\[ H(X) = \lim_{n \to \infty} \frac{1}{n} H(X^n). \]

5. A source with binary alphabet \( \mathcal{X} = \{0, 1\} \) emits a sequence of random variables \( \{X_n\}_{n=1}^{\infty} \). Let \( \{Z_n\}_{n=1}^{\infty} \) be a binary independent and identically distributed (i.i.d.) sequence of random variables such that \( \Pr\{Z_n = 1\} = \Pr\{Z_n = 0\} \). We assume that \( \{X_n\}_{n=1}^{\infty} \) is generated according to the equation

\[ X_n = X_{n-1} \oplus X_{n-2} \oplus Z_n, \quad n = 1, 2, \ldots \]

where \( \oplus \) denotes addition modulo-2, and \( X_0 = X_{-1} = 0 \). Find the entropy rate of \( \{X_n\}_{n=1}^{\infty} \).
6. For each of the following codes, either prove unique decodability or give an ambiguous concatenated sequence of codewords:
   (a) \{1, 0, 0\}.
   (b) \{1, 0, 0\}.
   (c) \{1, 10, 0\}.
   (d) \{1, 10, 0\}.
   (e) \{0, 0\}.
   (f) \{0, 0, 1, 0\}.

7. We know the fact that the average code rate of all non-singular uniquely decodable codes for a DMS must be no less than source entropy. But this is not necessarily true for non-singular codes. Give an example of a non-singular code in which the average code rate is less than entropy.

8. Under what condition does the average code rate of a uniquely decodable binary first-order variable-length code for a DMS equal the source entropy? 
   \textit{Hint:} See the proof of Theorem 3.17.

9. Binary Markov Source: Consider the binary homogeneous Markov source: 
   \( \{X_n\}_{n=1}^\infty, X_n \in \mathcal{X} = \{0, 1\} \), with 
   \[
   \Pr\{X_{n+1} = j | X_n = i\} = \begin{cases} 
   \frac{\rho}{1+\delta} & \text{if } i = 0 \text{ and } j = 1 \\
   \frac{\rho+\delta}{1+\delta} & \text{if } i = 1 \text{ and } j = 1 , 
   \end{cases}
   \]
   where \( n \geq 1, 0 \leq \rho \leq 1 \) and \( \delta \geq 0 \).
   (a) Find the initial state distribution \( \Pr\{X_1 = 0\}, \Pr\{X_1 = 1\} \) required to make the source \( \{X_n\}_{n=1}^\infty \) stationary.
   Assume in the next questions that the source is stationary.
   (b) Find the entropy rate of \( \{X_n\}_{n=1}^\infty \) in terms of \( \rho \) and \( \delta \).
   (c) If \( \delta = 1 \) and \( \rho = 1/2 \), compute the source redundancies \( \rho_d \), \( \rho_m \) and \( \rho_t \).
   (d) Suppose that \( \rho = 1 \). Is \( \{X_n\}_{n=1}^\infty \) irreducible? What is the value of the entropy rate in this case?
   (e) If \( \delta = 0 \), show that \( \{X_n\}_{n=1}^\infty \) is a discrete memoryless source and compute its entropy rate in terms of \( \rho \).
   (f) If \( \rho = 1/2 \) and \( \delta = 3/2 \), design first-, second-, and third-order binary Huffman codes for this source. Determine in each case the average code rate and compare it to the entropy rate.

10. Suppose random variables \( Z_1 \) and \( Z_2 \) are independent from each other and have the same distribution as \( Z \) with
   \[
   \begin{align*}
   &\Pr[Z = e_1] = 0.4; \\
   &\Pr[Z = e_2] = 0.3; \\
   &\Pr[Z = e_3] = 0.2; \\
   &\Pr[Z = e_4] = 0.1. 
   \end{align*}
   \]
(a) Design a first-order binary Huffman code \( f : \{e_1, e_2, e_3, e_4\} \rightarrow \{0, 1\}^* \) for \( Z \).

(b) Applying the Huffman code in (a) to \( Z_1 \) and \( Z_2 \) and concatenating \( f(Z_1) \) with \( f(Z_2) \) yields an overall codeword for the pair \( (Z_1, Z_2) \) given by

\[
f(Z_1, Z_2) \triangleq (f(Z_1), f(Z_2)) = (U_1, U_2, \ldots, U_k),
\]

where \( k \) ranges from 2 to 6, depending on the outcomes of \( Z_1 \) and \( Z_2 \). Are \( U_1 \) and \( U_2 \) independent? Justify your answer.

*Hint:* Examine \( \Pr[U_2 = 0|U_1 = u_1] \) for different values of \( u_1 \).

(c) Is the average code rate equal to the entropy given by

\[
0.4 \log_2 \frac{1}{0.4} + 0.3 \log_2 \frac{1}{0.3} + 0.2 \log_2 \frac{1}{0.2} + 0.1 \log_2 \frac{1}{0.1} = 1.84644 \text{ bits/letter}?
\]

Justify your answer.

(d) Now if we apply the Huffman code in (a) sequentially to the i.i.d. sequence \( Z_1, Z_2, Z_3, \ldots \) with the same marginal distribution as \( Z \), and yield the output \( U_1, U_2, U_3, \ldots \), can \( U_1, U_2, U_3, \ldots \) be further compressed?

If your answer to this question is NO, prove the i.i.d. uniformity of \( U_1, U_2, U_3, \ldots \). If your answer to this question is YES, then explain why the optimal Huffman code does not give an i.i.d. uniform output.

*Hint:* Examine whether the average code rate can achieve source entropy.

11. In the second part of Theorem 3.22, it is shown that there exists a \( D \)-ary prefix code with

\[
\bar{R}_n = \frac{1}{n} \sum_{x \in \mathcal{X}} P_X(x) \ell(c_x) \leq H_D(X) + \frac{1}{n},
\]

where \( c_x \) is the codeword for the source symbol \( x \) and \( \ell(c_x) \) is the length of codeword \( c_x \). Show that the upper bound can be improved to:

\[
\bar{R}_n < H_D(X) + \frac{1}{n}.
\]

*Hint:* Replace \( \ell(c_x) = \lceil -\log_D P_X(x) \rceil + 1 \) by a new assignment.

12. Let \( X_1, X_2, X_3, \cdots \) be an i.i.d. random variables with common alphabet \( \{x_1, x_2, x_3, \cdots\} \), and assume that \( P_X(x_i) > 0 \) for every \( i \).
(a) Prove that the average code rate of the first-order (single-letter) binary Huffman code is equal to $H(X)$ iff $P_X(x_i)$ is equal to $2^{-n_i}$ for every $i$, where $\{n_i\}_{i \geq 1}$ is a sequence of positive integers.

*Hint:* The if-part can be proved by the new bound in Problem 11, and the only-if-part can be proved by modifying the proof of Theorem 3.17.

(b) What is the sufficient and necessary condition under which the average code rate of the first-order (single-letter) ternary Huffman code equals $H_3(X)$?

(c) Prove that the average code rate of the second-order (two-letter) binary Huffman code cannot be equal to $H(X) + 1/2$ bits?

*Hint:* Use the new bound in Problem 11.

13. Decide whether each of the following statements is true or false. Prove the validity of those that are true and give counterexamples or arguments based on known facts to disprove those that are false.

(a) Every Huffman code for a discrete memoryless source (DMS) has a corresponding suffix code with the same average code rate.

(b) Consider a DMS $\{X_n\}_{n=1}^\infty$ with alphabet $\mathcal{X} = \{a_1, a_2, a_3, a_4, a_5, a_6\}$ and probability distribution

$$[p_1, p_2, p_3, p_4, p_5, p_6] = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}\right],$$

where $p_i \triangleq \Pr\{X = a_i\}, i = 1, \cdots, 6$. The Shannon-Fano-Elias code $f : \mathcal{X} \to \{0, 1\}^*$ for the source is optimal.
Chapter 4

Data Transmission and Channel Capacity

4.1 Principles of data transmission

A noisy communication channel is an input-output medium in which the output is not completely or deterministically specified by the input. The channel is indeed stochastically modeled, where given channel input $x$, the channel output $y$ is governed by a transition (conditional) probability distribution denoted by $P_{Y|X}(y|x)$. Since two different inputs may give rise to the same output, the receiver, upon receipt of an output, needs to guess the most probable sent input. In general, words of length $n$ are sent and received over the channel; in this case, the channel is characterized by a sequence of $n$-dimensional transition distributions $P_{Y^n|X^n}(y^n|x^n)$, for $n = 1, 2, \cdots$. A block diagram depicting a data transmission or channel coding system (with no feedback\(^1\)) is given in Figure 4.1.

![Block Diagram](image)

**Figure 4.1:** A data transmission system, where $W$ represents the message for transmission, $X^n$ denotes the codeword corresponding to message $W$, $Y^n$ represents the received word due to channel input $X^n$, and $\hat{W}$ denotes the reconstructed message from $Y^n$.

---

\(^1\)The capacity of channels with (output) feedback will be studied in Part II of the book.
The designer of a data transmission (or channel) code needs to carefully select codewords from the set of channel input words (of a given length) so that a minimal ambiguity is obtained at the channel receiver. For example, suppose that a channel has binary input and output alphabets and that its transition probability distribution induces the following conditional probability on its output symbols given that input words of length 2 are sent:

\[
\begin{align*}
P_{Y|X^2}(y = 0|x^2 = 00) &= 1 \\
P_{Y|X^2}(y = 0|x^2 = 01) &= 1 \\
P_{Y|X^2}(y = 1|x^2 = 10) &= 1 \\
P_{Y|X^2}(y = 1|x^2 = 11) &= 1,
\end{align*}
\]

which can be graphically depicted as

```
00  1  0
\  \\
  01
  \  \\
10  1  1
\  \\
  11
```

and a binary message (either event $A$ or event $B$) is required to be transmitted from the sender to the receiver. Then the data transmission code with (codeword 00 for event $A$, codeword 10 for event $B$) obviously induces less ambiguity at the receiver than the code with (codeword 00 for event $A$, codeword 01 for event $B$).

In short, the objective in designing a data transmission (or channel) code is to transform a noisy channel into a reliable medium for sending messages and recovering them at the receiver with minimal loss. To achieve this goal, the designer of a data transmission code needs to take advantage of the common parts between the sender and the receiver sites that are least affected by the channel noise. We will see that these common parts are probabilistically captured by the mutual information between the channel input and the channel output.

As illustrated in the previous example, if a “least-noise-affected” subset of the channel input words is appropriately selected as the set of codewords, the messages intended to be transmitted can be reliably sent to the receiver with arbitrarily small error. One then raises the question:

*What is the maximum amount of information (per channel use) that can be reliably transmitted over a given noisy channel?*
In the above example, we can transmit a binary message error-free, and hence the amount of information that can be reliably transmitted is at least 1 bit per channel use (or channel symbol). It can be expected that the amount of information that can be reliably transmitted for a highly noisy channel should be less than that for a less noisy channel. But such a comparison requires a good measure of the “noisiness” of channels.

From an information theoretic viewpoint, “channel capacity” provides a good measure of the noisiness of a channel; it represents the maximal amount of informational messages (per channel use) that can be transmitted via a data transmission code over the channel and recovered with arbitrarily small probability of error at the receiver. In addition to its dependence on the channel transition distribution, channel capacity also depends on the coding constraint imposed on the channel input, such as “only block (fixed-length) codes are allowed.” In this chapter, we will study channel capacity for block codes (namely, only block transmission code can be used).\(^2\) Throughout the chapter, the noisy channel is assumed to be memoryless (as defined in the next section).

### 4.2 Discrete memoryless channels

**Definition 4.1 (Discrete channel)** A discrete communication channel is characterized by

- A finite input alphabet \(\mathcal{X}\).
- A finite output alphabet \(\mathcal{Y}\).
- A sequence of \(n\)-dimensional transition distributions

\[
\{P_{Y^n|X^n}(y^n|x^n)\}_{n=1}^{\infty}
\]

such that \(\sum_{y^n \in \mathcal{Y}^n} P_{Y^n|X^n}(y^n|x^n) = 1\) for every \(x^n \in \mathcal{X}^n\), where \(x^n = (x_1, \cdots, x_n) \in \mathcal{X}^n\) and \(y^n = (y_1, \cdots, y_n) \in \mathcal{Y}^n\). We assume that the above sequence of \(n\)-dimensional distribution is consistent, i.e.,

\[
P_{Y^i|X^i}(y^i|x^i) = \frac{\sum_{x_{i+1} \in \mathcal{X}} \sum_{y_{i+1} \in \mathcal{Y}} P_{X_{i+1}|X^i}(x_{i+1}) P_{Y_{i+1}|X_{i+1}}(y_{i+1}|x_{i+1}) P_{Y^i|X^i}(y^i|x^i)} {\sum_{x_{i+1} \in \mathcal{X}} P_{X_{i+1}|X^i}(x_{i+1})}
\]

\[
= \sum_{x_{i+1} \in \mathcal{X}} \sum_{y_{i+1} \in \mathcal{Y}} P_{X_{i+1}|X^i}(x_{i+1}|x^i) P_{Y_{i+1}|X_{i+1}}(y_{i+1}|x_{i+1})
\]

for every \(x^i, y^i, P_{X_{i+1}|X^i}\), and \(i = 1, 2, \cdots\).

\(^2\)See [48] for recent results regarding channel capacity when no coding constraints are applied on the channel input (so that variable-length codes can be employed).
In general, real-world communications channels exhibit statistical memory in the sense that current channel outputs statistically depend on past outputs as well as past, current and (possibly) future inputs. However, for the sake of simplicity, we restrict our attention in this chapter to the class of memoryless channels (channels with memory will later be treated in Part II).

**Definition 4.2 (Discrete memoryless channel)** A discrete memoryless channel (DMC) is a channel whose sequence of transition distributions \( P_{Y^n|X^n} \) satisfies

\[
P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i) \tag{4.2.1}
\]

for every \( n = 1, 2, \ldots, x^n \in X^n \) and \( y^n \in Y^n \). In other words, a DMC is fully described by the channel’s transition distribution matrix \( Q \triangleq [p_{x,y}] \) of size \( |X| \times |Y| \), where

\[
p_{x,y} \triangleq P_{Y|X}(y|x)
\]

for \( x \in X, y \in Y \). Furthermore, the matrix \( Q \) is stochastic; i.e., the sum of the entries in each of its rows is equal to 1 (since \( \sum_{y \in Y} p_{x,y} = 1 \) for all \( x \in X \)).

**Observation 4.3** We note that the DMC’s condition (4.2.1) is actually equivalent to the following two sets of conditions:

\[
\begin{align*}
P_{Y_n|X_n,Y_{n-1}}(y_n|x^n, y^{n-1}) &= P_{Y|X}(y_n|x_n) \quad \forall n = 1, 2, \ldots, x^n, y^n; \\
P_{Y_{n-1}|X_n}(y^{n-1}|x^n) &= P_{Y_{n-1}|X_{n-1}}(y^{n-1}|x^{n-1}) \quad \forall n = 2, 3, \ldots, x^n, y^{n-1}. 
\end{align*}
\]

(4.2.2a) (4.2.2b)

\[
\begin{align*}
P_{X_n|X_n,Y_{n-1}}(x_n|x^n, y^{n-1}) &= P_{Y|X}(y_n|x_n) \quad \forall n = 1, 2, \ldots, x^n, y^n; \\
P_{X_{n-1}|X_n,Y_{n-1}}(x_{n-1}|x^{n-1}, y^{n-1}) &= P_{X_{n-1}|X_{n-1}}(x_{n-1}|x^{n-1}) \quad \forall n = 1, 2, \ldots, x^n, y^{n-1}. 
\end{align*}
\]

(4.2.3a) (4.2.3b)

Condition (4.2.2a) (also, (4.2.3a)) implies that the current output \( Y_n \) only depends on the current input \( X_n \) but not on past inputs \( X^{n-1} \) and outputs \( Y^{n-1} \). Condition (4.2.2b) indicates that the past outputs \( Y^{n-1} \) do not depend on the current input \( X_n \). These two conditions together give

\[
P_{Y^n|X^n}(y^n|x^n) = P_{Y_{n-1}|X_n}(y^{n-1}|x^n) P_{Y_n|X_n,Y_{n-1}}(y_n|x^n, y^{n-1})
\]

\[
= P_{Y_{n-1}|X_{n-1}}(y^{n-1}|x^{n-1}) P_{Y|X}(y_n|x_n);
\]

hence, (4.2.1) holds recursively on \( n = 1, 2, \ldots \). The converse (i.e., (4.2.1) implies both (4.2.2a) and (4.2.2b)) is a direct consequence of

\[
P_{Y_n|X_n,Y_{n-1}}(y_n|x^n, y^{n-1}) = \frac{P_{Y^n|X^n}(y^n|x^n)}{\sum_{y_n \in Y} P_{Y^n|X^n}(y^n|x^n)}
\]
and

\[ P_{Y^n-1|X^n}(y^{n-1}|x^n) = \sum_{y_n \in Y} P_{Y^n|X^n}(y^n|x^n). \]

Similarly, (4.2.3b) states that the current input \( X_n \) is independent of past outputs \( Y^{n-1} \), which together with (4.2.3a) implies again

\[ P_{Y^n|X^n}(y^n|x^n) = P_{X^n,Y^n}(x^n,y^n) \]

\[ = \frac{P_{X^n,Y^n-1}(x^n,y^{n-1})P_{X^n,Y^n-1}(x^n,y^{n-1})P_{Y^n|X^n,Y^n-1}(y_n|x^n,y^{n-1})}{P_{X^n-1}(x^{n-1})P_{X^n|X^n-1}(x_n|x^{n-1})} \]

\[ = P_{Y^n-1|X^n-1}(y^{n-1}|x^{n-1})P_{Y|X}(y_n|x_n), \]

hence, recursively yielding (4.2.1). The converse for (4.2.3b)—i.e., (4.2.1) implying (4.2.3b)—can be analogously proved by noting that

\[ P_{X^n,Y^n-1}(x_n|x^{n-1},y^{n-1}) = \frac{P_{X^n}(x^n) \sum_{y_n \in Y} P_{Y^n|X^n}(y^n|x^n)}{P_{X^n-1}(x^{n-1})P_{Y^n-1|X^n-1}(y^{n-1}|x^{n-1})}. \]

Note that the above definition of DMC in (4.2.1) prohibits the use of channel feedback, as feedback allows the current channel input to be a function of past channel outputs (therefore, conditions (4.2.2b) and (4.2.3b) cannot hold with feedback). Instead, a causality condition generalizing condition (4.2.2a) (e.g., see Definition 7.4 in [53]) will be needed to define a channel with feedback (feedback will be considered in Part II of this book).

Examples of DMCs:

1. **Identity (noiseless) channels**: An identity channel has equal-size input and output alphabets (|\( \mathcal{X} \)| = |\( \mathcal{Y} \)|) and channel transition probability satisfying

\[ P_{Y|X}(y|x) = \begin{cases} 1 & \text{if } y = x \\ 0 & \text{if } y \neq x. \end{cases} \]

This is a noiseless or perfect channel as the channel input is received error-free at the channel output.

2. **Binary symmetric channels**: A binary symmetric channel (BSC) is a channel with binary input and output alphabets such that each input has a
(conditional) probability given by $\varepsilon$ for being received inverted at the output, where $\varepsilon \in [0, 1]$ is called the channel’s crossover probability or bit error rate. The channel’s transition distribution matrix is given by

$$Q = [p_{x,y}] = \begin{bmatrix} p_{0,0} & p_{0,1} \\ p_{1,0} & p_{1,1} \end{bmatrix} = \begin{bmatrix} P_{Y|X}(0|0) & P_{Y|X}(1|0) \\ P_{Y|X}(0|1) & P_{Y|X}(1|1) \end{bmatrix} = \begin{bmatrix} 1 - \varepsilon & \varepsilon \\ \varepsilon & 1 - \varepsilon \end{bmatrix}$$

and can be graphically represented via a transition diagram as shown in Figure 4.2.

![Figure 4.2: Binary symmetric channel.](image)

If we set $\varepsilon = 0$, then the BSC reduces to the binary identity (noiseless) channel. The channel is called “symmetric” since $P_{Y|X}(1|0) = P_{Y|X}(0|1)$; i.e., it has the same probability for flipping an input bit into a 0 or a 1. A detailed discussion of DMCs with various symmetry properties will be discussed at the end of this chapter.

Despite its simplicity, the BSC is rich enough to capture most of the complexity of coding problems over more general channels. For example, it can exactly model the behavior of practical channels with additive memoryless Gaussian noise used in conjunction of binary symmetric modulation and hard-decision demodulation (e.g., see [51, p. 240].) It is also worth pointing out that the BSC can be explicitly represented via a binary modulo-2 additive noise channel whose output at time $i$ is the modulo-2 sum of its input and noise variables:

$$Y_i = X_i \oplus Z_i \quad \text{for } i = 1, 2, \cdots$$
where $\oplus$ denotes addition modulo-2, $Y_i$, $X_i$ and $Z_i$ are the channel output, input and noise, respectively, at time $i$, the alphabets $X = Y = Z = \{0, 1\}$ are all binary, and it is assumed that $X_i$ and $Z_j$ are independent from each other for any $i, j = 1, 2, \cdots$, and that the noise process is a Bernoulli($\epsilon$) process – i.e., a binary i.i.d. process with $\Pr[Z = 1] = \epsilon$.

3. **Binary erasure channels:** In the BSC, some input bits are received perfectly and others are received corrupted (flipped) at the channel output. In some channels however, some input bits are lost during transmission instead of being received corrupted (for example, packets in data networks may get dropped or blocked due to congestion or bandwidth constraints). In this case, the receiver knows the exact location of these bits in the received bitstream or codeword, but not their actual value. Such bits are then declared as “erased” during transmission and are called “erasures.” This gives rise to the so-called binary erasure channel (BEC) as illustrated in Figure 4.3, with input alphabet $X = \{0, 1\}$ and output alphabet $Y = \{0, E, 1\}$, where $E$ represents an erasure, and channel transition matrix given by

$$Q = [p_{x,y}] = \begin{bmatrix} p_{0,0} & p_{0,E} & p_{0,1} \\ p_{1,0} & p_{1,E} & p_{1,1} \end{bmatrix} = \begin{bmatrix} P_{Y|X}(0|0) & P_{Y|X}(E|0) & P_{Y|X}(1|0) \\ P_{Y|X}(0|1) & P_{Y|X}(E|1) & P_{Y|X}(1|1) \end{bmatrix} = \begin{bmatrix} 1 - \alpha & \alpha & 0 \\ 0 & \alpha & 1 - \alpha \end{bmatrix}$$

(4.2.5)

where $0 \leq \alpha \leq 1$ is called the channel’s erasure probability.

4. **Binary channels with errors and erasures:** One can combine the BSC with the BEC to obtain a binary channel with both errors and erasures, as shown in Figure 4.4. We will call such channel the binary symmetric erasure channel (BSEC). In this case, the channel’s transition matrix is given by

$$Q = [p_{x,y}] = \begin{bmatrix} p_{0,0} & p_{0,E} & p_{0,1} \\ p_{1,0} & p_{1,E} & p_{1,1} \end{bmatrix} = \begin{bmatrix} 1 - \varepsilon - \alpha & \alpha & \varepsilon \\ \varepsilon & \alpha & 1 - \varepsilon - \alpha \end{bmatrix}$$

(4.2.6)

where $\varepsilon, \alpha \in [0, 1]$ are the channel’s crossover and erasure probabilities, respectively. Clearly, setting $\alpha = 0$ reduces the BSEC to the BSC, and setting $\varepsilon = 0$ reduces the BSEC to the BEC.

More generally, the channel needs not have a symmetric property in the sense of having identical transition distributions when inputs bits 0 or
1 are sent. For example, the channel’s transition matrix can be given by

\[
Q = [p_{x,y}] = \begin{bmatrix}
p_{0,0} & p_{0,E} & p_{0,1} \\
p_{1,0} & p_{1,E} & p_{1,1}
\end{bmatrix} = \begin{bmatrix}
1 - \varepsilon - \alpha & \alpha & \varepsilon \\
\varepsilon' & \alpha' & 1 - \varepsilon' - \alpha'
\end{bmatrix}
\]  

(4.2.7)

where the probabilities \( \varepsilon \neq \varepsilon' \) and \( \alpha \neq \alpha' \) in general. We call such channel, an asymmetric channel with errors and erasures (this model might be useful to represent practical channels using asymmetric or non-uniform modulation constellations).

5. \( q \)-ary symmetric channels: Given an integer \( q \geq 2 \), the \( q \)-ary symmetric channel is a non-binary extension of the BSC; it has alphabets \( \mathcal{X} = \mathcal{Y} = \)
\{0, 1, \cdots, q - 1\} of size \(q\) and channel transition matrix given by

\[
Q = [p_{x,y}]
= \begin{bmatrix}
p_{0,0} & p_{0,1} & \cdots & p_{0,q-1} \\
p_{1,0} & p_{1,1} & \cdots & p_{1,q-1} \\
\vdots & \vdots & \ddots & \vdots \\
p_{q-1,0} & p_{q-1,1} & \cdots & p_{q-1,q-1}
\end{bmatrix}
\]

(4.2.8)

where \(0 \leq \varepsilon \leq 1\) is the channel’s symbol error rate (or probability). When \(q = 2\), the channel reduces to the BSC with bit error rate \(\varepsilon\), as expected.

As the BSC, the \(q\)-ary symmetric channel can be expressed as a modulo-\(q\) additive noise channel with common input, output and noise alphabets \(X = Y = Z = \{0, 1, \cdots, q - 1\}\) and whose output \(Y_i\) at time \(i\) is given by \(Y_i = X_i \oplus_q Z_i\), for \(i = 1, 2, \cdots\), where \(\oplus_q\) denotes addition modulo-\(q\), and \(X_i\) and \(Z_i\) are the channel’s input and noise variables, respectively, at time \(i\). Here, the noise process \(\{Z_n\}_{n=1}^{\infty}\) is assumed to be an i.i.d. process with distribution

\[
\Pr[Z = 0] = 1 - \varepsilon \quad \text{and} \quad \Pr[Z = a] = \frac{\varepsilon}{q-1} \quad \forall a \in \{1, \cdots, q - 1\}.
\]

It is also assumed that the input and noise processes are independent from each other.

6. \(q\)-ary erasure channels: Given an integer \(q \geq 2\), one can also consider a non-binary extension of the BEC, yielding the so called \(q\)-ary erasure channel. Specifically, this channel has input and output alphabets given by \(\mathcal{X} = \mathcal{Y} = Z = \{0, 1, \cdots, q - 1\}\) and \(\mathcal{Y} = \{0, 1, \cdots, q - 1, E\}\), respectively, where \(E\) denotes an erasure, and channel transition distribution given by

\[
P_{Y|X}(y|x) = \begin{cases}
1 - \alpha & \text{if } y = x, x \in \mathcal{X} \\
\alpha & \text{if } y = E, x \in \mathcal{X} \\
0 & \text{if } y \neq x, x \in \mathcal{X}
\end{cases}
\]

(4.2.9)

where \(0 \leq \alpha \leq 1\) is the erasure probability. As expected, setting \(q = 2\) reduces the channel to the BEC.
4.3 Block codes for data transmission over DMCs

Definition 4.4 (Fixed-length data transmission code) Given positive integers \(n\) and \(M\), and a discrete channel with input alphabet \(X\) and output alphabet \(Y\), a fixed-length data transmission code (or block code) for this channel with blocklength \(n\) and rate \(\frac{1}{n} \log_2 M\) message bits per channel symbol (or channel use) is denoted by \(\mathcal{C}_n = (n, M)\) and consists of:

1. \(M\) information messages intended for transmission.
2. An encoding function
   \[
   f : \{1, 2, \ldots, M\} \to X^n
   \]
   yielding codewords \(f(1), f(2), \ldots, f(M) \in X^n\), each of length \(n\). The set of these \(M\) codewords is called the codebook and we also usually write \(\mathcal{C}_n = \{f(1), f(2), \ldots, f(M)\}\) to list the codewords.
3. A decoding function \(g : Y^n \to \{1, 2, \ldots, M\}\).

The set \(\{1, 2, \ldots, M\}\) is called the message set and we assume that a message \(W\) follows a uniform distribution over the set of messages: \(\Pr[W = w] = \frac{1}{M}\) for all \(w \in \{1, 2, \ldots, M\}\). A block diagram for the channel code is given at the beginning of this chapter; see Figure 4.1. As depicted in the diagram, to convey message \(W\) over the channel, the encoder sends its corresponding codeword \(X^n = f(W)\) at the channel input. Finally, \(Y^n\) is received at the channel output (according to the memoryless channel distribution \(P_{Y^n|X^n}\)) and the decoder yields \(\hat{W} = g(Y^n)\) as the message estimate.

Definition 4.5 (Average probability of error) The average probability of error for a channel block code \(\mathcal{C}_n = (n, M)\) code with encoder \(f(\cdot)\) and decoder \(g(\cdot)\) used over a channel with transition distribution \(P_{Y^n|X^n}\) is defined as

\[
P_e(\mathcal{C}_n) \triangleq \frac{1}{M} \sum_{w=1}^{M} \lambda_w(\mathcal{C}_n),
\]

where

\[
\lambda_w(\mathcal{C}_n) \triangleq \Pr[\hat{W} \neq W | W = w] = \Pr[g(Y^n) \neq w | X^n = f(w)]
\]

\[
= \sum_{y^n \in Y^n : g(y^n) \neq w} P_{Y^n|X^n}(y^n | f(w))
\]

is the code’s conditional probability of decoding error given that message \(w\) is sent over the channel.
Note that, since we have assumed that the message $W$ is drawn uniformly from the set of messages, we have that

$$P_e(C) = \Pr[\hat{W} \neq W].$$

**Observation 4.6** Another more conservative error criterion is the so-called *maximal probability of error*

$$\lambda(C) \triangleq \max_{w \in \{1, 2, \ldots, M\}} \lambda_w(C).$$

Clearly, $P_e(C) \leq \lambda(C)$; so one would expect that $P_e(C)$ behaves differently than $\lambda(C)$. However it can be shown that from a code $C = (n, M)$ with arbitrarily small $P_e(C)$, one can construct (by throwing away from $C$ half of its codewords with largest conditional probability of error) a code $C' = (n, \frac{M}{2})$ with arbitrarily small $\lambda(C')$ at essentially the same code rate as $n$ grows to infinity (e.g., see [13, p. 204], [53, p. 163]). Hence, we will only use $P_e(C)$ as our criterion when evaluating the “goodness” or reliability of channel block codes.

Our target is to find a good channel block code (or to show the existence of a good channel block code). From the perspective of the (weak) law of large numbers, a good choice is to draw the code’s codewords based on the *jointly typical* set between the input and the output of the channel, since all the probability mass is ultimately placed on the jointly typical set. The decoding failure then occurs only when the channel input-output pair does not lie in the jointly typical set, which implies that the probability of decoding error is ultimately small. We next define the jointly typical set.

**Definition 4.7 (Jointly typical set)** The set $\mathcal{F}_n(\delta)$ of jointly typical $n$-tuple pairs $(x^n, y^n)$ with respect to the memoryless distribution $P_{X^n,Y^n}(x^n, y^n) = \prod_{i=1}^n P_{X,Y}(x_i, y_i)$ is defined by

$$\mathcal{F}_n(\delta) \triangleq \left\{(x^n, y^n) \in X^n \times Y^n : \ldots \right\}.$$
\[
\left\{-\frac{1}{n} \log_2 P_{X^n}(x^n) - H(X) < \delta, \left\{-\frac{1}{n} \log_2 P_{Y^n}(y^n) - H(Y) < \delta, \right. \right. \\
\text{and } \left. \left. -\frac{1}{n} \log_2 P_{X^n,Y^n}(x^n,y^n) - H(X,Y) < \delta \right\}\right. .
\]

In short, a pair \((x^n, y^n)\) generated by independently drawing \(n\) times under \(P_{X,Y}\) is jointly \(\delta\)-typical if its joint and marginal empirical entropies are respectively \(\delta\)-close to the true joint and marginal entropies.

With the above definition, we directly obtain the joint AEP theorem.

**Theorem 4.8 (Joint AEP)** If \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), \ldots\) are i.i.d., i.e., \(\{(X_i, Y_i)\}_{i=1}^{\infty}\) is a dependent pair of DMSs, then

\[
-\frac{1}{n} \log_2 P_{X^n}(X_1, X_2, \ldots, X_n) \to H(X) \quad \text{in probability,}
\]

\[
-\frac{1}{n} \log_2 P_{Y^n}(Y_1, Y_2, \ldots, Y_n) \to H(Y) \quad \text{in probability,}
\]

and

\[
-\frac{1}{n} \log_2 P_{X^n,Y^n}((X_1, Y_1), \ldots, (X_n, Y_n)) \to H(X,Y) \quad \text{in probability}
\]

as \(n \to \infty\).

**Proof:** By the weak law of large numbers, we have the desired result. \(\square\)

**Theorem 4.9 (Shannon-McMillan theorem for pairs)** Given a dependent pair of DMSs with joint entropy \(H(X,Y)\) and any \(\delta\) greater than zero, we can choose \(n\) big enough so that the jointly \(\delta\)-typical set satisfies:

1. \(P_{X^n,Y^n}(F_n^c(\delta)) < \delta\) for sufficiently large \(n\).

2. The number of elements in \(F_n(\delta)\) is at least \((1-\delta)2^{n(H(X,Y)-\delta)}\) for sufficiently large \(n\), and at most \(2^{n(H(X,Y)+\delta)}\) for every \(n\).

3. If \((x^n, y^n) \in F_n(\delta)\), its probability of occurrence satisfies

\[
2^{-n(H(X,Y)+\delta)} < P_{X^n,Y^n}(x^n, y^n) < 2^{-n(H(X,Y)-\delta)} .
\]
Proof: The proof is quite similar to that of the Shannon-McMillan theorem for a single memoryless source presented in the previous chapter; we hence leave it as an exercise. \(\square\)

We herein arrive at the main result of this chapter, Shannon’s channel coding theorem for DMCs. It basically states that a quantity \(C\), termed as channel capacity and defined as the maximum of the channel’s mutual information over the set of its input distributions (see below), is the supremum of all “achievable” channel block code rates; i.e., it is the supremum of all rates for which there exists a sequence of block codes for the channel with asymptotically decaying (as the blocklength grows to infinity) probability of decoding error. In other words, for a given DMC, its capacity \(C\), which can be calculated by solely using the channel’s transition matrix \(Q\), constitutes the largest rate at which one can reliably transmit information via a block code over this channel. Thus, it is possible to communicate reliably over an inherently noisy DMC at a fixed rate (without decreasing it) as long as this rate is below \(C\) and the code’s blocklength is allowed to be large.

**Theorem 4.10 (Shannon’s channel coding theorem)** Consider a DMC with finite input alphabet \(X\), finite output alphabet \(Y\) and transition distribution probability \(P_{Y|X}(y|x)\), \(x \in X\) and \(y \in Y\). Define the channel capacity\(^5\)

\[
C \triangleq \max_{P_X} I(X;Y) = \max_{P_X} I(P_X, P_{Y|X})
\]

where the maximum is taken over all input distributions \(P_X\). Then the following hold.

- **Forward part (achievability)**: For any \(0 < \varepsilon < 1\), there exist \(\gamma > 0\) and a sequence of data transmission block codes \(\{C_n = (n, M_n)\}_{n=1}^{\infty}\) with

\[
\liminf_{n \to \infty} \frac{1}{n} \log_2 M_n \geq C - \gamma
\]

\(^5\)First note that the mutual information \(I(X;Y)\) is actually a function of the input statistics \(P_X\) and the channel statistics \(P_{Y|X}\). Hence, we may write it as

\[
I(P_X, P_{Y|X}) = \sum_{x \in X} \sum_{y \in Y} P_X(x) P_{Y|X}(y|x) \log_2 \frac{P_{Y|X}(y|x)}{\sum_{x' \in X} P_X(x') P_{Y|X}(y|x')}.
\]

Such an expression is more suitable for calculating the channel capacity.

Note also that the channel capacity \(C\) is well-defined since, for a fixed \(P_{Y|X}\), \(I(P_X, P_{Y|X})\) is concave and continuous in \(P_X\) (with respect to both the variational distance and the Euclidean distance (i.e., \(L_2\)-distance) \([53, \text{Chapter 2}]\)), and since the set of all input distributions \(P_X\) is a compact (closed and bounded) subset of \(\mathbb{R}^{|X|}\) due to the finiteness of \(X\). Hence there exists a \(P_X\) that achieves the supremum of the mutual information and the maximum is attainable.
and

\[ P_e(\mathcal{C}_n) < \varepsilon \quad \text{for sufficiently large } n, \]

where \( P_e(\mathcal{C}_n) \) denotes the (average) probability of error for block code \( \mathcal{C}_n \).

- **Converse part:** For any \( 0 < \varepsilon < 1 \), any sequence of data transmission block codes \( \{\mathcal{C}_n = (n, M_n)\}_{n=1}^{\infty} \) with

\[ \liminf_{n \to \infty} \frac{1}{n} \log_2 M_n > C \]

satisfies

\[ P_e(\mathcal{C}_n) > (1 - \epsilon)\mu \quad \text{for sufficiently large } n, \quad (4.3.1) \]

where

\[ \mu = 1 - \frac{C}{\liminf_{n \to \infty} \frac{1}{n} \log_2 M_n} > 0, \]

i.e., the codes’ probability of error is bounded away from zero for all \( n \) sufficiently large.\(^6\)

**Proof of the forward part:** It suffices to prove the existence of a good block code sequence (satisfying the rate condition, i.e., \( \liminf_{n \to \infty} (1/n) \log_2 M_n \geq C - \gamma \) for some \( \gamma > 0 \)) whose average error probability is ultimately less than \( \varepsilon \). Since the forward part holds trivially when \( C = 0 \) by setting \( M_n = 1 \), we assume in the sequel that \( C > 0 \).

We will use Shannon’s original random coding proof technique in which the good block code sequence is not deterministically constructed; instead, its existence is implicitly proven by showing that for a class (ensemble) of block code sequences \( \{\mathcal{C}_n\}_{n=1}^{\infty} \) and a code-selecting distribution \( \Pr[\mathcal{C}_n] \) over these block code sequences, the expectation value of the average error probability, evaluated under the code-selecting distribution on these block code sequences, can be made smaller than \( \varepsilon \) for \( n \) sufficiently large:

\[ E_{\mathcal{C}}[P_e(\mathcal{C}_n)] = \sum_{\mathcal{C}_n} \Pr[\mathcal{C}_n] P_e(\mathcal{C}_n) \to 0 \quad \text{as } n \to \infty. \]

Hence, there must exist at least one such a desired good code sequence \( \{\mathcal{C}^*_n\}_{n=1}^{\infty} \) among them (with \( P_e(\mathcal{C}^*_n) \to 0 \) as \( n \to \infty \)).

\(^6\)\((4.3.1)\) actually implies that \( \liminf_{n \to \infty} P_e(\mathcal{C}_n) \geq \lim_{\epsilon \downarrow 0} (1 - \epsilon)\mu = \mu \), where the error probability lower bound is nothing to do with \( \epsilon \). Here we state the converse of Theorem 4.10 in a form in parallel to the converse statements in Theorems 3.5 and 3.13.
Fix $\varepsilon \in (0, 1)$ and some $\gamma$ in $(0, \min\{4\varepsilon, C\})$. Observe that there exists $N_0$ such that for $n > N_0$, we can choose an integer $M_n$ with

$$C - \frac{\gamma}{2} \geq \frac{1}{n} \log_2 M_n > C - \gamma.$$  \hfill (4.3.2)

(Since we are only concerned with the case of “sufficient large $n$,” it suffices to consider only those $n$’s satisfying $n > N_0$, and ignore those $n$’s for $n \leq N_0$.)

Define $\delta \triangleq \gamma/8$. Let $\hat{P}_X$ be the probability distribution achieving the channel capacity:

$$C \triangleq \max_{P_X} I(P_X, P_{Y|X}) = I(\hat{P}_X, P_{Y|X}).$$

Denote by $P_{Y^n}$ the channel output distribution due to channel input product distribution $\hat{P}_{X^n}$ (with $\hat{P}_{X^n}(x^n) = \prod_{i=1}^n \hat{P}_X(x_i)$), i.e.,

$$P_{Y^n}(y^n) = \sum_{x^n \in \mathcal{X}^n} \hat{P}_{X^n, Y^n}(x^n, y^n)$$

where

$$\hat{P}_{X^n, Y^n}(x^n, y^n) \triangleq \hat{P}_{X^n}(x^n) P_{Y^n|X^n}(y^n|x^n)$$

for all $x^n \in \mathcal{X}^n$ and $y^n \in \mathcal{Y}^n$. Note that since $\hat{P}_{X^n}(x^n) = \prod_{i=1}^n \hat{P}_X(x_i)$ and the channel is memoryless, the resulting joint input-output process $\{(\hat{X}_i, \hat{Y}_i)\}_{i=1}^\infty$ is also memoryless with

$$\hat{P}_{X^n, Y^n}(x^n, y^n) = \prod_{i=1}^n \hat{P}_{X, Y}(x_i, y_i)$$

and

$$\hat{P}_{X, Y}(x, y) = \hat{P}_{X}(x) P_{Y|X}(y|x) \quad \text{for} \quad x \in \mathcal{X}, y \in \mathcal{Y}.$$

We next present the proof in three steps.

**Step 1: Code construction.**

For any blocklength $n$, independently select $M_n$ channel inputs with replacement\(^7\) from $\mathcal{X}^n$ according to the distribution $\hat{P}_{X^n}(x^n)$. For the selected $M_n$ channel inputs yielding codebook $\mathcal{C}_n \triangleq \{c_1, c_2, \ldots, c_{M_n}\}$, define the encoder $f_n(\cdot)$ and decoder $g_n(\cdot)$, respectively, as follows:

$$f_n(m) = c_m \quad \text{for} \quad 1 \leq m \leq M_n,$$

\(^7\)Here, the channel inputs are selected with replacement. That means it is possible and acceptable that all the selected $M_n$ channel inputs are identical.
and
\[
g_n(y^n) = \begin{cases} 
m, & \text{if } c_m \text{ is the only codeword in } C_n \text{ satisfying } (c_m, y^n) \in \mathcal{F}_n(\delta); \\
\text{any one in } \{1, 2, \ldots, M_n\}, & \text{otherwise,}
\end{cases}
\]
where \(\mathcal{F}_n(\delta)\) is defined in Definition 4.7 with respect to distribution \(P_{X^n, Y^n}\).
(We evidently assume that the codebook \(C_n\) and the channel distribution \(P_{Y|X}\) are known at both the encoder and the decoder.) Hence, the code \(C_n\) operates as follows. A message \(W\) is chosen according to the uniform distribution from the set of messages. The encoder \(f_n\) then transmits the \(W\)th codeword \(c_W\) in \(C_n\) over the channel. Then \(Y^n\) is received at the channel output and the decoder guesses the sent message via \(\hat{W} = g_n(Y^n)\).

Note that there is a total \(|X|^n M_n\) possible randomly generated codebooks \(C_n\) and the probability of selecting each codebook is given by
\[
\Pr[C_n] = \prod_{m=1}^{M_n} P_{X^n}(c_m).
\]

**Step 2: Conditional error probability.**

For each (randomly generated) data transmission code \(C_n\), the conditional probability of error given that message \(m\) was sent, \(\lambda_m(C_n)\), can be upper bounded by:
\[
\lambda_m(C_n) \leq \sum_{\substack{y^n \in Y^n: (c_m, y^n) \not\in \mathcal{F}_n(\delta)}} P_{Y^n|X^n}(y^n|c_m)
+ \sum_{m' = 1}^{M_n} \sum_{\substack{y^n \in Y^n: (c_{m'}, y^n) \in \mathcal{F}_n(\delta) \\ m' \neq m}} P_{Y^n|X^n}(y^n|c_m)P_{Y^n|X^n}(y^n|c_{m'}),
\]
where the first term in (4.3.3) considers the case that the received channel output \(y^n\) is not jointly \(\delta\)-typical with \(c_m\), (and hence, the decoding rule \(g_n(\cdot)\) would possibly result in a wrong guess), and the second term in (4.3.3) reflects the situation when \(y^n\) is jointly \(\delta\)-typical with not only the transmitted codeword \(c_m\), but also with another codeword \(c_{m'}\) (which may cause a decoding error).

By taking expectation in (4.3.3) with respect to the \(m^{th}\) codeword-selecting distribution \(P_{X^n}(c_m)\), we obtain
\[
\sum_{c_m \in X^n} P_{X^n}(c_m) \lambda_m(C_n) \leq \sum_{c_m \in X^n} \sum_{y^n \not\in \mathcal{F}_n(\delta|c_m)} P_{X^n}(c_m)P_{Y^n|X^n}(y^n|c_m)
\]

105
+ \sum_{c_m \in X^n} \sum_{m' = 1}^{M_n} P_{X^n}(c_m) P_{Y^n|X^n}(y^n|c_m)

= P_{X^n,Y^n}(\mathcal{F}_n^c(\delta))

+ \sum_{m' = 1}^{M_n} \sum_{c_m \in X^n} \sum_{y^n \in \mathcal{F}_n(\delta|c_m')} P_{X^n,Y^n}(c_m, y^n),

(4.3.4)

where

\[ \mathcal{F}_n(\delta|x^n) \triangleq \{ y^n : (x^n, y^n) \in \mathcal{F}_n(\delta) \}. \]

**Step 3: Average error probability.**

We now can analyze the expectation of the average error probability

\[ E_{\mathcal{C}_n}[P_e(\mathcal{C}_n)] \]

over the ensemble of all codebooks \( \mathcal{C}_n \) generated at random according to \( \text{Pr}[\mathcal{C}_n] \) and show that it asymptotically vanishes as \( n \) grows without bound. We obtain the following series of inequalities.

\[
E_{\mathcal{C}_n}[P_e(\mathcal{C}_n)] = \sum_{\mathcal{C}_n} \text{Pr}[\mathcal{C}_n] P_e(\mathcal{C}_n)
\]

\[
= \sum_{c_1 \in X^n} \cdots \sum_{c_{M_n} \in X^n} P_{X^n}(c_1) \cdots P_{X^n}(c_{M_n}) \left( \frac{1}{M_n} \sum_{m=1}^{M_n} \lambda_m(\mathcal{C}_n) \right)
\]

\[
= \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{c_1 \in X^n} \cdots \sum_{c_{M_n} \in X^n} \sum_{c_{m-1} \in X^n} \sum_{c_{m+1} \in X^n} \lambda_m(\mathcal{C}_n)
\]

\[
P_{X^n}(c_1) \cdots P_{X^n}(c_{m-1}) P_{X^n}(c_{m+1}) \cdots P_{X^n}(c_{M_n})
\]

\[
\times \left( \sum_{c_m \in X^n} P_{X^n}(c_m) \lambda_m(\mathcal{C}_n) \right)
\]

\[
\leq \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{c_1 \in X^n} \cdots \sum_{c_{M_n} \in X^n} \sum_{c_{m-1} \in X^n} \sum_{c_{m+1} \in X^n} \lambda_m(\mathcal{C}_n)
\]

\[
P_{X^n}(c_1) \cdots P_{X^n}(c_{m-1}) P_{X^n}(c_{m+1}) \cdots P_{X^n}(c_{M_n})
\]

\[
\times P_{X^n,Y^n}(\mathcal{F}_n^c(\delta))
\]

\[
+ \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{c_1 \in X^n} \cdots \sum_{c_{M_n} \in X^n} \sum_{c_{m-1} \in X^n} \sum_{c_{m+1} \in X^n} \lambda_m(\mathcal{C}_n)
\]

\[
P_{X^n}(c_1) \cdots P_{X^n}(c_{m-1}) P_{X^n}(c_{m+1}) \cdots P_{X^n}(c_{M_n})
\]

\[
\times \sum_{c_m \in X^n} \lambda_m(\mathcal{C}_n)
\]

\[
= 1
\]

106
where (4.3.5) follows from (4.3.4), and the last step holds since $P_{X_n, Y_n}(F_n^c(\delta))$ is a constant independent of $c_1, \ldots, c_M$, and $m$. Observe that for $n > N_0$,
\[
\leq \sum_{m' = 1}^{M_n} |\mathcal{F}_n(\delta)| 2^{-n(H(\hat{X}) - \delta)} 2^{-n(H(\hat{Y}) - \delta)}
\]
\[
\leq \sum_{m' = 1}^{M_n} 2^{n(H(\hat{X}, \hat{Y}) + \delta)} 2^{-n(H(\hat{X}) - \delta)} 2^{-n(H(\hat{Y}) - \delta)}
\]
\[
= (M_n - 1) 2^{n(H(\hat{X}, \hat{Y}) + \delta)} 2^{-n(H(\hat{X}) - \delta)} 2^{-n(H(\hat{Y}) - \delta)}
\]
\[
< M_n \cdot 2^{n(H(\hat{X}, \hat{Y}) + \delta)} 2^{-n(H(\hat{X}) - \delta)} 2^{-n(H(\hat{Y}) - \delta)}
\]
\[
\leq 2^{n(C - 4\delta)} \cdot 2^{-n(I(\hat{X}; \hat{Y}) - 3\delta)} = 2^{-n\delta},
\]

where the first inequality follows from the definition of the jointly typical set \(\mathcal{F}_n(\delta)\), the second inequality holds by the Shannon-McMillan theorem for pairs (Theorem 4.9), the last inequality follows since \(C = I(\hat{X}; \hat{Y})\) by definition of \(\hat{X}\) and \(\hat{Y}\), and since \((1/n) \log_2 M_n \leq C - (\gamma/2) = C - 4\delta\). Consequently,

\[
E_{\mathcal{C}_n}[P_e(\mathcal{C}_n)] \leq P_{X^n, Y^n}(\mathcal{F}_n^c(\delta)) + 2^{-n\delta},
\]

which for sufficiently large \(n\) (and \(n > N_0\)), can be made smaller than \(2\delta = \gamma/4 < \varepsilon\) by the Shannon-McMillan theorem for pairs.

Before proving the converse part of the channel coding theorem, let us recall Fano’s inequality in a channel coding context. Consider an \((n, M_n)\) channel block code \(\mathcal{C}_n\) with encoding and decoding functions given by

\[
f_n : \{1, 2, \cdots, M_n\} \rightarrow X^n
\]

and

\[
g_n : Y^n \rightarrow \{1, 2, \cdots, M_n\},
\]

respectively. Let message \(W\), which is uniformly distributed over the set of messages \(\{1, 2, \cdots, M_n\}\), be sent via codeword \(X^n(W) = f_n(W)\) over the DMC, and let \(Y^n\) be received at the channel output. At the receiver, the decoder estimates the sent message via \(\hat{W} = g_n(Y^n)\) and the probability of estimation error is given by the code’s average error probability:

\[
\Pr[W \neq \hat{W}] = P_e(\mathcal{C}_n)
\]

since \(W\) is uniformly distributed. Then Fano’s inequality (2.5.2) yields

\[
H(W|Y^n) \leq 1 + P_e(\mathcal{C}_n) \log_2 (M_n - 1)
\]
\[
< 1 + P_e(\mathcal{C}_n) \log_2 M_n.
\]

(4.3.6)
We next proceed with the proof of the converse part.

**Proof of the converse part:** For any \((n, M_n)\) block channel code \(\mathcal{C}_n\) as described above, we have that \(W \rightarrow X^n \rightarrow Y^n\) form a Markov chain; we thus obtain by the data processing inequality that

\[
I(W; Y^n) \leq I(X^n; Y^n). \tag{4.3.7}
\]

We can also upper bound \(I(X^n; Y^n)\) in terms of the channel capacity \(C\) as follows

\[
I(X^n; Y^n) \leq \max_{P_{X^n}} I(X^n; Y^n) \leq \sum_{i=1}^{n} \max_{P_{X_i}} I(X_i; Y_i) \leq nC. \tag{4.3.8}
\]

Consequently, code \(\mathcal{C}_n\) satisfies the following:

\[
\log_2 M_n = H(W) \quad \text{(since } W \text{ is uniformly distributed)}
= H(W|Y^n) + I(W; Y^n)
\leq H(W|Y^n) + I(X^n; Y^n) \quad \text{(by (4.3.7))}
\leq H(W|Y^n) + nC \quad \text{(by (4.3.8))}
< 1 + P_e(\mathcal{C}_n) \cdot \log_2 M_n + nC. \quad \text{(by (4.3.6))}
\]

This implies that

\[
P_e(\mathcal{C}_n) > 1 - \frac{C}{(1/n) \log_2 M_n} - \frac{1}{\log_2 M_n} = 1 - \frac{C + 1/n}{(1/n) \log_2 M_n}.
\]

So if \(\liminf_{n \to \infty} (1/n) \log_2 M_n = \frac{C}{1 - \mu}\), then for any \(0 < \varepsilon < 1\), there exists an integer \(N\) such that for \(n \geq N\),

\[
\frac{1}{n} \log_2 M_n \geq \frac{C + 1/n}{1 - (1 - \varepsilon)\mu}. \tag{4.3.9}
\]

because, otherwise, (4.3.9) would be violated for infinitely many \(n\), implying a contradiction that

\[
\liminf_{n \to \infty} \frac{1}{n} \log_2 M_n \leq \liminf_{n \to \infty} \frac{C + 1/n}{1 - (1 - \varepsilon)\mu} = \frac{C}{1 - (1 - \varepsilon)\mu}.
\]

109
Hence, for \( n \geq N \),
\[
P_e(\mathcal{E}_n) > 1 - [1 - (1 - \varepsilon)\mu] \frac{C + 1/n}{C + 1/n} = (1 - \varepsilon)\mu > 0;
\]
i.e., \( P_e(\mathcal{E}_n) \) is bounded away from zero for \( n \) sufficiently large. \( \square \)

**Observation 4.11** The results of the above channel coding theorem is illustrated in Figure 4.5,\(^8\) where \( R = \liminf_{n \to \infty} (1/n) \log_2 M_n \) (measured in message bits/channel use) is usually called the *ultimate* (or *asymptotic*) coding rate of channel block codes. As indicated in the figure, the ultimate rate of any good block code for the DMC must be smaller than or equal to the channel capacity \( C \).\(^9\) Conversely, any block code with (ultimate) rate greater than \( C \), will have its probability of error bounded away from zero. Thus for a DMC, its capacity \( C \) is the supremum of all “achievable” channel block coding rates; i.e., it is the supremum of all rates for which there exists a sequence of channel block codes with asymptotically vanishing (as the blocklength goes to infinity) probability of error.

\[
\begin{array}{c|c}
\limsup_{n \to \infty} P_e = 0 & \limsup_{n \to \infty} P_e > 0 \\
for the best channel block code & for all channel block codes \\
\hline
C & R
\end{array}
\]

Figure 4.5: Ultimate channel coding rate \( R \) versus channel capacity \( C \) and behavior of the probability of error as blocklength \( n \) goes to infinity for a DMC.

Shannon’s channel coding theorem, established in 1948 [43], provides the ultimate limit for reliable communication over a noisy channel. However, it does not provide an explicit efficient construction for good codes since searching for

---

\(^8\)Note that Theorem 4.10 actually indicates \( \lim_{n \to \infty} P_e = 0 \) for \( R < C \) and \( \liminf_{n \to \infty} P_e > 0 \) for \( R > C \) rather than the behavior of \( \limsup_{n \to \infty} P_e \) indicated in Figure 4.5; these however only holds for a DMC. For a more general channel, three partitions instead of two may result, i.e., \( R < C, C < R < \bar{C} \) and \( R > \bar{C} \), which respectively correspond to \( \limsup_{n \to \infty} P_e = 0 \) for the best block code, \( \limsup_{n \to \infty} P_e > 0 \) but \( \liminf_{n \to \infty} P_e = 0 \) for the best block code, and \( \liminf_{n \to \infty} P_e > 0 \) for all channel code codes, where \( \bar{C} \) is named the optimistic channel capacity. Since \( \bar{C} = C \) for DMCs, the three regions are reduced to two. \( \bar{C} \) will later be introduced in Part II.

\(^9\)It can be seen from the theorem that \( C \) can be achieved as an ultimate transmission rate as long as \((1/n) \log_2 M_n\) approaches \( C \) from below with increasing \( n \) (see (4.3.2)).
a good code from the ensemble of randomly generated codes is prohibitively complex, as its size grows double-exponentially with blocklength (see Step 1 of the proof of the forward part). It thus spurred the entire area of coding theory, which flourished over the last 65 years with the aim of constructing powerful error-correcting codes operating close to the capacity limit. Particular advances were made for the class of linear codes (also known as group codes) whose rich yet elegantly simple algebraic structures made them amenable for efficient practically-implementable encoding and decoding. Examples of such codes include Hamming codes, Golay codes, BCH and Reed-Solomon codes, and convolutional codes. In 1993, the so-called Turbo codes were introduced by Berrou et al. [3, 4] and shown experimentally to perform close to the channel capacity limit for the class of memoryless channels. Similar near-capacity achieving linear codes were later established with the re-discovery of Gallager’s low-density parity-check codes [17, 18, 33, 34]. Many of these codes are used with increased sophistication in today’s ubiquitous communication, information and multimedia technologies. For detailed studies on coding theory, see the following texts [8, 10, 26, 31, 35, 41, 51].

4.4 Calculating channel capacity

Given a DMC with finite input alphabet $\mathcal{X}$, finite input alphabet $\mathcal{Y}$ and channel transition matrix $Q = [p_{x,y}]$ of size $|\mathcal{X}| \times |\mathcal{Y}|$, where $p_{x,y} \triangleq P_{Y|X}(y|x)$, for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we would like to calculate

$$C \triangleq \max_{P_X} I(X; Y)$$

where the maximization (which is well-defined) is carried over the set of input distributions $P_X$, and $I(X; Y)$ is the mutual information between the channel’s input and output.

Note that $C$ can be determined numerically via non-linear optimization techniques – such as the iterative algorithms developed by Arimoto [1] and Blahut [7, 9], see also [15] and [53, Chap. 9]. In general, there are no closed-form (single-letter) analytical expressions for $C$. However, for many “simplified” channels, it is possible to analytically determine $C$ under some “symmetry” properties of their channel transition matrix.

\[\text{Indeed, there exist linear codes that can achieve the capacity of memoryless channels with additive noise (e.g., see [14, p. 114]). Such channels include the BSC and the } q\text{-ary symmetric channel.}\]
4.4.1 Symmetric, weakly-symmetric and quasi-symmetric channels

**Definition 4.12** A DMC with finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$ and channel transition matrix $Q = [p_{x,y}]$ of size $|\mathcal{X}| \times |\mathcal{Y}|$ is said to be symmetric if the rows of $Q$ are permutations of each other and the columns of $Q$ are permutations of each other. The channel is said to be weakly-symmetric if the rows of $Q$ are permutations of each other and all the column sums in $Q$ are equal.

It directly follows from the definition that symmetry implies weak-symmetry. Examples of symmetric DMCs include the BSC, the $q$-ary symmetric channel and the following ternary channel with $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$ and transition matrix

$$Q = \begin{bmatrix}
P_{Y|X}(0|0) & P_{Y|X}(1|0) & P_{Y|X}(2|0) \\
P_{Y|X}(0|1) & P_{Y|X}(1|1) & P_{Y|X}(2|1) \\
P_{Y|X}(0|2) & P_{Y|X}(1|2) & P_{Y|X}(2|2)
\end{bmatrix} = \begin{bmatrix}
0.4 & 0.1 & 0.5 \\
0.5 & 0.4 & 0.1 \\
0.1 & 0.5 & 0.4
\end{bmatrix}.$$

The following DMC with $|\mathcal{X}| = |\mathcal{Y}| = 4$ and

$$Q = \begin{bmatrix}
0.5 & 0.25 & 0.25 & 0 \\
0.5 & 0.25 & 0.25 & 0 \\
0 & 0.25 & 0.25 & 0.5 \\
0 & 0.25 & 0.25 & 0.5
\end{bmatrix} \quad (4.4.1)$$

is weakly-symmetric (but not symmetric). Noting that all above channels involve square transition matrices, we emphasize that $Q$ can be rectangular while satisfying the symmetry or weak-symmetry properties. For example, the DMC with $|\mathcal{X}| = 2, |\mathcal{Y}| = 4$ and

$$Q = \begin{bmatrix}
\frac{1-\varepsilon}{2} & \frac{\varepsilon}{2} & \frac{1-\varepsilon}{2} & \frac{\varepsilon}{2} \\
\frac{\varepsilon}{2} & \frac{1-\varepsilon}{2} & \frac{\varepsilon}{2} & \frac{1-\varepsilon}{2}
\end{bmatrix} \quad (4.4.2)$$

is symmetric (where $\varepsilon \in [0, 1]$), while the DMC with $|\mathcal{X}| = 2, |\mathcal{Y}| = 3$ and

$$Q = \begin{bmatrix}
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3}
\end{bmatrix}$$

is weakly-symmetric.

**Lemma 4.13** The capacity of a weakly-symmetric channel $Q$ is achieved by a uniform input distribution and is given by

$$C = \log_2 |\mathcal{Y}| - H(q_1, q_2, \cdots, q_{|\mathcal{Y}|}) \quad (4.4.3)$$
where \((q_1, q_2, \ldots, q_{|Y|})\) denotes any row of \(Q\) and

\[
H(q_1, q_2, \ldots, q_{|Y|}) \triangleq -\sum_{i=1}^{|Y|} q_i \log_2 q_i
\]

is the row entropy.

**Proof:** The mutual information between the channel’s input and output is given by

\[
I(X;Y) = H(Y) - H(Y|X)
\]

\[
= H(Y) - \sum_{x \in \mathcal{X}} P_X(x) H(Y|X = x)
\]

where \(H(Y|X = x) = -\sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \log_2 P_{Y|X}(y|x) = -\sum_{y \in \mathcal{Y}} p_{x,y} \log_2 p_{x,y}.\)

Noting that every row of \(Q\) is a permutation of every other row, we obtain that \(H(Y|X = x)\) is independent of \(x\) and can be written as

\[
H(Y|X = x) = H(q_1, q_2, \ldots, q_{|Y|})
\]

where \((q_1, q_2, \ldots, q_{|Y|})\) is any row of \(Q\). Thus

\[
H(Y|X) = \sum_{x \in \mathcal{X}} P_X(x) H(q_1, q_2, \ldots, q_{|Y|})
\]

\[
= H(q_1, q_2, \ldots, q_{|Y|}) \left( \sum_{x \in \mathcal{X}} P_X(x) \right)
\]

\[
= H(q_1, q_2, \ldots, q_{|Y|}).
\]

This implies

\[
I(X;Y) = H(Y) - H(q_1, q_2, \ldots, q_{|Y|})
\]

\[
\leq \log_2 |\mathcal{Y}| - H(q_1, q_2, \ldots, q_{|Y|})
\]

with equality achieved iff \(Y\) is uniformly distributed over \(\mathcal{Y}\). We next show that choosing a uniform input distribution, \(P_X(x) = \frac{1}{|\mathcal{X}|} \forall x \in \mathcal{X}\), yields a uniform output distribution, hence maximizing mutual information. Indeed, under a uniform input distribution, we obtain that for any \(y \in \mathcal{Y},\)

\[
P_Y(y) = \sum_{x \in \mathcal{X}} P_X(x) P_{Y|X}(y|x) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} p_{x,y} = \frac{A}{|\mathcal{X}|}
\]

113
where \( A \triangleq \sum_{x \in \mathcal{X}} p_{x,y} \) is a constant given by the sum of the entries in any column of \( Q \), since by the weak-symmetry property all column sums in \( Q \) are identical. Note that \( \sum_{y \in \mathcal{Y}} P_Y(y) = 1 \) yields that

\[
\sum_{y \in \mathcal{Y}} \frac{A}{|\mathcal{Y}|} = 1
\]

and hence

\[
A = \frac{|\mathcal{X}|}{|\mathcal{Y}|} \quad (4.4.4)
\]

Accordingly,

\[
P_Y(y) = \frac{A}{|\mathcal{X}|} \frac{1}{|\mathcal{Y}|} = \frac{1}{|\mathcal{Y}|}
\]

for any \( y \in \mathcal{Y} \); thus the uniform input distribution induces a uniform output distribution and achieves channel capacity as given by (4.4.3).

Observation 4.14 Note that if the weakly-symmetric channel has a square (i.e., with \( |\mathcal{X}| = |\mathcal{Y}| \)) transition matrix \( Q \), then \( Q \) is a doubly-stochastic matrix; i.e., both its row sums and its column sums are equal to 1. Note however that having a square transition matrix does not necessarily make a weakly-symmetric channel symmetric; e.g., see (4.4.1).

Example 4.15 (Capacity of the BSC) Since the BSC with crossover probability (or bit error rate) \( \varepsilon \) is symmetric, we directly obtain from Lemma 4.13 that its capacity is achieved by a uniform input distribution and is given by

\[
C = \log_2(2) - H(1 - \varepsilon, \varepsilon) = 1 - h_b(\varepsilon) \quad (4.4.5)
\]

where \( h_b(\cdot) \) is the binary entropy function.

Example 4.16 (Capacity of the \( q \)-ary symmetric channel) Similarly, the \( q \)-ary symmetric channel with symbol error rate \( \varepsilon \) described in (4.2.8) is symmetric; hence, by Lemma 4.13, its capacity is given by

\[
C = \log_2 q - H \left( 1 - \varepsilon, \frac{\varepsilon}{q-1}, \ldots, \frac{\varepsilon}{q-1} \right) = \log_2 q + \varepsilon \log_2 \frac{\varepsilon}{q-1} + (1-\varepsilon) \log_2 (1-\varepsilon)
\]

Note that when \( q = 2 \), the channel capacity is equal to that of the BSC, as expected. Furthermore, when \( \varepsilon = 0 \), the channel reduces to the identity (noiseless) \( q \)-ary channel and its capacity is given by \( C = \log_2 q \).

We next note that one can further weaken the weak-symmetry property and define a class of “quasi-symmetric” channels for which the uniform input distribution still achieves capacity and yields a simple closed-form formula for capacity.
Definition 4.17 A DMC with finite input alphabet $\mathcal{X}$, finite output alphabet $\mathcal{Y}$ and channel transition matrix $Q = \left[p_{x,y}\right]$ of size $|\mathcal{X}| \times |\mathcal{Y}|$ is said to be quasi-symmetric\footnote{This notion of “quasi-symmetry” is slightly more general than Gallager’s notion [19, p. 94], as we herein allow each sub-matrix to be weakly-symmetric (instead of symmetric as in [19]).} if $Q$ can be partitioned along its columns into $m$ weakly-symmetric sub-matrices $Q_1, Q_2, \cdots, Q_m$ for some integer $m \geq 1$, where each $Q_i$ sub-matrix has size $|\mathcal{X}| \times |\mathcal{Y}_i|$ for $i = 1, 2, \cdots, m$ with $\mathcal{Y}_1 \cup \cdots \cup \mathcal{Y}_m = \mathcal{Y}$ and $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset \ \forall i \neq j, i, j = 1, 2, \cdots, m$.

Hence, quasi-symmetry is our weakest symmetry notion, since a weakly-symmetric channel is clearly quasi-symmetric (just set $m = 1$ in the above definition); we thus have: symmetry $\Rightarrow$ weak-symmetry $\Rightarrow$ quasi-symmetry.

Lemma 4.18 The capacity of a quasi-symmetric channel $Q$ as defined above is achieved by a uniform input distribution and is given by

$$C = \sum_{i=1}^{m} a_i C_i \quad (4.4.6)$$

where

$$a_i \triangleq \sum_{y \in \mathcal{Y}_i} p_{x,y} = \text{sum of any row in } Q_i, \quad i = 1, \cdots, m,$$

and

$$C_i = \log_2 |\mathcal{Y}_i| - H\left(\text{any row in the matrix } \frac{1}{a_i} Q_i\right), \quad i = 1, \cdots, m$$

is the capacity of the $i$th weakly-symmetric “sub-channel” whose transition matrix is obtained by multiplying each entry of $Q_i$ by $\frac{1}{a_i}$ (this normalization renders sub-matrix $Q_i$ into a stochastic matrix and hence a channel transition matrix).

Proof: We first observe that for each $i = 1, \cdots, m$, $a_i$ is independent of the input value $x$, since sub-matrix $i$ is weakly-symmetric (so any row in $Q_i$ is a permutation of any other row); and hence $a_i$ is the sum of any row in $Q_i$.

For each $i = 1, \cdots, m$, define

$$P_{Y_i|X}(y|x) \triangleq \begin{cases} \frac{p_{x,y}}{a_i} & \text{if } y \in \mathcal{Y}_i \text{ and } x \in \mathcal{X}; \\ 0 & \text{otherwise} \end{cases}$$

where $Y_i$ is a random variable taking values in $\mathcal{Y}_i$. It can be easily verified that $P_{Y_i|X}(y|x)$ is a legitimate conditional distribution. Thus $[P_{Y_i|X}(y|x)] = \frac{1}{a_i} Q_i$ is the transition matrix of the weakly-symmetric “sub-channel” $i$ with input...
alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}_i$. Let $I(X; Y_i)$ denote its mutual information. Since each such sub-channel $i$ is weakly-symmetric, we know that its capacity $C_i$ is given by

$$C_i = \max_{P_X} I(X; Y_i) = \log_2 |\mathcal{Y}_i| - H \left( \text{any row in the matrix } \frac{1}{a_i} \mathbb{Q}_i \right),$$

where the maximum is achieved by a uniform input distribution.

Now, the mutual information between the input and the output of our original quasi-symmetric channel $\mathbb{Q}$ can be written as

$$I(X; Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P_X(x) \frac{p_{x,y}}{\sum_{x' \in \mathcal{X}} P_X(x')} \log_2 \frac{p_{x,y}}{\sum_{x' \in \mathcal{X}} P_X(x') p_{x',y}}$$

$$= \sum_{i=1}^m \sum_{y \in \mathcal{Y}_i} \sum_{x \in \mathcal{X}} a_i P_X(x) \frac{p_{x,y}}{a_i} \log_2 \frac{p_{x,y}}{\sum_{x' \in \mathcal{X}} P_X(x') p_{x',y}}$$

$$= \sum_{i=1}^m a_i \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P_X(x) P_{Y_i|X}(y|x) \log_2 \frac{P_{Y_i|X}(y|x)}{\sum_{x' \in \mathcal{X}} P_X(x') P_{Y_i|X}(y|x')}$$

$$= \sum_{i=1}^m a_i I(X; Y_i).$$

Therefore, the capacity of channel $\mathbb{Q}$ is

$$C = \max_{P_X} I(X; Y)$$

$$= \max_{P_X} \sum_{i=1}^m a_i I(X; Y_i)$$

$$= \sum_{i=1}^m a_i \max_{P_X} I(X; Y_i) \quad \text{(as the same uniform } P_X \text{ maximizes each } I(X; Y_i))$$

$$= \sum_{i=1}^m a_i C_i.$$

\[ \square \]

**Example 4.19 (Capacity of the BEC)** The BEC with erasure probability $\alpha$ as given in (4.2.5) is quasi-symmetric (but neither weakly-symmetric nor symmetric). Indeed, its transition matrix $\mathbb{Q}$ can be partitioned along its columns into two symmetric (hence weakly-symmetric) sub-matrices

$$\mathbb{Q}_1 = \begin{bmatrix} 1 - \alpha & 0 \\ 0 & 1 - \alpha \end{bmatrix}$$

116
and
\[ Q_2 = \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}. \]

Thus applying the capacity formula for quasi-symmetric channels of Lemma 4.18 yields that the capacity of the BEC is given by
\[ C = a_1 C_1 + a_2 C_2 \]
where \( a_1 = 1 - \alpha \), \( a_2 = \alpha \),
\[ C_1 = \log_2(2) - H \left( \frac{1 - \alpha}{1 - \alpha}, \frac{\alpha}{1 - \alpha} \right) = 1 - H(1, 0) = 1 - 0 = 1, \]
and
\[ C_2 = \log_2(1) - H \left( \frac{\alpha}{\alpha} \right) = 0 - 0 = 0. \]

Therefore, the BEC capacity is given by
\[ C = (1 - \alpha)(1) + (\alpha)(0) = 1 - \alpha. \quad (4.4.7) \]

**Example 4.20 (Capacity of the BSEC)** Similarly, the BSEC with crossover probability \( \varepsilon \) and erasure probability \( \alpha \) as described in (4.2.6) is quasi-symmetric; its transition matrix can be partitioned along its columns into two symmetric sub-matrices
\[ Q_1 = \begin{bmatrix} 1 - \varepsilon - \alpha & \varepsilon \\ \varepsilon & 1 - \varepsilon - \alpha \end{bmatrix} \]
and
\[ Q_2 = \begin{bmatrix} \alpha \\ \alpha \end{bmatrix}. \]

Hence by Lemma 4.18, the channel capacity is given by
\[ C = a_1 C_1 + a_2 C_2 \]
where \( a_1 = 1 - \alpha \), \( a_2 = \alpha \),
\[ C_1 = \log_2(2) - H \left( \frac{1 - \varepsilon - \alpha}{1 - \alpha}, \frac{\varepsilon}{1 - \alpha} \right) = 1 - h_b \left( \frac{1 - \varepsilon - \alpha}{1 - \alpha} \right), \]
and
\[ C_2 = \log_2(1) - H \left( \frac{\alpha}{\alpha} \right) = 0. \]

We thus obtain that
\[ C = (1 - \alpha) \left[ 1 - h_b \left( \frac{1 - \varepsilon - \alpha}{1 - \alpha} \right) \right] + (\alpha)(0) \]
\[ = (1 - \alpha) \left[ 1 - h_b \left( \frac{1 - \varepsilon - \alpha}{1 - \alpha} \right) \right]. \quad (4.4.8) \]

As already noted, the BSEC is a combination of the BSC with bit error rate \( \varepsilon \) and the BEC with erasure probability \( \alpha \). Indeed, setting \( \alpha = 0 \) in (4.4.8) yields that \( C = 1 - h_b(1 - \varepsilon) = 1 - h_b(\varepsilon) \) which is the BSC capacity. Furthermore, setting \( \varepsilon = 0 \) results in \( C = 1 - \alpha \), the BEC capacity.
4.4.2 Karuch-Kuhn-Tucker condition for channel capacity

When the channel does not satisfy any symmetry property, the following necessary and sufficient Karuch-Kuhn-Tucker (KKT) condition (e.g., cf. Appendix B.10, [19, pp. 87-91] or [5, 11]) for calculating channel capacity can be quite useful.

Definition 4.21 (Mutual information for a specific input symbol) The mutual information for a specific input symbol is defined as:

\[ I(x; Y) \triangleq \sum_{y \in Y} P_{Y|X}(y|x) \log_2 \frac{P_{Y|X}(y|x)}{P_Y(y)}. \]

From the above definition, the mutual information becomes:

\[
I(X; Y) = \sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \log_2 \frac{P_{Y|X}(y|x)}{P_Y(y)} = \sum_{x \in \mathcal{X}} P_X(x) I(x; Y).
\]

Lemma 4.22 (KKT condition for channel capacity) For a given DMC, an input distribution \( P_X \) achieves its channel capacity iff there exists a constant \( C \) such that

\[
\begin{align*}
I(x : Y) &= C \quad \forall x \in \mathcal{X} \text{ with } P_X(x) > 0; \\
I(x : Y) &\leq C \quad \forall x \in \mathcal{X} \text{ with } P_X(x) = 0. 
\end{align*}
\]

Furthermore, the constant \( C \) is the channel capacity (justifying the choice of notation).

Proof: The forward (if) part holds directly; hence, we only prove the converse (only-if) part.

Without loss of generality, we assume that \( P_X(x) < 1 \) for all \( x \in \mathcal{X} \), since \( P_X(x) = 1 \) for some \( x \) implies that \( I(X; Y) = 0 \). The problem of calculating the channel capacity is to maximize

\[
I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log_2 \frac{P_{Y|X}(y|x)}{\sum_{x' \in \mathcal{X}} P_X(x') P_{Y|X}(y|x')} 
\]

subject to the condition

\[
\sum_{x \in \mathcal{X}} P_X(x) = 1
\]
for a given channel distribution \( P_{Y|X} \). By using the Lagrange multiplier method (e.g., see Appendix B.10 or [5]), maximizing (4.4.10) subject to (4.4.11) is equivalent to maximize:

\[
f(P_X) \triangleq \sum_{x \in \mathcal{X}} P_X(x) P_{Y|X}(y|x) \log_2 \frac{P_{Y|X}(y|x)}{\sum_{x' \in \mathcal{X}} P_X(x') P_{Y|X}(y|x')} + \lambda \left( \sum_{x \in \mathcal{X}} P_X(x) - 1 \right).
\]

We then take the derivative of the above quantity with respect to \( P_X(x'') \), and obtain that\(^{12}\)

\[
\frac{\partial f(P_X)}{\partial P_X(x'')} = I(x''; Y) - \log_2(e) + \lambda.
\]

By Property 2 of Lemma 2.46, \( I(X; Y) = I(P_X, P_{Y|X}) \) is a concave function in \( P_X \) (for a fixed \( P_{Y|X} \)). Therefore, the maximum of \( I(P_X, P_{Y|X}) \) occurs for a zero derivative when \( P_X(x) \) does not lie on the boundary, namely \( 1 > P_X(x) > 0 \). For those \( P_X(x) \) lying on the boundary, i.e., \( P_X(x) = 0 \), the maximum occurs if a displacement from the boundary to the interior decreases the quantity, which implies a non-positive derivative, namely

\[
I(x; Y) \leq -\lambda + \log_2(e), \quad \text{for those } x \text{ with } P_X(x) = 0.
\]

To summarize, if an input distribution \( P_X \) achieves the channel capacity, then

\[
\begin{cases}
I(x''; Y) = -\lambda + \log_2(e), & \text{for } P_X(x'') > 0; \\
I(x''; Y) \leq -\lambda + \log_2(e), & \text{for } P_X(x'') = 0.
\end{cases}
\]

\(^{12}\)The details for taking the derivative are as follows:

\[
\begin{align*}
\frac{\partial}{\partial P_X(x'')} & \left\{ \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log_2 P_{Y|X}(y|x) \\
& \quad - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \log_2 \left[ \sum_{x' \in \mathcal{X}} P_X(x') P_{Y|X}(y|x') \right] \right\} \\
& = \sum_{y \in \mathcal{Y}} P_{Y|X}(y|''x) \log_2 P_{Y|X}(y|''x) - \left( \sum_{y \in \mathcal{Y}} P_{Y|X}(y|''x) \log_2 \left[ \sum_{x' \in \mathcal{X}} P_X(x') P_{Y|X}(y|x') \right] \right) \\
& \quad + \log_2(e) \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) \frac{P_{Y|X}(y|''x)}{\sum_{x' \in \mathcal{X}} P_X(x') P_{Y|X}(y|x')} + \lambda \\
& = I(x''; Y) - \log_2(e) \sum_{y \in \mathcal{Y}} \left[ \sum_{x \in \mathcal{X}} P_X(x) P_{Y|X}(y|x) \right] \frac{P_{Y|X}(y|''x)}{\sum_{x' \in \mathcal{X}} P_X(x') P_{Y|X}(y|x')} + \lambda \\
& = I(x''; Y) - \log_2(e) \sum_{y \in \mathcal{Y}} P_{Y|X}(y|''x) + \lambda \\
& = I(x''; Y) - \log_2(e) + \lambda.
\end{align*}
\]
for some $\lambda$. With the above result, setting $C = -\lambda + 1$ yields (4.4.9). Finally, multiplying both sides of each equation in (4.4.9) by $P_X(x)$ and summing over $x$ yields that $\max_{P_X} I(X;Y)$ on the left and the constant $C$ on the right, thus proving that the constant $C$ is indeed the channel’s capacity.

**Example 4.23 (Quasi-symmetric channels)** For a quasi-symmetric channel, one can directly verify that the uniform input distribution satisfies the KKT condition of Lemma 4.22 and yields that the channel capacity is given by (4.4.6); this is left as an exercise. As we already saw, the BSC, the $q$-ary symmetric channel, the BEC and the BSEC are all quasi-symmetric.

**Example 4.24** Consider a DMC with a ternary input alphabet $X = \{0, 1, 2\}$, binary output alphabet $Y = \{0, 1\}$ and the following transition matrix

$$Q = \begin{bmatrix}
1 & 0 \\
\frac{1}{2} & \frac{1}{2} \\
0 & 1
\end{bmatrix}.$$

This channel is not quasi-symmetric. However, one may guess that the capacity of this channel is achieved by the input distribution $(P_X(0), P_X(1), P_X(2)) = (\frac{1}{2}, 0, \frac{1}{2})$ since the input $x = 1$ has an equal conditional probability of being received as 0 or 1 at the output. Under this input distribution, we obtain that $I(x = 0; Y) = I(x = 2; Y) = 1$ and that $I(x = 1; Y) = 0$. Thus the KKT condition of (4.4.9) is satisfied; hence confirming that the above input distribution achieves channel capacity and that channel capacity is equal to 1 bit.

**Observation 4.25 (Capacity achieved by a uniform input distribution)** We close this chapter by noting that there is a class of DMC’s that is larger than that of quasi-symmetric channels for which the uniform input distribution achieves capacity. It concerns the class of so-called “$T$-symmetric” channels [40, Section V, Definition 1] for which

$$T(x) \triangleq I(x; Y) - \log_2 |X| = \sum_{y \in Y} P_{Y|X}(y|x) \log_2 \frac{P_{Y|X}(y|x)}{\sum_{x' \in X} P_{Y|X}(y|x')},$$

is a constant function of $x$ (i.e., independent of $x$), where $I(x; Y)$ is the mutual information for input $x$ under a uniform input distribution. Indeed the $T$-symmetry condition is equivalent to the property of having the uniform input distribution achieve capacity. This directly follows from the KKT condition of Lemma 4.22. An example of a $T$-symmetric channel that is not quasi-symmetric is the binary-input ternary-output channel with the following transition matrix

$$Q = \begin{bmatrix}
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\frac{1}{3} & \frac{1}{3} & \frac{1}{3}
\end{bmatrix}.$$
Hence its capacity is achieved by the uniform input distribution. See [40, Figure 2] for (infinitely-many) other examples of $T$-symmetric channels. However, unlike quasi-symmetric channels, $T$-symmetric channels do not admit in general a simple closed-form expression for their capacity (such as the one given in (4.4.6)).

**Problems**

1. Prove the Shannon-McMillan theorem for pairs (Theorem 4.9).

2. The proof of Shannon’s channel coding theorem is based on the random coding technique. What is the codeword selecting distribution of the random codebook? What is the decoding rule in the proof?

3. Show that processing the output of a DMC (via a given function) does not strictly increase its capacity.

4. Consider the system shown in the block diagram below. Can the channel capacity between channel input $X$ and channel output $Z$ be strictly larger than the channel capacity between channel input $X$ and channel output $Y$? Which lemma or theorem is your answer based on?

5. Consider a DMC with input $X$ and output $Y$. Assume that the input alphabet is $\mathcal{X} = \{1, 2\}$, the output alphabet is $\mathcal{Y} = \{0, 1, 2, 3\}$, and the transition probability is given by

$$P_{Y|X}(y|x) = \begin{cases} 1 - 2\epsilon, & \text{if } x = y; \\ \epsilon, & \text{if } |x - y| = 1; \\ 0, & \text{otherwise}, \end{cases}$$

where $0 < \epsilon < 1/2$.

(a) Determine the channel probability transition matrix $Q \equiv [P_{Y|X}(y, x)]$.

(b) Compute the capacity of this channel. What is the maximizing input distribution that achieves capacity?
6. Find the channel capacity of the DMC modeled as $Y = X + Z$, where $P_Z(0) = P_Z(a) = 1/2$. The alphabet for channel input $X$ is $\mathcal{X} = \{0, 1\}$. Assume that $Z$ is independent of $X$. Discuss the dependence of the channel capacity on the value of $a$.

7. Find the capacity of the $Z$-channel described by the following transition diagram.

8. A DMC has identical input and output alphabets given by $\{0, 1, 2, 3, 4\}$. Let $X$ be the channel input, and $Y$ be the channel output. Suppose that

$$ P_{Y|X}(i|i) = \frac{1}{2} \quad \forall \ i \in \{0, 1, 2, 3, 4\}. $$

(a) Find the channel transition matrix that maximizes $H(Y|X)$.

(b) Using the channel transition matrix obtained in (a), evaluate the channel capacity.

9. Binary Channel: Consider a binary memoryless channel with the following probability transition matrix:

$$ Q = \begin{bmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{bmatrix}, $$

where $\alpha > 0$, $\beta > 0$ and $\alpha + \beta < 1$.

(a) Determine the capacity $C$ of this channel in terms of $\alpha$ and $\beta$.

(b) What does the expression of $C$ reduces to if $\alpha = \beta$ ?

10. Find the capacity of the asymmetric binary channel with errors and erasures described in (4.2.7). Verify that the channel capacity reduces to that of the BSEC when setting $\varepsilon' = \varepsilon$ and $\alpha' = \alpha$.

11. Find the capacity of the binary-input quaternary-output DMC given in (4.4.2). For what values of $\varepsilon$ is capacity maximized, and for what values of $\varepsilon$ is capacity minimized ?

12. Find the capacity of the $q$-ary erasure channel described in (4.2.9) and compare the result with the capacity of the BEC.
13. Consider two DMCs

\[(X_1, P_{Y_1|X_1}, Y_1) \text{ and } (X_2, P_{Y_2|X_2}, Y_2)\]

with capacity \(C_1\) and \(C_2\) respectively. A new channel \((X_1 \times X_2, P_{Y_1|X_1 \times X_2}, Y_1 \times Y_2)\) is formed in which \(x_1 \in X_1\) and \(x_2 \in X_2\) are simultaneously sent, resulting in \(Y_1, Y_2\). Find the capacity of this channel.

14. The Sum Channel:

(a) Let \((X_1, P_{Y_1|X_1}, Y_1)\) be a DMC with finite input alphabet \(X_1\), finite output alphabet \(Y_1\), transition distribution \(P_{Y_1|X_1}(y|x)\) and capacity \(C_1\). Similarly, let \((X_2, P_{Y_2|X_2}, Y_2)\) be another DMC with capacity \(C_2\). Assume that \(X_1 \cap X_2 = \emptyset\) and that \(Y_1 \cap Y_2 = \emptyset\).

Now let \((X, P_{Y|X}, Y)\) be the sum of these two channels where \(X = X_1 \cup X_2\), \(Y = Y_1 \cup Y_2\) and

\[P_{Y|X}(y|x) = \begin{cases} 
P_{Y_1|X_1}(y|x) & \text{if } x \in X_1, y \in Y_1 \\
P_{Y_2|X_2}(y|x) & \text{if } x \in X_2, y \in Y_2 \\
0 & \text{otherwise.} \end{cases}\]

Show that the capacity of the sum channel is given by

\[C_{sum} = \log_2 \left( 2^{C_1} + 2^{C_2} \right) \text{ bits/channel use.}\]

*Hint:* Introduce a Bernoulli random variable \(Z\) with \(\Pr[Z = 1] = \alpha\) such that \(Z = 1\) if \(X \in X_1\) (when the first channel is used), and \(Z = 2\) if \(X \in X_2\) (when the second channel is used). Then show that

\[I(X; Y) = I(X, Z; Y) = h_b(\alpha) + \alpha I(X_1; Y_1) + (1 - \alpha) I(X_2; Y_2),\]

where \(h_b(\cdot)\) is the binary entropy function, and \(I(X_i; Y_i)\) is the mutual information for channel \(P_{Y_i|X_i}(y|x), i = 1, 2\). Then maximize (jointly) over the input distribution and \(\alpha\).

(b) Compute \(C_{sum}\) above if the first channel is a BSC with crossover probability 0.11, and the second channel is a BEC with erasure probability 0.5.

15. Prove that the quasi-symmetric channel satisfies the KKT condition of Lemma 4.22 and yields the channel capacity given by (4.4.6).

16. Let the channel transition probability \(P_{Y|X}\) of a DMC be defined as the following figure, where \(0 < \epsilon < 0.5\).
(a) Is the channel weakly symmetric? Is the channel symmetric?

(b) Determine the channel capacity of this channel. Also, indicate the input distribution that achieves the channel capacity.

17. Let the relation between the channel input \( \{X_n\}_{n=1}^{\infty} \) and channel output \( \{Y_n\}_{n=1}^{\infty} \) be given by

\[
Y_n = (\alpha_n \times X_n) \oplus N_n \text{ for each } n,
\]

where \( \alpha_n, X_n, Y_n \) and \( N_n \) all take values from \( \{0, 1\} \), and "\( \oplus \)" represents the modulo-2 addition operation. Assume that the attenuation \( \{\alpha_n\}_{n=1}^{\infty} \), channel input \( \{X_n\}_{n=1}^{\infty} \) and noise \( \{N_n\}_{n=1}^{\infty} \) processes are independent from each other. Also, \( \{\alpha_n\}_{n=1}^{\infty} \) and \( \{N_n\}_{n=1}^{\infty} \) are i.i.d. with

\[
\Pr[\alpha_n = 1] = \Pr[\alpha_n = 0] = \frac{1}{2}
\]

and

\[
\Pr[N_n = 1] = 1 - \Pr[N_n = 0] = \varepsilon \in (0, 1/2).
\]

(a) Derive the channel transition probability matrix

\[
\begin{bmatrix}
P_{Yj|Xj}(0|0) & P_{Yj|Xj}(1|0) \\
P_{Yj|Xj}(0|1) & P_{Yj|Xj}(1|1)
\end{bmatrix}.
\]

(b) The channel is apparently a DMC. Determine its channel capacity \( C \).

*Hint:* Use the KKT condition for channel capacity (Lemma 4.22).
(c) Suppose that \( \alpha^n \) is known, and consists of \( k \) 1’s. Find the maximum 
\( I(X^n;Y^n) \) for the same channel with known \( \alpha^n \).

*Hint:* For known \( \alpha^n \), \( \{(X_j,Y_j)\}_{j=1}^n \) are independent. Recall 
\( I(X^n;Y^n) \leq \sum_{j=1}^n I(X_j;Y_j) \).

(d) Some researchers attempt to derive the capacity of the channel in (b) 
in terms of the following steps:

- Derive the maximum mutual information between channel input 
  \( X^n \) and output \( Y^n \) for a given \( \alpha^n \) (namely the solution in (c));
- Calculate the expectation value of the maximum mutual informa-
  tion obtained from the previous step according to the statistics 
  of \( \alpha^n \).
- Then the capacity of the channel is equal to this “expected value” 
  divided by \( n \).

Does this “expected capacity” \( \bar{C} \) coincide with that in (b)?

18. Suppose that blocklength \( n = 2 \) and code size \( M = 2 \). Assume each code 
bit is either 0 or 1.

(a) What is the number of all possible codebook designs? (Note: This 
number includes those lousy code designs, such as \( \{00,00\} \).)

(b) Suppose that one randomly draws one of these possible code designs 
according to a uniform distribution and applies the selected code to 
BSC with crossover probability \( \varepsilon \). Then what is the expected error 
probability, if the decoder simply selects the codeword whose Ham-
ming distance\(^\text{13}\) to the received vector is the smallest? (When both 
codewords have the same Hamming distance to the received vector, 
the decoder chooses one of them at random as the transmitted code-
word.)

(c) Explain why the error in (b) does not vanish as \( \varepsilon \downarrow 0 \).

*Hint:* The error of random \((n,M)\) code is lower bounded by the error 
of random \((n,2)\) code for \( M \geq 2 \).

19. Assume that the alphabets for variables \( X \) and \( Y \) are both \( \{1,2,3,4,5\} \). 
Let \( \hat{x} = g(y) \) be an estimate of \( x \) from observing \( y \). Define the probability 
of estimation error as \( P_e = \Pr\{g(Y) \neq X\} \). Then, Fano’s inequality gives 
a lower bound for \( P_e \) as \( h_b(P_e) + 2P_e \geq H(X|Y) \), where \( h_b(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p} \) is the binary entropy function. The curve for \( h_b(P_e) + 2P_e = H(X|Y) \) is plotted below.

\(^{\text{13}}\)The Hamming distance between two codewords of identical length is the number of places 
in which they differ.
(a) Point A on the above figure shows that if $H(X|Y) = 0$, zero estimation error, namely, $P_e = 0$, can be achieved. In this case, characterize the distribution $P_{X|Y}$. Also, give an estimator $g(\cdot)$ that achieves $P_e = 0$.

_Hint:_ Think what kind of relation between $X$ and $Y$ can render $H(X|Y) = 0$.

(b) Point B on the above figure indicates that when $H(X|Y) = \log_2(5)$, the estimation error can only be equal to 0.8. In this case, characterize the distributions $P_{X|Y}$ and $P_X$. Prove that at $H(X|Y) = \log_2(5)$, all estimators yield $P_e = 0.8$.

_Hint:_ Think what kind of relation between $X$ and $Y$ can result in $H(X|Y) = \log_2(5)$.

(c) Point C on the above figure hints that when $H(X|Y) = 2$, the estimation error can be as worse as 1. Give an estimator $g(\cdot)$ that leads to $P_e = 1$, if $P_{X|Y}(x|y) = 1/4$ for $x \neq y$, and $P_{X|Y}(x|y) = 0$ for $x = y$.

(d) Similarly, point D on the above figure hints that when $H(X|Y) = 0$, the estimation error can be as worse as 1. Give an estimator $g(\cdot)$ that leads to $P_e = 1$ at $H(X|Y) = 0$.

20. Decide whether the following statement is _true_ or _false_. Consider a discrete memoryless channel with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$ and transition distribution $P_{Y|X}(y|x) \triangleq \Pr\{Y = y|X = x\}$. Let $P_{X_1}(\cdot)$ and $P_{X_2}(\cdot)$ be two possible input distributions, and $P_{Y_1}(\cdot)$ and $P_{Y_2}(\cdot)$ be the corresponding output distributions; i.e., $\forall y \in \mathcal{Y}$, $P_{Y_1}(y) = \sum_{x \in \mathcal{X}} P_{Y|X}(y|x)P_{X_1}(x)$,
$i = 1, 2$. Then

$$D(P_{X_i} || P_{X_2}) \geq D(P_{Y_i} || P_{Y_2}).$$
Chapter 5

Differential Entropy and Gaussian Channels

We have so far examined information measures and their operational characterization for discrete-time discrete-alphabet systems. In this chapter, we turn our focus to continuous-alphabet (real-valued) systems. Except for a brief interlude with the continuous-time (waveform) Gaussian channel, we consider discrete-time systems, as treated throughout the book.

We first recall that a real-valued (continuous) random variable $X$ is described by its cumulative distribution function (cdf)

$$F_X(x) \triangleq \Pr[X \leq x]$$

for $x \in \mathbb{R}$, the set of real numbers. The distribution of $X$ is called absolutely continuous (with respect to the Lebesgue measure) if a probability density function (pdf) $f_X(\cdot)$ exists such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

where $f_X(t) \geq 0 \ \forall t$ and $\int_{-\infty}^{+\infty} f_X(t)dt = 1$. If $F_X(\cdot)$ is differentiable everywhere, then the pdf $f_X(\cdot)$ exists and is given by the derivative of $F_X(\cdot)$: $f_X(t) = \frac{dF_X(t)}{dt}$. The support of a random variable $X$ with pdf $f_X(\cdot)$ is denoted by $S_X$ and can be conveniently given as

$$S_X = \{x \in \mathbb{R} : f_X(x) > 0\}.$$ 

We will deal with random variables that admit a pdf.\(^1\)

\(^1\)A rigorous (measure-theoretic) study for general continuous systems, initiated by Kolmogorov [28], can be found in [38, 25].
5.1 Differential entropy

Recall that the definition of entropy for a discrete random variable $X$ representing a DMS is

$$H(X) \triangleq \sum_{x \in \mathcal{X}} -P_X(x) \log_2 P_X(x) \text{ (in bits)}.$$ 

As already seen in Shannon’s source coding theorem, this quantity is the minimum average code rate achievable for the lossless compression of the DMS. But if the random variable takes on values in a continuum, the minimum number of bits per symbol needed to losslessly describe it must be infinite. This is illustrated in the following example, where we take a discrete approximation (quantization) of a random variable uniformly distributed on the unit interval and study the entropy of the quantized random variable as the quantization becomes finer and finer.

**Example 5.1** Consider a real-valued random variable $X$ that is uniformly distributed on the unit interval, i.e., with pdf given by

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0, 1); \\ 0 & \text{otherwise.} \end{cases}$$

Given a positive integer $m$, we can discretize $X$ by uniformly quantizing it into $m$ levels by partitioning the support of $X$ into equal-length segments of size $\Delta = \frac{1}{m}$ ($\Delta$ is called the quantization step-size) such that:

$$q_m(X) = \frac{i}{m}, \quad \text{if } \frac{i-1}{m} \leq X < \frac{i}{m},$$

for $1 \leq i \leq m$. Then the entropy of the quantized random variable $q_m(X)$ is given by

$$H(q_m(X)) = -\sum_{i=1}^{m} \frac{1}{m} \log_2 \left( \frac{1}{m} \right) = \log_2 m \text{ (in bits)}.$$ 

Since the entropy $H(q_m(X))$ of the quantized version of $X$ is a lower bound to the entropy of $X$ (as $q_m(X)$ is a function of $X$) and satisfies in the limit

$$\lim_{m \to \infty} H(q_m(X)) = \lim_{m \to \infty} \log_2 m = \infty,$$

we obtain that the entropy of $X$ is infinite.

The above example indicates that to compress a continuous source without incurring any loss or distortion indeed requires an infinite number of bits. Thus
when studying continuous sources, the entropy measure is limited in its effectiveness and the introduction of a new measure is necessary. Such new measure is indeed obtained upon close examination of the entropy of a uniformly quantized real-valued random-variable minus the quantization accuracy as the accuracy increases without bound.

**Lemma 5.2** Consider a real-valued random variable $X$ with support $[a, b)$ and pdf $f_X$ such that $-f_X \log_2 f_X$ is integrable\(^2\) (where $-\int_a^b f_X(x) \log_2 f_X(x) dx$ is finite). Then a uniform quantization of $X$ with an $n$-bit accuracy (i.e., with a quantization step-size of $\Delta = 2^{-n}$) yields an entropy approximately equal to $-\int_a^b f_X(x) \log_2 f_X(x) dx + n$ bits for $n$ sufficiently large. In other words,

$$
\lim_{n \to \infty} [H(q_n(X)) - n] = -\int_a^b f_X(x) \log_2 f_X(x) dx
$$

where $q_n(X)$ is the uniformly quantized version of $X$ with quantization step-size $\Delta = 2^{-n}$.

**Proof:**

**Step 1:** Mean-value theorem. Let $\Delta = 2^{-n}$ be the quantization step-size, and let

$$
t_i \triangleq \begin{cases} 
  a + i\Delta, & i = 0, 1, \cdots, j - 1 \\
  b, & i = j
\end{cases}
$$

where $j = [(b - a)2^n]$. From the mean-value theorem (e.g., cf. [36]), we can choose $x_i \in [t_{i-1}, t_i]$ for $1 \leq i \leq j$ such that

$$
p_i \triangleq \int_{t_{i-1}}^{t_i} f_X(x) dx = f_X(x_i)(t_i - t_{i-1}) = \Delta \cdot f_X(x_i).
$$

**Step 2:** Definition of $h^{(n)}(X)$. Let

$$
h^{(n)}(X) \triangleq -\sum_{i=1}^{j} [f_X(x_i) \log_2 f_X(x_i)]2^{-n}.
$$

Since $-f_X(x) \log_2 f_X(x)$ is integrable,

$$
h^{(n)}(X) \to -\int_a^b f_X(x) \log_2 f_X(x) dx \quad \text{as } n \to \infty.
$$

Therefore, given any $\varepsilon > 0$, there exists $N$ such that for all $n > N$,

$$
\left| -\int_a^b f_X(x) \log_2 f_X(x) dx - h^{(n)}(X) \right| < \varepsilon.
$$

\(^2\)By integrability, we mean the usual Riemann integrability (e.g., see [42]).
Step 3: Computation of $H(q_n(X))$. The entropy of the (uniformly) quantized version of $X$, $q_n(X)$, is given by

$$H(q_n(X)) = - \sum_{i=1}^{j} p_i \log_2 p_i$$

$$= - \sum_{i=1}^{j} (f_X(x_i) \Delta) \log_2 (f_X(x_i) \Delta)$$

$$= - \sum_{i=1}^{j} (f_X(x_i) 2^{-n}) \log_2 (f_X(x_i) 2^{-n})$$

where the $p_i$’s are the probabilities of the different values of $q_n(X)$.

Step 4: $H(q_n(X)) - h^{(n)}(X)$.

From Steps 2 and 3,

$$H(q_n(X)) - h^{(n)}(X) = - \sum_{i=1}^{j} [f_X(x_i) 2^{-n}] \log_2 (2^{-n})$$

$$= n \sum_{i=1}^{j} \int_{t_{i-1}}^{t_i} f_X(x) dx$$

$$= n \int_a^b f_X(x) dx = n.$$

Hence, we have that for $n > N$,

$$\left[ - \int_a^b f_X(x) \log_2 f_X(x) dx + n \right] - \varepsilon < H(q_n(X))$$

$$= h^{(n)}(X) + n$$

$$< \left[ - \int_a^b f_X(x) \log_2 f_X(x) dx + n \right] + \varepsilon,$$

yielding that

$$\lim_{n \to \infty} [H(q_n(X)) - n] = - \int_a^b f_X(x) \log_2 f_X(x) dx.$$

More generally, the following result due to Rényi [39] can be shown for (absolutely continuous) random variables with arbitrary support.
Theorem 5.3 [39, Theorem 1] For any real-valued random variable with pdf $f_X$, if $- \sum_{i=1}^j p_i \log_2 p_i$ is finite, where the $p_i$’s are the probabilities of the different values of uniformly quantized $q_n(X)$ over support $S_X$, then

$$\lim_{n \to \infty} [H(q_n(X)) - n] = - \int_{S_X} f_X(x) \log_2 f_X(x) dx$$

provided the integral on the right-hand side exists.

In light of the above results, we can define the following information measure.

**Definition 5.4 (Differential entropy)** The differential entropy (in bits) of a continuous random variable $X$ with pdf $f_X$ and support $S_X$ is defined as

$$h(X) \triangleq - \int_{S_X} f_X(x) \cdot \log_2 f_X(x) dx = E[- \log_2 f_X(X)],$$

when the integral exists.

Thus the differential entropy $h(X)$ of a real-valued random variable $X$ has an operational meaning in the following sense. Since $H(q_n(X))$ is the minimum average number of bits needed to losslessly describe $q_n(X)$, we thus obtain that $h(X) + n$ is approximately needed to describe $X$ when uniformly quantizing it with an $n$-bit accuracy. Therefore, we may conclude that the larger $h(X)$ is, the larger is the average number of bits required to describe a uniformly quantized $X$ within a fixed accuracy.

**Example 5.5** A continuous random variable $X$ with support $S_X = [0, 1)$ and pdf $f_X(x) = 2x$ for $x \in S_X$ has differential entropy equal to

$$\int_0^1 -2x \cdot \log_2(2x) dx = \left. \frac{x^2(\log_2 e - 2 \log_2(2x))}{2} \right|_0^1 = \frac{1}{2 \ln 2} - \log_2(2) = -0.278652 \text{ bits.}$$

We herein illustrate Lemma 5.2 by uniformly quantizing $X$ to an $n$-bit accuracy and computing the entropy $H(q_n(X))$ and $H(q_n(X)) - n$ for increasing values of $n$, where $q_n(X)$ is the quantized version of $X$.

We have that $q_n(X)$ is given by

$$q_n(X) = \frac{i}{2^n} \quad \text{if} \quad \frac{i-1}{2^n} \leq X < \frac{i}{2^n}.$$
Table 5.1: Quantized random variable $q_n(X)$ under an $n$-bit accuracy: $H(q_n(X))$ and $H(q_n(X)) - n$ versus $n$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$H(q_n(X))$</th>
<th>$H(q_n(X)) - n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.811278 bits</td>
<td>-0.188722 bits</td>
</tr>
<tr>
<td>2</td>
<td>1.748999 bits</td>
<td>-0.251000 bits</td>
</tr>
<tr>
<td>3</td>
<td>2.729560 bits</td>
<td>-0.270440 bits</td>
</tr>
<tr>
<td>4</td>
<td>3.723726 bits</td>
<td>-0.276275 bits</td>
</tr>
<tr>
<td>5</td>
<td>4.722023 bits</td>
<td>-0.277977 bits</td>
</tr>
<tr>
<td>6</td>
<td>5.721399 bits</td>
<td>-0.278600 bits</td>
</tr>
<tr>
<td>7</td>
<td>6.721361 bits</td>
<td>-0.278638 bits</td>
</tr>
<tr>
<td>8</td>
<td>7.721351 bits</td>
<td>-0.278648 bits</td>
</tr>
</tbody>
</table>

for $1 \leq i \leq 2^n$. Hence,

$$\Pr \left\{ q_n(X) = \frac{i}{2^n} \right\} = \frac{(2i - 1)}{2^{2n}},$$

which yields

$$H(q_n(X)) = -\sum_{i=1}^{2^n} \frac{2i - 1}{2^{2n}} \log_2 \left( \frac{2i - 1}{2^{2n}} \right) = \left[ -\frac{1}{2^{2n}} \sum_{i=1}^{2^n} (2i - 1) \log_2(2i - 1) + 2 \log_2(2^n) \right].$$

As shown in Table 5.1, we indeed observe that as $n$ increases, $H(q_n(X))$ tends to infinity while $H(q_n(X)) - n$ converges to $h(X) = -0.278652$ bits.

Thus a continuous random variable $X$ contains an infinite amount of information; but we can measure the information contained in its $n$-bit quantized version $q_n(X)$ as: $H(q_n(X)) \approx h(X) + n$ (for $n$ large enough).

**Example 5.6** Let us determine the minimum average number of bits required to describe the uniform quantization with 3-digit accuracy of the decay time (in years) of a radium atom assuming that the half-life of the radium (i.e., the median of the decay time) is 80 years and that its pdf is given by $f_X(x) = \lambda e^{-\lambda x}$, where $x > 0$.

Since the median of the decay time is 80, we obtain:

$$\int_0^{80} \lambda e^{-\lambda x} \, dx = 0.5,$$
which implies that $\lambda = 0.00866$. Also, 3-digit accuracy is approximately equivalent to $\log_2 999 = 9.96 \approx 10$ bits accuracy. Therefore, by Theorem 5.3, the number of bits required to describe the quantized decay time is approximately

$$h(X) + 10 = \log_2 \frac{e}{\lambda} + 10 = 18.29 \text{ bits.}$$

We close this section by computing the differential entropy for two common real-valued random variables: the uniformly distributed random variable and the Gaussian distributed random variable.

**Example 5.7 (Differential entropy of a uniformly distributed random variable)** Let $X$ be a continuous random variable that is uniformly distributed over the interval $(a, b)$, where $b > a$; i.e., its pdf is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in (a, b); \\ 0 & \text{otherwise.} \end{cases}$$

So its differential entropy is given by

$$h(X) = - \int_{a}^{b} \frac{1}{b-a} \log_2 \frac{1}{b-a} = \log_2 (b-a) \text{ bits.}$$

Note that if $(b-a) < 1$ in the above example, then $h(X)$ is negative, unlike entropy. The above example indicates that although differential entropy has a form analogous to entropy (in the sense that summation and pmf for entropy are replaced by integration and pdf, respectively, for differential entropy), differential entropy does not retain all the properties of entropy (one such operational difference was already highlighted in the previous lemma and theorem).\(^3\)

**Example 5.8 (Differential entropy of a Gaussian random variable)** Let $X \sim \mathcal{N}(\mu, \sigma^2)$; i.e., $X$ is a Gaussian (or normal) random variable with finite mean $\mu$, variance $\text{Var}(X) = \sigma^2 > 0$ and pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for $x \in \mathbb{R}$. Then its differential entropy is given by

$$h(X) = \int_{\mathbb{R}} f_X(x) \left[ \frac{1}{2} \log_2(2\pi\sigma^2) + \frac{(x-\mu)^2}{2\sigma^2} \log_2 e \right] \, dx$$

\(^3\)By contrast, entropy and differential entropy are sometimes called *discrete entropy* and *continuous entropy*, respectively.
\[ h(X) = \frac{1}{2} \log_2(2\pi e\sigma^2) \] bits.

Note that for a Gaussian random variable, its differential entropy is only a function of its variance \( \sigma^2 \) (it is independent from its mean \( \mu \)). This is similar to the differential entropy of a uniform random variable, which only depends on difference \( (b - a) \) but not the mean \( (a + b)/2 \).

### 5.2 Joint and conditional differential entropies, divergence and mutual information

**Definition 5.9 (Joint differential entropy)** If \( X^n = (X_1, X_2, \cdots, X_n) \) is a continuous random vector of size \( n \) (i.e., a vector of \( n \) continuous random variables) with joint pdf \( f_{X^n} \) and support \( S_{X^n} \subseteq \mathbb{R}^n \), then its joint differential entropy is defined as

\[
h(X^n) \triangleq -\int_{S_{X^n}} f_{X^n}(x_1, x_2, \cdots, x_n) \log_2 f_{X^n}(x_1, x_2, \cdots, x_n) \, dx_1 \, dx_2 \cdots \, dx_n = E[-\log_2 f_{X^n}(X^n)]
\]

when the \( n \)-dimensional integral exists.

**Definition 5.10 (Conditional differential entropy)** Let \( X \) and \( Y \) be two jointly distributed continuous random variables with joint pdf \( f_{X,Y} \) and support \( S_{X,Y} \subseteq \mathbb{R}^2 \) such that the conditional pdf of \( Y \) given \( X \), given by

\[
f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)},
\]

is well defined for all \( (x,y) \in S_{X,Y} \), where \( f_X \) is the marginal pdf of \( X \). Then the conditional entropy of \( Y \) given \( X \) is defined as

\[
h(Y|X) \triangleq -\int_{S_{X,Y}} f_{X,Y}(x,y) \log_2 f_{Y|X}(y|x) \, dx \, dy = E[-\log_2 f_{Y|X}(Y|X)],
\]

when the integral exists.

Note that as in the case of (discrete) entropy, the chain rule holds for differential entropy:

\[
h(X, Y) = h(X) + h(Y|X) = h(Y) + h(X|Y).
\]
Definition 5.11 (Divergence or relative entropy) Let $X$ and $Y$ be two continuous random variables with marginal pdfs $f_X$ and $f_Y$, respectively, such that their supports satisfy $S_X \subseteq S_Y \subseteq \mathbb{R}$. Then the divergence (or relative entropy or Kullback-Leibler distance) between $X$ and $Y$ is written as $D(X\|Y)$ or $D(f_X\|f_Y)$ and defined by

$$D(X\|Y) \triangleq \int_{S_X} f_X(x) \log_2 \frac{f_X(x)}{f_Y(x)} \, dx = E \left[ \frac{f_X(X)}{f_Y(X)} \right]$$

when the integral exists. The definition carries over similarly in the multivariate case: for $X^n = (X_1, X_2, \cdots, X_n)$ and $Y^n = (Y_1, Y_2, \cdots, Y_n)$ two random vectors with joint pdfs $f_{X^n}$ and $f_{Y^n}$, respectively, and supports satisfying $S_{X^n} \subseteq S_{Y^n} \subseteq \mathbb{R}^n$, the divergence between $X^n$ and $Y^n$ is defined as

$$D(X^n\|Y^n) \triangleq \int_{S_{X^n}} f_{X^n}(x_1, x_2, \cdots, x_n) \log_2 \frac{f_{X^n}(x_1, x_2, \cdots, x_n)}{f_{Y^n}(x_1, x_2, \cdots, x_n)} \, dx_1 \, dx_2 \cdots \, dx_n$$

when the integral exists.

Definition 5.12 (Mutual information) Let $X$ and $Y$ be two jointly distributed continuous random variables with joint pdf $f_{X,Y}$ and support $S_{X,Y} \subseteq \mathbb{R}^2$. Then the mutual information between $X$ and $Y$ is defined by

$$I(X;Y) \triangleq D(f_{X,Y}\|f_X f_Y) = \int_{S_{X,Y}} f_{X,Y}(x,y) \log_2 \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} \, dx \, dy,$$

assuming the integral exists, where $f_X$ and $f_Y$ are the marginal pdfs of $X$ and $Y$, respectively.

Observation 5.13 For two jointly distributed continuous random variables $X$ and $Y$ with joint pdf $f_{X,Y}$, support $S_{X,Y} \subseteq \mathbb{R}^2$ and joint differential entropy

$$h(X,Y) = -\int_{S_{X,Y}} f_{X,Y}(x,y) \log_2 f_{X,Y}(x,y) \, dx \, dy,$$

then as in Lemma 5.2 and the ensuing discussion, one can write

$$H(q_n(X), q_m(Y)) \approx h(X,Y) + n + m$$

for $n$ and $m$ sufficiently large, where $q_k(Z)$ denotes the (uniformly) quantized version of random variable $Z$ with a $k$-bit accuracy.

On the other hand, for the above continuous $X$ and $Y$,

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

136
\[
\approx [h(X) + n] + [h(Y) + m] - [h(X, Y) + n + m] = h(X) + h(Y) - h(X, Y) = \int_{S_{X,Y}} f_{X,Y}(x,y) \log_2 \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} \, dx \, dy
\]

for \( n \) and \( m \) sufficiently large; in other words,

\[
\lim_{n,m \to \infty} I(q_n(X);q_m(Y)) = h(X) + h(Y) - h(X, Y).
\]

Furthermore, it can be shown that

\[
\lim_{n \to \infty} D(q_n(X)||q_n(Y)) = \int_{S_X} f_X(x) \log_2 \frac{f_X(x)}{f_Y(x)} \, dx.
\]

Thus mutual information and divergence can be considered as the \textit{true} tools of Information Theory, as they retain the same operational characteristics and properties for both discrete and continuous probability spaces (as well as general spaces where they can be defined in terms of Radon-Nikodym derivatives (e.g., cf. [25]).\footnote{This justifies using identical notations for both \( I(\cdot;\cdot) \) and \( D(\cdot\|\cdot) \) as opposed to the discerning notations of \( H(\cdot) \) for entropy and \( h(\cdot) \) for differential entropy.}

The following lemma illustrates that for continuous systems, \( I(\cdot;\cdot) \) and \( D(\cdot\|\cdot) \) keep the same properties already encountered for discrete systems, while differential entropy (as already seen with its possibility if being negative) satisfies some different properties from entropy. The proof is left as an exercise.

**Lemma 5.14** The following properties hold for the information measures of continuous systems.

1. **Non-negativity of divergence:** Let \( X \) and \( Y \) be two continuous random variables with marginal pdfs \( f_X \) and \( f_Y \), respectively, such that their supports satisfy \( S_X \subseteq S_Y \subseteq \mathbb{R} \). Then

   \[
   D(f_X\|f_Y) \geq 0
   \]

   with equality iff \( f_X(x) = f_Y(x) \) for all \( x \in S_X \) except in a set of \( f_X \)-measure zero (i.e., \( X = Y \) almost surely).

2. **Non-negativity of mutual information:** For any two continuous random variables \( X \) and \( Y \),

   \[
   I(X;Y) \geq 0
   \]

   with equality iff \( X \) and \( Y \) are independent.
3. **Conditioning never increases differential entropy:** For any two continuous random variables $X$ and $Y$ with joint pdf $f_{X,Y}$ and well-defined conditional pdf $f_{X|Y}$,

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

with equality iff $X$ and $Y$ are independent.

4. **Chain rule for differential entropy:** For a continuous random vector $X^n = (X_1, X_2, \cdots, X_n)$,

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, X_2, \ldots, X_{i-1}),$$

where $h(X_i|X_1, X_2, \ldots, X_{i-1}) \triangleq h(X_1)$ for $i = 1$.

5. **Chain rule for mutual information:** For continuous random vector $X^n = (X_1, X_2, \cdots, X_n)$ and random variable $Y$ with joint pdf $f_{X^n,Y}$ and well-defined conditional pdfs $f_{X_i,Y|X_i-1}, f_{X_i|X_{i-1}}$ and $f_{Y|X_{i-1}}$ for $i = 1, \cdots, n$, we have that

$$I(X_1, X_2, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \cdots, X_1),$$

where $I(X_i; Y|X_{i-1}, \cdots, X_1) \triangleq I(X_1; Y)$ for $i = 1$.

6. **Data processing inequality:** For continuous random variables $X$, $Y$ and $Z$ such that $X \to Y \to Z$, i.e., $X$ and $Z$ are conditional independent given $Y$ (cf. Appendix B),

$$I(X; Y) \geq I(X; Z).$$

7. **Independence bound for differential entropy:** For a continuous random vector $X^n = (X_1, X_2, \cdots, X_n)$,

$$h(X^n) \leq \sum_{i=1}^{n} h(X_i)$$

with equality iff all the $X_i$'s are independent from each other.

8. **Invariance of differential entropy under translation:** For continuous random variables $X$ and $Y$ with joint pdf $f_{X,Y}$ and well-defined conditional pdf $f_{X|Y}$,

$$h(X + c) = h(X)$$

for any constant $c \in \mathbb{R}$.
Joint differential entropy under linear mapping:

10. Joint differential entropy under nonlinear mapping:

9. Differential entropy under scaling: For any continuous random variable $X$ and any non-zero real constant $a$,

$$h(aX) = h(X) + \log_2 |a|.$$ 

10. Joint differential entropy under linear mapping: Consider the random (column) vector $\underline{X} = (X_1, X_2, \ldots, X_n)^T$ with joint pdf $f_{\underline{X}^n}$, where $T$ denotes transposition, and let $\underline{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ be a random (column) vector obtained from the linear transformation $\underline{Y} = A\underline{X}$, where $A$ is an invertible (non-singular) $n \times n$ real-valued matrix. Then

$$h(\underline{Y}) = h(Y_1, Y_2, \ldots, Y_n) = h(X_1, X_2, \ldots, X_n) + \log_2 |\text{det}(A)|,$$

where $\text{det}(A)$ is the determinant of the square matrix $A$.

11. Joint differential entropy under nonlinear mapping: Consider the random (column) vector $\underline{X} = (X_1, X_2, \ldots, X_n)^T$ with joint pdf $f_{\underline{X}^n}$, and let $\underline{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ be a random (column) vector obtained from the nonlinear transformation \( \underline{Y} = g(\underline{X}) \triangleq (g_1(X_1), g_2(X_2), \ldots, g_n(X_n))^T \), where each $g_i : \mathbb{R} \to \mathbb{R}$ is a differentiable function, $i = 1, 2, \ldots, n$. Then

$$h(\underline{Y}) = h(Y_1, Y_2, \ldots, Y_n)$$

$$= h(X_1, \ldots, X_n) + \int_{\mathbb{R}^n} f_{\underline{X}^n}(x_1, \ldots, x_n) \log_2 |\text{det}(J)| \, dx_1 \cdots dx_n,$$

where $J$ is the $n \times n$ Jacobian matrix given by

$$J \triangleq \begin{bmatrix}
\frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\
\frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial g_n}{\partial x_1} & \frac{\partial g_n}{\partial x_2} & \cdots & \frac{\partial g_n}{\partial x_n}
\end{bmatrix}.$$
Observation 5.15 Property 9 of the above Lemma indicates that for a continuous random variable $X$, $h(X) \neq h(aX)$ (except for the trivial case of $a = 1$) and hence differential entropy is not in general invariant under invertible maps. This is in contrast to entropy, which is always invariant under invertible maps: given a discrete random variable $X$ with alphabet $\mathcal{X}$,

$$H(f(X)) = H(X)$$

for all invertible maps $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is a discrete set; in particular $H(aX) = H(X)$ for all non-zero reals $a$.

On the other hand, for both discrete and continuous systems, mutual information and divergence are invariant under invertible maps:

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

and

$$D(X||Y) = D(g(X)||g(Y))$$

for all invertible maps $g$ and $h$ properly defined on the alphabet/support of the concerned random variables. This reinforces the notion that mutual information and divergence constitute the true tools of Information Theory.

Definition 5.16 (Multivariate Gaussian) A continuous random vector $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$ is called a size-$n$ (multivariate) Gaussian random vector with a finite mean vector $\mu \triangleq (\mu_1, \mu_2, \cdots, \mu_n)^T$, where $\mu_i \triangleq E[X_i] < \infty$ for $i = 1, 2, \cdots, n$, and an $n \times n$ invertible (real-valued) covariance matrix

$$K_\mathbf{X} = [K_{i,j}] \triangleq E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \begin{bmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Cov}(X_n, X_n) \end{bmatrix},$$

where $K_{i,j} = \text{Cov}(X_i, X_j) \triangleq E[(X_i - \mu_i)(X_j - \mu_j)]$ is the covariance between $X_i$ and $X_j$ for $i, j = 1, 2, \cdots, n$, if its joint pdf is given by the multivariate Gaussian pdf

$$f_{\mathbf{X}}(x_1, x_2, \cdots, x_n) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\text{det}(K_\mathbf{X})}} e^{-\frac{1}{2}(\mathbf{x} - \mu)^TK_\mathbf{X}^{-1}(\mathbf{x} - \mu)}$$

for any $(x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$, where $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$. As in the scalar case (i.e., for $n = 1$), we write $\mathbf{X} \sim \mathcal{N}_n(\mu, K_\mathbf{X})$ to denote that $\mathbf{X}$ is a size-$n$ Gaussian random vector with mean vector $\mu$ and covariance matrix $K_\mathbf{X}$.

Note that the diagonal components of $K_\mathbf{X}$ yield the variance of the different random variables: $K_{i,i} = \text{Cov}(X_i, X_i) = \text{Var}(X_i) = \sigma_{X_i}^2$, $i = 1, \cdots, n$. 

140
Observation 5.17 In light of the above definition, we make the following remarks.

1. Note that a covariance matrix $K$ is always symmetric (i.e., $K^T = K$) and positive-semidefinite. But as we require $K_X$ to be invertible in the definition of the multivariate Gaussian distribution above, we will hereafter assume that the covariance matrix of Gaussian random vectors is positive-definite (which is equivalent to having all the eigenvalues of $K_X$ positive), thus rendering the matrix invertible.

2. If a random vector $\mathbf{X} = (X_1, X_2, \cdots, X_n)^T$ has a diagonal covariance matrix $K_X$ (i.e., all the off-diagonal components of $K_X$ are zero: $K_{i,j} = 0$ for all $i \neq j$, $i, j = 1, \cdots, n$), then all its component random variables are uncorrelated but not necessarily independent. However, if $\mathbf{X}$ is Gaussian and have a diagonal covariance matrix, then all its component random variables are independent from each other.

3. Any linear transformation of a Gaussian random vector yields another Gaussian random vector. Specifically, if $\mathbf{X} \sim \mathcal{N}_n(\mu, K_X)$ is a size-$n$ Gaussian random vector with mean vector $\mu$ and covariance matrix $K_X$, and if $\mathbf{Y} = A_{mn} \mathbf{X}$, where $A_{mn}$ is a given $m \times n$ real-valued matrix, then

$$\mathbf{Y} \sim \mathcal{N}_m(A_{mn}\mu, A_{mn}K_X A_{mn}^T)$$

is a size-$m$ Gaussian random vector with mean vector $A_{mn}\mu$ and covariance matrix $A_{mn}K_X A_{mn}^T$.

More generally, any affine transformation of a Gaussian random vector yields another Gaussian random vector: if $\mathbf{X} \sim \mathcal{N}_n(\mu, K_X)$ and $\mathbf{Y} = A_{mn} \mathbf{X} + b_m$, where $A_{mn}$ is a $m \times n$ real-valued matrix and $b_m$ is a size-$m$ real-valued vector, then

$$\mathbf{Y} \sim \mathcal{N}_m(A_{mn}\mu + b_m, A_{mn}K_X A_{mn}^T).$$

---

6 An $n \times n$ real-valued symmetric matrix $K$ is positive-semidefinite (e.g., cf. [16]) if for every real-valued vector $\mathbf{z} = (x_1, x_2, \cdots, x_n)^T$,

$$\mathbf{z}^T K \mathbf{z} = (x_1, \cdots, x_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \geq 0,$$

with equality holding only when $x_i = 0$ for $i = 1, 2, \cdots, n$. Furthermore, the matrix is positive-definite if $\mathbf{z}^T K \mathbf{z} > 0$ for all real-valued vectors $\mathbf{z} \neq \mathbf{0}$, where $\mathbf{0}$ is the all-zero vector of size $n$. 

141
Theorem 5.18 (Joint differential entropy of the multivariate Gaussian)

If \( \mathbf{X} \sim \mathcal{N}_n(\mu, \mathbf{K}_X) \) is a Gaussian random vector with mean vector \( \mu \) and (positive-definite) covariance matrix \( \mathbf{K}_X \), then its joint differential entropy is given by

\[
h(\mathbf{X}) = h(X_1, X_2, \ldots, X_n) = \frac{1}{2} \log_2 [(2\pi e)^n \det(\mathbf{K}_X)]. \tag{5.2.1}
\]

In particular, in the univariate case of \( n = 1 \), (5.2.1) reduces to (5.1.1).

**Proof:** Without loss of generality we assume that \( \mathbf{X} \) has a zero mean vector since its differential entropy is invariant under translation by Property 8 of Lemma 5.14:

\[
h(\mathbf{X}) = h(\mathbf{X} - \mu);
\]
so we assume that \( \mu = 0 \).

Since the covariance matrix \( \mathbf{K}_X \) is a real-valued symmetric matrix, then it is orthogonally diagonalizable; i.e., there exists a square \( (n \times n) \) orthogonal matrix \( \mathbf{A} \) (i.e., satisfying \( \mathbf{A}^T = \mathbf{A}^{-1} \)) such that \( \mathbf{A} \mathbf{K}_X \mathbf{A}^T \) is a diagonal matrix whose entries are given by the eigenvalues of \( \mathbf{K}_X \) (\( \mathbf{A} \) is constructed using the eigenvectors of \( \mathbf{K}_X \); e.g., see [16]). As a result the linear transformation \( \mathbf{Y} = \mathbf{A} \mathbf{X} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{A} \mathbf{K}_X \mathbf{A}^T) \) is a Gaussian vector with the diagonal covariance matrix \( \mathbf{K}_Y = \mathbf{A} \mathbf{K}_X \mathbf{A}^T \) and has therefore independent components (as noted in Observation 5.17). Thus

\[
h(\mathbf{Y}) = h(Y_1, Y_2, \cdots, Y_n)
\]
\[
= h(Y_1) + h(Y_2) + \cdots + h(Y_n) \tag{5.2.2}
\]
\[
= \sum_{i=1}^{n} \frac{1}{2} \log_2 [2\pi e \text{Var}(Y_i)] \tag{5.2.3}
\]
\[
= \frac{n}{2} \log_2 (2\pi e) + \frac{1}{2} \log_2 \left[ \prod_{i=1}^{n} \text{Var}(Y_i) \right]
\]
\[
= \frac{n}{2} \log_2 (2\pi e) + \frac{1}{2} \log_2 \left[ \det(\mathbf{K}_Y) \right] \tag{5.2.4}
\]
\[
= \frac{1}{2} \log_2 (2\pi e)^n + \frac{1}{2} \log_2 \left[ \det(\mathbf{K}_X) \right] \tag{5.2.5}
\]
\[
= \frac{1}{2} \log_2 [(2\pi e)^n \det(\mathbf{K}_X)], \tag{5.2.6}
\]

where (5.2.2) follows by the independence of the random variables \( Y_1, \ldots, Y_n \) (e.g., see Property 7 of Lemma 5.14), (5.2.3) follows from (5.1.1), (5.2.4) holds since the matrix \( \mathbf{K}_Y \) is diagonal and hence its determinant is given by the product of its diagonal entries, and (5.2.5) holds since

\[
\det(\mathbf{K}_Y) = \det(\mathbf{A} \mathbf{K}_X \mathbf{A}^T)
\]

142
\[
\begin{align*}
&= \det(A) \det(K_X) \det(A^T) \\
&= \det(A)^2 \det(K_X) \\
&= \det(K_X),
\end{align*}
\]
where the last equality holds since \((\det(A))^2 = 1\), as the matrix \(A\) is orthogonal \((A^T = A^{-1} \Rightarrow \det(A) = \det(A^T) = 1/|\det(A)|); \) thus, \(\det(A)^2 = 1\). \\
Now invoking Property 10 of Lemma 5.14 and noting that \(|\det(A)| = 1\) yield that \\
\[
\begin{align*}
h(Y_1, Y_2, \cdots, Y_n) &= h(X_1, X_2, \cdots, X_n) + \log_2 |\det(A)| = h(X_1, X_2, \cdots, X_n).
\end{align*}
\]
We therefore obtain using (5.2.6) that \\
\[
\begin{align*}
h(X_1, X_2, \cdots, X_n) &= \frac{1}{2} \log_2 [(2\pi e)^n \det(K_X)],
\end{align*}
\]
hence completing the proof.

An alternate (but rather mechanical) proof to the one presented above consists of directly evaluating the joint differential entropy of \(X\) by integrating \\
\[-f_{X^n}(x^n) \log_2 f_{X^n}(x^n)\] over \(\mathbb{R}^n\); it is left as an exercise. \(\Box\)

**Corollary 5.19 (Hadamard’s inequality)** For any real-valued \(n \times n\) positive-definite matrix \(K = [K_{i,j}]_{i,j=1,\ldots,n}\), \\
\[
\det(K) \leq \prod_{i=1}^{n} K_{i,i}
\]
with equality iff \(K\) is a diagonal matrix, where \(K_{i,i}\) are the diagonal entries of \(K\).

**Proof:** Since every positive-definite matrix is a covariance matrix (e.g., see [22]), let \(X = (X_1, X_2, \cdots, X_n)^T \sim \mathcal{N}_n(\underline{0}, K)\) be a jointly Gaussian random vector with zero mean vector and covariance matrix \(K\). Then \\
\[
\begin{align*}
\frac{1}{2} \log_2 [(2\pi e)^n \det(K)] &= h(X_1, X_2, \cdots, X_n) \\
&\leq \sum_{i=1}^{n} h(X_i) \\
&= \sum_{i=1}^{n} \frac{1}{2} \log_2 [2\pi e \text{Var}(X_i)]
\end{align*}
\]
\[ \det(K) \leq \prod_{i=1}^{n} K_{i,i}, \]

with equality iff the jointly Gaussian random variables \( X_1, X_2, \ldots, X_n \) are independent from each other, or equivalently iff the covariance matrix \( K \) is diagonal. \( \square \)

The next theorem states that among all real-valued size-\( n \) random vectors (of support \( \mathbb{R}^n \)) with identical mean vector and covariance matrix, the Gaussian random vector has the largest differential entropy.

**Theorem 5.20 (Maximal differential entropy for real-valued random vectors)** Let \( X = (X_1, X_2, \ldots, X_n)^T \) be a real-valued random vector with support \( S_X = \mathbb{R}^n \), mean vector \( \mu \) and covariance matrix \( K_X \). Then

\[ h(X_1, X_2, \ldots, X_n) \leq \frac{1}{2} \log_2 \left( (2\pi e)^n \det(K_X) \right), \]

with equality iff \( X \) is Gaussian; i.e., \( X \sim \mathcal{N}(\mu, K_X) \).

**Proof:** We will present the proof in two parts: the scalar or univariate case, and the multivariate case.

(i) **Scalar case** \((n = 1)\): For a real-valued random variable with support \( S_X = \mathbb{R} \), mean \( \mu \) and variance \( \sigma^2 \), let us show that

\[ h(X) \leq \frac{1}{2} \log_2 \left( 2\pi e \sigma^2 \right), \]

with equality iff \( X \sim \mathcal{N}(\mu, \sigma^2) \).

For a Gaussian random variable \( Y \sim \mathcal{N}(\mu, \sigma^2) \), using the non-negativity of divergence, can write

\[ 0 \leq D(X\|Y) \]
\[
\begin{align*}
&= \int_\mathbb{R} f_X(x) \log_2 \frac{f_X(x)}{\sqrt{2\pi\sigma^2} e^{-(x-\mu)^2/2\sigma^2}} \, dx \\
&= -h(X) + \int_\mathbb{R} f_X(x) \left[ \log_2 \left( \sqrt{2\pi\sigma^2} \right) + \frac{(x-\mu)^2}{2\sigma^2} \log_2 e \right] \, dx \\
&= -h(X) + \frac{1}{2} \log_2 (2\pi\sigma^2) + \frac{1}{2} \log_2 e \int_\mathbb{R} (x-\mu)^2 f_X(x) \, dx \\
&= -h(X) + \frac{1}{2} \log_2 \left( 2\pi e \sigma^2 \right).
\end{align*}
\]

Thus
\[
h(X) \leq \frac{1}{2} \log_2 \left( 2\pi e \sigma^2 \right),
\]
with equality iff \(X = Y\) (almost surely); i.e., \(X \sim \mathcal{N}(\mu, \sigma^2)\).

(ii). Multivariate case \((n > 1)\): As in the proof of Theorem 5.18, we can use an orthogonal square matrix \(A\) (i.e., satisfying \(A^T = A^{-1}\) and hence \(|\det(A)| = 1\)) such that \(AK_X A^T\) is diagonal. Therefore, the random vector generated by the linear map
\[
Z = AX
\]
will have a covariance matrix given by \(K_Z = AK_X A^T\) and hence have uncorrelated (but not necessarily independent) components. Thus
\[
h(X) = h(Z) - \log_2 |\det(A)| = 0 \quad (5.2.13)
\]
\[
\leq \sum_{i=1}^n h(Z_i) \quad (5.2.14)
\]
\[
\leq \sum_{i=1}^n \frac{1}{2} \log_2 \left( 2\pi e \text{Var}(Z_i) \right) \quad (5.2.15)
\]
\[
= \frac{n}{2} \log_2 (2\pi e) + \frac{1}{2} \log_2 \left[ \prod_{i=1}^n \text{Var}(Z_i) \right] \quad (5.2.16)
\]
\[
= \frac{1}{2} \log_2 (2\pi e)^n + \frac{1}{2} \log_2 \left[ |\det(K_Z)| \right] \quad (5.2.17)
\]
\[
= \frac{1}{2} \log_2 \left( (2\pi e)^n |\det(K_X)| \right),
\]
where (5.2.13) holds by Property 10 of Lemma 5.14 and since \(|\det(A)| = 1\), (5.2.14) follows from Property 7 of Lemma 5.14, (5.2.15) follows from (5.2.12)
(the scalar case above), (5.2.16) holds since $K_Z$ is diagonal, and (5.2.17) follows from the fact that $\det(K_Z) = \det(K_X)$ (as $A$ is orthogonal). Finally, equality is achieved in both (5.2.14) and (5.2.15) iff the random variables $Z_1, Z_2, \ldots, Z_n$ are Gaussian and independent from each other, or equivalently iff $\overline{X} \sim \mathcal{N}_n(\mu, K_X)$.

**Observation 5.21** The following two results can also be shown (the proof is left as an exercise):

1. Among all continuous random variables admitting a pdf with support the interval $(a, b)$, where $b > a$ are real numbers, the uniformly distributed random variable maximizes differential entropy.

2. Among all continuous random variables admitting a pdf with support the interval $[0, \infty)$ and finite mean $\mu$, the exponential distribution with parameter (or rate parameter) $\lambda = 1/\mu$ maximizes differential entropy.

A systematic approach to finding distributions that maximize differential entropy subject to various support and moments constraints can be found in [13, 53].

### 5.3 AEP for continuous memoryless sources

The AEP theorem and its consequence for discrete memoryless (i.i.d.) sources reveal to us that the number of elements in the typical set is approximately $2^{nH(X)}$, where $H(X)$ is the source entropy, and that the typical set carries almost all the probability mass asymptotically (see Theorems 3.3 and 3.4). An extension of this result from discrete to continuous memoryless sources by just counting the number of elements in a continuous (typical) set defined via a law-of-large-numbers argument is not possible, since the total number of elements in a continuous set is infinite. However, when considering the volume of that continuous typical set (which is a natural analog to the size of a discrete set), such an extension, with differential entropy playing a similar role as entropy, becomes straightforward.

**Theorem 5.22 (AEP for continuous memoryless sources)** Let $\{X_i\}_{i=1}^{\infty}$ be a continuous memoryless source (i.e., an infinite sequence of continuous i.i.d. random variables) with pdf $f_X(\cdot)$ and differential entropy $h(X)$. Then

$$-\frac{1}{n} \log f_X(X_1, \ldots, X_n) \to E[-\log f_X(X)] = h(X) \quad \text{in probability}.$$  

**Proof:** The proof is an immediate result of the law of large numbers (e.g., see Theorem 3.3).
Definition 5.23 (Typical set) For $\delta > 0$ and any $n$ given, define the typical set for the above continuous source as

$$\mathcal{F}_n(\delta) \triangleq \left\{ x^n \in \mathbb{R}^n : \left| -\frac{1}{n} \log_2 f_X(X_1, \ldots, X_n) - h(X) \right| < \delta \right\}.$$

Definition 5.24 (Volume) The volume of a set $A \subset \mathbb{R}^n$ is defined as

$$\text{Vol}(A) \triangleq \int_A dx_1 \cdots dx_n.$$

Theorem 5.25 (Consequence of the AEP for continuous memoryless sources) For a continuous memoryless source $\{X_i\}_{i=1}^\infty$ with differential entropy $h(X)$, the following hold.

1. For $n$ sufficiently large, $P_{X^n} \{\mathcal{F}_n(\delta)\} > 1 - \delta$.
2. $\text{Vol}(\mathcal{F}_n(\delta)) \leq 2^{n(h(X) + \delta)}$ for all $n$.
3. $\text{Vol}(\mathcal{F}_n(\delta)) \geq (1 - \delta)2^{n(h(X) - \delta)}$ for $n$ sufficiently large.

Proof: The proof is quite analogous to the corresponding theorem for discrete memoryless sources (Theorem 3.4) and is left as an exercise.

5.4 Capacity and channel coding theorem for the discrete-time memoryless Gaussian channel

We next study the fundamental limits for error-free communication over the discrete-time memoryless Gaussian channel, which is the most important continuous-alphabet channel and is widely used to model real-world wired and wireless channels. We first state the definition of discrete-time continuous-alphabet memoryless channels.

Definition 5.26 (Discrete-time continuous memoryless channels) Consider a discrete-time channel with continuous input and output alphabets given by $\mathcal{X} \subseteq \mathbb{R}$ and $\mathcal{Y} \subseteq \mathbb{R}$, respectively, and described by a sequence of $n$-dimensional transition (conditional) pdfs $\{f_{Y^n|X^n}(y^n|x^n)\}_{n=1}^\infty$ that govern the reception of $y^n = (y_1, y_2, \ldots, y_n) \in \mathcal{Y}^n$ at the channel output when $x^n = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ is sent as the channel input.

The channel (without feedback) is said to be memoryless with a given (marginal) transition pdf $f_{Y^n|X^n}$ if its sequence of transition pdfs $f_{Y^n|X^n}$ satisfies

$$f_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^n f_{Y|X}(y_i|x_i) \quad (5.4.1)$$
for every \( n = 1, 2, \cdots, x^n \in X^n \) and \( y^n \in Y^n \).

In practice, the real-valued input to a continuous channel satisfies a certain constraint or limitation on its amplitude or power; otherwise, one would have a realistically implausible situation where the input can take on any value from the uncountably infinite set of real numbers. We will thus impose an *average cost constraint* \((t, P)\) on any input \( n \)-tuple \( x^n = (x_1, x_2, \cdots, x_n) \) transmitted over the channel by requiring that

\[
\frac{1}{n} \sum_{i=1}^{n} t(x_i) \leq P,
\]

where \( t(\cdot) \) is a given non-negative real-valued function describing the cost for transmitting an input symbol, and \( P \) is a given positive number representing the maximal average amount of available resources per input symbol.

**Definition 5.27** The capacity (or capacity-cost function) of a discrete-time continuous memoryless channel with input average cost constraint \((t, P)\) is denoted by \( C(P) \) and defined as

\[
C(P) \triangleq \sup_{F_X : E[t(X)] \leq P} I(X; Y) \quad \text{(in bits/channel use)}
\]

where the supremum is over all input distributions \( F_X \).

**Lemma 5.28 (Concavity of capacity)** If \( C(P) \) as defined in (5.4.3) is finite for any \( P > 0 \), then it is concave, continuous and strictly increasing in \( P \).

**Proof:** Fix \( P_1 > 0 \) and \( P_2 > 0 \). Then since \( C(P) \) is finite for any \( P > 0 \), then by the 3rd property in Property A.4, there exist two input distributions \( F_{X_1} \) and \( F_{X_2} \) such that for all \( \epsilon > 0 \),

\[
I(X_i; Y_i) \geq C(P_i) - \epsilon
\]

and

\[
E[t(X_i)] \leq P_i
\]

where \( X_i \) denotes the input with distribution \( F_{X_i} \) and \( Y_i \) is the corresponding channel output for \( i = 1, 2 \). Now, for \( 0 \leq \lambda \leq 1 \), let \( X_\lambda \) be a random variable with distribution \( F_{X_\lambda} \triangleq \lambda F_{X_1} + (1 - \lambda) F_{X_2} \). Then by (5.4.5)

\[
E_{X_\lambda}[t(X)] = \lambda E_{X_1}[t(X)] + (1 - \lambda) E_{X_2}[t(X)] \leq \lambda P_1 + (1 - \lambda) P_2.
\]

Furthermore,

\[
C(\lambda P_1 + (1 - \lambda) P_2) = \sup_{F_X : E[t(X)] \leq \lambda P_1 + (1 - \lambda) P_2} I(F_X, f_{Y|X})
\]
\[ \begin{align*}
\geq & \quad I(F_{X_1}, f_{Y_1}|X) \\
\geq & \quad \lambda I(F_{X_1}, f_{Y_1}|X) + (1 - \lambda)I(F_{X_2}, f_{Y_1}|X) \\
= & \quad \lambda I(X_1; Y_1) + (1 - \lambda)I(X_2; Y_2) \\
\geq & \quad \lambda C(P_1) - \epsilon + (1 - \lambda)C(P_2) - \epsilon, \\
\end{align*} \]

where the first inequality holds by (5.4.6), the second inequality follows from the concavity of the mutual information with respect to its first argument (cf. Lemma 2.46) and the third inequality follows from (5.4.4). Letting \( \epsilon \to 0 \) yields that

\[ C(\lambda P_1 + (1 - \lambda)P_2) \geq \lambda C(P_1) + (1 - \lambda)C(P_2) \]

and hence \( C(P) \) is concave in \( P \).

Finally, it can directly be seen by definition that \( C(\cdot) \) is non-decreasing, which, together with its concavity, imply that it is continuous and strictly increasing. □

The most commonly used cost function is the power cost function, \( t(x) = x^2 \), resulting in the average power constraint \( P \) for each transmitted input \( n \)-tuple:

\[ \frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P. \]  

(5.4.7)

Throughout this chapter, we will adopt this average power constraint on the channel input.

We herein focus on the discrete-time memoryless Gaussian channel with average input power constraint \( P \) and establish an operational meaning for the channel capacity \( C(P) \) as the largest coding rate for achieving reliable communication over the channel. The channel is described by the following additive noise equation:

\[ Y_i = X_i + Z_i, \quad \text{for} \quad i = 1, 2, \cdots, \]  

(5.4.8)

where \( Y_i, X_i \) and \( Z_i \) are the channel output, input and noise at time \( i \). The input and noise processes are assumed to be independent from each other and the noise source \( \{Z_i\}_{i=1}^{\infty} \) is i.i.d. Gaussian with each \( Z_i \) having mean zero and variance \( \sigma^2 \), \( Z_i \sim \mathcal{N}(0, \sigma^2) \). Since the noise process is i.i.d, we directly get that the channel satisfies (5.4.1) and is hence memoryless, where the channel transition pdf is explicitly given in terms of the noise pdf as follows:

\[ f_{Y|X}(y|x) = f_Z(y - x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-x)^2}{2\sigma^2}}. \]

As mentioned above, we impose the average power constraint (5.4.7) on the channel input.
Observation 5.29 The memoryless Gaussian channel is a good approximating model for many practical channels such as radio, satellite and telephone line channels. The additive noise is usually due to a multitude of causes, whose cumulative effect can be approximated via the Gaussian distribution. This is justified by the Central Limit Theorem which states that for an i.i.d. process \( \{U_i\}_{i=1}^{\infty} \) with mean \( \mu \) and variance \( \sigma^2 \), 
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (U_i - \mu) \text{ converges in distribution as } n \to \infty \text{ to a Gaussian distributed random variable with mean zero and variance } \sigma^2 \text{ (see Appendix B)}.
\]

Before proving the channel coding theorem for the above memoryless Gaussian channel with input power constraint \( P \), we first show that its capacity \( C(P) \) as defined in (5.4.3) with \( t(x) = x^2 \) admits a simple expression in terms of \( P \) and the channel noise variance \( \sigma^2 \). Indeed, we can write the channel mutual information \( I(X; Y) \) between its input and output as follows:

\[
I(X; Y) = h(Y) - h(Y|X) = h(Y) - h(X + Z|X) = h(Y) - h(Z|X) = h(Y) - h(Z)
\]

(5.4.11)

\[
= h(Y) - \frac{1}{2} \log_2 (2\pi e \sigma^2),
\]

(5.4.12)

where (5.4.9) follows from (5.4.8), (5.4.10) holds since differential entropy is invariant under translation (see Property 8 of Lemma 5.14), (5.4.11) follows from the independence of \( X \) and \( Z \), and (5.4.12) holds since \( Z \sim \mathcal{N}(0, \sigma^2) \) is Gaussian (see (5.1.1)). Now since \( Y = X + Z \), we have that

\[
E[Y^2] = E[X^2] + E[Z^2] + 2E[X]E[Z] = E[X^2] + \sigma^2 + 2E[X](0) \leq P + \sigma^2
\]

since the input in (5.4.3) is constrained to satisfy \( E[X^2] \leq P \). Thus the variance of \( Y \) satisfies \( \text{Var}(Y) \leq E[Y^2] \leq P + \sigma^2 \), and

\[
h(Y) \leq \frac{1}{2} \log_2 (2\pi e \text{Var}(Y)) \leq \frac{1}{2} \log_2 (2\pi e (P + \sigma^2))
\]

where the first inequality follows by Theorem 5.20 since \( Y \) is real-valued (with support \( \mathbb{R} \)). Noting that equality holds in the first inequality above iff \( Y \) is Gaussian and in the second inequality iff \( \text{Var}(Y) = P + \sigma^2 \), we obtain that choosing the input \( X \) as \( X \sim \mathcal{N}(0, P) \) yields \( Y \sim \mathcal{N}(0, P + \sigma^2) \) and hence maximizes \( I(X; Y) \) over all inputs satisfying \( E[X^2] \leq P \). Thus, the capacity of the discrete-time memoryless Gaussian channel with input average power constraint \( P \) and noise variance (or power) \( \sigma^2 \) is given by

\[
C(P) = \frac{1}{2} \log_2 (2\pi e (P + \sigma^2)) - \frac{1}{2} \log_2 (2\pi e \sigma^2)
\]

150
\[ = \frac{1}{2} \log_2 \left( 1 + \frac{P}{\sigma^2} \right). \tag{5.4.13} \]

**Definition 5.30** Given positive integers \( n \) and \( M \), and a discrete-time memoryless Gaussian channel with input average power constraint \( P \), a fixed-length data transmission code (or block code) \( \mathcal{C}_n = (n, M) \) for this channel with blocklength \( n \) and rate \( \frac{1}{n} \log_2 M \) message bits per channel symbol (or channel use) consists of:

1. \( M \) information messages intended for transmission.
2. An encoding function
   \[ f : \{1, 2, \ldots, M\} \to \mathbb{R}^n \]
   yielding real-valued codewords \( c_1 = f(1), c_2 = f(2), \ldots, c_M = f(M) \), where each codeword \( c_m = (c_{m1}, \ldots, c_{mn}) \) is of length \( n \) and satisfies the power constraint \( P \)
   \[ \frac{1}{n} \sum_{i=1}^{n} c_i^2 \leq P, \]
   for \( m = 1, 2, \ldots, M \). The set of these \( M \) codewords is called the codebook and we usually write \( \mathcal{C}_n = \{c_1, c_2, \ldots, c_M\} \) to list the codewords.
3. A decoding function \( g : \mathbb{R}^n \to \{1, 2, \ldots, M\} \).

As in Chapter 4, we assume that a message \( W \) follows a uniform distribution over the set of messages: \( \Pr[W = w] = \frac{1}{M} \) for all \( w \in \{1, 2, \ldots, M\} \). Similarly, to convey message \( W \) over the channel, the encoder sends its corresponding codeword \( X^n = f(W) \in \mathcal{C}_n \) at the channel input. Finally, \( Y^n \) is received at the channel output and the decoder yields \( \hat{W} = g(Y^n) \) as the message estimate. Also, the average probability of error for this block code used over the memoryless Gaussian channel is defined as

\[ P_e(\mathcal{C}_n) \triangleq \frac{1}{M} \sum_{w=1}^{M} \lambda_w(\mathcal{C}_n), \]

where

\[ \lambda_w(\mathcal{C}_n) \triangleq \Pr[\hat{W} \neq W | W = w] = \Pr[g(Y^n) \neq w | X^n = f(w)] = \int_{y^n \in \mathbb{R}^n : g(y^n) \neq w} f_{Y^n|X^n}(y^n | f(w)) \, dy^n \]
is the code’s conditional probability of decoding error given that message \( w \) is sent over the channel. Here \( f_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} f_{Y|X}(y_i|x_i) \) as the channel is memoryless, where \( f_{Y|X} \) is the channel’s transition pdf.

We next prove that for a memoryless Gaussian channel with input average power constraint \( P \), its capacity \( C(P) \) has an operational meaning in the sense that it is the supremum of all rates for which there exists a sequence of data transmission block codes satisfying the power constraint and having a probability of error that vanishes with increasing blocklength.

**Theorem 5.31 (Shannon’s coding theorem for the memoryless Gaussian channel)** Consider a discrete-time memoryless Gaussian channel with input average power constraint \( P \), channel noise variance \( \sigma^2 \) and capacity \( C(P) \) as given by (5.4.13).

- **Forward part (achievability):** For any \( \varepsilon \in (0, 1) \), there exist \( 0 < \gamma < 2\varepsilon \) and a sequence of data transmission block code \( \{\mathcal{C}_n = (n, M_n)\}_{n=1}^{\infty} \) satisfying
  \[
  \frac{1}{n} \log_2 M_n > C(P) - \gamma
  \]
  with each codeword \( c = (c_1, c_2, \ldots, c_n) \) in \( \mathcal{C}_n \) satisfying
  \[
  \frac{1}{n} \sum_{i=1}^{n} c_i^2 \leq P
  \]  \tag{5.4.14}
  such that the probability of error \( P_e(\mathcal{C}_n) < \varepsilon \) for sufficiently large \( n \).

- **Converse part:** If for any sequence of data transmission block codes \( \{\mathcal{C}_n = (n, M_n)\}_{n=1}^{\infty} \) whose codewords satisfy (5.4.14), we have that
  \[
  \liminf_{n \to \infty} \frac{1}{n} \log_2 M_n > C(P),
  \]
  then the codes’ probability of error \( P_e(\mathcal{C}_n) \) is bounded away from zero for all \( n \) sufficiently large.

**Proof of the forward part:** The theorem holds trivially when \( C(P) = 0 \) because we can choose \( M_n = 1 \) for every \( n \) and have \( P_e(\mathcal{C}_n) = 0 \). Hence, we assume without loss of generality \( C(P) > 0 \).

**Step 0:**

Take a positive \( \gamma \) satisfying \( \gamma < \min\{2\varepsilon, C(P)\} \). Pick \( \xi > 0 \) small enough such that \( 2[C(P) - C(P - \xi)] < \gamma \), where the existence of such \( \xi \) is assured.
by the strictly increasing property of $C(P)$. Hence, we have $C(P - \xi) - \gamma/2 > C(P) - \gamma > 0$. Choose $M_n$ to satisfy
\[
C(P - \xi) - \frac{\gamma}{2} > \frac{1}{n} \log_2 M_n > C(P) - \gamma,
\]
for which the choice should exist for all sufficiently large $n$. Take $\delta = \gamma/8$.

Let $F_X$ be the distribution that achieves $C(P - \xi)$, where $C(P)$ is given by (5.4.13). In this case, $F_X$ is the Gaussian distribution with mean zero and variance $P - \xi$ and admits a pdf $f_X$. Hence, $E[X^2] \leq P - \xi$ and $I(X; Y) = C(P - \xi)$.

**Step 1: Random coding with average power constraint.**

Randomly draw $M_n$ codewords according to pdf $f_{X^n}$ with
\[
f_{X^n}(x^n) = \prod_{i=1}^{n} f_X(x_i).
\]

By law of large numbers, each randomly selected codeword
\[
c_m = (c_{m1}, \ldots, c_{mn})
\]
satisfies
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} c_{mi}^2 = E[X^2] \leq P - \xi
\]
for $m = 1, 2, \ldots, M_n$.

**Step 2: Code construction.**

For $M_n$ selected codewords $\{c_1, \ldots, c_{M_n}\}$, replace the codewords that violate the power constraint (i.e., (5.4.14)) by an all-zero (default) codeword $0$. Define the encoder as
\[
f_n(m) = c_m \quad \text{for} \quad 1 \leq m \leq M_n.
\]

Given a received output sequence $y^n$, the decoder $g_n(\cdot)$ is given by
\[
g_n(y^n) = \begin{cases} 
  m, & \text{if } (c_m, y^n) \in \mathcal{F}_n(\delta) \\
  \text{arbitrary}, & \text{otherwise},
\end{cases}
\]
where the set
\[
\mathcal{F}_n(\delta) \triangleq \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left| -\frac{1}{n} \log_2 f_{X^nY^n}(x^n, y^n) - h(X, Y) \right| < \delta \}.
\]
Step 3: Conditional probability of error.

Let $\lambda_m$ denote the conditional error probability given codeword $m$ is transmitted. Define

$$E_0 \triangleq \left\{ x^n \in X^n : \frac{1}{n} \sum_{i=1}^{n} x_i^2 > P \right\}.$$

Then by following similar argument as (4.3.4), we get:

$$E[\lambda_m] \leq P_{X^n}(E_0) + P_{X^n,Y^n}(F_n(\delta))$$
$$+ \sum_{m'=1}^{M_n} \int_{c_m \in X^n} \int_{y^n \in F_n(\delta|c_{m'})} f_{X^n,Y^n}(c_m, y^n) \, dy^n \, dc_m,$$

(5.4.15)

---

7In this proof, specifically, (4.3.3) becomes:

$$\lambda_m(E_n) \leq \int y^n \notin F_n(\delta|c_m) f_{Y^n|X^n}(y^n|c_m) \, dy^n + \sum_{m'=1}^{M_n} \int_{y^n \in F_n(\delta|c_{m'})} f_{Y^n|X^n}(y^n|c_m) \, dy^n.$$

By taking expectation with respect to the $m^{th}$ codeword-selecting distribution $f_{X^n}(c_m)$, we obtain

$$E[\lambda_m] = \int_{c_m \in X^n} f_{X^n}(c_m) \lambda_m(c_m) \, dc_m$$
$$= \int_{c_m \in X^n \cap E_0} f_{X^n}(c_m) \lambda_m(c_m) \, dc_m + \int_{c_m \in X^n \cap E_0^c} f_{X^n}(c_m) \lambda_m(c_m) \, dc_m$$
$$\leq \int_{c_m \in E_0} f_{X^n}(c_m) \, dc_m + \int_{c_m \in X^n} f_{X^n}(c_m) \lambda_m(c_m) \, dc_m$$
$$\leq P_{X^n}(E_0) + \int_{c_m \in X^n} \int y^n \notin F_n(\delta|c_m) f_{X^n}(c_m) f_{Y^n|X^n}(y^n|c_m) \, dy^n \, dc_m$$
$$+ \sum_{m'=1}^{M_n} \int_{y^n \in F_n(\delta|c_{m'})} f_{X^n}(c_m) f_{Y^n|X^n}(y^n|c_m) \, dy^n \, dc_m.$$
where 
\[ F_n(\delta|x^n) \triangleq \{ y^n \in Y^n : (x^n, y^n) \in F_n(\delta) \}. \]

Note that the additional term \( P_{X^n}(E_0) \) in (5.4.15) is to cope with the errors due to all-zero codeword replacement, which will be less than \( \delta \) for all sufficiently large \( n \) by the law of large numbers. Finally, by carrying out a similar procedure as in the proof of the channel coding theorem for discrete channels (cf. page 108), we obtain:

\[
E[P_n(C_n)] \leq P_{X^n}(E_0) + P_{X^n,Y^n}(F_n(\delta)) + M_n \cdot 2^{n(h(X,Y)+\delta)} 2^{-n(h(X)-\delta)} 2^{-n(h(Y)-\delta)} \\
\leq P_{X^n}(E_0) + P_{X^n,Y^n}(F_n(\delta)) + 2^{n(C-\gamma-4\delta)} 2^{-n(I(X;Y)-3\delta)} \\
= P_{X^n}(E_0) + P_{X^n,Y^n}(F_n(\delta)) + 2^{-n\delta}.
\]

Accordingly, we can make the average probability of error, \( E[P_n(C_n)] \), less than \( 3\delta = 3\gamma/8 < 3\varepsilon/4 < \varepsilon \) for all sufficiently large \( n \).

\[ \square \]

**Proof of the converse part:** Consider an \((n, M_n)\) block data transmission code satisfying the power constraint (5.4.14) with encoding function

\[ f_n : \{1, 2, \ldots, M_n\} \rightarrow X^n \]

and decoding function

\[ g_n : Y^n \rightarrow \{1, 2, \ldots, M_n\}. \]

Since the message \( W \) is uniformly distributed over \( \{1, 2, \ldots, M_n\} \), we have \( H(W) = \log_2 M_n \). Since \( W \rightarrow X^n = f_n(W) \rightarrow Y^n \) forms a Markov chain (as \( Y^n \) only depends on \( X^n \)), we obtain by the data processing lemma that \( I(W;Y^n) \leq I(X^n;Y^n) \). We can also bound \( I(X^n;Y^n) \) by \( C(P) \) as follows:

\[
I(X^n;Y^n) \leq \sup_{F_{X^n} : (1/n) \sum_{i=1}^n E[X_i^2] \leq P} I(X^n;Y^n) \\
\leq \sup_{F_{X^n} : (1/n) \sum_{i=1}^n E[X_i^2] \leq P} \sum_{j=1}^n I(X_j;Y_j) \quad \text{(by Theorem 2.21)} \\
= \sup_{(P_1, P_2, \ldots, P_n) : (1/n) \sum_{i=1}^n P_i = P} \sup_{F_{X^n} : (\forall i) \ E[X_i^2] \leq P_i} \sum_{j=1}^n I(X_j;Y_j) \\
\leq \sup_{(P_1, P_2, \ldots, P_n) : (1/n) \sum_{i=1}^n P_i = P} \sum_{j=1}^n \sup_{F_{X_j} : E[X_j^2] \leq P_j} I(X_j;Y_j) \\
= \sup_{(P_1, P_2, \ldots, P_n) : (1/n) \sum_{i=1}^n P_i = P} \sum_{j=1}^n \sup_{F_{X_j} : E[X_j^2] \leq P_j} I(X_j;Y_j)
\]

155
\[
\begin{align*}
&= \sup_{(P_1,P_2,...,P_n):(1/n)\sum_{j=1}^n P_j = P} \frac{1}{n} \sum_{j=1}^n C(P_j) \\
&= \sup_{(P_1,P_2,...,P_n):(1/n)\sum_{j=1}^n P_j = P} \frac{1}{n} \sum_{j=1}^n C(P_j) \\
&\leq \sup_{(P_1,P_2,...,P_n):(1/n)\sum_{j=1}^n P_j = P} nC\left(\frac{1}{n} \sum_{j=1}^n P_j\right) \quad \text{(by concavity of } C(P)) \\
&= nC(P).
\end{align*}
\]

Consequently, recalling that \( P_e(\mathcal{C}_n) \) is the average error probability incurred by guessing \( W \) from observing \( Y^n \) via the decoding function \( g_n : Y^n \to \{1, 2, \ldots, M_n\} \), we get

\[
\log_2 M_n = H(W) = H(W|Y^n) + I(W;Y^n) \leq H(W|Y^n) + I(X^n;Y^n) \leq h_b(P_e(\mathcal{C}_n)) + P_e(\mathcal{C}_n) \cdot \log_2(|W| - 1) + nC(P)
\]

(by Fano's inequality)

\[
\leq 1 + P_e(\mathcal{C}_n) \cdot \log_2(M_n - 1) + nC(P),
\]

(by the fact that \( \forall t \in [0,1] \) \( h_b(t) \leq 1 \))

\[
< 1 + P_e(\mathcal{C}_n) \cdot \log_2 M_n + nC(P),
\]

which implies that

\[
P_e(\mathcal{C}_n) > 1 - \frac{C(P)}{(1/n) \log_2 M_n} - \frac{1}{\log_2 M_n}.
\]

So if \( \liminf_{n \to \infty} (1/n) \log_2 M_n > C(P) \), then there exists \( \delta > 0 \) and an integer \( N \) such that for \( n \geq N \),

\[
\frac{1}{n} \log_2 M_n > C(P) + \delta.
\]

Hence, for \( n \geq N_0 \equiv \max\{N, 2/\delta\} \),

\[
P_e(\mathcal{C}_n) \geq 1 - \frac{C(P)}{C(P) + \delta} - \frac{1}{n(C(P) + \delta)} \geq \frac{\delta}{2(C(P) + \delta)}.
\]

\[
\Box
\]

We next show that among all power-constrained continuous memoryless channels with additive noise admitting a pdf, choosing a Gaussian distributed noise yields the smallest channel capacity. In other words, the memoryless Gaussian model results in the most pessimistic (smallest) capacity within the class of additive-noise continuous memoryless channels.
Theorem 5.32 (Gaussian noise minimizes capacity of additive-noise channels)Every discrete-time continuous memoryless channel with additive noise (admitting a pdf) of mean zero and variance $\sigma^2$ and input average power constraint $P$ has its capacity $C(P)$ lower bounded by the capacity of the memoryless Gaussian channel with identical input constraint and noise variance:

$$C(P) \geq \frac{1}{2} \log_2 \left( 1 + \frac{P}{\sigma^2} \right).$$

**Proof:** Let $f_{Y|X}$ and $f_{Y_g|X_g}$ denote the transition pdfs of the additive-noise channel and the Gaussian channel, respectively, where both channels satisfy input average power constraint $P$. Let $Z$ and $Z_g$ respectively denote their zero-mean noise variables of identical variance $\sigma^2$. Writing the mutual information in terms of the channel’s transition pdf and input distribution as in Lemma 2.46, then for any Gaussian input with pdf $f_{X_g}$ with corresponding outputs $Y$ and $Y_g$ when applied to channels $f_{Y|X}$ and $f_{Y_g|X_g}$, respectively, we have that

$$I(f_{Y_g}, f_{Y|X}) - I(f_{X_g}, f_{Y_g|X_g})$$

$$= \int_X \int_Y f_{X_g}(x) f_Z(y - x) \log_2 \frac{f_Z(y) - x}{f_Y(y)} dy dx$$

$$- \int_X \int_Y f_{X_g}(x) f_{Z_g}(y - x) \log_2 \frac{f_{Z_g}(y) - x}{f_{Y_g}(y)} dy dx$$

$$= \int_X \int_Y f_{X_g}(x) f_Z(y - x) \log_2 \frac{f_Z(y) - x}{f_Y(y)} dy dx$$

$$- \int_X \int_Y f_{X_g}(x) f_{Z_g}(y - x) \log_2 \frac{f_{Z_g}(y) - x}{f_{Y_g}(y)} dy dx$$

$$\geq \int_X \int_Y f_{X_g}(x) f_Z(y - x) \log_2 (1 - \frac{f_{Z_g}(y - x) f_Y(y)}{f_Z(y) f_{Y_g}(y)}) dy dx$$

$$= (\log_2 e) \left[ 1 - \int_Y f_Y(y) \left( \int_X f_{X_g}(x) f_{Z_g}(y - x) dx \right) dy \right]$$

$$= 0,$$

with equality holding in the inequality iff $f_Y(y) / f_{Y_g}(y) = f_Z(y - x) / f_{Z_g}(y - x)$ for all $x$. Therefore,

$$\frac{1}{2} \log_2 \left( 1 + \frac{P}{\sigma^2} \right) = \sup_{F_X : E[X^2] \leq P} I(F_X, f_{Y_g|X_g})$$

$$= I(f_{X_g}, f_{Y_g|X_g})$$

157
\[ I(f_{X_y}; f_{Y|x}) \]
\[ \leq \sup_{F_X : E[X^2] \leq P} I(F_X, f_{Y|x}) \]
\[ = C(P). \]

Observation 5.33 (Channel coding theorem for continuous memoryless channels) We close this section by noting that Theorem 5.31 can be generalized to a wide class of discrete-time continuous memoryless channels with input cost constraint (5.4.2) where the cost function \( t(\cdot) \) is arbitrary, by showing that \( C(P) \triangleq \sup_{F_X : E[t(X)] \leq P} I(X; Y) \) is the largest rate for which there exist block codes for the channel satisfying (5.4.2) which are reliably good (i.e., with asymptotically vanishing error probability). The proof is quite similar to that of Theorem 5.31, except that some modifications are needed in the forward part as for a general (non-Gaussian) channel, the input distribution \( F_X \) used to construct the random code may not admit a pdf (e.g., cf. [19, Chapter 7], [53, Theorem 11.14]).

5.5 Capacity of uncorrelated parallel Gaussian channels: The water-filling principle

Consider a network of \( k \) mutually-independent discrete-time memoryless Gaussian channels with respective positive noise powers (variances) \( \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2 \). If one wants to transmit information using these channels simultaneously (in parallel), what will be the system’s channel capacity, and how should the signal powers for each channel be apportioned given a fixed overall power budget? The answer to the above question lies in the so-called water-filling or water-pouring principle.

Theorem 5.34 (Capacity of uncorrelated parallel Gaussian channels) The capacity of \( k \) uncorrelated parallel Gaussian channels under an overall input power constraint \( P \) is given by

\[ C(P) = \sum_{i=1}^{k} \frac{1}{2} \log_2 \left( 1 + \frac{P_i}{\sigma_i^2} \right), \]

where \( \sigma_i^2 \) is the noise variance of channel \( i \),

\[ P_i = \max\{0, \theta - \sigma_i^2\}, \]

and \( \theta \) is chosen to satisfy \( \sum_{i=1}^{k} P_i = P \). This capacity is achieved by a tuple of independent Gaussian inputs \( (X_1, X_2, \ldots, X_k) \), where \( X_i \sim \mathcal{N}(0, P_i) \) is the input to channel \( i \), for \( i = 1, 2, \ldots, k \).
Proof: By definition,

\[
C(P) = \sup_{F_{X^k}} I(X^k; Y^k). 
\]

Since the noise random variables \(Z_1, \ldots, Z_k\) are independent from each other,

\[
I(X^k; Y^k) = h(Y^k) - h(Y^k | X^k) \\
= h(Y^k) - h(Z^k + X^k | X^k) \\
= h(Y^k) - h(Z^k | X^k) \\
= h(Y^k) - h(Z^k) \\
= h(Y^k) - \sum_{i=1}^{k} h(Z_i) \\
\leq \sum_{i=1}^{k} h(Y_i) - \sum_{i=1}^{k} h(Z_i) \\
\leq \sum_{i=1}^{k} \frac{1}{2} \log_2 \left(1 + \frac{P_i}{\sigma_i^2}\right)
\]

where the first inequality follows from the chain rule for differential entropy and the fact that conditioning cannot increase differential entropy, and the second inequality holds since output \(Y_i\) of channel \(i\) due to input \(X_i\) with \(E[X_i^2] = P_i\) has its differential entropy maximized if it is Gaussian distributed with zero-mean and variance \(P_i + \sigma_i^2\). Equalities hold above if all the \(X_i\) inputs are independent of each other with each input \(X_i \sim \mathcal{N}(0, P_i)\) such that \(\sum_{i=1}^{k} P_i = P\).

Thus the problem is reduced to finding the power allotment that maximizes the overall capacity subject to the constraint \(\sum_{i=1}^{k} P_i = P\) with \(P_i \geq 0\). By using the Lagrange multiplier technique and verifying the KKT condition (see Example B.20 in Appendix B.10), the maximizer \((P_1, \ldots, P_k)\) of

\[
\max \left\{ \sum_{i=1}^{k} \frac{1}{2} \log_2 \left(1 + \frac{P_i}{\sigma_i^2}\right) + \sum_{i=1}^{k} \lambda_i P_i - \nu \left( \sum_{i=1}^{k} P_i - P \right) \right\}
\]

can be found by taking the derivative of the above equation (with respect to \(P_i\)) and setting it to zero, which yields

\[
\lambda_i = \begin{cases} 
-\frac{1}{2 \ln(2)} \frac{1}{P_i + \sigma_i^2} + \nu = 0, & \text{if } P_i > 0; \\
-\frac{1}{2 \ln(2)} \frac{1}{P_i + \sigma_i^2} + \nu \geq 0, & \text{if } P_i = 0.
\end{cases}
\]

Hence,

\[
P_i = \theta - \sigma_i^2, \quad \text{if } P_i > 0; \\
P_i \geq \theta - \sigma_i^2, \quad \text{if } P_i = 0,
\]

(equivalently, \(P_i = \max\{0, \theta - \sigma_i^2\}\)).
where $\theta \triangleq \log_2 e/(2\nu)$ is chosen to satisfy $\sum_{i=1}^{k} P_i = P$. □

We illustrate the above result in Figure 5.1 and elucidate why the $P_i$ power allotments form a water-filling (or water-pouting) scheme. In the figure, we have a vessel where the height of each of the solid bins represents the noise power of each channel (while the width is set to unity so that the area of each bin yields the noise power of the corresponding Gaussian channel). We can thus visualize the system as a vessel with an uneven bottom where the optimal input signal allocation $P_i$ to each channel is realized by pouring an amount $P$ units of water into the vessel (with the resulting overall area of filled water equal to $P$). Since the vessel has an uneven bottom, water is unevenly distributed among the bins: noisier channels are allotted less signal power (note that in this example, channel 3, whose noise power is largest, is given no input power at all and is hence not used).

Figure 5.1: The water-pouring scheme for uncorrelated parallel Gaussian channels. The horizontal dashed line, which indicates the level where the water rises to, indicates the value of $\theta$ for which $\sum_{i=1}^{k} P_i = P$.

**Observation 5.35** Although it seems reasonable to allocate more power to less noisy channels for the optimization of the channel capacity according to the water-filling principle, the Gaussian inputs required for achieving channel capacity do not fit the digital communication system in practice. One may wonder what an optimal power allocation scheme will be when the channel inputs are dictated to be practically discrete in values, such as binary phase-shift keying (BPSK), quadrature phase-shift keying (QPSK), or 16 quadrature-amplitude modulation (16-QAM). Surprisingly, a different procedure from the water-filling principle results under certain conditions. By characterizing the relationship between mutual information and minimum mean square error (MMSE) [23], the
optimal power allocation for parallel AWGN channels with inputs constrained to be discrete is established, resulting in a new graphical power allocation interpretation called the mercury/water-filling principle [32]: i.e., mercury of proper amounts [32, eq. (43)] must be individually poured into each bin before the water of amount $P = \sum_{i=1}^{k} P_i$ is added to the vessel. It is hence named (i.e., the mercury/water-filling principle) because mercury is heavier than water and does not dissolve in it, and so can play the role of a pre-adjustment of bin heights. This research concludes that when the total transmission power $P$ is small, the strategy that maximizes the capacity follows approximately the equal signal-to-noise ratio (SNR) principle, i.e., a larger power should be allotted to a noisier channel to optimize the capacity. Very recently, it was found that when the additive noise is no longer Gaussian, the mercury adjustment fails to interpret the optimal power allocation scheme. For additive Gaussian noise with arbitrary discrete inputs, the pre-adjustment before the water pouring step is always upward; hence, a mercury-filling scheme is proposed to materialize the lifting of heights of bin bases. However, since the pre-adjustment of heights of bin bases generally can be in both up and down directions (see Example 1 in [49] for quaternary-input additive Laplacian noise channels), the use of the name mercury/water filling becomes inappropriate. For this reason, the graphical interpretation of the optimal power allocation principle is simply named two-phase water-filling principle in [49]. We end this observation by emphasizing that the true measure for a digital communication system is of no doubt the effective transmission rate subject to an acceptably good transmission error (e.g., an overall bit error rate $\leq 10^{-5}$). In order to make the analysis tractable, researchers adopt the channel capacity as the design criterion instead, in the hope of obtaining a simple reference scheme for practical systems. Further study is thus required to examine the feasibility of such an approach (even though the water-filling principle is certainly a theoretically interesting result).

5.6 Capacity of correlated parallel Gaussian channels

In the previous section, we considered a network of $k$ parallel discrete-time memoryless Gaussian channels in which the noise samples from different channels are independent from each other. We found out that the power allocation strategy that maximizes the system’s capacity is given by the water-filling scheme. We next study a network of $k$ parallel memoryless Gaussian channels where the noise variables from different channels are correlated. Surprisingly, we obtain that water-filling provides also the optimal power allotment policy.

Let $K_Z$ denote the covariance matrix of the noise tuple $(Z_1, Z_2, \ldots, Z_k)$, and let $K_X$ denote the covariance matrix of the system input $(X_1, \ldots, X_k)$, where we assume (without loss of the generality) that each $X_i$ has zero mean. We assume
that $K_Z$ is positive definite. The input power constraint becomes

$$\sum_{i=1}^{k} E[X_i^2] = \text{tr}(K_X) \leq P,$$

where $\text{tr}(\cdot)$ denotes the trace of the $k \times k$ matrix $K_X$. Since in each channel, the input and noise variables are independent from each other, we have

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

$$= h(Y^k) - h(Z^k + X^k | X^k)$$

$$= h(Y^k) - h(Z^k | X^k)$$

$$= h(Y^k) - h(Z^k).$$

Since $h(Z^k)$ is not determined by the input, determining the system’s capacity reduces to maximizing $h(Y^k)$ over all possible inputs $(X_1, \ldots, X_k)$ satisfying the power constraint.

Now observe that the covariance matrix of $Y^k$ is equal to $K_Y = K_X + K_Z$, which implies by Theorem 5.20 that the differential entropy of $Y^k$ is upper bounded by

$$h(Y^k) \leq \frac{1}{2} \log_2 \left( (2\pi e)^k \text{det}(K_X + K_Z) \right),$$

with equality iff $Y^k$ Gaussian. It remains to find out whether we can find inputs $(X_1, \ldots, X_k)$ satisfying the power constraint which achieve the above upper bound and maximize it.

As in the proof of Theorem 5.18, we can orthogonally diagonalize $K_Z$ as

$$K_Z = A \Lambda A^T,$$

where $AA^T = I_k$ (and thus $\text{det}(A)^2 = 1$), $I_k$ is the $k \times k$ identity matrix, and $\Lambda$ is a diagonal matrix with positive diagonal components consisting of the eigenvalues of $K_Z$ (as $K_Z$ is positive definite). Then

$$\text{det}(K_X + K_Z) = \text{det}(K_X + A \Lambda A^T)$$

$$= \text{det}(AA^T K_X AA^T + A \Lambda A^T)$$

$$= \text{det}(A) \cdot \text{det}(A^T K_X A + \Lambda) \cdot \text{det}(A^T)$$

$$= \text{det}(A^T K_X A + \Lambda)$$

$$= \text{det}(B + \Lambda),$$

where $B \triangleq A^T K_X A$. Since for any two matrices $C$ and $D$, $\text{tr}(CD) = \text{tr}(DC)$, we have that $\text{tr}(B) = \text{tr}(A^T K_X A) = \text{tr}(AA^T K_X) = \text{tr}(I_k K_X) = \text{tr}(K_X)$. Thus the capacity problem is further transformed to maximizing $\text{det}(B + \Lambda)$ subject to $\text{tr}(B) \leq P$. 

162
By observing that $B + \Lambda$ is positive definite (because $\Lambda$ is positive definite) and using Hadamard’s inequality given in Corollary 5.19, we have

$$\det(B + \Lambda) \leq \prod_{i=1}^{k} (B_{ii} + \lambda_i),$$

where $\lambda_i$ is the component of matrix $\Lambda$ locating at $i^{th}$ row and $i^{th}$ column, which is exactly the $i$-th eigenvalue of $K_Z$. Thus, the maximum value of $\det(B + \Lambda)$ under $\text{tr}(B) \leq P$ is realized by a diagonal matrix $B$ (to achieve equality in Hadamard’s inequality) with

$$\sum_{i=1}^{k} B_{ii} = P.$$

Finally, as in the proof of Theorem 5.34, we obtain a water-filling allotment for the optimal diagonal elements of $B$:

$$B_{ii} = \max\{0, \theta - \lambda_i\},$$

where $\theta$ is chosen to satisfy $\sum_{i=1}^{k} B_{ii} = P$. We summarize this result in the next theorem.

**Theorem 5.36 (Capacity of correlated parallel Gaussian channels)** The capacity of $k$ correlated parallel Gaussian channels with positive-definite noise covariance matrix $K_Z$ under overall input power constraint $P$ is given by

$$C(P) = \sum_{i=1}^{k} \frac{1}{2} \log_2 \left( 1 + \frac{P_i}{\lambda_i} \right),$$

where $\lambda_i$ is the $i$-th eigenvalue of $K_Z$,

$$P_i = \max\{0, \theta - \lambda_i\},$$

and $\theta$ is chosen to satisfy $\sum_{i=1}^{k} P_i = P$. This capacity is achieved by a tuple of zero-mean Gaussian inputs $(X_1, X_2, \cdots, X_k)$ with covariance matrix $K_X$ having the same eigenvectors as $K_Z$, where the $i$-th eigenvalue of $K_X$ is $P_i$, for $i = 1, 2, \cdots, k$.

### 5.7 Non-Gaussian discrete-time memoryless channels

If a discrete-time channel has an additive but non-Gaussian memoryless noise and an input power constraint, then it is often hard to calculate its capacity. Hence, in this section, we introduce an upper bound and a lower bound on the capacity of such a channel (we assume that the noise admits a pdf).
**Definition 5.37 (Entropy power)** For a continuous random variable $Z$ with
(well-defined) differential entropy $h(Z)$ (measured in bits), its entropy power is
denoted by $Z_e$ and defined as

$$Z_e \triangleq \frac{1}{2\pi e} 2^{2h(Z)}.$$

**Lemma 5.38** For a discrete-time continuous-alphabet memoryless additive-noise
channel with input power constraint $P$ and noise variance $\sigma^2$, its capacity satisfies

$$\frac{1}{2} \log_2 \frac{P + \sigma^2}{Z_e} \geq C(P) \geq \frac{1}{2} \log_2 \frac{P + \sigma^2}{\sigma^2}. \quad (5.7.1)$$

**Proof:** The lower bound in (5.7.1) is already proved in Theorem 5.32. The upper bound follows from

$$I(X; Y) = h(Y) - h(Z) \leq \frac{1}{2} \log_2 [2\pi e (P + \sigma^2)] - \frac{1}{2} \log_2 [2\pi e Z_e].$$

The entropy power of $Z$ can be viewed as the variance of a corresponding
Gaussian random variable with the same differential entropy as $Z$. Indeed, if $Z$
is Gaussian, then its entropy power is equal to

$$Z_e = \frac{1}{2\pi e} 2^{2h(Z)} = \text{Var}(Z),$$

as expected.

Whenever two independent Gaussian random variables, $Z_1$ and $Z_2$, are added,
the power (variance) of the sum is equal to the sum of the powers (variances) of
$Z_1$ and $Z_2$. This relationship can then be written as

$$2^{2h(Z_1 + Z_2)} = 2^{2h(Z_1)} + 2^{2h(Z_2)},$$
or equivalently

$$\text{Var}(Z_1 + Z_2) = \text{Var}(Z_1) + \text{Var}(Z_2).$$

However, when two independent random variables are non-Gaussian, the relationship becomes

$$2^{2h(Z_1 + Z_2)} \geq 2^{2h(Z_1)} + 2^{2h(Z_2)}, \quad (5.7.2)$$
or equivalently

$$Z_e(Z_1 + Z_2) \geq Z_e(Z_1) + Z_e(Z_2). \quad (5.7.3)$$

Inequality (5.7.2) (or equivalently (5.7.3)), whose proof can be found in [13,
Section 17.8] or [9, Theorem 7.10.4], is called the **entropy-power inequality**. It
reveals that the sum of two independent random variables may introduce more
entropy power than the sum of each individual entropy power, except in the
Gaussian case.
5.8 Capacity of the band-limited white Gaussian channel

We have so far considered discrete-time channels (with discrete or continuous alphabets). We close this chapter by briefly presenting the capacity expression of the continuous-time (waveform) band-limited channel with additive white Gaussian noise. The reader is referred to [52], [19, Chapter 8], [2, Sections 8.2 and 8.3] and [25, Chapter 6] for rigorous and detailed treatments (including coding theorems) of waveform channels.

The continuous-time band-limited channel with additive white Gaussian noise is a common model for a radio network or a telephone line. For such a channel, illustrated in Figure 5.2, the output waveform is given by

\[ Y(t) = (X(t) + Z(t)) * h(t), \quad t \geq 0, \]

where “\(*\)” represents the convolution operation (recall that the convolution between two signals \( a(t) \) and \( b(t) \) is defined as \( a(t) * b(t) = \int_{-\infty}^{\infty} a(\tau)b(t-\tau)d\tau \)). Here \( X(t) \) is the channel input waveform with average power constraint

\[ \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} E[X^2(t)]dt \leq P \]  \hspace{1cm} (5.8.1)

and bandwidth \( W \) cycles per second or Hertz (Hz); i.e., its spectrum or Fourier transform \( X(f) \triangleq \mathcal{F}[X(t)] = \int_{-\infty}^{\infty} X(t)e^{-j2\pi ft}dt = 0 \) for all frequencies \( |f| > W \), where \( j = \sqrt{-1} \) is the imaginary unit number. \( Z(t) \) is the noise waveform of a zero-mean stationary white Gaussian process with power spectral density \( N_0/2 \); i.e., its power spectral density \( \text{PSD}_Z(f) \), which is the Fourier transform of the process covariance (equivalently, correlation) function \( K_Z(\tau) \triangleq E[Z(s)Z(s+\tau)] \), \( s, \tau \in \mathbb{R} \), is given by

\[ \text{PSD}_Z(f) = \mathcal{F}[K_Z(t)] = \int_{-\infty}^{\infty} K_Z(t)e^{-j2\pi ft}dt = \frac{N_0}{2} \quad \forall f. \]

Finally, \( h(t) \) is the impulse response of an ideal bandpass filter with cutoff frequencies at \( \pm W \) Hz:

\[ H(f) = \mathcal{F}[(h(t)] = \begin{cases} 1 & \text{if } -W \leq f \leq W, \\ 0 & \text{otherwise.} \end{cases} \]

Recall that one can recover \( h(t) \) by taking the inverse Fourier transform of \( H(f) \); this yields

\[ h(t) = \mathcal{F}^{-1}[H(f)] = \int_{-\infty}^{\infty} H(f)e^{j2\pi ft}df = 2W\text{sinc}(2Wt), \]
where

\[ \text{sinc}(t) \triangleq \frac{\sin(\pi t)}{\pi t} \]

is the sinc function and is defined to equal 1 at \( t = 0 \) by continuity.

![Waveform channel diagram](image)

Figure 5.2: Band-limited waveform channel with additive white Gaussian noise.

Note that we can write the channel output as

\[ Y(t) = X(t) + \tilde{Z}(t) \]

where \( \tilde{Z}(t) \triangleq Z(t) * h(t) \) is the filtered noise waveform. The input \( X(t) \) is not affected by the ideal unit-gain bandpass filter since it has an identical bandwidth as \( h(t) \). Note also that the power spectral density of the filtered noise is given by

\[
\text{PSD}_{\tilde{Z}}(f) = \text{PSD}_Z(f)|H(f)|^2 = \begin{cases} \frac{N_0}{2} & \text{if } -W \leq f \leq W, \\ 0 & \text{otherwise.} \end{cases}
\]

Taking the inverse Fourier transform of \( \text{PSD}_{\tilde{Z}}(f) \) yields the covariance function of the filtered noise process:

\[
K_{\tilde{Z}}(\tau) = \mathcal{F}^{-1}[\text{PSD}_{\tilde{Z}}(f)] = N_0 W \text{sinc}(2W\tau) \quad \tau \in \mathbb{R}. \tag{5.8.2}
\]

To determine the capacity (in bits per second) of this continuous-time band-limited white Gaussian channel with parameters, \( P \), \( W \) and \( N_0 \), we convert it to an “equivalent” discrete-time channel with power constraint \( P \) by using the well-known Sampling theorem (due to Nyquist, Kotelnikov and Shannon), which states that sampling a band-limited signal with bandwidth \( W \) at a rate of \( 1/(2W) \) is sufficient to reconstruct the signal from its samples. Since \( X(t) \), \( \tilde{Z}(t) \) and \( Y(t) \) are all band-limited to \([-W; W]\), we can thus represent these signals by
their samples taken $\frac{1}{2W}$ seconds apart and model the channel by a discrete-time channel described by:

$$Y_n = X_n + \tilde{Z}_n, \quad n = 0, \pm 1, \pm 2, \ldots,$$

where $X_n \triangleq X\left(\frac{n}{2W}\right)$ are the input samples and $\tilde{Z}_n = Z\left(\frac{n}{2W}\right)$ and $Y_n = Y\left(\frac{n}{2W}\right)$ are the random samples of the noise $\tilde{Z}(t)$ and output $Y(t)$ signals, respectively.

Since $\tilde{Z}(t)$ is a filtered version of $Z(t)$, which is a zero-mean stationary Gaussian process, we obtain that $\tilde{Z}(t)$ is also zero-mean, stationary and Gaussian. This directly implies that the samples $\tilde{Z}_n$, $n = 1, 2, \ldots$, are zero-mean Gaussian identically distributed random variables. Now an examination of the expression $K_{\tilde{Z}}(\tau)$ in (5.8.2) reveals that $K_{\tilde{Z}}(\tau) = 0$ for $\tau = \frac{n}{2W}$, $n = 1, 2, \ldots$, since $\text{sinc}(t) = 0$ for all non-zero integer values of $t$. Hence, the random variables $\tilde{Z}_n$, $n = 1, 2, \ldots$, are uncorrelated and hence independent (since they are Gaussian) and their variance is given by $E[\tilde{Z}_n^2] = K_{\tilde{Z}}(0) = N_0W$. We conclude that the discrete-time process $\{\tilde{Z}_n\}_{n=1}^\infty$ is i.i.d. Gaussian with each $\tilde{Z}_n \sim N(0, N_0W)$. As a result, the above discrete-time channel is a discrete-time memoryless Gaussian channel with power constraint $P$ and noise variance $N_0W$; thus the capacity of the band-limited white Gaussian channel in bits per channel use is given using (5.4.13) by

$$\frac{1}{2} \log_2 \left( 1 + \frac{P}{N_0W} \right) \quad \text{bits/channel use.}$$

Given that we are using the channel (with inputs $X_n$) every $\frac{1}{2W}$ seconds, we obtain that the capacity in bits/second of the band-limited white Gaussian channel is given by

$$C(P) = W \log_2 \left( 1 + \frac{P}{N_0W} \right) \quad \text{bits/second,} \quad (5.8.3)$$

where $\frac{P}{N_0W}$ is typically referred to as the signal-to-noise ratio (SNR).\(^8\)

We emphasize that the above derivation of (5.8.3) is heuristic as we have not rigorously shown the equivalence between the original band-limited Gaussian

\(^8\) (5.8.3) is achieved by zero-mean i.i.d. Gaussian $\{X_n\}_{n=-\infty}^\infty$ with $E[X_n^2] = P$, which can be obtained by sampling a zero-mean, stationary and Gaussian $X(t)$ with

$$\text{PSD}_X(f) = \begin{cases} \frac{P}{2W} & \text{if } -W \leq f \leq W, \\ 0 & \text{otherwise.} \end{cases}$$

Examining this $X(t)$ confirms that it satisfies (5.8.1):

$$\frac{1}{T} \int_{-T/2}^{T/2} E[X^2(t)] dt = E[X^2(t)] = K_X(0) = P \cdot \text{sinc}(2W \cdot 0) = P.$$
channel and its discrete-time version and we have not established a coding theorem for the original channel. We point the reader to the references mentioned at the beginning of the section for a full development of this subject.

**Example 5.39 (Telephone line channel)** Suppose telephone signals are bandlimited to 4 kHz. Given an SNR of 40 decibels (dB) – i.e., $10 \log_{10} \frac{P}{N_0 W} = 40$ dB – then from (5.8.3), we calculate that the capacity of the telephone line channel (when modeled via the band-limited white Gaussian channel) is given by

$$4000 \log_2 (1 + 10000) = 53151.4 \text{ bits/second}.$$

**Example 5.40 (Infinite bandwidth white Gaussian channel)** As the channel bandwidth $W$ grows without bound, we obtain from (5.8.3) that

$$\lim_{W \to \infty} C(P) = \frac{P}{N_0} \log_2 e \text{ bits/second},$$

which indicates that in the infinite-bandwidth regime, capacity grows linearly with power.

**Observation 5.41 (Band-limited colored Gaussian channel)** If the above band-limited channel has a stationary colored (non-white) additive Gaussian noise, then it can be shown (e.g., see [19]) that the capacity of this channel becomes

$$C(P) = \frac{1}{2} \int^W_{-W} \max \left[ 0, \log_2 \frac{\theta}{\text{PSD}_Z(f)} \right] df,$$

where $\theta$ is the solution of

$$P = \int^W_{-W} \max [0, \theta - \text{PSD}_Z(f)] df.$$

The above capacity formula is indeed reminiscent of the water-pouring scheme we saw in Sections 5.5 and 5.6, albeit it is herein applied in the spectral domain. In other words, we can view the curve of $\text{PSD}_Z(f)$ as a bowl, and water is imagined being poured into the bowl up to level $\theta$ under which the area of the water is equal to $P$ (see Figure 5.3.(a)). Furthermore, the distributed water indicates the shape of the optimum transmission power spectrum (see Figure 5.3.(b)).
(a) The spectrum of $\text{PSD}_Z(f)$ where the horizontal line represents $\theta$, the level at which water rises to.

(b) The input spectrum that achieves capacity.

Figure 5.3: Water-pouring for the band-limited colored Gaussian channel.
Chapter 6

Lossy Data Compression and Transmission

6.1 Fundamental concept on lossy data compression

6.1.1 Motivations

In operation, one sometimes needs to compress a source in a rate less than entropy, which is known to be the minimum code rate for lossless data compression. In such case, some sort of data loss is inevitable. People usually refer the resultant codes as lossy data compression code.

Some of the examples for requiring lossy data compression are made below.

Example 6.1 (Digitization or quantization of continuous signals) The information content of continuous signals, such as voice or multi-dimensional images, is usually infinity. It may require an infinite number of bits to digitize such a source without data loss, which is not feasible. Therefore, a lossy data compression code must be used to reduce the output of a continuous source to a finite number of bits.

Example 6.2 (Constraint on channel capacity) To transmit a source through a channel with capacity less than the source entropy is challenging. As stated in channel coding theorem, it always introduces certain error for rate above channel capacity. Recall that Fano’s inequality only provides a lower bound on the amount of the decoding error, and does not tell us how large the error is. Hence, the error could go beyond control when one desires to convey source that generates information at rates above channel capacity.

In order to have a good control of the transmission error, another approach is to first reduce the data rate with manageable distortion, and then transmit the source data at rates less than the channel capacity. With such approach,
the transmission error is only introduced at the (lossy) data compression step since the error due to transmission over channel can be made arbitrarily small (cf. Figure 6.1).

Example 6.3 (Extracting useful information) In some situation, some of the information is not useful for operational objective. A quick example will be the hypothesis testing problem in which case the system designer concerns only the likelihood ratio of the null hypothesis distribution against the alternative hypothesis distribution. Therefore, any two distinct source letters which produce the same likelihood ratio should not be encoded into different codewords. The resultant code is usually a lossy data compression code since reconstruction of the source from certain code is usually impossible.

6.1.2 Distortion measures

A source is modelled as a random process $Z_1, Z_2, \ldots, Z_n$. For simplicity, we assume that the source discussed in this section is memoryless and with finite generic alphabet. Our objective is to compress the source with rate less than entropy under a pre-specified criterion. In general, the criterion is given by a distortion measure as defined below.

Definition 6.4 (Distortion measure) A distortion measure is a mapping

$$\rho : \mathcal{Z} \times \hat{\mathcal{Z}} \rightarrow \mathbb{R}^+,$$
where $\mathcal{Z}$ is the source alphabet, $\hat{\mathcal{Z}}$ is the reproduction alphabet for the compressed code, and $\mathbb{R}^+$ is the set of non-negative real number.

From the above definition, the distortion measure $\rho(z, \hat{z})$ can be viewed as a cost of representing the source symbol $z$ by a code symbol $\hat{z}$. It is then expected to choose a certain number of (typical) reproduction letters in $\hat{\mathcal{Z}}$ to represent the source letters, which cost least.

When $\hat{\mathcal{Z}} = \mathcal{Z}$, the selection of typical reproduction letters is similar to divide the source letters into several groups, and then choose one element in each group to represent the rest of the members in the same group. For example, suppose that $\hat{\mathcal{Z}} = \mathcal{Z} = \{1, 2, 3, 4\}$. Due to some constraints, we need to reduce the number of outcomes to 2, and the resultant expected cost can not be larger than 0.5.\(^1\) The source is uniformly distributed. Given a distortion measure by a matrix as:

\[
\left[ \rho(i, j) \right] \triangleq \begin{bmatrix}
0 & 1 & 2 & 2 \\
1 & 0 & 2 & 2 \\
2 & 2 & 0 & 1 \\
2 & 2 & 1 & 0
\end{bmatrix},
\]

the resultant two groups which cost least should be $\{1, 2\}$ and $\{3, 4\}$. We may choose respectively 1 and 3 as the typical elements for these two groups (cf. Figure 6.2). The expected cost of such selection is

\[
\frac{1}{4}\rho(1, 1) + \frac{1}{4}\rho(2, 1) + \frac{1}{4}\rho(3, 3) + \frac{1}{4}\rho(4, 3) = \frac{1}{2}.
\]

Note that the entropy of the source is reduced from 2 bits to 1 bit.

Sometimes, it is convenient to have $|\hat{\mathcal{Z}}| = |\mathcal{Z}| + 1$. For example,

$|\mathcal{Z} = \{1, 2, 3\}| = 3, \quad |\hat{\mathcal{Z}} = \{1, 2, 3, E\}| = 4,$

where $E$ can be regarded as the erasure symbol, and the distortion measure is defined by

\[
\left[ \rho(i, j) \right] \triangleq \begin{bmatrix}
0 & 2 & 2 & 0.5 \\
2 & 0 & 2 & 0.5 \\
2 & 2 & 0 & 0.5
\end{bmatrix}.
\]

The source is again uniformly distributed. In this example, to denote source letters by distinct letters in $\{1, 2, 3\}$ will cost four times than to represent them by $E$. Therefore, if only 2 outcomes are allowed, and the expected distortion

---

\(^1\)Note that the constraints for lossy data compression code are usually specified on the resultant entropy and expected distortion. Here, instead of putting constraints on entropy, we adopt the number-of-outcome constraint simply because it is easier to understand, especially for those who are not familiar with this subject.
cannot be greater than 1/3, then employing typical elements 1 and E to represent source \{1\} and \{2, 3\} is an optimal choice. The resultant entropy is reduced from \(\log_2(3)\) bits to \(\log_2(3) - 2/3\) bits.

It needs to be pointed out that to have \(|\hat{Z}| > |Z| + 1\) is usually not advantageous. Indeed, it has been proved that under some reasonable assumptions on the distortion measure, to have larger reproduction alphabet than \(|Z| + 1\) will not perform better.

6.1.3 Frequently used distortion measures

Example 6.5 (Hamming distortion measure) Let source alphabet and reproduction alphabet be the same, i.e., \(Z = \hat{Z}\). Then the Hamming distortion measure is given by

\[
\rho(z, \hat{z}) \triangleq \begin{cases} 
0, & \text{if } z = \hat{z}; \\
1, & \text{if } z \neq \hat{z}.
\end{cases}
\]

It is also named the probability-of-error distortion measure because

\[
E[\rho(Z, \hat{Z})] = \Pr(Z \neq \hat{Z}).
\]

Example 6.6 (Squared error distortion) Let source alphabet and reproduction alphabet be the same, i.e., \(Z = \hat{Z}\). Then the squared error distortion is given by

\[
\rho(z, \hat{z}) \triangleq (z - \hat{z})^2.
\]

The squared error distortion measure is perhaps the most popular distortion measure used for continuous alphabets.

The squared error distortion has the advantages of simplicity and having close form solution for most cases of interest, such as using least squares prediction.
Yet, such distortion measure has been criticized as an unhumanized criterion. For example, two speech waveforms in which one is a slightly time-shifted version of the other may have large square error distortion; however, they sound very similar to human.

The above definition for distortion measures can be viewed as a single-letter distortion measure since they consider only one random variable \( Z \) which draws a single letter. For sources modelled as a sequence of random variables \( Z_1, \ldots, Z_n \), some extension needs to be made. A straightforward extension is the additive distortion measure.

**Definition 6.7 (Additive distortion measure)** The additive distortion measure \( \rho_n \) between sequences \( z^n \) and \( \hat{z}^n \) is defined by

\[
\rho_n(z^n, \hat{z}^n) = \sum_{i=1}^{n} \rho(z_i, \hat{z}_i).
\]

Another example that is also based on a per-symbol distortion is the maximum distortion measure:

**Definition 6.8 (Maximum distortion measure)**

\[
\rho_n(z^n, \hat{z}^n) = \max_{1 \leq i \leq n} \rho(z_i, \hat{z}_i).
\]

After defining the distortion measures for source sequences, a natural question to ask is whether to reproduce source sequence \( z^n \) by sequence \( \hat{z}^n \) of the same length is a must or not. To be more precise, can we use \( \hat{z}^k \) to represent \( z^n \) for \( k \neq n \)? The answer is certainly yes if a distortion measure for \( z^n \) and \( \hat{z}^k \) is defined. A quick example will be that the source is a ternary sequence of length \( n \), while the (fixed-length) data compression result is a set of binary indexes of length \( k \), which is taken as small as possible subject to some given constrains. Hence, \( k \) is not necessarily equal to \( n \). One of the problems for taking \( k \neq n \) is that the distortion measure for sequences can no longer be defined based on per-letter distortions, and hence a per-letter formula for the best lossy data compression rate cannot be rendered.

In order to alleviate the aforementioned (\( k \neq n \)) problem, we claim that for most cases of interest, it is reasonable to assume \( k = n \). This is because one can actually implement the lossy data compression from \( Z^n \) to \( \hat{Z}^k \) in two steps: the first step corresponds to lossy compression mapping \( h_n : Z^n \to \hat{Z}^k \), and the second step performs indexing \( h_n(Z^n) \) into \( \hat{Z}^k \). For ease of understanding, these two steps are illustrated below.
Step 1: Find the data compression code

\[ h_n : \mathcal{Z}^n \to \hat{\mathcal{Z}}^n \]

for which the pre-specified distortion constraint and rate constraint are both satisfied.

Step 2: Derive the (asymptotically) lossless data compression block code for source \( h_n(\mathcal{Z}^n) \). When \( n \) is sufficiently large, the existence of such code with block length

\[ k > H(h(\mathcal{Z}^n)) \quad \text{(equivalently, } R = \frac{k}{n} > \frac{1}{n} H(h(\mathcal{Z}^n)) \text{)} \]

is guaranteed by Shannon’s source coding theorem.

Through the above two steps, a lossy data compression code from

\[ \mathcal{Z}^n \to \hat{\mathcal{Z}}^n \to \{0, 1\}^k \]

is established. Since the second step is already discussed in (asymptotically) lossless data compression context, we can say that the theorem regarding the lossy data compression is basically a theorem on the first step.

### 6.2 Fixed-length lossy data compression codes

Similar to lossless source coding theorem, the objective of information theorists is to find the theoretical limit of the compression rate for lossy data compression codes. Before introducing the main theorem, we need to first define lossy data compression codes.

**Definition 6.9 (Fixed-length lossy data compression code subject to average distortion constraint)** An \((n, M, D)\) fixed-length lossy data compression code for source alphabet \( \mathcal{Z}^n \) and reproduction alphabet \( \hat{\mathcal{Z}}^n \) consists of a compression function

\[ h : \mathcal{Z}^n \to \hat{\mathcal{Z}}^n \]

with the size of the codebook (i.e., the image \( h(\mathcal{Z}^n) \)) being \( |h(\mathcal{Z}^n)| = M \), and the average distortion satisfying

\[ E \left[ \frac{1}{n} \rho_n(Z^n, h(Z^n)) \right] \leq D. \]
Since the size of the codebook is $M$, it can be binary-indexed by $k = \lceil \log_2 M \rceil$ bits. Therefore, the average rate of such code is $(1/n) \lceil \log_2 M \rceil \approx (1/n) \log_2 M$ bits/sourceword.

Note that a parallel definition for variable-length source compression code can also be defined. However, there is no conclusive results for the bound of such code rate, and hence we omit it for the moment. This is also an interesting open problem to research on.

After defining fixed-length lossy data compression codes, we are ready to define the achievable rate-distortion pair.

**Definition 6.10 (Achievable rate-distortion pair)** For a given sequence of distortion measures $\{\rho_n\}_{n \geq 1}$, a rate distortion pair $(R, D)$ is achievable if there exists a sequence of fixed-length lossy data compression codes $(n, M_n, D_n)$ with ultimate code rate $\limsup_{n \to \infty} (1/n) \log M_n \leq R$.

With the achievable rate-distortion region, we define the rate-distortion function as follows.

**Definition 6.11 (Rate-distortion function)** The rate-distortion function, denoted by $R(D)$, is equal to

$$R(D) \triangleq \inf \{ \hat{R} \in \mathbb{R} : (\hat{R}, D) \text{ is an achievable rate-distortion pair} \}.$$

### 6.3 Rate-distortion function for discrete memoryless sources

The main result here is based on the discrete memoryless source (DMS) and bounded additive distortion measure. “Boundedness” of a distortion measure means that

$$\max_{(z, \hat{z}) \in \mathcal{Z} \times \hat{\mathcal{Z}}} \rho(z, \hat{z}) < \infty.$$

The basic idea of choosing the data compression codewords from the set of source input symbols under DMS is to draw the codewords from the distortion typical set. This set is defined similarly as the joint typical set for channels.

**Definition 6.12 (Distortion typical set)** For a memoryless distribution with generic marginal $P_{Z, \hat{Z}}$ and a bounded additive distortion measure $\rho_n(\cdot, \cdot)$, the distortion $\delta$-typical set is defined by

$$D_n(\delta) \triangleq \left\{ (z^n, \hat{z}^n) \in \mathcal{Z}^n \times \hat{\mathcal{Z}}^n : \right\}$$

176
\[-\frac{1}{n} \log P_Z^n(z^n) - H(Z) \leq \delta,\]
\[-\frac{1}{n} \log P_{\hat{Z}}^n(\hat{z}^n) - H(\hat{Z}) \leq \delta,\]
\[-\frac{1}{n} \log P_{Z,\hat{Z}}^n(z^n, \hat{z}^n) - H(Z, \hat{Z}) < \delta,\]
and \[\frac{1}{n} \rho_n(z^n, \hat{z}^n) - E[\rho(Z, \hat{Z})] < \delta\}.

Note that this is the definition of the jointly typical set with additional constraint on the distortion being close to the expected value. Since the additive distortion measure between two joint i.i.d. random sequences is actually the sum of i.i.d. random variable, i.e.,
\[\rho_n(Z^n, \hat{Z}^n) = \sum_{i=1}^{n} \rho(Z_i, \hat{Z}_i),\]
the (weak) law-of-large-number holds for the distortion typical set. Therefore, an AEP-like theorem can be derived for distortion typical set.

**Theorem 6.13** If \((Z_1, \hat{Z}_1), (Z_2, \hat{Z}_2), \ldots, (Z_n, \hat{Z}_n), \ldots\) are i.i.d., and \(\rho_n\) are bounded additive distortion measure, then
\[-\frac{1}{n} \log P_{Z^n}(Z_1, Z_2, \ldots, Z_n) \to H(Z) \text{ in probability; }\]
\[-\frac{1}{n} \log P_{\hat{Z}^n}(\hat{Z}_1, \hat{Z}_2, \ldots, \hat{Z}_n) \to H(\hat{Z}) \text{ in probability; }\]
\[-\frac{1}{n} \log P_{Z^n,\hat{Z}^n}((Z_1, \hat{Z}_1), \ldots, (Z_n, \hat{Z}_n)) \to H(Z, \hat{Z}) \text{ in probability; }\]
and
\[\frac{1}{n} \rho_n(Z^n, \hat{Z}^n) \to E[\rho(Z, \hat{Z})] \text{ in probability.}\]

**Proof:** Functions of independent random variables are also independent random variables. Thus by the weak law of large numbers, we have the desired result.

It needs to be pointed out that without boundedness assumption, the normalized sum of an i.i.d. sequence is not necessary convergence in probability to a finite mean. That is why an additional condition, “boundedness” on distortion measure, is imposed, which guarantees the required convergence.

**Theorem 6.14 (AEP for distortion measure)** Given a discrete memoryless sources \(Z\), a single-letter conditional distribution \(P_{Z|Z}\), and any \(\delta > 0\), the weakly distortion \(\delta\)-typical set satisfies
1. $P_{Z^n, \hat{Z}^n}(D_n(\delta)) < \delta$ for $n$ sufficiently large;

2. for all $(z^n, \hat{z}^n)$ in $D_n(\delta)$,

$$P_{Z^n}(\check{Z}^n) \geq P_{Z^n|Z^n}(\check{z}^n|z^n)e^{-n[I(Z;\hat{Z})+3\delta]}.$$  \hspace{1cm} (6.3.1)

**Proof:** The first one follows from the definition. The second one can be proved by

$$P_{Z^n|Z^n}(\check{z}^n|z^n) = \frac{P_{Z^n, \check{Z}^n}(z^n, \hat{z}^n)}{P_{Z^n}(z^n)} \leq \frac{P_{Z^n}(\check{z}^n)P_{Z^n, \check{Z}^n}(z^n, \hat{z}^n)}{P_{Z^n}(z^n)P_{Z^n}(\check{z}^n)} \leq \frac{P_{Z^n}(\check{z}^n)e^{-n[H(Z, \hat{Z})-\delta]}}{e^{-n[H(Z)+\delta]}e^{-n[H(\hat{Z})+\delta]}} = \frac{P_{Z^n}(\check{z}^n)e^{n[I(Z;\hat{Z})]}}{e^{-n[I(Z;\hat{Z})+3\delta]}}.$$
From the above derivation, we know that equality holds for (6.3.2) if, and only if,

\[(x = 0) \lor (x = 1) \lor (y = 0) \land [(x = 0) \lor (y = 0)] \land [x = 1] = (x = 1, y = 0).
\]

(Note that \((x = 0)\) represents \(\{(x, y) \in \mathbb{R}^2 : x = 0 \text{ and } y \in [0, 1]\}\). Similar definition applies to the other sets.)

**Theorem 6.16 (Rate distortion theorem)** For DMS and bounded additive distortion measure (namely,

\[\rho_{\text{max}} \triangleq \max_{(z, \hat{z}) \in \mathcal{Z} \times \mathcal{\hat{Z}}} \rho(z, \hat{z}) < \infty \text{ and } \rho_n(z^n, \hat{z}^n) = \sum_{i=1}^n \rho(z_i, \hat{z}_i),\]

the rate-distortion function is

\[R(D) = \min_{P_{\hat{Z}|Z}} I(Z; \hat{Z}).\]

**Proof:** Denote \(f(D) \triangleq \min_{P_{\hat{Z}|Z} : E[\rho(Z, \hat{Z})] \leq D} I(Z; \hat{Z}).\) Then we shall show that \(R(D)\) defined in Definition 6.11 equals \(f(D)\).

1. **Achievability** (i.e., \(R(D + \epsilon) \leq f(D) + 4\epsilon\) for arbitrarily small \(\epsilon > 0\)): We need to show that for any \(\epsilon > 0\), there exist \(0 < \gamma < 4\epsilon\) and a sequence of lossy data compression codes \(\{(n, M_n, D + \epsilon)\}_{n=1}^{\infty}\) with

\[\limsup_{n \to \infty} \frac{1}{n} \log M_n \leq f(D) + \gamma.\]

The proof is as follows.

**Step 1: Optimizer.** Let \(P_{\hat{Z}|Z}\) be the optimizer of \(f(D)\), i.e.,

\[f(D) = \min_{P_{\hat{Z}|Z} : E[\rho(Z, \hat{Z})] \leq D} I(Z; \hat{Z}) = I(Z; \hat{Z}).\]

Then

\[E[\rho(Z, \hat{Z})] = \frac{1}{n} E[\rho_n(Z^n, \hat{Z}^n)] \leq D.\]

Choose \(M_n\) to satisfy

\[f(D) + \frac{1}{2} \gamma \leq \frac{1}{n} \log M_n \leq f(D) + \gamma\]

for some \(\gamma\) in \((0, 4\epsilon)\), for which the choice should exist for all sufficiently large \(n > N_0\) for some \(N_0\). Define

\[\delta \triangleq \min \left\{ \frac{\gamma}{8} \frac{\epsilon}{1 + 2\rho_{\text{max}}} \right\}.\]
Step 2: Random coding. Independently select $M_n$ codewords from $\tilde{Z}^n$ according to

$$P_{\tilde{Z}^n}(\tilde{z}^n) = \prod_{i=1}^{n} P_{\tilde{Z}}(\tilde{z}_i),$$

and denote this random codebook by $C_{\sim} n$, where

$$P_{\tilde{Z}}(\tilde{z}) = \sum_{z \in Z} P_{Z}(z) P_{\tilde{Z}|Z}(\tilde{z}|z).$$

Step 3: Encoding rule. Define a subset of $Z^n$ as

$$J(C_{\sim} n) \triangleq \{ z^n \in Z^n : \exists \tilde{z}^n \in C_{\sim} n \text{ such that } (z^n, \tilde{z}^n) \in D_n(\delta) \},$$

where $D_n(\delta)$ is defined under $P_{Z|\tilde{Z}}$. Based on the codebook

$$C_{\sim} n = \{ c_1, c_2, \ldots, c_{M_n} \},$$

define the encoding rule as:

$$h_n(z^n) = \begin{cases} c_m, & \text{if } (z^n, c_m) \in D_n(\delta); \\ 0, & \text{otherwise.} \end{cases}$$

(when more than one satisfying the requirement, just pick any.)

Note that when $z^n \in J(C_{\sim} n)$, we have $(z^n, h_n(z^n)) \in D_n(\delta)$ and

$$\frac{1}{n} \rho_n(z^n, h_n(z^n)) \leq E[\rho(Z, \tilde{Z})] + \delta \leq D + \delta.$$

Step 4: Calculation of probability outside $J(C_{\sim} n)$. Let $N_1$ satisfying that for $n > N_1$,

$$P_{Z^n, \tilde{Z}^n}(D_{c}^n(\delta)) < \delta.$$

Let

$$\Omega \triangleq P_{Z^n}(J^c(C_{\sim} n)).$$

Then by random coding argument,

$$E[\Omega] = \sum_{C_{\sim} n} P_{Z^n}(C_{\sim} n) \left( \sum_{z^n \in J^c(C_{\sim} n)} P_{Z^n}(z^n) \right)$$

$$= \sum_{z^n \in Z^n} P_{Z^n}(z^n) \left( \sum_{C_{\sim} n : z^n \notin J^c(C_{\sim} n)} P_{Z^n}(C_{\sim} n) \right).$$

180
For any $z^n$ given, to select a codebook $\mathcal{C}_n$ satisfying $z^n \notin J(\mathcal{C}_n)$ is equivalent to independently draw $M_n$ $n$-tuple from $\tilde{Z}^n$ which is not distortion joint typical with $z^n$. Hence,

$$\sum_{\mathcal{C}_n \notin J(\mathcal{C}_n)} P_{\tilde{Z}^n}(\mathcal{C}_n) = \left( \Pr \left[ (z^n, \tilde{Z}^n) \notin D_n(\delta) \right] \right)^{M_n}.$$ 

For convenience, we let $K(z^n, \tilde{z}^n)$ be the indicator function of $D_n(\delta)$, i.e.,

$$K(z^n, \tilde{z}^n) = \begin{cases} 1, & \text{if } (z^n, \tilde{z}^n) \in D_n(\delta); \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\sum_{\mathcal{C}_n \notin J(\mathcal{C}_n)} P_{\tilde{Z}^n}(\mathcal{C}_n) = \left( 1 - \sum_{\tilde{z}^n \in \tilde{Z}^n} P_{\tilde{Z}^n}(\tilde{z}^n) K(z^n, \tilde{z}^n) \right)^{M_n}.$$ 

Continuing the computation of $E[\Omega]$, we get

$$E[\Omega] = \sum_{z^n \in Z^n} P_{Z^n}(z^n) \left( 1 - \sum_{\tilde{z}^n \in \tilde{Z}^n} P_{\tilde{Z}^n}(\tilde{z}^n) K(z^n, \tilde{z}^n) \right)^{M_n}$$

$$\leq \sum_{z^n \in Z^n} P_{Z^n}(z^n) \left( 1 - \sum_{\tilde{z}^n \in \tilde{Z}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n) e^{-n(I(Z;\tilde{Z}) + 3\delta)} K(z^n, \tilde{z}^n) \right)^{M_n}$$

(by (6.3.1))

$$= \sum_{z^n \in Z^n} P_{Z^n}(z^n) \left( 1 - e^{-n(I(Z;\tilde{Z}) + 3\delta)} \sum_{\tilde{z}^n \in \tilde{Z}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n) K(z^n, \tilde{z}^n) \right)^{M_n}$$

$$\leq \sum_{z^n \in Z^n} P_{Z^n}(z^n) \left( 1 - \sum_{\tilde{z}^n \in \tilde{Z}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n) K(z^n, \tilde{z}^n) \right. \right.$$ 

$$+ \exp \left\{ -M_n \cdot e^{-n(I(Z;\tilde{Z}) + 3\delta)} \right\}$$

(from (6.3.2))

$$\leq \sum_{z^n \in Z^n} P_{Z^n}(z^n) \left( 1 - \sum_{\tilde{z}^n \in \tilde{Z}^n} P_{\tilde{Z}^n|Z^n}(\tilde{z}^n|z^n) K(z^n, \tilde{z}^n) \right.$$ 

$$+ \exp \left\{ -e^{n(f(D) + \gamma/2)} \cdot e^{-n(I(Z;\tilde{Z}) + 3\delta)} \right\} ,$$

(for $f(D) + \gamma/2 < (1/n) \log M_n$)

$$\leq 1 - P_{Z^n,\tilde{Z}^n}(D_n(\delta)) + \exp \left\{ -e^{nd} \right\} ,$$

181
(for $f(D) = I(Z; \hat{Z})$ and $\delta \leq \gamma/8$)

\[
= P_{Z^n, \hat{Z}^n}(D^c_n(\delta)) + \exp \left\{-e^{nt} \right\} \\
\leq \delta + \delta = 2\delta, \text{ for } n > N \triangleq \max \left\{ N_0, N_1, \frac{1}{\delta} \log \log \left( \frac{1}{\min\{\delta, 1\}} \right) \right\}.
\]

Since $E[\Omega] = E[P_{Z^n}(J^c(\mathcal{C}_n))] \leq 2\delta$, there must exists a codebook $\mathcal{C}_n^*$ such that $P_{Z^n}(J_c(C^*_n))$ is no greater than $2\delta$.

**Step 5: Calculation of distortion.** For the optimal codebook $\mathcal{C}_n^*$ (from the previous step) at $n > N$, its distortion is:

\[
\frac{1}{n} E[\rho_n(Z^n, h_n(Z^n))] = \sum_{z^n \in J(\mathcal{C}_n^*)} P_{Z^n}(z^n) \frac{1}{n} \rho_n(z^n, h_n(z^n)) \\
+ \sum_{z^n \not\in J(\mathcal{C}_n^*)} P_{Z^n}(z^n) \frac{1}{n} \rho_n(z^n, h_n(z^n)) \\
\leq \sum_{z^n \in J(\mathcal{C}_n^*)} P_{Z^n}(z^n) (D + \delta) + \sum_{z^n \not\in J(\mathcal{C}_n^*)} P_{Z^n}(z^n) \rho_{\max} \\
\leq (D + \delta) + 2\delta \cdot \rho_{\max} \\
\leq D + \delta (1 + 2\rho_{\max}) \\
\leq D + \varepsilon.
\]

2. **Converse Part** (i.e., $R(D + \varepsilon) \geq f(D)$ for arbitrarily small $\varepsilon > 0$ and any $D \in \{D \geq 0 : f(D) > 0\}$): We need to show that for any sequence of $\{(n, M_n, D_n)\}_{n=1}^{\infty}$ code with

\[
\limsup_{n \to \infty} \frac{1}{n} \log M_n < f(D),
\]

there exists $\varepsilon > 0$ such that

$$
D_n = \frac{1}{n} E[\rho_n(Z^n, h_n(Z^n))] > D + \varepsilon
$$

for $n$ sufficiently large.

The proof is as follows.

**Step 1: Convexity of mutual information.** By the convexity of mutual information $I(Z; \hat{Z})$ with respective to $P_{\hat{Z}|Z}$,

$$
I(Z; \hat{Z}_\lambda) \leq \lambda \cdot I(Z; \hat{Z}_1) + (1 - \lambda) \cdot I(Z; \hat{Z}_2),
$$

where $\lambda \in [0, 1]$, and

$$
P_{Z|\hat{Z}}(\hat{z}|z) \triangleq \lambda P_{Z|\hat{Z}}(\hat{z}|z) + (1 - \lambda) P_{Z|\hat{Z}}(\hat{z}|z).
$$

Step 2: Convexity of \( f(D) \). Let \( P_{\hat{Z}_1|Z} \) and \( P_{\hat{Z}_2|Z} \) be two distributions achieving \( f(D_1) \) and \( f(D_2) \), respectively. Since

\[
E[\rho(Z, \hat{Z}_\lambda)] = \sum_{z \in \mathcal{Z}} P_Z(z) \sum_{\hat{z} \in \hat{Z}} P_{\hat{Z}_1|Z}(\hat{z}|z) \rho(z, \hat{z}) = \sum_{z \in \mathcal{Z}} P_Z(z) \sum_{\hat{z} \in \hat{Z}} \left[ \lambda P_{\hat{Z}_1|Z}(\hat{z}|z) + (1 - \lambda) P_{\hat{Z}_2|Z}(\hat{z}|z) \right] \rho(z, \hat{z}) = \lambda D_1 + (1 - \lambda) D_2,
\]

we have

\[
f(\lambda D_1 + (1 - \lambda) D_2) \leq I(Z; \hat{Z}_\lambda) \leq \lambda I(Z; \hat{Z}_1) + (1 - \lambda) I(Z; \hat{Z}_2) = \lambda f(D_1) + (1 - \lambda) f(D_2).
\]

Therefore, \( f(D) \) is a convex function.

Step 3: Strictly decreasingness and continuity of \( f(D) \).

By definition, \( f(D) \) is non-increasing in \( D \). Also, \( f(D) = 0 \) for

\[
D \geq \min_{P_Z} \sum_{z \in \mathcal{Z}} P_Z(z) P_{\hat{Z}_\lambda}(\hat{z}|z) \rho(z, \hat{z})
\]

(which is finite from boundedness of distortion measure). Together with its convexity, the strictly decreasingness and continuity of \( f(D) \) over \( \{D \geq 0 : f(D) > 0\} \) is proved.

Step 4: Main proof.

\[
\log M_n \geq H(h_n(Z^n)) = H(h_n(Z^n)) - H(h_n(Z^n)|Z^n), \quad \text{since } H(h_n(Z^n)|Z^n) = 0;
\]

\[
= I(Z^n; h_n(Z^n)) = H(Z^n) - H(Z^n|h_n(Z^n)) = \sum_{i=1}^n H(Z_i) - \sum_{i=1}^n H(Z_i|h_n(Z^n), Z_1, \ldots, Z_{i-1})
\]

by the independence of \( Z^n \), and chain rule for conditional entropy;

\[
\geq \sum_{i=1}^n H(Z_i) - \sum_{i=1}^n H(Z_i|\hat{Z}_i)
\]

where \( \hat{Z}_i \) is the \( i^{th} \) component of \( h_n(Z^n) \);
\[ \sum_{i=1}^{n} I(Z_i; \hat{Z}_i) \geq \sum_{i=1}^{n} f(D_i), \quad \text{where } D_i \triangleq E[\rho(Z_i, \hat{Z}_i)]; \]
\[ = n \sum_{i=1}^{n} \frac{1}{n} f(D_i) \geq n f \left( \frac{1}{n} \sum_{i=1}^{n} D_i \right), \quad \text{by convexity of } f(D); \]
\[ = n f \left( \frac{1}{n} E[\rho_n(Z^n, h_n(Z^n))] \right), \]

where the last step follows since the distortion measure is additive. Finally, 
\[ \limsup_{n \to \infty} (1/n) \log M_n < f(D) \] implies the existence of \( N \) and \( \gamma > 0 \) such that \( (1/n) \log M_n < f(D) - \gamma \) for all \( n > N \). Therefore, for \( n > N \),
\[ f \left( \frac{1}{n} E[\rho_n(Z^n, h_n(Z^n))] \right) < f(D) - \gamma, \]
which, together with the strictly decreasing of \( f(D) \), implies
\[ \frac{1}{n} E[\rho_n(Z^n, h_n(Z^n))] > D + \epsilon \]
for some \( \epsilon = \epsilon(\gamma) > 0 \) and for all \( n > N \).

3. Summary: For \( D \in \{ D \geq 0 : f(D) > 0 \} \), the achievability and converse parts jointly imply that \( f(D) + 4\epsilon \geq R(D + \epsilon) \geq f(D) \) for arbitrarily small \( \epsilon > 0 \). Together with the continuity of \( f(D) \), we obtain that \( R(D) = f(D) \) for \( D \in \{ D \geq 0 : f(D) > 0 \} \).

For \( D \in \{ D \geq 0 : f(D) = 0 \} \), the achievability part gives us \( f(D) + 4\epsilon = 4\epsilon \geq R(D + \epsilon) \geq 0 \) for arbitrarily small \( \epsilon > 0 \). This immediately implies that \( R(D) = 0(= f(D)) \) as desired. \( \square \)

The formula of the rate-distortion function obtained in the previous theorem is also valid for the squared error distortion over real numbers, even if it is unbounded. Here, we put the boundedness assumption just to facilitate the exposition of the current proof. Readers may refer to Part II of the lecture notes for a more general proof.

The discussion on lossy data compression, especially on continuous sources, will be continued in Section 6.4. Examples of the calculation of rate-distortion functions will also be given in the same section.
After introducing Shannon’s source coding theorem for block codes, Shannon’s channel coding theorem for block codes and rate-distortion theorem for i.i.d. or stationary ergodic system setting, we would like to once again make clear the “key concepts” behind these lengthy proofs, that is, typical-set and random-coding. The typical-set argument (specifically, δ-typical set for source coding, joint δ-typical set for channel coding, and distortion typical set for rate-distortion) uses the law-of-large-number or AEP reasoning to claim the existence of a set with very high probability; hence, the respective information manipulation can just focus on the set with negligible performance loss. The random-coding argument shows that the expectation of the desired performance over all possible information manipulation schemes (randomly drawn according to some properly chosen statistics) is already acceptably good, and hence the existence of at least one good scheme that fulfills the desired performance index is validated. As a result, in situations where the two above arguments apply, a similar theorem can often be established. Question is “Can we extend the theorems to cases where the two arguments fail?” It is obvious that only when new proving technique (other than the two arguments) is developed can the answer be affirmative. We will further explore this issue in Part II of the lecture notes.

6.4 Property and calculation of rate-distortion functions

6.4.1 Rate-distortion function for binary sources and Hamming additive distortion measure

A specific application of rate-distortion function that is useful in practice is when binary alphabets and Hamming additive distortion measure are assumed. The Hamming additive distortion measure is defined as:

$$\rho_n(z^n, \hat{z}^n) = \sum_{i=1}^{n} z_i \oplus \hat{z}_i,$$

where “⊕” denotes modulo two addition. In such case, $\rho(z^n, \hat{z}^n)$ is exactly the number of bit errors or changes after compression. Therefore, the distortion bound $D$ becomes a bound on the average probability of bit error. Specifically, among $n$ compressed bits, it is expected to have $E[\rho(Z^n, \hat{Z}^n)]$ bit errors; hence, the expected value of bit-error-rate is $(1/n)E[\rho(Z^n, \hat{Z}^n)]$. The rate-distortion function for binary sources and Hamming additive distortion measure is given by the next theorem.

**Theorem 6.17** Fix a memoryless binary source

$$Z = \{Z^n = (Z_1, Z_2, \ldots, Z_n)\}_{n=1}^{\infty}$$
with marginal distribution \( P_Z(0) = 1 - P_Z(1) = p \). Assume the Hamming additive distortion measure is employed. Then the rate-distortion function

\[
R(D) = \begin{cases} 
\frac{1}{2} \cdot \log_2(p) - \frac{1}{2} \cdot \log_2(1 - p), & \text{if } 0 \leq D < \min\{p, 1 - p\}; \\
0, & \text{if } D \geq \min\{p, 1 - p\},
\end{cases}
\]

where \( h_b \triangleq -p \cdot \log_2(p) - (1 - p) \cdot \log_2(1 - p) \) is the binary entropy function.

**Proof:** Assume without loss of generality that \( p \leq 1/2 \).

We first prove the theorem under \( 0 \leq D < \min\{p, 1 - p\} = p \). Observe that

\[
H(Z|\hat{Z}) = H(Z \oplus \hat{Z}|\hat{Z}).
\]

Also observe that

\[
E[\rho(Z, \hat{Z})] \leq D \implies \Pr\{Z \oplus \hat{Z} = 1\} \leq D.
\]

Then

\[
I(Z; \hat{Z}) = H(Z) - H(Z|\hat{Z}) \\
= h_b(p) - H(Z \oplus \hat{Z}|\hat{Z}) \\
\geq h_b(p) - H(Z \oplus \hat{Z}) \quad \text{(conditioning never increase entropy)} \\
\geq h_b(p) - h_b(D),
\]

where the last inequality follows since the binary entropy function \( h_b(x) \) is increasing for \( x \leq 1/2 \), and \( \Pr\{Z \oplus \hat{Z} = 1\} \leq D \). Since the above derivation is true for any \( P_{\hat{Z}|Z} \), we have

\[
R(D) \geq h_b(p) - h_b(D).
\]

It remains to show that the lower bound is achievable by some \( P_{\hat{Z}|Z} \), or equivalently, \( H(Z|\hat{Z}) = h_b(D) \) for some \( P_{\hat{Z}|Z} \). By defining \( P_{Z|\hat{Z}}(0|0) = P_{Z|\hat{Z}}(1|1) = 1 - D \), we immediately obtain \( H(Z|\hat{Z}) = h_b(D) \). The desired \( P_{\hat{Z}|Z} \) can be obtained by solving

\[
1 = P_{\hat{Z}}(0) + P_{\hat{Z}}(1) \\
= \frac{P_{z}(0)}{P_{z|\hat{z}}(0|0)} P_{z|\hat{z}}(0|0) + \frac{P_{z}(0)}{P_{z|\hat{z}}(0|1)} P_{z|\hat{z}}(1|0) \\
= \frac{p}{1 - D} P_{z|\hat{z}}(0|0) + \frac{p}{D} (1 - P_{z|\hat{z}}(0|0))
\]

and

\[
1 = P_{\hat{Z}}(0) + P_{\hat{Z}}(1)
\]
\[
\frac{P_Z(1)}{P_{\hat{Z}|Z}(1|0)} P_{Z|\hat{Z}(0|0)} + \frac{P_Z(1)}{P_{\hat{Z}|Z}(1|1)} P_{Z|\hat{Z}(1|1)} = \frac{1 - p}{D} (1 - P_{\hat{Z}|Z}(1|1)) + \frac{1 - p}{1 - D} P_{\hat{Z}|Z}(1|1),
\]

and yield

\[
P_{\hat{Z}|Z}(0|0) = \frac{1 - D}{1 - 2D} \left(1 - \frac{D}{p}\right) \quad \text{and} \quad P_{\hat{Z}|Z}(1|1) = \frac{1 - D}{1 - 2D} \left(1 - \frac{D}{1 - p}\right).
\]

This completes the proof for \(0 \leq D < \min\{p, 1 - p\} = p\).

Now in the case of \(p \leq D < 1 - p\), we can let \(P_{\hat{Z}|Z}(1|0) = P_{\hat{Z}|Z}(1|1) = 1\) to obtain \(I(Z; \hat{Z}) = 0\) and

\[
E[\rho(Z, \hat{Z})] = \sum_{z=0}^{1} \sum_{\hat{z}=0}^{1} P_Z(z) P_{\hat{Z}|Z}(\hat{z}|z) \rho(z, \hat{z}) = p \leq D.
\]

Similarly, in the case of \(D \geq 1 - p\), we let \(P_{\hat{Z}|Z}(0|0) = P_{\hat{Z}|Z}(0|1) = 1\) to obtain \(I(Z; \hat{Z}) = 0\) and

\[
E[\rho(Z, \hat{Z})] = \sum_{z=0}^{1} \sum_{\hat{z}=0}^{1} P_Z(z) P_{\hat{Z}|Z}(\hat{z}|z) \rho(z, \hat{z}) = 1 - p \leq D.
\]

\[\square\]

### 6.4.2 Rate-distortion function for Gaussian source and squared error distortion measure

In this subsection, we actually show that the Gaussian source maximizes the rate-distortion function under square error distortion measure. The rate-distortion function for Gaussian source and squared error distortion measure is thus an upper bound for all other continuous sources of the same variance as summarized in the next theorem. The result parallels Theorem 5.20 which states that the Gaussian source maximizes the differential entropy among all continuous sources of the same covariance matrix. In a sense, one may draw the conclusion that the Gaussian source is rich in content and is the most difficult one to be compressed (subject to squared-error distortion measure).

**Theorem 6.18** Under the squared error distortion measure, namely

\[
\rho(z, \hat{z}) = (z - \hat{z})^2,
\]

\[187\]
the rate-distortion function for continuous source $Z$ with zero mean and variance $\sigma^2$ satisfies

$$R(D) \leq \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D}, & \text{for } 0 \leq D \leq \sigma^2; \\ 0, & \text{for } D > \sigma^2. \end{cases}$$

Equality holds when $Z$ is Gaussian.

**Proof:** By Theorem 6.16 (extended to “unbounded” squared-error distortion measure),

$$R(D) = \min_{p_{\hat{Z}|Z}} I(Z; \hat{Z}).$$

So for any $p_{\hat{Z}|Z}$ satisfying the distortion constraint,

$$R(D) \leq I(p_Z; p_{\hat{Z}|Z}).$$

For $0 \leq D \leq \sigma^2$, choose a dummy Gaussian random variable $W$ with zero mean and variance $aD$, where $a = 1 - D/\sigma^2$, and is independent of $Z$. Let $\hat{Z} = aZ + W$. Then

$$E[(Z - \hat{Z})^2] = E[(1 - a)^2 Z^2 + E[W^2] = (1 - a)^2 \sigma^2 + aD = D$$

which satisfies the distortion constraint. Note that the variance of $\hat{Z}$ is equal to $E[a^2Z^2] + E[W^2] = \sigma^2 - D$. Consequently,

$$R(D) \leq I(Z; \hat{Z}) = h(\hat{Z}) - h(\hat{Z}|Z) = h(\hat{Z}) - h(W + aZ|Z) = h(\hat{Z}) - h(W) \quad \text{(By Lemma 5.14)}$$

$$= h(\hat{Z}) - h \left( \frac{1}{2} \log(2\pi e(aD)) \right) \leq \frac{1}{2} \log \left( 2\pi e(\sigma^2 - D) \right) - \frac{1}{2} \log \left( 2\pi e(aD) \right) = \frac{1}{2} \log \frac{\sigma^2}{D}.$$  

For $D > \sigma^2$, let $\hat{Z}$ satisfy $\Pr\{\hat{Z} = 0\} = 1$, and be independent of $Z$. Then

$$E[(Z - \hat{Z})^2] = E[Z^2] + E[\hat{Z}^2] - 2E[Z]E[\hat{Z}] = \sigma^2 < D, \text{ and } I(Z; \hat{Z}) = 0.$$ 

Hence, $R(D) = 0$ for $D > \sigma^2$.  

188
The achievability of this upper bound by Gaussian source can be proved by showing that under the Gaussian source, \((1/2) \log(\sigma^2/D)\) is a lower bound to \(R(D)\) for \(0 \leq D \leq \sigma^2\).

For Gaussian source \(Z\) with \(E[(Z - \hat{Z})^2] \leq D\),

\[
I(Z; \hat{Z}) = h(Z) - h(Z|\hat{Z}) = \frac{1}{2} \log(2\pi e\sigma^2) - h(Z|\hat{Z}) \quad \text{(Lemma 5.14)}
\]

\[
\geq \frac{1}{2} \log(2\pi e\sigma^2) - h(Z - \hat{Z}) \quad \text{(Lemma 5.14)}
\]

\[
\geq \frac{1}{2} \log(2\pi e\sigma^2) - \frac{1}{2} \log \left(2\pi e \text{Var}[(Z - \hat{Z})]\right) \quad \text{(Theorem 5.20)}
\]

\[
\geq \frac{1}{2} \log(2\pi e\sigma^2) - \frac{1}{2} \log \left(2\pi e E[(Z - \hat{Z})^2]\right)
\]

\[
\geq \frac{1}{2} \log(2\pi e\sigma^2) - \frac{1}{2} \log (2\pi e D)
\]

\[
= \frac{1}{2} \log \frac{\sigma^2}{D}.
\]

\[
\square
\]

### 6.5 Joint source-channel information transmission

At the end of Part I of the lecture notes, we continue from Fig. 6.1 the discussion about whether a source can be transmitted via a noisy channel of capacity \(C(S)\) within distortion \(D\). Note that if the source entropy rate is less than the capacity, reliable transmission is achievable and hence the answer to the above query is affirmative for arbitrary \(D\). So the query is mostly concerned when the capacity is not large enough to reliably convey the source (i.e., the source entropy rate exceeds the capacity). This query then leads to the following theorem.

**Theorem 6.19 (Joint source-channel coding theorem)** Fix a distortion measure. A DMS can be reproduced at the output of a channel with distortion less than \(D\) (by taking sufficiently large blocklength), if

\[
\frac{R(D)}{T_s} < \frac{C(S)}{T_c},
\]

where \(T_s\) and \(T_c\) represent the durations per source letter and per channel input, respectively. Conversely, all data transmission codes will have average distortion larger than \(D\) for sufficiently large blocklength, if

\[
\frac{R(D)}{T_s} > \frac{C(S)}{T_c}.
\]
Note that the units of $R(D)$ and $C(S)$ should be the same, i.e., they should be measured both in nats (by taking natural logarithm), or both in bits (by taking base-2 logarithm).

The theorem is simply a consequence of Shannon’s channel coding theorem (i.e., Theorem 4.10) and Rate-distortion theorem (i.e., Theorem 6.16) and hence we omit its proof but focus on examples that can reveal its significance.

**Example 6.20 (Binary-input additive white Gaussian noise (AWGN) channel)** Assume that the discrete-time binary source to be transmitted is memoryless with uniform marginal distribution, and the discrete-time channel has binary input alphabet and real-line output alphabet with Gaussian transition probability. Denote by $P_b$ the probability of bit error.

From Theorem 6.17, the rate-distortion function for binary input and Hamming additive distortion measure is

$$R(D) = \begin{cases} 1 - h_b(D), & \text{if } 0 \leq D \leq \frac{1}{2}; \\ 0, & \text{if } D > \frac{1}{2}, \end{cases}$$

where $h_b(x) = -x \log_2(x) - (1 - x) \log_2(1 - x)$ is the binary entropy function. Notably, the distortion bound $D$ is exactly a bound on bit error rate $P_b$ since Hamming additive distortion measure is used.

According to [12], the channel capacity-cost function for antipodal binary-input AWGN channel is

$$C(S) = \frac{S}{\sigma^2} \log_2(e) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \log_2 \left[ \cosh \left( \frac{S}{\sigma^2} + y \sqrt{\frac{S}{\sigma^2}} \right) \right] dy$$

$$= \frac{E_b T_c / T_s}{N_0/2} \log_2(e) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \log_2 \left[ \cosh \left( \frac{E_b E_c / T_s}{N_0/2} + y \sqrt{\frac{E_b T_c / T_s}{N_0/2}} \right) \right] dy$$

$$= 2R\gamma_b \log_2(e) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \log_2(\cosh(2R\gamma_b + y\sqrt{2R\gamma_b}))dy,$$

where $R = T_c / T_s$ is the code rate for data transmission and is measured in the unit of source letter/channel usage (or information bit/channel bit), and $\gamma_b$ (often denoted by $E_b / N_0$) is the signal-to-noise ratio per information bit.

Then from the joint source-channel coding theorem, good codes exist when

$$R(D) < \frac{T_s}{T_c} C(S),$$
or equivalently
\[
1 - h_b(P_b) < \frac{1}{R} \left[ 2R\gamma_b \log_2(e) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \log_2[\cosh(2R\gamma_b + y\sqrt{2R\gamma_b})] dy \right].
\]

By re-formulating the above inequality as
\[
h_b(P_b) > 1 - 2\gamma_b \log_2(e) + \frac{1}{R\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \log_2[\cosh(2R\gamma_b + y\sqrt{2R\gamma_b})] dy,
\]
a lower bound on the bit error probability as a function of \(\gamma_b\) is established. This is the Shannon limit for any code to achieve binary-input Gaussian channel (cf. Fig. 6.3).

![Shannon Limit](image)

Figure 6.3: The Shannon limits for (2, 1) and (3, 1) codes under antipodal binary-input AWGN channel.

The result in the above example becomes important due to the invention of the Turbo coding [3, 4], for which a near-Shannon-limit performance is first obtained. Specifically, the half-rate turbo coding system proposed in [3, 4] can approach the bit error rate of \(10^{-5}\) at \(\gamma_b = 0.9\) dB, which is only 0.7 dB away from the Shannon limit 0.19 dB. This implies that a near-optimal channel code has been constructed, since in principle, no codes can perform better than the Shannon limit.
Example 6.21 (AWGN channel with real input) Assume that the binary source is memoryless with uniform marginal distribution, and the channel has both real-line input and real-line output alphabets with Gaussian transition probability. Denote by $P_b$ the probability of bit error.

Again, the rate-distortion function for binary input and Hamming additive distortion measure is

$$R(D) = \begin{cases} 
1 - h_b(D), & \text{if } 0 \leq D \leq \frac{1}{2}; \\
0, & \text{if } D > \frac{1}{2},
\end{cases}$$

In addition, the channel capacity-cost function for real-input AWGN channel is

$$C(S) = \frac{1}{2} \log_2 \left( 1 + \frac{S}{\sigma^2} \right)$$

$$= \frac{1}{2} \log_2 \left( 1 + \frac{E_b T_c / T_s}{N_0 / 2} \right)$$

$$= \frac{1}{2} \log_2 (1 + 2R\gamma_b) \text{ bits/channel symbol,}$$

where $R = T_c / T_s$ is the code rate for data transmission and is measured in the unit of information bit/channel usage, and $\gamma_b = E_b / N_0$ is the signal-to-noise ratio per information bit.

Then from the joint source-channel coding theorem, good codes exist when

$$R(D) < \frac{T_s}{T_c} C(S),$$

or equivalently

$$1 - h_b(P_b) < \frac{1}{R} \left[ \frac{1}{2} \log_2 (1 + 2R\gamma_b) \right].$$

By re-formulating the above inequality as

$$h_b(P_b) > 1 - \frac{1}{2R} \log_2 (1 + 2R\gamma_b),$$

a lower bound on the bit error probability as a function of $\gamma_b$ is established. This is the Shannon limit for any code to achieve for real input Gaussian channel (cf. Fig. 6.4).

192
Figure 6.4: The Shannon limits for (2, 1) and (3, 1) codes under continuous-input AWGN channels.
Appendix A

Overview on Suprema and Limits

We herein review basic results on suprema and limits which are useful for the development of information theoretic coding theorems; they can be found in standard real analysis texts (e.g., see [36, 50]).

A.1 Supremum and maximum

Throughout, we work on subsets of \( \mathbb{R} \), the set of real numbers.

Definition A.1 (Upper bound of a set) A real number \( u \) is called an upper bound of a non-empty subset \( A \) of \( \mathbb{R} \) if every element of \( A \) is less than or equal to \( u \); we say that \( A \) is bounded above. Symbolically, the definition becomes:

\[
A \subset \mathbb{R} \text{ is bounded above } \iff (\exists \ u \in \mathbb{R}) \text{ such that } (\forall \ a \in A), a \leq u.
\]

Definition A.2 (Least upper bound or supremum) Suppose \( A \) is a non-empty subset of \( \mathbb{R} \). Then we say that a real number \( s \) is a least upper bound or supremum of \( A \) if \( s \) is an upper bound of the set \( A \) and if \( s \leq s' \) for each upper bound \( s' \) of \( A \). In this case, we write \( s = \text{sup} \ A \); other notations are \( s = \text{sup}_{x \in A} x \) and \( s = \text{sup}\{x : x \in A\} \).

Completeness Axiom: (Least upper bound property) Let \( A \) be a non-empty subset of \( \mathbb{R} \) that is bounded above. Then \( A \) has a least upper bound.

It follows directly that if a non-empty set in \( \mathbb{R} \) has a supremum, then this supremum is unique. Furthermore, note that the empty set \( (\emptyset) \) and any set not bounded above do not admit a supremum in \( \mathbb{R} \). However, when working in the set of extended real numbers given by \( \mathbb{R} \cup \{\pm\infty\} \), we can define the
supremum of the empty set as $-\infty$ and that of a set not bounded above as $\infty$. These extended definitions will be adopted in the text.

We now distinguish between two situations: (i) the supremum of a set $\mathcal{A}$ belongs to $\mathcal{A}$, and (ii) the supremum of a set $\mathcal{A}$ does not belong to $\mathcal{A}$. It is quite easy to create examples for both situations. A quick example for (i) involves the set $(0, 1]$, while the set $(0, 1)$ can be used for (ii). In both examples, the supremum is equal to 1; however, in the former case, the supremum belongs to the set, while in the latter case it does not. When a set contains its supremum, we call the supremum the *maximum* of the set.

**Definition A.3 (Maximum)** If $\sup \mathcal{A} \in \mathcal{A}$, then $\sup \mathcal{A}$ is also called the *maximum* of $\mathcal{A}$, and is denoted by $\max \mathcal{A}$. However, if $\sup \mathcal{A} \not\in \mathcal{A}$, then we say that the maximum of $\mathcal{A}$ does not exist.

**Property A.4 (Properties of the supremum)**

1. The supremum of any set in $\mathbb{R} \cup \{-\infty, \infty\}$ always exists.

2. $(\forall a \in \mathcal{A}) \ a \leq \sup \mathcal{A}$.

3. If $-\infty < \sup \mathcal{A} < \infty$, then $(\forall \ \varepsilon > 0)(\exists \ a_0 \in \mathcal{A}) \ a_0 > \sup \mathcal{A} - \varepsilon$.

(The existence of $a_0 \in (\sup \mathcal{A} - \varepsilon, \sup \mathcal{A}]$ for any $\varepsilon > 0$ under the condition of $|\sup \mathcal{A}| < \infty$ is called the *approximation property for the supremum*.)

4. If $\sup \mathcal{A} = \infty$, then $(\forall L \in \mathbb{R})(\exists B_0 \in \mathcal{A}) \ B_0 > L$.

5. If $\sup \mathcal{A} = -\infty$, then $\mathcal{A}$ is empty.

**Observation A.5** In Information Theory, a typical channel coding theorem establishes that a (finite) real number $\alpha$ is the supremum of a set $\mathcal{A}$. Thus, to prove such a theorem, one must show that $\alpha$ satisfies both properties 3 and 2 above, i.e.,

$$(\forall \ \varepsilon > 0)(\exists \ a_0 \in \mathcal{A}) \ a_0 > \alpha - \varepsilon$$  \hspace{1cm} (A.1.1)

and

$$(\forall \ a \in \mathcal{A}) \ a \leq \alpha,$$  \hspace{1cm} (A.1.2)

where (A.1.1) and (A.1.2) are called the *achievability* (or *forward*) part and the *converse* part, respectively, of the theorem. Specifically, (A.1.2) states that $\alpha$ is an upper bound of $\mathcal{A}$, and (A.1.1) states that no number less than $\alpha$ can be an upper bound for $\mathcal{A}$.
Property A.6 (Properties of the maximum)

1. \((\forall a \in \mathcal{A}) a \leq \max \mathcal{A}\), if \(\max \mathcal{A}\) exists in \(\mathbb{R} \cup \{-\infty, \infty\}\).

2. \(\max \mathcal{A} \in \mathcal{A}\).

From the above property, in order to obtain \(\alpha = \max \mathcal{A}\), one needs to show that \(\alpha\) satisfies both

\[(\forall a \in \mathcal{A}) a \leq \alpha \quad \text{and} \quad \alpha \in \mathcal{A}\.

A.2 Infimum and minimum

The concepts of infimum and minimum are dual to those of supremum and maximum.

Definition A.7 (Lower bound of a set) A real number \(\ell\) is called a lower bound of a non-empty subset \(\mathcal{A}\) in \(\mathbb{R}\) if every element of \(\mathcal{A}\) is greater than or equal to \(\ell\); we say that \(\mathcal{A}\) is bounded below. Symbolically, the definition becomes:

\[\mathcal{A} \subset \mathbb{R} \text{ is bounded below} \iff (\exists \ell \in \mathbb{R}) \text{ such that } (\forall a \in \mathcal{A}) a \geq \ell.\]

Definition A.8 (Greatest lower bound or infimum) Suppose \(\mathcal{A}\) is a non-empty subset of \(\mathbb{R}\). Then we say that a real number \(\ell\) is a greatest lower bound or infimum of \(\mathcal{A}\) if \(\ell\) is a lower bound of \(\mathcal{A}\) and if \(\ell \geq \ell'\) for each lower bound \(\ell'\) of \(\mathcal{A}\). In this case, we write \(\ell = \inf \mathcal{A}\); other notations are \(\ell = \inf_{x \in \mathcal{A}} x\) and \(\ell = \inf\{x : x \in \mathcal{A}\}\).

Completeness Axiom: (Greatest lower bound property) Let \(\mathcal{A}\) be a non-empty subset of \(\mathbb{R}\) that is bounded below. Then \(\mathcal{A}\) has a greatest lower bound.

As for the case of the supremum, it directly follows that if a non-empty set in \(\mathbb{R}\) has an infimum, then this infimum is unique. Furthermore, working in the set of extended real numbers, the infimum of the empty set is defined as \(\infty\) and that of a set not bounded below as \(-\infty\).

Definition A.9 (Minimum) If \(\inf \mathcal{A} \in \mathcal{A}\), then \(\inf \mathcal{A}\) is also called the minimum of \(\mathcal{A}\), and is denoted by \(\min \mathcal{A}\). However, if \(\inf \mathcal{A} \notin \mathcal{A}\), we say that the minimum of \(\mathcal{A}\) does not exist.
Property A.10 (Properties of the infimum)

1. The infimum of any set in \( \mathbb{R} \cup \{-\infty, \infty\} \) always exists.

2. \((\forall a \in \mathcal{A})\ a \geq \inf \mathcal{A}\).

3. If \( \infty > \inf \mathcal{A} > -\infty \), then \((\forall \varepsilon > 0)(\exists a_0 \in \mathcal{A})\ a_0 < \inf \mathcal{A} + \varepsilon\).

   (The existence of \( a_0 \in [\inf \mathcal{A}, \inf \mathcal{A}+\varepsilon) \) for any \( \varepsilon > 0 \) under the assumption of \( |\inf \mathcal{A}| < \infty \) is called the approximation property for the infimum.)

4. If \( \inf \mathcal{A} = -\infty \), then \((\forall A \in \mathbb{R})(\exists B_0 \in \mathcal{A})B_0 < L\).

5. If \( \inf \mathcal{A} = \infty \), then \( \mathcal{A} \) is empty.

Observation A.11 Analogously to Observation A.5, a typical source coding theorem in Information Theory establishes that a (finite) real number \( \alpha \) is the infimum of a set \( \mathcal{A} \). Thus, to prove such a theorem, one must show that \( \alpha \) satisfies both properties 3 and 2 above, i.e.,

\[
(\forall \varepsilon > 0)(\exists a_0 \in \mathcal{A})\ a_0 < \alpha + \varepsilon
\]

(A.2.1)

and

\[
(\forall a \in \mathcal{A})\ a \geq \alpha.
\]

(A.2.2)

Here, (A.2.1) is called the achievability or forward part of the coding theorem; it specifies that no number greater than \( \alpha \) can be a lower bound for \( \mathcal{A} \). Also, (A.2.2) is called the converse part of the theorem; it states that \( \alpha \) is a lower bound of \( \mathcal{A} \).

Property A.12 (Properties of the minimum)

1. \((\forall a \in \mathcal{A})\ a \geq \min \mathcal{A} \), if \( \min \mathcal{A} \) exists in \( \mathbb{R} \cup \{-\infty, \infty\} \).

2. \( \min \mathcal{A} \in \mathcal{A} \).

A.3 Boundedness and suprema operations

Definition A.13 (Boundedness) A subset \( \mathcal{A} \) of \( \mathbb{R} \) is said to be bounded if it is both bounded above and bounded below; otherwise it is called unbounded.

Lemma A.14 (Condition for boundedness) A subset \( \mathcal{A} \) of \( \mathbb{R} \) is bounded iff \((\exists k \in \mathbb{R})\) such that \((\forall a \in \mathcal{A})\ |a| \leq k\).
Lemma A.15 (Monotone property) Suppose that $A$ and $B$ are non-empty subsets of $\mathbb{R}$ such that $A \subset B$. Then

1. $\sup A \leq \sup B$.
2. $\inf A \geq \inf B$.

Lemma A.16 (Supremum for set operations) Define the “addition” of two sets $A$ and $B$ as

$$A + B \triangleq \{ c \in \mathbb{R} : c = a + b \text{ for some } a \in A \text{ and } b \in B \}.$$  

Define the “scaler multiplication” of a set $A$ by a scalar $k \in \mathbb{R}$ as

$$k \cdot A \triangleq \{ c \in \mathbb{R} : c = k \cdot a \text{ for some } a \in A \}.$$  

Define the “negation” of a set $A$ as

$$-A \triangleq \{ c \in \mathbb{R} : c = -a \text{ for some } a \in A \}.$$  

Then the following hold.

1. If $A$ and $B$ are both bounded above, then $A + B$ is also bounded above and $\sup(A + B) = \sup A + \sup B$.
2. If $0 < k < \infty$ and $A$ is bounded above, then $k \cdot A$ is also bounded above and $\sup(k \cdot A) = k \cdot \sup A$.
3. $\sup A = -\inf(-A)$ and $\inf A = -\sup(-A)$.

Property 1 does not hold for the “product” of two sets, where the “product” of sets $A$ and $B$ is defined as as

$$A \cdot B \triangleq \{ c \in \mathbb{R} : c = ab \text{ for some } a \in A \text{ and } b \in B \}.$$  

In this case, both of these two situations can occur:

$$\sup(A \cdot B) > (\sup A) \cdot (\sup B)$$
$$\sup(A \cdot B) = (\sup A) \cdot (\sup B).$$
Lemma A.17 (Supremum/infimum for monotone functions)

1. If \( f : \mathbb{R} \rightarrow \mathbb{R} \) is a non-decreasing function, then
\[
\sup\{x \in \mathbb{R} : f(x) < \varepsilon\} = \inf\{x \in \mathbb{R} : f(x) \geq \varepsilon\}
\]
and
\[
\sup\{x \in \mathbb{R} : f(x) \leq \varepsilon\} = \inf\{x \in \mathbb{R} : f(x) > \varepsilon\}.
\]
2. If \( f : \mathbb{R} \rightarrow \mathbb{R} \) is a non-increasing function, then
\[
\sup\{x \in \mathbb{R} : f(x) > \varepsilon\} = \inf\{x \in \mathbb{R} : f(x) \leq \varepsilon\}
\]
and
\[
\sup\{x \in \mathbb{R} : f(x) \geq \varepsilon\} = \inf\{x \in \mathbb{R} : f(x) < \varepsilon\}.
\]

The above lemma is illustrated in Figure A.1.

A.4 Sequences and their limits

Let \( \mathbb{N} \) denote the set of “natural numbers” (positive integers) 1, 2, 3, \( \cdot \cdot \cdot \). A sequence drawn from a real-valued function is denoted by
\[
f : \mathbb{N} \rightarrow \mathbb{R}.
\]

In other words, \( f(n) \) is a real number for each \( n = 1, 2, 3, \cdot \cdot \cdot \). It is usual to write \( f(n) = a_n \), and we often indicate the sequence by any one of these notations
\[
\{a_1, a_2, a_3, \cdot \cdot \cdot, a_n, \cdot \cdot \cdot\} \quad \text{or} \quad \{a_n\}_{n=1}^{\infty}.
\]

One important question that arises with a sequence is what happens when \( n \) gets large. To be precise, we want to know that when \( n \) is large enough, whether or not every \( a_n \) is close to some fixed number \( L \) (which is the limit of \( a_n \)).

Definition A.18 (Limit) The limit of \( \{a_n\}_{n=1}^{\infty} \) is the real number \( L \) satisfying:
\[
(\forall \varepsilon > 0)(\exists N) \text{ such that } (\forall n > N) \quad |a_n - L| < \varepsilon.
\]
In this case, we write \( L = \lim_{n \rightarrow \infty} a_n \). If no such \( L \) satisfies the above statement, we say that the limit of \( \{a_n\}_{n=1}^{\infty} \) does not exist.
Property A.19 If \( \{a_n\}_{n=1}^{\infty} \) and \( \{b_n\}_{n=1}^{\infty} \) both have a limit in \( \mathbb{R} \), then the following hold.

1. \( \lim_{n \to \infty} (a_n + b_n) = \lim_{n \to \infty} a_n + \lim_{n \to \infty} b_n. \)

2. \( \lim_{n \to \infty} (\alpha \cdot a_n) = \alpha \cdot \lim_{n \to \infty} a_n. \)

3. \( \lim_{n \to \infty} (a_n b_n) = (\lim_{n \to \infty} a_n)(\lim_{n \to \infty} b_n). \)

Note that in the above definition, \(-\infty\) and \(\infty\) cannot be a legitimate limit for any sequence. In fact, if \((\forall \ L)(\exists \ N)\) such that \((\forall \ n > N) \ a_n > L\), then we
say that \( a_n \) \textit{diverges} to \( \infty \) and write \( a_n \to \infty \). A similar argument applies to \( a_n \) diverging to \( -\infty \). For convenience, we will work in the set of extended real numbers and thus state that a sequence \( \{a_n\}_{n=1}^{\infty} \) that diverges to either \( \infty \) or \( -\infty \) has a limit in \( \mathbb{R} \cup \{-\infty, \infty\} \).

**Lemma A.20 (Convergence of monotone sequences)** If \( \{a_n\}_{n=1}^{\infty} \) is non-decreasing in \( n \), then \( \lim_{n \to \infty} a_n \) exists in \( \mathbb{R} \cup \{-\infty, \infty\} \). If \( \{a_n\}_{n=1}^{\infty} \) is also bounded from above – i.e., \( a_n \leq L \) \( \forall \) \( n \) for some \( L \) in \( \mathbb{R} \) – then \( \lim_{n \to \infty} a_n \) exists in \( \mathbb{R} \).

Likewise, if \( \{a_n\}_{n=1}^{\infty} \) is non-increasing in \( n \), then \( \lim_{n \to \infty} a_n \) exists in \( \mathbb{R} \cup \{-\infty, \infty\} \). If \( \{a_n\}_{n=1}^{\infty} \) is also bounded from below – i.e., \( a_n \geq L \) \( \forall \) \( n \) for some \( L \) in \( \mathbb{R} \) – then \( \lim_{n \to \infty} a_n \) exists in \( \mathbb{R} \).

As stated above, the limit of a sequence may not exist. For example, \( a_n = (-1)^n \). Then \( a_n \) will be close to either \(-1\) or \(1\) for \( n \) large. Hence, more generalized definitions that can describe the general limiting behavior of a sequence is required.

**Definition A.21 (limsup and liminf)** The \textit{limit supremum} of \( \{a_n\}_{n=1}^{\infty} \) is the extended real number in \( \mathbb{R} \cup \{-\infty, \infty\} \) defined by

\[
\limsup_{n \to \infty} a_n \triangleq \lim_{n \to \infty} \left( \sup_{k \geq n} a_k \right),
\]

and the \textit{limit infimum} of \( \{a_n\}_{n=1}^{\infty} \) is the extended real number defined by

\[
\liminf_{n \to \infty} a_n \triangleq \lim_{n \to \infty} \left( \inf_{k \geq n} a_k \right).
\]

Some also use the notations \( \lim \sup \) and \( \lim \inf \) to denote limsup and liminf, respectively.

Note that the limit supremum and the limit infimum of a sequence is always defined in \( \mathbb{R} \cup \{-\infty, \infty\} \), since the sequences \( \sup_{k \geq n} a_k = \sup\{a_k : k \geq n\} \) and \( \inf_{k \geq n} a_k = \inf\{a_k : k \geq n\} \) are monotone in \( n \) (cf. Lemma A.20). An immediate result follows from the definitions of limsup and liminf.

**Lemma A.22 (Limit)** For a sequence \( \{a_n\}_{n=1}^{\infty} \),

\[
\lim_{n \to \infty} a_n = L \iff \limsup_{n \to \infty} a_n = \liminf_{n \to \infty} a_n = L.
\]

Some properties regarding the limsup and liminf of sequences (which are parallel to Properties A.4 and A.10) are listed below.
Property A.23 (Properties of the limit supremum)

1. The limit supremum always exists in $\mathbb{R} \cup \{-\infty, \infty\}$.

2. If $|\limsup_{m \to \infty} a_m| < \infty$, then $(\forall \varepsilon > 0)(\exists N)$ such that $(\forall n > N)$ $a_n < \limsup_{m \to \infty} a_m + \varepsilon$. (Note that this holds for every $n > N$.)

3. If $|\limsup_{m \to \infty} a_m| < \infty$, then $(\forall \varepsilon > 0$ and integer $K)(\exists N > K)$ such that $a_N > \limsup_{m \to \infty} a_m - \varepsilon$. (Note that this holds only for one $N$, which is larger than $K$.)

Property A.24 (Properties of the limit infimum)

1. The limit infimum always exists in $\mathbb{R} \cup \{-\infty, \infty\}$.

2. If $|\liminf_{m \to \infty} a_m| < \infty$, then $(\forall \varepsilon > 0$ and integer $K)(\exists N > K)$ such that $a_N < \liminf_{m \to \infty} a_m + \varepsilon$. (Note that this holds only for one $N$, which is larger than $K$.)

3. If $|\liminf_{m \to \infty} a_m| < \infty$, then $(\forall \varepsilon > 0)(\exists N)$ such that $(\forall n > N)$ $a_n > \liminf_{m \to \infty} a_m - \varepsilon$. (Note that this holds for every $n > N$.)

The last two items in Properties A.23 and A.24 can be stated using the terminology of sufficiently large and infinitely often, which is often adopted in Information Theory.

Definition A.25 (Sufficiently large) We say that a property holds for a sequence $\{a_n\}_{n=1}^{\infty}$ almost always or for all sufficiently large $n$ if the property holds for every $n > N$ for some $N$.

Definition A.26 (Infinitely often) We say that a property holds for a sequence $\{a_n\}_{n=1}^{\infty}$ infinitely often or for infinitely many $n$ if for every $K$, the property holds for one (specific) $N$ with $N > K$.

Then properties 2 and 3 of Property A.23 can be respectively re-phrased as: if $|\limsup_{m \to \infty} a_m| < \infty$, then $(\forall \varepsilon > 0)$

$$a_n < \limsup_{m \to \infty} a_m + \varepsilon \quad \text{for all sufficiently large } n$$

and

$$a_n > \limsup_{m \to \infty} a_m - \varepsilon \quad \text{for infinitely many } n.$$
Similarly, properties 2 and 3 of Property A.24 becomes: if \( |\liminf_{m \to \infty} a_m| < \infty \), then (\( \forall \varepsilon > 0 \))

\[
a_n < \liminf_{m \to \infty} a_n + \varepsilon \quad \text{for infinitely many } n
\]

and

\[
a_n > \liminf_{m \to \infty} a_n - \varepsilon \quad \text{for all sufficiently large } n.
\]

Lemma A.27

1. \( \liminf_{n \to \infty} a_n \leq \limsup_{n \to \infty} a_n \).

2. If \( a_n \leq b_n \) for all sufficiently large \( n \), then

\[
\liminf_{n \to \infty} a_n \leq \liminf_{n \to \infty} b_n \quad \text{and} \quad \limsup_{n \to \infty} a_n \leq \limsup_{n \to \infty} b_n.
\]

3. \( \limsup_{n \to \infty} a_n < r \implies a_n < r \) for all sufficiently large \( n \).

4. \( \limsup_{n \to \infty} a_n > r \implies a_n > r \) for infinitely many \( n \).

5.

\[
\liminf_{n \to \infty} a_n + \liminf_{n \to \infty} b_n \leq \liminf_{n \to \infty} (a_n + b_n) \\
\leq \limsup_{n \to \infty} a_n + \limsup_{n \to \infty} b_n \\
\leq \limsup_{n \to \infty} (a_n + b_n) \\
\leq \limsup_{n \to \infty} a_n + \limsup_{n \to \infty} b_n.
\]

6. If \( \lim_{n \to \infty} a_n \) exists, then

\[
\liminf_{n \to \infty} (a_n + b_n) = \lim_{n \to \infty} a_n + \liminf_{n \to \infty} b_n
\]

and

\[
\limsup_{n \to \infty} (a_n + b_n) = \lim_{n \to \infty} a_n + \limsup_{n \to \infty} b_n.
\]

Finally, one can also interpret the limit supremum and limit infimum in terms of the concept of clustering points. A clustering point is a point that a sequence \( \{a_n\}_{n=1}^{\infty} \) approaches (i.e., belonging to a ball with arbitrarily small radius and that point as center) infinitely many times. For example, if \( a_n = \sin(n\pi/2) \), then \( \{a_n\}_{n=1}^{\infty} = \{1, 0, -1, 0, 1, 0, -1, 0, \ldots\} \). Hence, there are three clustering points in this sequence, which are \(-1, 0\) and \(1\). Then the limit supremum of the sequence is nothing but its largest clustering point, and its limit infimum is exactly its smallest clustering point. Specifically, \( \limsup_{n \to \infty} a_n = 1 \) and \( \liminf_{n \to \infty} a_n = -1 \). This approach can sometimes be useful to determine the limsup and liminf quantities.
A.5 Equivalence

We close this appendix by providing some equivalent statements that are often used to simplify proofs. For example, instead of directly showing that quantity \( x \) is less than or equal to quantity \( y \), one can take an arbitrary constant \( \varepsilon > 0 \) and prove that \( x < y + \varepsilon \). Since \( y + \varepsilon \) is a larger quantity than \( y \), in some cases it might be easier to show \( x < y + \varepsilon \) than proving \( x \leq y \). By the next theorem, any proof that concludes that “\( x < y + \varepsilon \) for all \( \varepsilon > 0 \)” immediately gives the desired result of \( x \leq y \).

**Theorem A.28** For any \( x, y \) and \( a \) in \( \mathbb{R} \),

1. \( x < y + \varepsilon \) for all \( \varepsilon > 0 \) iff \( x \leq y \);
2. \( x < y - \varepsilon \) for some \( \varepsilon > 0 \) iff \( x < y \);
3. \( x > y - \varepsilon \) for all \( \varepsilon > 0 \) iff \( x \geq y \);
4. \( x > y + \varepsilon \) for some \( \varepsilon > 0 \) iff \( x > y \);
5. \( |a| < \varepsilon \) for all \( \varepsilon > 0 \) iff \( a = 0 \).
Appendix B

Overview in Probability and Random Processes

This appendix presents a quick overview of basic concepts from probability theory and the theory of random processes. The reader can consult comprehensive texts on these subjects for a thorough study (see, e.g., [2, 6, 22]). We close the appendix with a brief discussion of Jensen’s inequality and the Lagrange multipliers technique for the optimization of convex functions [5, 11].

B.1 Probability space

Definition B.1 (σ-Fields) Let \( F \) be a collection of subsets of a non-empty set \( \Omega \). Then \( F \) is called a σ-field (or σ-algebra) if the following hold:

1. \( \Omega \in F \).
2. Closedness of \( F \) under complementation: If \( A \in F \), then \( A^c \in F \), where \( A^c = \{ \omega \in \Omega : \omega \notin A \} \).
3. Closedness of \( F \) under countable union: If \( A_i \in F \) for \( i = 1, 2, 3, \ldots \), then \( \bigcup_{i=1}^{\infty} A_i \in F \).

It directly follows that the empty set \( \emptyset \) is also an element of \( F \) (as \( \Omega^c = \emptyset \)) and that \( F \) is closed under countable intersection since

\[
\bigcap_{i=1}^{\infty} A_i^c = \left( \bigcup_{i=1}^{\infty} A_i \right)^c.
\]

The largest σ-field of subsets of a given set \( \Omega \) is the collection of all subsets of \( \Omega \) (i.e., its powerset), while the smallest σ-field is given by \( \{\Omega, \emptyset\} \). Also, if \( A \) is
a proper (strict) non-empty subset of $\Omega$, then the smallest $\sigma$-field containing $A$ is given by $\{\Omega, \emptyset, A, A^c\}$.

**Definition B.2 (Probability space)** A probability space is a triple $(\Omega, \mathcal{F}, P)$, where $\Omega$ is a given set called sample space containing all possible outcomes (usually observed from an experiment), $\mathcal{F}$ is a $\sigma$-field of subsets of $\Omega$ and $P$ is a probability measure $P : \mathcal{F} \rightarrow [0,1]$ on the $\sigma$-field satisfying the following:

1. $0 \leq P(A) \leq 1$ for all $A \in \mathcal{F}$.
2. $P(\Omega) = 1$.
3. Countable additivity: If $A_1, A_2, \ldots$ is a sequence of disjoint sets (i.e., $A_i \cap A_j = \emptyset$ for all $i \neq j$) in $\mathcal{F}$, then
   \[
P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k).
   \]

It directly follows from properties 1-3 of the above definition that $P(\emptyset) = 0$. Usually, the $\sigma$-field $\mathcal{F}$ is called the event space and its elements (which are subsets of $\Omega$ satisfying the properties of Definition B.1) are called events.

### B.2 Random variable and random process

A random variable $X$ defined over probability space $(\Omega, \mathcal{F}, P)$ is a real-valued function $X : \Omega \rightarrow \mathbb{R}$ that is measurable (or $\mathcal{F}$-measurable), i.e., satisfying the property that

\[
X^{-1}((-\infty, t]) \triangleq \{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F}
\]

for each real $t$.\(^1\)

A random process is a collection of random variables that arise from the same probability space. It can be mathematically represented by the collection

\[
\{X_t, t \in I\},
\]

\(^1\)One may question why bother defining random variables based on some abstract probability space. One may continue “A random variable $X$ can simply be defined based on its distribution $P_X$,” which is indeed true (cf. Section B.3).

A perhaps easier way to understand the abstract definition of random variables is that the underlying probability space $(\Omega, \mathcal{F}, P)$ on which the random variable is defined is what truly occurs internally, but it is possibly non-observable. In order to infer which of the non-observable $\omega$ occurs, an experiment that results in observable $x$ that is a function of $\omega$ is performed. Such an experiment results in the random variable $X$ whose probability is so defined over the probability space $(\Omega, \mathcal{F}, P)$. 

206
where $X_t$ denotes the $t$-th random variable in the process, and the index $t$ runs over an index set $I$ which is arbitrary. The index set $I$ can be uncountably infinite — e.g., $I = \mathbb{R}$ — in which case we are effectively dealing with a continuous-time process. We will however exclude such a case in this chapter for the sake of simplicity. To be precise, we will only consider the following cases of index set $I$:

- **case a)** $I$ consists of one index only.
- **case b)** $I$ is finite.
- **case c)** $I$ is countably infinite.

### B.3 Distribution functions versus probability space

In applications, we are perhaps more interested in the distribution functions of the random variables and random processes than the underlying probability space on which the random variables and random processes are defined. It can be proved [6, Thm. 14.1] that given a real-valued non-negative function $F(\cdot)$ that is non-decreasing and right-continuous and satisfies $\lim_{x \downarrow -\infty} F(x) = 0$ and $\lim_{x \uparrow \infty} F(x) = 1$, there exist a random variable and an underlying probability space such that the cumulative distribution function (cdf) of the random variable defined over the probability space is equal to $F(\cdot)$. This result releases us from the burden of referring to a probability space before defining the random variable. In other words, we can define a random variable $X$ directly by its cdf, i.e., $\Pr[X \leq x]$, without bothering to refer to its underlying probability space. Nevertheless, it is better to keep in mind (and learn) that, formally, random variables and random processes are defined over underlying probability spaces.

### B.4 Relation between a source and a random process

In statistical communication, a *discrete-time source* $(X_1, X_2, X_3, \ldots, X_n) \triangleq X^n$ consists of a sequence of random quantities, where each quantity usually takes on values from a source *generic alphabet* $\mathcal{X}$, namely

$$(X_1, X_2, \ldots, X_n) \in \mathcal{X} \times \mathcal{X} \times \cdots \times \mathcal{X} \triangleq \mathcal{X}^n.$$  

Another merit in defining random variables based on an abstract probability space can be observed from the extension definition of random variables to random processes. With the underlying probability space, any finite dimensional distribution of $\{X_t, t \in I\}$ is well-defined. For example,

$$\Pr[X_1 \leq x_1, X_5 \leq x_5, X_9 \leq x_9] = P\{\omega \in \Omega : X_1(\omega) \leq x_1, X_5(\omega) \leq x_5, X_9(\omega) \leq x_9\}.$$
The elements in $X$ are usually called *letters*.

### B.5 Statistical properties of random sources

For a random process $X = \{X_1, X_2, \ldots\}$ with alphabet $\mathcal{X}$ (i.e., $\mathcal{X} \subseteq \mathbb{R}$ is the range of each $X_i$) defined over probability space $(\Omega, \mathcal{F}, P)$, consider $\mathcal{X}^\infty$, the set of all sequences $x \triangleq (x_1, x_2, x_3, \ldots)$ of real numbers in $\mathcal{X}$. An event $E$ in $\mathcal{F}_X$, the smallest $\sigma$-field generated by all open sets of $\mathcal{X}^\infty$ (i.e., the Borel $\sigma$-field of $\mathcal{X}^\infty$), is said to be $T$-invariant with respect to the left-shift (or shift transformation) $T : \mathcal{X}^\infty \rightarrow \mathcal{X}^\infty$ if

$$TE \subseteq E,$$

where

$$TE \triangleq \{Tx : x \in E\} \quad \text{and} \quad Tx \triangleq T(x_1, x_2, x_3, \ldots) = (x_2, x_3, \ldots).$$

In other words, $T$ is equivalent to “chopping the first component.” For example, applying $T$ onto an event $E_1$ defined below,

$$E_1 \triangleq \{(x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1, \ldots), (x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, \ldots),$$

$$(x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1, \ldots)\}, \quad (B.5.1)$$

yields

$$TE_1 = \{(x_1 = 1, x_2 = 1, x_3 = 1, \ldots), (x_1 = 1, x_2 = 1, x_3 = 1 \ldots),$$

$$(x_1 = 0, x_2 = 1, x_3 = 1, \ldots)\}$$

$$= \{(x_1 = 1, x_2 = 1, x_3 = 1, \ldots), (x_1 = 0, x_2 = 1, x_3 = 1, \ldots)\}.$$ 

We then have $TE_1 \subseteq E_1$, and hence $E_1$ is $T$-invariant.

It can be proved$^3$ that if $TE \subseteq E$, then $T^2E \subseteq TE$. By induction, we can further obtain

$$\cdots \subseteq T^3E \subseteq T^2E \subseteq TE \subseteq E. \quad (B.5.2)$$

Thus, if an element say $(1, 0, 0, 1, 0, 0, \ldots)$ is in a $T$-invariant set $E$, then all its left-shift counterparts (i.e., $(0, 0, 1, 0, 0, 1 \ldots)$ and $(0, 1, 0, 0, 1, 0, \ldots)$) should be contained in $E$. As a result, for a $T$-invariant set $E$, an element and all its left-shift counterparts are either all in $E$ or all outside $E$, but cannot be partially inside $E$. Hence, a “$T$-invariant group” such as one containing $(1, 0, 0, 1, 0, 0, \ldots)$, $^2$Note that by definition, the events $E \in \mathcal{F}_X$ are measurable, and $P_X(E) = P(A)$, where $P_X$ is the distribution of $X$ induced from the underlying probability measure $P$.

$^3$If $A \subseteq B$, then $TA \subseteq TB$. Thus $T^2E \subseteq TE$ holds whenever $TE \subseteq E$. 

208
(0, 0, 1, 0, 0, 1, 0, ... ) and (0, 1, 0, 0, 0, 0, 0, ...) should be treated as an indecomposable group in $T$-invariant sets.

Although we are in particular interested in these “$T$-invariant indecomposable groups” (especially when defining an ergodic random process), it is possible that some single “transient” element, such as (0, 0, 1, 1, ...) in (B.5.1), is included in a $T$-invariant set, and will be excluded after applying left-shift operation $T$. This however can be resolved by introducing the inverse operation $T^{-1}$. Note that $T$ is a many-to-one mapping, so its inverse operation does not exist in general.

Similar to taking the closure of an open set, the definition adopted below [46, p.3] allows us to “enlarge” the $T$-invariant set such that all right-shift counterparts of the single “transient” element are included.

$$T^{-1}E \triangleq \{ x \in X^\infty : Tx \in E \}.$$ We then notice from the above definition that if

$$T^{-1}E = E,$$ (B.5.3)

then

$$TE = T(T^{-1}E) = E,$$

and hence $E$ is constituted only by the $T$-invariant groups because

$$\cdots = T^{-2}E = T^{-1}E = E = TE = T^2E = \cdots .$$

The sets that satisfy (B.5.3) are sometimes referred to as ergodic sets because as time goes by (the left-shift operator $T$ can be regarded as a shift to a future time), the set always stays in the state that it has been before. A quick example of an ergodic set for $X = \{0, 1\}$ is one that consists of all binary sequences that contain finitely many 0’s.

We now classify several useful statistical properties of random process $X = \{X_1, X_2, \ldots \}.$

---

4The proof of $T(T^{-1}E) = E$ is as follows. If $y \in T(T^{-1}E) = T(\{ x \in X^\infty : Tx \in E \})$, then there must exist an element $x \in \{ x \in X^\infty : Tx \in E \}$ such that $y = Tx$. Since $Tx \in E$, we have $y \in E$ and $T(T^{-1}E) \subseteq E$. On the contrary, if $y \in E$, all $x$’s satisfying $Tx = y$ belong to $T^{-1}E$. Thus, $y \in T(T^{-1}E)$, which implies $E \subseteq T(T^{-1}E)$.

5As the textbook only deals with one-sided random processes, the discussion on $T$-invariance only focuses on sets of one-sided sequences. When a two-sided random process $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$ is considered, the left-shift operation $T$ of a two-sided sequence actually has a unique inverse. Hence, $TE \subseteq E$ implies $TE = E$. Also, $TE = E$ if, and only if, $T^{-1}E = E$. Ergodicity for two-sided sequences can therefore be directly defined using $TE = E$. 

209
• **Memoryless**: A random process is said to be memoryless if the sequence of random variables $X_i$ is independent and identically distributed (i.i.d.).

• **First-order stationary**: A process is first-order stationary if the marginal distribution is unchanged for every time instant.

• **Second-order stationary**: A process is second-order stationary if the joint distribution of any two (not necessarily consecutive) time instances is invariant before and after left (time) shift.

• **Weakly stationary process**: A process is weakly stationary (or wide-sense stationary or stationary in the weak or wide sense) if the mean and auto-correlation function are unchanged by a left (time) shift.

• **Stationary process**: A process is stationary (or strictly stationary) if the probability of every sequence or event is unchanged by a left (time) shift.

• **Ergodic process**: A process is ergodic if any ergodic set (satisfying (B.5.3)) in $\mathcal{F}_X$ has probability either 1 or 0. This definition is not very intuitive, but some interpretations and examples may shed some light.

  Observe that the definition has nothing to do with stationarity. It simply states that events that are unaffected by time-shifting (both left- and right-shifting) must have probability either zero or one.

  Ergodicity implies that all convergent sample averages\(^6\) converge to a constant (but not necessarily to the ensemble average), and stationarity assures that the time average converges to a random variable; hence, it is reasonably to expect that they jointly imply the ultimate time average equals the ensemble average. This is validated by the well-known ergodic theorem by Birkhoff and Khinchin.

Theorem B.3 (Pointwise ergodic theorem) Consider a discrete-time stationary random process, $X = \{X_n\}^\infty_{n=1}$. For real-valued function $f(\cdot)$ on $\mathbb{R}$ with finite mean (i.e., $|E[f(X_n)]| < \infty$), there exists a random variable $Y$ such that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = Y \quad \text{with probability 1.}$$

If, in addition to stationarity, the process is also ergodic, then

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = E[f(X_1)] \quad \text{with probability 1.}$$

\(^6\)Two alternative names for sample average are time average and Cesàro mean. In this book, these names will be used interchangeably.
Example B.4  Consider the process \( \{X_i\}_{i=1}^{\infty} \) consisting of a family of i.i.d. binary random variables (obviously, it is stationary and ergodic). Define the function \( f(\cdot) \) by \( f(0) = 0 \) and \( f(1) = 1 \). Hence,\(^7\)

\[
E[f(X_n)] = P_{X_n}(0)f(0) + P_{X_n}(1)f(1) = P_{X_n}(1)
\]
is finite. By the pointwise ergodic theorem, we have

\[
\lim_{n \to \infty} \frac{f(X_1) + f(X_2) + \cdots + f(X_n)}{n} = \lim_{n \to \infty} \frac{X_1 + X_2 + \cdots + X_n}{n} = P_X(1).
\]

As seen in the above example, one of the important consequences that the pointwise ergodic theorem indicates is that the time average can ultimately replace the statistical average, which is a useful result in engineering. Hence, with stationarity and ergodicity, one, who observes

\[
X_1^{30} = 1543265433422563245644234443
\]
from the experiment of rolling a dice, can draw the conclusion that the true distribution of rolling the dice can be well approximated by:

\[
\Pr\{X_i = 1\} \approx \frac{1}{30}, \quad \Pr\{X_i = 2\} \approx \frac{6}{30}, \quad \Pr\{X_i = 3\} \approx \frac{7}{30}
\]
\[
\Pr\{X_i = 4\} \approx \frac{9}{30}, \quad \Pr\{X_i = 5\} \approx \frac{4}{30}, \quad \Pr\{X_i = 6\} \approx \frac{3}{30}
\]
Such result is also known by the law of large numbers. The relation between ergodicity and the law of large numbers will be further explored in Section B.7.

We close the discussion on ergodicity by remarking that in communications theory, one may assume that the source is stationary or the source is stationary ergodic. But it is rare to see the assumption of the source being ergodic but non-stationary. This is perhaps because an ergodic but non-stationary source not only does not facilitate the analytical study of communications problems, but may have limited application in practice. From this, we note that assumptions are made either to facilitate the analytical study of communications problems or to fit a specific need of applications. Without these two objectives, an assumption becomes of minor interest. This, to some extent, justifies that the ergodicity assumption usually comes after stationarity assumption. A specific example is the pointwise ergodic theorem, where the random processes considered is presumed to be stationary.

\(^7\)In terms of notation, we use \( P_{X_n}(0) \) to denote \( \Pr\{X_n = 0\} \). The above two representations will be used alternatively throughout the book.
First-order Markov chain: Three random variables $X$, $Y$ and $Z$ are said to form a Markov chain or a first-order Markov chain if

$$P_{X,Y,Z}(x, y, z) = P_X(x) \cdot P_{Y|X}(y|x) \cdot P_{Z|Y}(z|y);$$  \hspace{1cm} (B.5.4)

i.e., $P_{Z|X,Y}(z|x,y) = P_{Z|Y}(z|y)$. This is usually denoted by $X \rightarrow Y \rightarrow Z$. $X \rightarrow Y \rightarrow Z$ is sometimes read as “$X$ and $Z$ are conditionally independent given $Y$” because it can be shown that (B.5.4) is equivalent to

$$P_{X,Z|Y}(x,z|y) = P_{X|Y}(x|y) \cdot P_{Z|Y}(z|y).$$

Therefore, $X \rightarrow Y \rightarrow Z$ is equivalent to $Z \rightarrow Y \rightarrow X$. Accordingly, the Markovian notation is sometimes expressed as $X \leftrightarrow Y \leftrightarrow Z$.

Markov chain for random sequences: The random variables $X_1$, $X_2$, $X_3$, ... are said to form a $k$-th order Markov chain if for all $k < n$,

$$\Pr[X_n = x_n|X_{n-1} = x_{n-1}, \ldots, X_1 = x_1] = \Pr[X_n = x_n|X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}].$$

Each $x_{n-k}^{n-1} \in \mathcal{X}^k$ is called the state at time $n$.

A Markov chain is irreducible if with some probability, we can go from any state in $\mathcal{X}^k$ to another state in a finite number of steps, i.e., for all $x^k, y^k \in \mathcal{X}^k$ there exists $j \geq 1$ such that

$$\Pr \left\{ X_{j+k}^{j-1} = x^k \mid X_1^k = y^k \right\} > 0.$$

A Markov chain is said to be time-invariant or homogeneous, if for every $n > k$,

$$\Pr[X_n = x_n|X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k}] = \Pr[X_{k+1} = x_{k+1}|X_k = x_k, \ldots, X_1 = x_1].$$

Therefore, a homogeneous first-order Markov chain can be defined through its transition probability:

$$\left[ \Pr \{X_2 = x_2|X_1 = x_1\} \right]_{\mathcal{X}^{1}|\mathcal{X}},$$

and its initial state distribution $P_{X_1}(x)$. A distribution $\pi(x)$ on $\mathcal{X}$ is said to be a stationary distribution for a homogeneous first-order Markov chain, if for every $y \in \mathcal{X}$,

$$\pi(y) = \sum_{x \in \mathcal{X}} \pi(x) \Pr\{X_2 = y|X_1 = x\}.$$  

If the initial state distribution is equal to the stationary distribution, then the homogeneous first-order Markov chain is a stationary process.
The general relations among i.i.d. sources, Markov sources, stationary sources and ergodic sources are depicted in Figure B.1.

**B.6 Convergence of sequences of random variables**

In this section, we will discuss modes in which a random process $X_1, X_2, \ldots$ converges to a limiting random variable $X$. Recall that a random variable is a real-valued function from $\Omega$ to $\mathbb{R}$, where $\Omega$ the sample space of the probability space over which the random variable is defined. So the following two expressions will be used interchangeably.

$$X_1(\omega), X_2(\omega), X_3(\omega), \ldots \equiv X_1, X_2, X_3, \ldots$$

for $\omega \in \Omega$. Note that the random variables in a random process are defined over the same probability space $(\Omega, \mathcal{F}, P)$.

**Definition B.5 (Convergence modes for random sequences)**

1. *Point-wise convergence on $\Omega$.*

---

8Although such mode of convergence is not used in probability theory, we introduce it herein to contrast it with the almost sure convergence mode (see Example B.6).
\( \{X_n\}_{n=1}^{\infty} \) is said to converge to \( X \) pointwisely on \( \Omega \) if
\[
\lim_{n \to \infty} X_n(\omega) = X(\omega) \quad \text{for all } \omega \in \Omega.
\]
This is usually denoted by \( X_n \overset{p.w.}{\to} X \).

2. **Almost sure convergence** or **convergence with probability 1**.
\( \{X_n\}_{n=1}^{\infty} \) is said to converge to \( X \) with probability 1, if
\[
P\{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\} = 1.
\]
This is usually denoted by \( X_n \overset{a.s.}{\to} X \).

3. **Convergence in probability**.
\( \{X_n\}_{n=1}^{\infty} \) is said to converge to \( X \) in probability, if for any \( \varepsilon > 0 \),
\[
\lim_{n \to \infty} P\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\} = \lim_{n \to \infty} \Pr\{|X_n - X| > \varepsilon\} = 0.
\]
This is usually denoted by \( X_n \overset{p}{\to} X \).

4. **Convergence in \( r \)th mean**.
\( \{X_n\}_{n=1}^{\infty} \) is said to converge to \( X \) in \( r \)th mean, if
\[
\lim_{n \to \infty} E[|X - X_n|^r] = 0.
\]
This is usually denoted by \( X_n \overset{L_r}{\to} X \).

5. **Convergence in distribution**.
\( \{X_n\}_{n=1}^{\infty} \) is said to converge to \( X \) in distribution, if
\[
\lim_{n \to \infty} F_{X_n}(x) = F_X(x),
\]
for every continuity point of \( F(x) \), where
\[
F_{X_n}(x) \triangleq \Pr\{X_n \leq x\} \quad \text{and} \quad F_X(x) \triangleq \Pr\{X \leq x\}.
\]
This is usually denoted by \( X_n \overset{d}{\to} X \).

An example that facilitates the understanding of pointwise convergence and almost sure convergence is as follows.
Example B.6 Give a probability space \((\Omega, 2^\Omega, P)\), where \(\Omega = \{0, 1, 2, 3\}\), and \(P(0) = P(1) = P(2) = 1/3\) and \(P(3) = 0\). Define a random variable as \(X_n(\omega) = \omega/n\). Then
\[
\Pr\{X_n = 0\} = \Pr\left\{X_n = \frac{1}{n}\right\} = \Pr\left\{X_n = \frac{2}{n}\right\} = \frac{1}{3}.
\]
It is clear that for every \(\omega\) in \(\Omega\), \(X_n(\omega)\) converges to \(X(\omega)\), where \(X(\omega) = 0\) for every \(\omega \in \Omega\); so
\[X_n \xrightarrow{p.w.} X.\]
Now let \(\tilde{X}(\omega) = 0\) for \(\omega = 0, 1, 2\) and \(\tilde{X}(\omega) = 1\) for \(\omega = 3\). Then both of the following statements are true:
\[X_n \xrightarrow{a.s.} X \quad \text{and} \quad X_n \xrightarrow{a.s.} \tilde{X},\]
since
\[
\Pr\left\{\lim_{n \to \infty} X_n = \tilde{X}\right\} = \sum_{\omega=0}^{3} P(\omega) \cdot \mathbf{1}\left\{\lim_{n \to \infty} X_n(\omega) = \tilde{X}(\omega)\right\} = 1,
\]
where \(\mathbf{1}\{\cdot\}\) represents the set indicator function. However, \(X_n\) does not converge to \(\tilde{X}\) pointwise because \(\lim_{n \to \infty} X_n(3) \neq \tilde{X}(3)\). So to speak, pointwise convergence requires “equality” even for those samples without probability mass, for which almost surely convergence does not take into consideration.

Observation B.7 (Uniqueness of convergence)

1. If \(X_n \xrightarrow{p.w.} X\) and \(X_n \xrightarrow{p.w.} Y\), then \(X = Y\) pointwisely. I.e., \((\forall \omega \in \Omega)\)
\(X(\omega) = Y(\omega)\).

2. If \(X_n \xrightarrow{a.s.} X\) and \(X_n \xrightarrow{a.s.} Y\) \((\text{or } X_n \xrightarrow{p} X \text{ and } X_n \xrightarrow{p} Y)\) \((\text{or } X_n \xrightarrow{L^r} X \text{ and } X_n \xrightarrow{L^r} Y)\), then \(X = Y\) with probability 1. I.e., \(\Pr\{X = Y\} = 1\).

3. \(X_n \xrightarrow{d} X\) and \(X_n \xrightarrow{d} Y\), then \(F_X(x) = F_Y(x)\) for all \(x\).

For ease of understanding, the relations of the five modes of convergence can be depicted as follows. As usual, a double arrow denotes implication.

\[
\begin{align*}
X_n & \xrightarrow{p.w.} X \\
\Downarrow & \\
X_n & \xrightarrow{a.s.} X \quad \text{Thm. B.9} \\
\Downarrow & \\
X_n & \xrightarrow{L^r} X \quad (r \geq 1) \quad \text{Thm. B.8} \\
\Downarrow & \\
X_n & \xrightarrow{p} X \\
\Downarrow & \\
X_n & \xrightarrow{d} X
\end{align*}
\]
There are some other relations among these five convergence modes that are also depicted in the above graph (via the dotted line); they are stated below.

**Theorem B.8 (Monotone convergence theorem [6])**

\[ X_n \xrightarrow{a.s.} X, \quad (\forall n) Y \leq X_n \leq X_{n+1}, \quad \text{and} \quad E[|Y|] < \infty \quad \Rightarrow \quad X_n \xrightarrow{L_1} X \]

\[ \Rightarrow \quad E[X_n] \rightarrow E[X]. \]

**Theorem B.9 (Dominated convergence theorem [6])**

\[ X_n \xrightarrow{a.s.} X, \quad (\forall n) |X_n| \leq Y, \quad \text{and} \quad E[|Y|] < \infty \quad \Rightarrow \quad X_n \xrightarrow{L_1} X \]

\[ \Rightarrow \quad E[X_n] \rightarrow E[X]. \]

The implication of \( X_n \xrightarrow{L_1} X \) to \( E[X_n] \rightarrow E[X] \) can be easily seen from

\[ |E[X_n] - E[X]| = |E[X_n - X]| \leq E[|X_n - X|]. \]

**B.7 Ergodicity and laws of large numbers**

**B.7.1 Laws of large numbers**

Consider a random process \( X_1, X_2, \ldots \) with common marginal mean \( \mu \). Suppose that we wish to estimate \( \mu \) on the basis of the observed sequence \( x_1, x_2, x_3, \ldots \). The weak and strong laws of large numbers ensure that such inference is possible (with reasonable accuracy), provided that the dependencies between \( X_n \)'s are suitably restricted: e.g., the weak law is valid for uncorrelated \( X_n \)'s, while the strong law is valid for independent \( X_n \)'s. Since independence is a more restrictive condition than the absence of correlation, one expects the strong law to be more powerful than the weak law. This is indeed the case, as the weak law states that the sample average

\[ \frac{X_1 + \cdots + X_n}{n} \]

converges to \( \mu \) in probability, while the strong law asserts that this convergence takes place with probability 1.

The following two inequalities will be useful in the discussion of this subject.

**Lemma B.10 (Markov's inequality)** For any integer \( k > 0 \), real number \( \alpha > 0 \) and any random variable \( X \),

\[ \Pr[|X| \geq \alpha] \leq \frac{1}{\alpha^k} E[|X|^k]. \]
Proof: Let \( F_X(\cdot) \) be the cdf of random variable \( X \). Then,

\[
E[|X|^k] = \int_{-\infty}^{\infty} |x|^k dF_X(x)
\]

\[
\geq \int_{\{x \in \mathbb{R} : |x| \geq \alpha\}} |x|^k dF_X(x)
\]

\[
\geq \alpha^k F_X(\alpha)
\]

\[
= \alpha^k \int_{\{x \in \mathbb{R} : |x| \geq \alpha\}} dF_X(x)
\]

\[
= \alpha^k \Pr[|X| \geq \alpha].
\]

Equality holds iff

\[
\int_{\{x \in \mathbb{R} : |x| < \alpha\}} |x|^k dF_X(x) = 0 \quad \text{and} \quad \int_{\{x \in \mathbb{R} : |x| > \alpha\}} |x|^k dF_X(x) = 0,
\]

namely,

\[
\Pr[X = 0] + \Pr[|X| = \alpha] = 1.
\]

\[\square\]

In the proof of Markov’s inequality, we use the general representation for integration with respect to a (cumulative) distribution function \( F_X(\cdot) \), i.e.,

\[
\int_A \cdot dF_X(x),
\]

(B.7.1)

which is named the \textit{Lebesgue-Stieltjes integral}. Such a representation can be applied for both discrete and continuous supports as well as the case that the probability density function does not exist. We use this notational convention to remove the burden of differentiating discrete random variables from continuous ones.

Lemma B.11 (Chebyshev’s inequality) For any random variable \( X \) and real number \( \alpha > 0 \),

\[
\Pr[|X - E[X]| \geq \alpha] \leq \frac{1}{\alpha^2} \text{Var}[X].
\]

Proof: By Markov’s inequality with \( k = 2 \), we have:

\[
\Pr[|X - E[X]| \geq \alpha] \leq \frac{1}{\alpha^2} E[|X - E[X]|^2].
\]

Equality holds iff

\[
\Pr[|X - E[X]| = 0] + \Pr[|X - E[X]| = \alpha] = 1,
\]
equivalently, there exists \( p \in \[0, 1\] \) such that
\[
\Pr \left[ X = E[X] + \alpha \right] = \Pr \left[ X = E[X] - \alpha \right] = p \quad \text{and} \quad \Pr[X = E[X]] = 1 - 2p.
\]

In the proofs of the above two lemmas, we also provide the condition under which equality holds. These conditions indicate that equality usually cannot be fulfilled. Hence in most cases, the two inequalities are strict.

**Theorem B.12 (Weak law of large numbers)** Let \( \{X_n\}_{n=1}^{\infty} \) be a sequence of uncorrelated random variables with common mean \( E[X_i] = \mu \). If the variables also have common variance, or more generally,
\[
\lim_{n \to \infty} \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[X_i] = 0, \quad \text{(equivalently, } \frac{X_1 + \cdots + X_n}{n} \xrightarrow{L^2} \mu \text{) }
\]
then the sample average
\[
\frac{X_1 + \cdots + X_n}{n}
\]
converges to the mean \( \mu \) in probability.

**Proof:** By Chebyshev’s inequality,
\[
\Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| \geq \varepsilon \right\} \leq \frac{1}{n^2 \varepsilon^2} \sum_{i=1}^{n} \text{Var}[X_i].
\]

Note that the right-hand side of the above Chebyshev’s inequality is just the second moment of the difference between the \( n \)-sample average and the mean \( \mu \). Thus the variance constraint is equivalent to the statement that \( X_n \xrightarrow{L^2} \mu \) implies \( X_n \xrightarrow{p} \mu \).

**Theorem B.13 (Kolmogorov’s strong law of large numbers)** Let \( \{X_n\}_{n=1}^{\infty} \) be an independent sequence of random variables with common mean \( E[X_n] = \mu \). If either
1. \( X_n \)'s are identically distributed; or
2. \( X_n \)'s are square-integrable\(^9\) with
\[
\sum_{i=1}^{\infty} \frac{\text{Var}[X_i]}{i^2} < \infty,
\]

\(^9\)A random variable \( X \) is said to be **square-integrable** if \( E[|X|^2] < \infty \).
then
\[ \frac{X_1 + \cdots + X_n}{n} \xrightarrow{a.s.} \mu. \]

Note that the above i.i.d. assumption does not exclude the possibility of \( \mu = \infty \) (or \( \mu = -\infty \)), in which case the sample average converges to \( \infty \) (or \( -\infty \)) with probability 1. Also note that there are cases of independent sequences to which the weak law applies, but the strong law does not. This is due to the fact that
\[
\sum_{i=1}^{n} \frac{\text{Var}[X_i]}{i^2} \geq \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[X_i].
\]

The final remark is that Kolmogorov’s strong law of large number can be extended to a function of an independent sequence of random variables:
\[
\frac{g(X_1) + \cdots + g(X_n)}{n} \xrightarrow{a.s.} E[g(X_1)].
\]
But such extension cannot be applied to the weak law of large number, since \( g(Y_i) \) and \( g(Y_j) \) can be correlated even if \( Y_i \) and \( Y_i \) are not.

### B.7.2 Ergodicity and law of large numbers

After the introduction of Kolmogorov’s strong law of large numbers, one may find that the pointwise ergodic theorem (Theorem B.3) actually indicates a similar result. In fact, the pointwise ergodic theorem can be viewed as another version of strong law of large numbers, which states that for stationary and ergodic processes, time averages converge with probability 1 to the ensemble expectation.

The notion of ergodicity is often misinterpreted, since the definition is not very intuitive. Some engineering texts may provide a definition that a stationary process satisfying the ergodic theorem is also ergodic.\(^\text{10}\) However, the ergodic

\(^\text{10}\)Here is one example. A stationary random process \( \{X_n\}_{n=1}^{\infty} \) is called ergodic if for arbitrary integer \( k \) and function \( f(\cdot) \) on \( \mathbb{X}^k \) of finite mean,
\[
\frac{1}{n} \sum_{i=1}^{n} f(X_{i+1}, \ldots, X_{i+k}) \xrightarrow{a.s.} E[f(X_1, \ldots, X_k)].
\]

As a result of this definition, a stationary ergodic source is the most general dependent random process for which the strong law of large numbers holds. This definition somehow implies that if a process is not stationary-ergodic, then the strong law of large numbers is violated (or the time average does not converge with probability 1 to its ensemble expectation). But this is not true. One can weaken the conditions of stationarity and ergodicity from its original mathematical definitions to asymptotic stationarity and ergodicity, and still make the strong law of large numbers hold. (Cf. the last remark in this section and also Figure B.2)
Theorem is indeed a consequence of the original mathematical definition of ergodicity in terms of the shift-invariant property (see Section B.5 and the discussion in [21, pp. 174-175]).

Ergodicity defined through
shift-invariance
property

Figure B.2: Relation of ergodic random processes respectively defined through time-shift invariance and ergodic theorem.

Let us try to clarify the notion of ergodicity by the following remarks.

- The concept of ergodicity does not require stationarity. In other words, a non-stationary process can be ergodic.

- Many perfectly good models of physical processes are not ergodic, yet they have a form of law of large numbers. In other words, non-ergodic processes can be perfectly good and useful models.

- There is no finite-dimensional equivalent definition of ergodicity as there is for stationarity. This fact makes it more difficult to describe and interpret ergodicity.

- I.i.d. processes are ergodic; hence, ergodicity can be thought of as a (kind of) generalization of i.i.d.

- As mentioned earlier, stationarity and ergodicity imply the time average converges with probability 1 to the ensemble mean. Now if a process is stationary but not ergodic, then the time average still converges, but possibly not to the ensemble mean.

For example, let \( \{A_n\}_{n=-\infty}^{\infty} \) and \( \{B_n\}_{n=-\infty}^{\infty} \) be two i.i.d. binary 0-1 random variables with \( \Pr\{A_n = 0\} = \Pr\{B_n = 1\} = 1/4 \). Suppose that \( X_n = A_n \) if \( U = 1 \), and \( X_n = B_n \) if \( U = 0 \), where \( U \) is equiprobable binary
random variable, and \( \{A_n\}_{n=1}^\infty \), \( \{B_n\}_{n=1}^\infty \), and \( U \) are independent. Then \( \{X_n\}_{n=1}^\infty \) is stationary. Is the process ergodic? The answer is negative. If the stationary process were ergodic, then from the pointwise ergodic theorem (Theorem B.3), its relative frequency would converge to

\[
\Pr(X_n = 1) = \Pr(U = 1) \Pr(X_n = 1|U = 1) + \Pr(U = 0) \Pr(X_n = 1|U = 0)
\]

\[
= \Pr(U = 1) \Pr(A_n = 1) + \Pr(U = 0) \Pr(B_n = 1) = \frac{1}{2}.
\]

However, one should observe that the outputs of \((X_1, \ldots, X_n)\) form a Bernoulli process with relative frequency of 1’s being either \(\frac{3}{4}\) or \(\frac{1}{4}\), depending on the value of \(U\). Therefore,

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_n \xrightarrow{a.s.} Y,
\]

where \(\Pr(Y = 1/4) = \Pr(Y = 3/4) = 1/2\), which contradicts to the ergodic theorem.

From the above example, the pointwise ergodic theorem can actually be made useful in such a stationary but non-ergodic case, since the estimate with stationary ergodic process (either \(\{A_n\}_{n=-\infty}^\infty \) or \(\{B_n\}_{n=-\infty}^\infty \)) is actually being observed by measuring the relative frequency (3/4 or 1/4). This renders a surprising fundamental result of random processes—ergodic decomposition theorem: under fairly general assumptions, any (not necessarily ergodic) stationary process is in fact a mixture of stationary ergodic processes, and hence one always observes a stationary ergodic outcome. As in the above example, one always observe either \(A_1, A_2, A_3, \ldots\) or \(B_1, B_2, B_3, \ldots\), depending on the value of \(U\), for which both sequences are stationary ergodic (i.e., the time-stationary observation \(X_n\) satisfies \(X_n = U \cdot A_n + (1 - U) \cdot B_n\)).

- The previous remark implies that ergodicity is not required for the strong law of large numbers to be useful. The next question is whether or not stationarity is required. Again the answer is negative. In fact, the main concern of the law of large numbers is the convergence of sample averages to its ensemble expectation. It should be reasonable to expect that random processes could exhibit transient behavior that violates the stationarity definition, yet the sample average still converges. One can then introduce the notion of asymptotically stationary to achieve the law of large numbers. For example, a finite-alphabet time-invariant (but not necessarily stationary) irreducible Markov chain satisfies the law of large numbers.
Accordingly, one should not take the notions of stationarity and ergodicity too seriously (if the main concern is the law of large numbers) since they can be significantly weakened and still have laws of large numbers holding (i.e., time averages and relative frequencies have desired and well-defined limits).

### B.8 Central limit theorem

**Theorem B.14 (Central limit theorem)** If \( \{X_n\}_{n=1}^{\infty} \) is a sequence of i.i.d. random variables with finite common marginal mean \( \mu \) and variance \( \sigma^2 \), then

\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} Z \sim \mathcal{N}(0, \sigma^2),
\]

where the convergence is in distribution (as \( n \to \infty \)) and \( Z \sim \mathcal{N}(0, \sigma^2) \) is a Gaussian distributed random variable with mean 0 and variance \( \sigma^2 \).

### B.9 Convexity, concavity and Jensen’s inequality

Jensen’s inequality provides a useful bound for the expectation of convex (or concave) functions.

**Definition B.15 (Convexity)** Consider a convex set\(^{11} \mathcal{O} \in \mathbb{R}^m \), where \( m \) is a fixed positive integer. Then a function \( f : \mathcal{O} \to \mathbb{R} \) is said to be convex over \( \mathcal{O} \) if for every \( \underline{x}, \underline{y} \) in \( \mathcal{O} \) and \( 0 \leq \lambda \leq 1 \),

\[
f(\lambda \underline{x} + (1 - \lambda) \underline{y}) \leq \lambda f(\underline{x}) + (1 - \lambda) f(\underline{y}).
\]

Furthermore, a function \( f \) is said to be strictly convex if equality holds only when \( \lambda = 0 \) or \( \lambda = 1 \).

Note that different from the usual notations \( x^n = (x_1, x_2, \ldots, x_n) \) or \( \underline{x} = (x_1, x_2, \ldots) \) throughout this book, we use \( \underline{x} \) to denote a column vector in this section.

**Definition B.16 (Concavity)** A function \( f \) is concave if \( -f \) is convex.

---

\(^{11}\)A set \( \mathcal{O} \in \mathbb{R}^m \) is said to be convex if for every \( \underline{x} = (x_1, x_2, \ldots, x_m)^T \) and \( \underline{y} = (y_1, y_2, \ldots, y_m)^T \) in \( \mathcal{O} \) (where \( T \) denotes transposition), and every \( 0 \leq \lambda \leq 1 \), \( \lambda \underline{x} + (1 - \lambda) \underline{y} \in \mathcal{O} \); in other words, the “convex combination” of any two “points” \( \underline{x} \) and \( \underline{y} \) in \( \mathcal{O} \) also belongs to \( \mathcal{O} \).
Note that when $O = (a, b)$ is an interval in $\mathbb{R}$ and function $f : O \rightarrow \mathbb{R}$ has a non-negative (resp. positive) second derivative over $O$, then the function is convex (resp. strictly convex). This can be shown via the Taylor series expansion of the function.

**Theorem B.17 (Jensen’s inequality)** If $f : O \rightarrow \mathbb{R}$ is convex over a convex set $O \subset \mathbb{R}^m$, and $\mathcal{X} = (X_1, X_2, \ldots, X_m)^T$ is an $m$-dimensional random vector with alphabet $\mathcal{X} \subset O$, then

$$E[f(\mathcal{X})] \geq f(E[\mathcal{X}]).$$

Moreover, if $f$ is strictly convex, then equality in the above inequality immediately implies $\mathcal{X} = E[\mathcal{X}]$ with probability 1.

**Note:** $O$ is a convex set; hence, $\mathcal{X} \subset O$ implies $E[\mathcal{X}] \in O$. This guarantees that $f(E[\mathcal{X}])$ is defined. Similarly, if $f$ is concave, then

$$E[f(\mathcal{X})] \leq f(E[\mathcal{X}]).$$

Furthermore, if $f$ is strictly concave, then equality in the above inequality immediately implies that $\mathcal{X} = E[\mathcal{X}]$ with probability 1.

**Proof:** Let $y = a^T x + b$ be a “support hyperplane” for $f$ with “slope” vector $a^T$ and affine parameter $b$ that passes through the point $(E[\mathcal{X}], f(E[\mathcal{X}]))$, where a support hyperplane\(^{12}\) for function $f$ at $x'$ is by definition a hyperplane passing through the point $(x', f(x'))$ and lying entirely below the graph of $f$ (see Figure B.9 for an illustration of a support line for a convex function over $\mathbb{R}$).

Thus,

$$\forall x \in \mathcal{X}, \quad a^T x + b \leq f(x).$$

By taking the expectation of both sides, we obtain

$$a^T E[\mathcal{X}] + b \leq E[f(\mathcal{X})],$$

but we know that $a^T E[\mathcal{X}] + b = f(E[\mathcal{X}])$. Consequently,

$$f(E[\mathcal{X}]) \leq E[f(\mathcal{X})].$$

\(\square\)

\(^{12}\) A hyperplane $y = a^T x + b$ is said to be a support hyperplane for a function $f$ with “slope” vector $a^T \in \mathbb{R}^m$ and affine parameter $b \in \mathbb{R}$ if among all hyperplanes of the same slope vector $a^T$, it is the largest one satisfying $a^T x + b \leq f(x)$ for every $x \in O$. A support hyperplane may not necessarily be made to pass through the desired point $(x', f(x'))$. Here, since we only consider convex functions, the validity of the support hyperplane passing $(x', f(x'))$ is therefore guaranteed. Note that when $x$ is one-dimensional (i.e., $m = 1$), a support hyperplane is simply referred to as a support line.
B.10 Lagrange multipliers technique and Karush-Kuhn-Tucker (KKT) condition

Optimization of a function \( f(\mathbf{x}) \) over \( \mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X} \subseteq \mathbb{R}^n \) subject to some inequality constraints \( g_i(\mathbf{x}) \leq 0 \) for \( 1 \leq i \leq m \) and equality constraints \( h_j(\mathbf{x}) = 0 \) for \( 1 \leq j \leq \ell \) is a central technique to problems in information theory. An immediate example is to maximize the mutual information subject to an “inequality” power constraint and an “equality” probability unity-sum constraint in order to determine the channel capacity. We can formulate such an optimization problem [11, Eq. (5.1)] mathematically as\(^{13}\)

\[
\min_{\mathbf{x} \in \mathcal{Q}} f(\mathbf{x}), \tag{B.10.1}
\]

where

\[
\mathcal{Q} \triangleq \{ \mathbf{x} \in \mathcal{X}: g_i(\mathbf{x}) \leq 0 \text{ for } 1 \leq i \leq m \text{ and } h_j(\mathbf{x}) = 0 \text{ for } 1 \leq j \leq \ell \}.
\]

In most cases, solving the constrained optimization problem defined in (B.10.1) is hard. Instead, one may introduce a dual optimization problem without constraints as

\[
L(\boldsymbol{\lambda}, \boldsymbol{\nu}) \triangleq \min_{\mathbf{x} \in \mathcal{X}} \left( f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{\ell} \nu_j h_j(\mathbf{x}) \right). \tag{B.10.2}
\]

\(^{13}\)Since maximization of \( f(\cdot) \) is equivalent to minimization of \(-f(\cdot)\), it suffices to discuss the KKT condition for the minimization problem defined in (B.10.1).
In the literature, $\lambda = (\lambda_1, \ldots, \lambda_m)$ and $\nu = (\nu_1, \ldots, \nu_\ell)$ are usually referred to as Lagrange multipliers, and $L(\lambda, \nu)$ is called the Lagrange dual function. Note that $L(\lambda, \nu)$ is a concave function of $\lambda$ and $\nu$ since it is the minimization of affine functions of $\lambda$ and $\nu$.

It can be verified that when $\lambda_i \geq 0$ for $1 \leq i \leq m$,

$$L(\lambda, \nu) \leq \min_{x \in \mathcal{Q}} \left( f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{\ell} \nu_j h_j(x) \right) \leq \min_{x \in \mathcal{Q}} f(x). \quad \text{(B.10.3)}$$

We are however interested in when the above inequality becomes equality (i.e., when the so-called **strong duality** holds) because if there exist non-negative $\tilde{\lambda}$ and $\tilde{\nu}$ that equate (B.10.3), then

$$f(x^*) = \min_{x \in \mathcal{X}} f(x) = L(\tilde{\lambda}, \tilde{\nu})$$

$$= \min_{x \in \mathcal{X}} \left[ f(x) + \sum_{i=1}^{m} \tilde{\lambda}_i g_i(x) + \sum_{j=1}^{\ell} \tilde{\nu}_j h_j(x) \right]$$

$$\leq f(x^*) + \sum_{i=1}^{m} \tilde{\lambda}_i g_i(x^*) + \sum_{j=1}^{\ell} \tilde{\nu}_j h_j(x^*)$$

$$\leq f(x^*)$$ \quad \text{(B.10.4)}

where (B.10.4) follows because the minimizer $x^*$ of (B.10.1) lies in $\mathcal{Q}$. Hence, if the strong duality holds, the same $x^*$ achieves both $\min_{x \in \mathcal{Q}} f(x)$ and $L(\tilde{\lambda}, \tilde{\nu})$, and $\tilde{\lambda}_i g_i(x^*) = 0$ for $1 \leq i \leq m$.\(^{14}\)

The strong duality does not in general hold. A situation that guarantees the validity of the strong duality has been determined by William Karush [27], and separately Harold W. Kuhn and Albert W. Tucker [30]. In particular, when $f(\cdot)$ and $\{g_i(\cdot)\}_{i=1}^{m}$ are both convex, and $\{h_j(\cdot)\}_{j=1}^{\ell}$ are affine, and these functions are all differentiable, they found that the strong duality holds if, and only if, the KKT condition is satisfied [11, p. 258].

**Definition B.18 (Karush-Kuhn-Tucker (KKT) condition)** Point $x = (x_1, \ldots, x_n)$ and multipliers $\lambda = (\lambda_1, \ldots, \lambda_m)$ and $\nu = (\nu_1, \ldots, \nu_\ell)$ are said to satisfy the KKT condition if

\[
\begin{align*}
  g_i(x) &\leq 0, \quad \lambda_i \geq 0, \quad \lambda_i g_i(x) = 0 \quad i = 1, \ldots, m \\
  h_j(x) &\leq 0 \quad j = 1, \ldots, \ell \\
  \frac{\partial f}{\partial x_k}(x) + \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_k}(x) + \sum_{j=1}^{\ell} \nu_j \frac{\partial h_j}{\partial x_k}(x) &\leq 0 \quad k = 1, \ldots, n
\end{align*}
\]

\(^{14}\)Equating (B.10.4) implies $\sum_{i=1}^{m} \tilde{\lambda}_i g_i(x^*) = 0$. It can then be easily verified from $\tilde{\lambda}_i g_i(x^*) \leq 0$ for every $1 \leq i \leq m$ that $\tilde{\lambda}_i g_i(x^*) = 0$ for $1 \leq i \leq m$. 225
Note that when \( f(\cdot) \) and constraints \( \{g_i(\cdot)\}_{i=1}^{m} \) and \( \{h_j(\cdot)\}_{j=1}^{\ell} \) are arbitrary functions, the KKT condition is only a necessary condition for the validity of the strong duality. In other words, for a non-convex optimization, we can only claim that if the strong duality holds, then the KKT condition is satisfied but not vice versa.

A case that is particularly useful in information theory is when \( x \) is restricted to be a probability distribution. In such case, apart from other problem-specific constraints, we have additionally \( n \) inequality constraints \( g_{m+i}(x) = -x_i \leq 0 \) for \( 1 \leq i \leq n \) and one equality constraint \( h_{\ell+1}(x) = \sum_{k=1}^{n} x_k - 1 = 0 \). Hence, the KKT condition becomes

\[
\begin{aligned}
g_i(x) &\leq 0, \quad \lambda_i \geq 0, \quad \lambda_i g_i(x) = 0 \quad &i = 1, \ldots, m \\
g_{m+k}(x) &= -x_k \leq 0, \quad \lambda_{m+k} \geq 0, \quad \lambda_{m+k} x_k = 0 \quad &k = 1, \ldots, n \\
h_j(x) &= 0 \quad &j = 1, \ldots, \ell \\
h_{\ell+1}(x) &= \sum_{k=1}^{n} x_k - 1 = 0 \\
\frac{\partial f}{\partial x_k}(x) + \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_k}(x) - \lambda_{m+k} + \sum_{j=1}^{\ell} \nu_j \frac{\partial h_j}{\partial x_k}(x) + \nu_{\ell+1} &= 0 \quad &k = 1, \ldots, n 
\end{aligned}
\]

From \( \lambda_{m+k} \geq 0 \) and \( \lambda_{m+k} x_k = 0 \), we can obtain the well-known relation below.

\[
\lambda_{m+k} = \begin{cases} 
\frac{\partial f}{\partial x_k}(x) + \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_k}(x) + \sum_{j=1}^{\ell} \nu_j \frac{\partial h_j}{\partial x_k}(x) + \nu_{\ell+1} = 0 & \text{if } x_k > 0 \\
\frac{\partial f}{\partial x_k}(x) + \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_k}(x) + \sum_{j=1}^{\ell} \nu_j \frac{\partial h_j}{\partial x_k}(x) + \nu_{\ell+1} \geq 0 & \text{if } x_k = 0.
\end{cases}
\]

The above relation is the mostly seen form of the KKT condition when it is used in problems of information theory.

**Example B.19** Suppose for non-negative \( \{q_{i,j}\}_{1 \leq i \leq n, 1 \leq j \leq n'} \) with \( \sum_{j=1}^{n'} q_{i,j} = 1 \),

\[
\begin{aligned}
f(x) &= - \sum_{i=1}^{n} \sum_{j=1}^{n'} x_i q_{i,j} \log \frac{q_{i,j}}{\sum_{i'=1}^{n} x_{i'} q_{i',j}} \\
g_i(x) &= -x_i \leq 0 \quad &i = 1, \ldots, n \\
h(x) &= \sum_{i=1}^{n} x_i - 1 = 0
\end{aligned}
\]

Then the KKT condition implies

\[
\begin{aligned}
x_i &\geq 0, \quad \lambda_i \geq 0, \quad \lambda_i x_i = 0 \quad &i = 1, \ldots, n \\
\sum_{i=1}^{n} x_i &= 1 \\
- \sum_{j=1}^{n'} q_{k,j} \log \frac{q_{k,j}}{\sum_{i'=1}^{n} x_{i'} q_{i',j}} + 1 - \lambda_k + \nu &= 0 \quad &k = 1, \ldots, n
\end{aligned}
\]

226
which further implies that

\[
\lambda_k = \begin{cases}
- \sum_{j=1}^{n'} q_{k,j} \log \left( \frac{q_{k,j}}{\sum_{l'=1}^{n} x_{l'} q_{l',j}} \right) + 1 + \nu = 0 & x_k > 0 \\
- \sum_{j=1}^{n'} q_{k,j} \log \left( \frac{q_{k,j}}{\sum_{l'=1}^{n} x_{l'} q_{l',j}} \right) + 1 + \nu \geq 0 & x_k = 0
\end{cases}
\]

By this, the input distributions that achieve the channel capacities of some channels such as BSC and BEC can be identified.

The next example shows the analogy of determining the channel capacity to the problem of optimal power allocation.

**Example B.20 (Water-filling)** Suppose with \( \sigma_i^2 > 0 \) for \( 1 \leq i \leq n \) and \( P > 0 \),

\[
\begin{align*}
f(x) &= - \sum_{i=1}^{n} \log \left( 1 + \frac{x_i}{\sigma_i^2} \right) \\
g_i(x) &= -x_i \leq 0 & i = 1, \ldots, n \\
h(x) &= \sum_{i=1}^{n} x_i - P = 0
\end{align*}
\]

Then the KKT condition implies

\[
\begin{align*}
x_i &\geq 0, \quad \lambda_i \geq 0, \quad \lambda_i x_i = 0 & i = 1, \ldots, n \\
\sum_{i=1}^{n} x_i &= P \\
- \frac{1}{\sigma_i^2 + x_i} - \lambda_i + \nu &= 0 & i = 1, \ldots, n
\end{align*}
\]

which further implies that

\[
\lambda_i = \begin{cases}
- \frac{1}{\sigma_i^2 + x_i} + \nu = 0 & x_i > 0 \\
- \frac{1}{\sigma_i^2 + x_i} + \nu \geq 0 & x_i = 0
\end{cases}
\]

equivalently \( x_i = \begin{cases}
\frac{1}{\nu} - \sigma_i^2 & \sigma_i^2 < \frac{1}{\nu} \\
0 & \sigma_i^2 \geq \frac{1}{\nu}
\end{cases} \)

This then gives the water-filling solution for the power allocation over continuous-input additive white Gaussian noise channels.
Bibliography


Index

δ-typical set, 51

AEP, 51, 146, 177
asymptotic equipartition property, 51
average probability of error, 99

block code, 49, 50
bound
Sharp, 24
Tight, 24

capacity
non-Gaussian channels, 164
parallel additive Gaussian channels, 158, 163

capacity-cost function
concavity, 148

central limit theorem, 222
Cessaro-mean theorem, 59

channel
quasi-symmetric, 114
symmetric, 112
weakly symmetric, 112

Chebyshev’s inequality, 217
code rate for data compression, 50
communication system, 3
general description, 2
concavity, 222
convexity, 222

data processing inequality, 20
for divergence, 28
system interpretation, 21

data processing lemma, see data processing inequality
differential entropy
definition, 132
Gaussian source, 134
operation characteristic, 130
uniform source, 134
discrete memoryless channel, 93
discrete memoryless source, 49
distortion measure, 171
additive, 174
Hamming distortion measure, 173
maximum, 174
squared error distortion, 173
distortion typical set, 176
divergence
additivity for independence, 36
bound from variational distance, 31
conditional, 34
convexity, 36
non-negativity, 27

DMC, 93
dominated convergence theorem, 216

data processing lemma, see data processing inequality
differential entropy

data processing inequality, 20
for divergence, 28
system interpretation, 21

data processing lemma, see data processing inequality
differential entropy
definition, 132
Gaussian source, 134
operation characteristic, 130
uniform source, 134
discrete memoryless channel, 93
discrete memoryless source, 49
distortion measure, 171
additive, 174
Hamming distortion measure, 173
maximum, 174
squared error distortion, 173
distortion typical set, 176
divergence
additivity for independence, 36
bound from variational distance, 31
conditional, 34
convexity, 36
non-negativity, 27

DMC, 93
dominated convergence theorem, 216

data processing lemma, see data processing inequality
differential entropy

232
Venn diagram, 17
entropy power, 163
entropy rate, 59
entropy stability property, 51
Fano’s inequality, 22
field
$\sigma$-field, 206
fixed-length code, 48
fixed-length data transmission code, 99
fixed-length lossy data compression code, 175
function
concavity, 222
convexity, 222
fundamental inequality (FI), 10
generalized AEP, 61

Huffman code, 74
adaptive, 80
sibling property, 81
hypothesis testing, 39
Bayes criterion, 40
Neyman Pearson criterion, 41
simple hypothesis testing, 39
type I error, 40
type II error, 40

infimum, 196, 196
approximation property for infimum, 197
equivalence to greatest lower bound, 196
monotone property, 197
property for monotone function, 199
set algebraic operations, 198
infinitely often, definition, 202
information-transmission theorem, 189
instantaneous code, 68

Jensen’s inequality, 223
joint AEP, 101
joint source-channel coding theorem, 189
jointly typical set, 100
Karush-Kuhn-Tucker (KKT) condition, 225
Kraft inequality, 66
Lagrange multipliers technique, 224
law of large numbers
strong law, 218, 219
weak law, 218
Lempel-Ziv code, 82
limit infimum, see liminf under sequence
limit supremum, see limsup under sequence
log-sum inequality, 11
Markov’s inequality, 216
maximum, 195, 195
minimum, 196, 197
modes of convergences
almost surely or with probability one, 214
in distribution, 214
in mean, 214
in probability, 214
pointwise, 213
uniqueness of convergence limit, 215
monotone convergence theorem, 216
monotone sequence
convergence, 201
mutual information, 16
bound for memoryless channel, 19
chain rule, 17, 19
conditional, 17, 21
convexity and concavity, 36
for continuous cases, 136
for specific input symbol, 118
Venn diagram, 17

Neyman-Pearson lemma, 41
non-singular code, 64

Pinsker’s inequality, 31
pointwise ergodic theorem, 210
prefix code, 68
probability space, 206
processing of distribution, 28
quasi-symmetric channel, 114
random coding argument, 103
random process, 206
rate-distortion theorem, 179
rate-distortion function, 176
binary source, 185
refinement of distribution, 28
self-information, 5
joint, 12
uniqueness, 6
sequence, 199
\( \text{liminf}, 201, 202, 203 \)
\( \text{limit}, 199, 201 \)
\( \text{limsup}, 201, 203 \)
set
boundedness, 197
Shannon’s channel coding theorem
direct part, 102
Shannon’s channel coding theorem for
the memoryless Gaussian channel, 152
Shannon’s source coding theorem
for DMS, 54
for stationary ergodic sources, 61
Shannon, Claude E., 1
Shannon-Fano-Elias code, 78
Shannon-McMillan theorem, 51
Shannon-McMillan theorem for continuous sources, 147
Shannon-McMillan theorem for pairs, 101
Shannon-McMillan-Breiman theorem, 61
shift-invariant transformation, 208
statistics of processes
  ergodic, 210, 219
  Markov
  \( k \)-th order, 212
  first-order, 212
  memoryless, 210
  stationary
    first-order, 210
    second-order, 210
    strictly sense, 210
    weakly stationary, 210
sufficiently large, definition, 202
supremum, 194, 195
  approximation property for supremum, 195
  completeness axiom, 195
  equivalence to least upper bound, 194
  monotone property, 197
  property for monotone function, 199
  set algebraic operations, 198
symmetric channel, 112
typical set, 147
uniquely decodability, 64
variable-length code, 48
variational distance, 30
  bound from divergence, 31
water-pouring scheme, 160
weakly \( \delta \)-typical set, 51
weakly symmetric channel, 112