

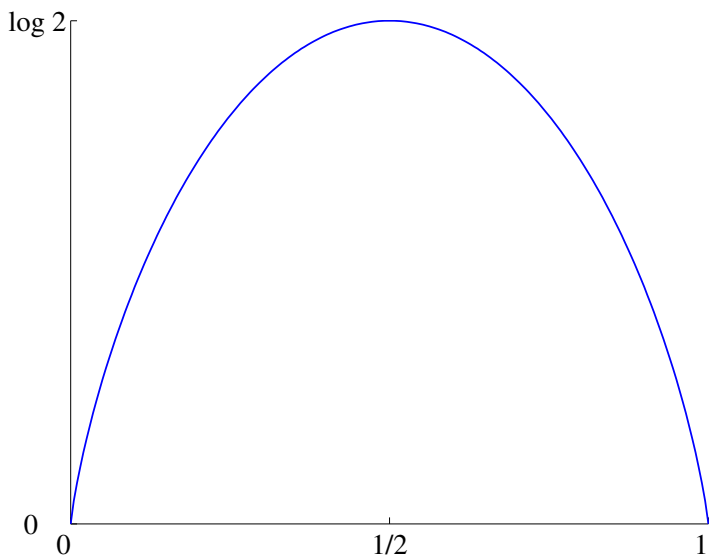
ECE 729 Course Notes

John A. Gubner

*Department of Electrical and Computer Engineering
University of Wisconsin–Madison*

January 23, 2022

with the scribing assistance of the Spring 2010 class



Contents

1	Introduction	1
1.1	Communication Systems	1
1.2	Digital Communication Systems	1
1.2.1	Source Encoders for Binary Data	2
1.2.2	Channel Codes	3
1.2.3	Source Decoder	3
1.2.4	Rate-Distortion Theory	3
1.2.5	Source-Channel Separation	4
1.3	Course Outline	4
	Problems	4
2	The Sampling Theorem and Quantization	6
2.1	The Sampling Theorem	6
2.2	Quantization	7
2.2.1	Vector Quantizers	7
2.2.2	The Lloyd–Max Algorithm	9
	Problems	9
3	Data Compression Using Variable-Length Codes	11
3.1	Introduction	11
3.2	Codewords	11
3.3	Codes	12
3.4	Decodability	13
3.5	Prefix-Free Codes and the Kraft Inequality	15
3.6	Unique Decodability and McMillan’s Theorem	18
3.7	Properties of Lengths that Satisfy the Kraft Inequality	19
3.8	Entropy	20
3.9	Expected Lengths of Codes	23
3.9.1	Shannon Codes under Mismatch [3]	26
3.10	Coding Sequences of Source Letters	26
3.10.1	A Simple Method	26
3.10.2	A Better Method	27
3.11	Optimal Variable-Length Prefix Codes (Huffman, 1952)	27
	Problems	30
4	Data Compression Using Fixed-Length Codes	36
4.1	Introduction	36
4.2	Achievable Rates for Source Codes	36
4.3	Shannon’s Source Coding Theorem	37

4.3.1	The Asymptotic Equipartition Property and Typical Sequences	40
4.3.2	Interpretation of Fixed-Length Codes for Data Compression	40
4.4	Properties of Achievable Rates	41
	Problems	42
5	Introduction to Channel Coding	45
5.1	Channels	45
5.2	Encoders and Decoders	48
5.3	Probabilistic Model	50
5.4	Probability of Error and Achievable Rates	50
5.5	Capacity Regions	53
5.5.1	Small Averages	54
5.5.2	New Codes from Old Codes by Throwing Away Codewords	54
5.5.3	Proof that $\mathcal{C}_a \subset \mathcal{C}_m$	55
	Problems	56
6	Converse Theorems for the DMC	58
6.1	Preliminaries	58
6.1.1	Notation	58
6.1.2	Mutual Information	58
6.1.3	Markov Chains	59
6.1.4	Data Processing Inequalities	61
6.1.5	Fano's Inequality	62
6.2	The Weak Converse	63
6.3	The Strong Converse	64
	Problems	66
7	General Techniques for Studying Channel Codes	69
7.1	More about Decoders	69
7.1.1	Decoders from Subsets	69
7.1.2	Decoder Properties	70
7.2	The Random Coding Argument	71
	Problems	72
8	Forward Theorem for the DMC	73
8.1	Joint Typicality	73
8.2	The Typicality Decoder for the DMC	76
8.3	Forward Theorem for the DMC	76
8.4	Generalization to Countably Infinite Sets	78
9	Evaluation of the DMC Capacity	80
9.1	Weakly Symmetric Channels	80
	Problems	83

10 Codeword Constraints	85
10.1 Capacity Regions	85
10.1.1 Capacity-Cost Function	86
10.2 Weak Converse	87
10.3 A Forward Theorem	88
Problems	92
11 Continuous Channels	93
11.1 Differential Entropy	93
11.2 Divergence for Densities	95
11.3 Divergence for Mixed Random Variables	97
11.4 The Discrete-Time Gaussian Channel	98
11.4.1 Parallel Gaussian Channels	99
11.5 Continuous-Time Channels	101
11.5.1 Projections	101
11.5.2 Approximately-Bandlimited Signaling Waveforms	102
11.5.3 White Noise	102
11.5.4 Gaussian Processes	104
11.5.5 Detection of Signals in Additive White Gaussian Noise	104
11.5.6 The Bandlimited Gaussian Channel	105
Notes	106
Problems	107
12 Rate-Distortion Theory	110
12.1 Introduction	110
12.2 Achievable Rate-Distortion Pairs	111
12.2.1 Properties of the Rate-Distortion Function for a DMS	112
12.3 Properties of $\tilde{R}(D)$	113
12.4 Preparation for the Converse	114
12.5 The Converse Theorem	116
12.6 Preparation for the Forward Theorem	117
12.6.1 The Joint AEP for Rate Distortion	117
12.7 The Forward Theorem	119
12.8 Computation of $\tilde{R}(D)$	121
12.9 The Shannon Lower Bound	124
Problems	126
13 Source Coding Revisited — Discrete Stationary Sources	127
13.1 Stationarity and the Entropy Rate	127
13.2 Ergodic Sources	128
13.3 The Shannon–McMillan–Breiman Theorem	128
13.4 Universal Data Compression	132
13.4.1 The Lempel–Ziv Algorithm	132

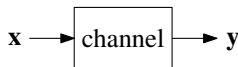
13.4.2	Length of Lempel–Ziv Codes	133
	Problems	134
14	The Discrete Memoryless Multiple-Access Channel	135
14.1	The Weak Converse	136
15	Analysis of ML Decoding	141
15.1	First Steps	141
15.2	ML Decoder Analysis	141
	Notes	144
A	Continuity: Pointwise vs. Uniform	149
A.1	Metrics	149
A.2	Pointwise Continuity	149
A.3	Uniform Continuity	150
	Bibliography	152
	Index	154

CHAPTER 1

Introduction

1.1. Communication Systems

Communication is the conveying of a message (or its approximation) from one place, called the **source location**, to another place, called the **destination** [20]. In order to accomplish this task, we have available a **channel** that accepts inputs $\mathbf{x}$ at the source location, and produces a related output $\mathbf{y}$ at the destination; the output $\mathbf{y}$ de-



pends on the input $\mathbf{x}$, usually in some random way. The first difficulty we have is that the message, call it U , is usually not the same type of object as the acceptable inputs to the channel. For example, the message may be a baseband voice waveform, but the channel may accept only waveforms in a certain passband. Or the channel may be a fiber-optic cable that accepts only pulses of light. So we usually need some sort of converter to transform a message object into a channel-input object. Depending on the context, this converter may be called an **encoder**, a **modulator**, or a **transmitter**. The second difficulty we have is that the channel output $\mathbf{y}$ may not be the same type of object as the message U that is desired at the destination. For this reason, we need a deconverter that transforms the channel output $\mathbf{y}$ into an approximation of U , which we denote by $\hat{U}$. Depending on the context, the deconverter may be called a **decoder**, a **demodulator**, or a **receiver**. The concatenation of the converter, the channel, and the deconverter constitutes a **communication system** as shown in Figure 1.1.

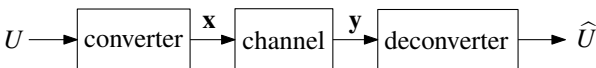


Figure 1.1. A communication system.

1.2. Digital Communication Systems

In a **digital communication system**, although there are many admissible values of the channel input $\mathbf{x}$, we intentionally use only a finite number of them, say N_c . This

means that the converter in Figure 1.1 can be broken down into two subsystems.^a The first one is a **source encoder** that takes a message and maps it into one of the integers $1, \dots, N_s$. The second subsystem is a **channel encoder** that maps each of these integers into one of our N_c selected channel inputs, say $\mathbf{x}_1, \dots, \mathbf{x}_{N_c}$. The deconverter can be similarly broken down into a **channel decoder** that takes channel outputs into the integers $1, \dots, N_c$ and a **source decoder** that takes these integers into message values or their approximations.



Figure 1.2. A digital communication system.

Another consequence of using only N_c channel inputs is that if the message U takes more than N_c distinct values, we cannot recover U in all cases. This situation can occur whether message is discrete or analog. Source encoders for binary data are discussed in the next subsection. As for analog data, for now we just mention that to approximate real-valued messages, we use **quantization**. If the message is a continuous-time waveform or a spatially continuous image, we sample it and then quantize the samples. Sampling is justified by the **sampling theorem** in the case of bandlimited data [20], [21]. As a practical matter, all sources are approximately bandlimited — a nice discussion of this can be found in the first three pages of Slepian [23] (easy reading — I recommend it).

1.2.1. Source Encoders for Binary Data

Consider an information source that generates an infinite stream of zeros and ones, which we denote by $U_1, U_2, \dots$. Suppose we group the information symbols into blocks of size k , e.g., $U := (U_1, \dots, U_k)$. Then there are *at most* 2^k possible values of U . A **source encoder** is a function that maps each possible k -tuple of bits into an integer $i \in \{1, \dots, N_s\}$, where $N_s \leq 2^k$. The **rate** of a such a source encoder is

$$R_s := \frac{\log_2 N_s}{k},$$

measured in *bits per source symbol*.

^aJust because we can break a system into two subsystems does not necessarily mean that it is a good idea to do so. We will say more about this shortly when we discuss the **Source-Channel Separation Theorem**.

Why Might We Use $N_s < 2^k$?

Suppose that $U_{2i} = 0$ for $i = 1, 2, \dots$. Then $(U_1, U_2, \dots, U_k)$ can take only $2^{k/2}$ possible values. More generally, not all length- k blocks have equal probability. Some may be so unlikely that we can ignore them and still have a very low probability of error. **Shannon's Source Coding Theorem** says that this will be the case if R_s greater than or equal to the **entropy** H of the source.

1.2.2. Channel Codes

A **channel encoder** is a function that maps each $i \in \{1, \dots, N_c\}$ into a vector $\mathbf{x}_i$ in a vector space X of dimension n . For example, we might have $X = \{0, 1\}^n$; i.e., the set of n -tuples of bits. In any case, if a source block $U = (U_1, \dots, U_k)$ is mapped into the integer i , then the signal $\mathbf{x}_i$ is transmitted over a channel, whose output $\mathbf{y}$ is determined by the transmitted signal $\mathbf{x}_i$ as well as noise. The **channel decoder** is a mapping that assigns to each possible channel output $\mathbf{y}$, an element in $\{1, \dots, N_c\}$. The **rate** of the channel code is

$$R_c := \frac{\log_2 N_c}{n},$$

measured in *bits per dimension* or *bits per channel use*, depending on the space X . **Shannon's Channel Coding Theorem** says that under suitable conditions, if R_c is less than or equal to the **capacity** C of the channel, then $P(i \neq \hat{i})$ can be made as small as we like by increasing n . In other words, the dashed box in Figure 1.2 can be made almost noiseless!

It is important to stress here that C does not depend on n . Since all we need is $(\log N_c)/n \leq C$, as n increases, we can let N_c increase as well!

1.2.3. Source Decoder

A **source decoder** is a mapping that assigns to each element in $\{1, \dots, N_s\}$, a unique binary k -tuple.

1.2.4. Rate-Distortion Theory

When the dashed box in Figure 1.2 is almost noiseless, what remains is the **rate-distortion** problem. If N_s is not large enough to code each message separately, and this is always the case for analog data, how can we make $\hat{U}$ the best approximation of U ?

1.2.5. Source-Channel Separation

The conditions $R_s \geq H$ and $R_c \leq C$ along with $N_s \leq N_c$ imply that

$$kH \leq \log_2 N_s \leq \log_2 N_c \leq nC.$$

If source symbols arrive at rate ρ_s per second and channel symbols are sent at rate ρ_c symbols (or dimensions) per second, then during a time window of duration T , $k = \rho_s T$ and $n = \rho_c T$. Hence, the above condition implies $\rho_s TH \leq \rho_c TC$, or $\rho_s H \leq \rho_c C$. If source symbols arrive at the same rate as channel symbols (or dimensions) leave, then we need $H \leq C$.

Under suitable assumptions, there is *no* loss in designing the source encoder and the channel encoder separately. This is known as the **Source-Channel Separation Theorem** [3].

1.3. Course Outline

As suggested by the foregoing discussion, key topics in the course include the sampling theorem, quantization, fixed-length source codes and Shannon's Source Coding Theorem along with Shannon's Channel Coding Theorem. In addition, we will study the sampling theorem, quantization, variable-length source codes, and rate-distortion theory. Various other topics may be included as time permits.

Problems

Remark. The problems below familiarize you with convex functions in general and the logarithm in particular (whose negative is convex). The logarithm is used later to define entropy and mutual information.

1. Let $g(x)$ be a real-valued function defined on an interval I of the real line. We say that g is **convex**^b if for all $x, y \in I$ and all $0 \leq \lambda \leq 1$,

$$g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y).$$

Let X be a discrete random variable taking values $x_1, \dots, x_n$ in I with probabilities $p_i = P(X = x_i)$. Derive **Jensen's inequality**,

$$E[g(X)] \geq g(E[X]).$$

^b A function is **concave** if its negative is convex. Hence, for *concave* functions h , Jensen's inequality says that $E[h(X)] \leq h(E[X])$. Some authors do not use the term "concave." Instead they write "convex $\cap$ " for "concave," and they write "convex $\cup$ " for "convex." In this context, the symbol $\cap$ is read "cap," and the symbol $\cup$ is read "cup."

Hint: Use **mathematical induction** on n : If $n = 2$, the result follows from the definition of a convex function if you take $\lambda = p_1$ and $(1 - \lambda) = p_2$. Now show that if the result holds for n equal to any $k \geq 2$, then the result holds for $n = k + 1$.

2. Evaluate $\lim_{t \downarrow 0} t \ln t$. *Hint:* Apply **L'Hôpital's rule** [19, p. 109] to $(\ln t)/(1/t)$.

3. Show that $\log_b x = (\ln x)/(\ln b)$.

Remark. Taking $x = e$ shows that $\log_b e = 1/(\ln b)$. It then follows that $\log_b x = (\log_b e) \ln x$.

4. Given any $\alpha > 0$, show that

$$\lim_{n \rightarrow \infty} (n\alpha)^{1/n} = 1.$$

Hint: First apply **L'Hôpital's rule** [19, p. 109] to $(\ln t)/t$ and make a note of the result.

CHAPTER 2

The Sampling Theorem and Quantization

2.1. The Sampling Theorem

Recall that a finite-energy waveform $x(t)$ is **bandlimited** if its **Fourier transform** $X(f)$ satisfies $X(f) = 0$ for $|f| > W$ for some positive **bandwidth** W . The **sampling theorem** says that such a waveform has the representation

$$x(t) = \sum_{k=-\infty}^{\infty} x(k/f_s) \operatorname{sinc}(f_s[t - k/f_s]),$$

where the **sampling frequency** f_s is at least twice the bandwidth; i.e., $f_s \geq 2W$. The **sinc function** is

$$\operatorname{sinc}(t) := \begin{cases} \frac{\sin(\pi t)}{\pi t}, & t \neq 0, \\ 1, & t = 0. \end{cases}$$

The point of the sampling theorem is that knowledge of the samples $x(k/f_s)$ allows us to reconstruct $x(t)$ for all t . Hence, in principle, the problem of communicating bandlimited waveforms has been reduced to the problem of communicating an infinite sequence of real numbers.

Derivation

To derive the sampling theorem, we start with the fact that the assumptions on x imply

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi f t} df = \int_{-W}^W X(f) e^{j2\pi f t} df.$$

Since X is zero outside of $[-W, W]$, X is also zero outside of $[-f_s/2, f_s/2]$ for any $f_s \geq 2W$. From the theory of Fourier series, we can write^a

$$X(f) = \sum_{k=-\infty}^{\infty} c_k e^{-j2\pi k f / f_s}, \quad |f| \leq f_s/2,$$

^aThe sum is understood to converge in mean square.

where the Fourier-series coefficients c_k are easily found to be $c_k = x(k/f_s)/f_s$. Now observe that^b

$$\begin{aligned}
 x(t) &= \int_{-\infty}^{\infty} X(f) e^{j2\pi f t} df \\
 &= \int_{-f_s/2}^{f_s/2} X(f) e^{j2\pi f t} df \\
 &= \int_{-f_s/2}^{f_s/2} \left[\sum_{k=-\infty}^{\infty} c_k e^{-j2\pi k f / f_s} \right] e^{j2\pi f t} df \\
 &= \sum_{k=-\infty}^{\infty} c_k \int_{-f_s/2}^{f_s/2} e^{-j2\pi k f / f_s} e^{j2\pi f t} df \\
 &= \sum_{k=-\infty}^{\infty} x(k/f_s)/f_s \cdot f_s \operatorname{sinc}(f_s[t - k/f_s]).
 \end{aligned}$$

2.2. Quantization

2.2.1. Vector Quantizers

Let $\mathbf{X} = (X_1, \dots, X_d)$ be an $\mathbb{R}^d$ -valued random vector. An N -level **vector quantizer** is any function q of the form^c

$$q(\mathbf{x}) := \sum_{i=1}^N \mathbf{c}_i \mathbf{1}_{V_i}(\mathbf{x}), \quad (2.1)$$

where the sets $V_1, \dots, V_N$ form a partition of $\mathbb{R}^d$ and the $\mathbf{c}_i \in \mathbb{R}^d$. Recall that a **partition** is a collection of nonempty sets that are pairwise disjoint and whose union is the whole space.

Optimizing the Partition – Voronoi Regions

If the $\mathbf{c}_i$ are given, what is the best choice of partition to minimize the mean-squared error $\mathbb{E}[\|\mathbf{X} - q(\mathbf{X})\|^2]$? Here $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$.

We claim that no partition can do better than the partition of **Voronoi regions** defined by^d

$$V_j^* := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{c}_j\|^2 \leq \|\mathbf{x} - \mathbf{c}_i\|^2 \text{ for all } i\}. \quad (2.2)$$

^bThe second integral can be viewed as an inner product, and so the interchange of infinite sum and integral is permissible by continuity of the inner product.

^cHere $\mathbf{1}_V$ denotes the **indicator function** of the set V ; i.e., $\mathbf{1}_V(\mathbf{x}) := 1$ if $\mathbf{x} \in V$ and $\mathbf{1}_V(\mathbf{x}) := 0$ otherwise.

^dTo be precise, we should let B_j denote the set in (2.2) and then put $V_1^* := B_1$ and $V_j^* := B_j \cap B_{j-1}^c \cap \dots \cap B_1^c$ for $j = 2, \dots, N$.

To establish this claim, let

$$q^*(\mathbf{x}) := \sum_{j=1}^N \mathbf{c}_j \mathbf{1}_{V_j^*}(\mathbf{x}).$$

Then write

$$\begin{aligned} \mathbb{E}[\|\mathbf{X} - q(\mathbf{X})\|^2] &= \mathbb{E}[\|\mathbf{X} - q(\mathbf{X})\|^2 \cdot \mathbf{1} \cdot \mathbf{1}] \\ &= \mathbb{E}\left[\|\mathbf{X} - q(\mathbf{X})\|^2 \left(\sum_{i=1}^N \mathbf{1}_{V_i}(\mathbf{X})\right) \left(\sum_{j=1}^N \mathbf{1}_{V_j^*}(\mathbf{X})\right)\right] \\ &= \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[\|\mathbf{X} - q(\mathbf{X})\|^2 \mathbf{1}_{V_i}(\mathbf{X}) \mathbf{1}_{V_j^*}(\mathbf{X})] \\ &= \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[\|\mathbf{X} - \mathbf{c}_i\|^2 \mathbf{1}_{V_i}(\mathbf{X}) \mathbf{1}_{V_j^*}(\mathbf{X})], \end{aligned}$$

since $q(\mathbf{x}) := \mathbf{c}_i$ for $\mathbf{x} \in V_i$. On the other hand, for $\mathbf{x} \in V_j^*$, we know that

$$\|\mathbf{x} - \mathbf{c}_j\|^2 \leq \|\mathbf{x} - \mathbf{c}_i\|^2, \quad \text{for all } i.$$

Hence,

$$\begin{aligned} \mathbb{E}[\|\mathbf{X} - q(\mathbf{X})\|^2] &\geq \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[\|\mathbf{X} - \mathbf{c}_j\|^2 \mathbf{1}_{V_i}(\mathbf{X}) \mathbf{1}_{V_j^*}(\mathbf{X})] \\ &= \sum_{i=1}^N \sum_{j=1}^N \mathbb{E}[\|\mathbf{X} - q^*(\mathbf{X})\|^2 \mathbf{1}_{V_i}(\mathbf{X}) \mathbf{1}_{V_j^*}(\mathbf{X})] \\ &= \mathbb{E}\left[\|\mathbf{X} - q^*(\mathbf{X})\|^2 \left(\sum_{i=1}^N \mathbf{1}_{V_i}(\mathbf{X})\right) \left(\sum_{j=1}^N \mathbf{1}_{V_j^*}(\mathbf{X})\right)\right] \\ &= \mathbb{E}[\|\mathbf{X} - q^*(\mathbf{X})\|^2]. \end{aligned}$$

Optimizing the Coefficients – Linear Estimation

If the sets V_i of an arbitrary quantizer q in (2.1) are given, what are the best $\mathbf{c}_i$ to use?

Let $\mathbf{Y} := [\mathbf{1}_{V_1}(\mathbf{X}), \dots, \mathbf{1}_{V_N}(\mathbf{X})]^\top$. Then $q(\mathbf{X}) = A\mathbf{Y}$, where A is the $d \times N$ matrix $A := [\mathbf{c}_1, \dots, \mathbf{c}_N]$. Hence, the mean-squared error is $\mathbb{E}[\|\mathbf{X} - A\mathbf{Y}\|^2]$. In other words, finding the best vectors $\mathbf{c}_i$ is equivalent to finding the best linear transformation (matrix) A . By the **orthogonality principle** [7, Section 8.4], the best matrix A is any solution of the **normal equations**

$$A R_{\mathbf{Y}} = R_{\mathbf{X}\mathbf{Y}}, \tag{2.3}$$

where R_Y is the correlation (not covariance) matrix $R_Y := E[YY^T]$ and R_{XY} is the cross-correlation matrix $R_{XY} := E[XY^T]$.

2.2.2. The Lloyd–Max Algorithm

The repeated alternating iteration of optimizing the partition and $\mathbf{c}_i$ is the **Lloyd–Max algorithm** [16], [17]. The scalar case, $d = 1$, is quite simple. First, if $c_1 < \dots < c_N$, then (2.2) reduces to

$$V_j^* = \left(\frac{c_{j-1} + c_j}{2}, \frac{c_j + c_{j+1}}{2} \right].$$

Second, since R_Y is diagonal, (2.3) is easily solved to yield

$$c_j^* = \frac{E[X \mathbf{1}_{V_j}(X)]}{P(X \in V_j)}.$$

If X has density function f , and if V_j is an interval, say $V_j = (a_j, b_j]$, then

$$c_j^* = \frac{\int_{a_j}^{b_j} x f(x) dx}{\int_{a_j}^{b_j} f(x) dx}.$$

Problems

1. **MATLAB.** The sampling theorem assumes that the signal is exactly bandlimited and that you have the samples $x(k/f_s)$ for *all* k . In this problem you will explore what happens when the bandlimited condition only holds approximately and when not all samples are available. Download and save the MATLAB M-files `smptm.m`, `lincmb.m`, and `sscript.m`. following parameter changes:

- (a) No changes.
- (b) `Tasamp = -1` and `Tbsamp = 3`.
- (c) `W = 0.3`, `Tasamp = -5` and `Tbsamp = 5`.

Turn in your plots in all three cases.

2. Show that the mean-squared error of an N -level quantizer of a scalar random variable X is

$$E[|X - q(X)|^2] = \sum_{j=1}^N E[(X - c_j)^2 \mathbf{1}_{V_j}(X)].$$

Now put

$$F_k(x) := E[X^k \mathbf{1}_{(-\infty, x]}(X)],$$

where F_0 is the ordinary cumulative distribution function. Show that if $V_j = (a_j, b_j]$, then the j th expectation in the above formula is given by

$$F_2(b_j) - F_2(a_j) - 2c_j[F_1(b_j) - F_1(a_j)] + c_j^2[F_0(b_j) - F_0(a_j)].$$

3. The formula for V_j^* given in the scalar case of the Lloyd–Max algorithm is valid only for $j = 2, \dots, N-1$. Determine formulas for V_1^* and V_N^* .

CHAPTER 3

Data Compression Using Variable-Length Codes

3.1. Introduction

Consider an **alphabet** of source letters, say $X = \{a, b, c, d\}$. Here are possible choices for a fixed-length binary code and a variable-length binary code.

symbol	codeword
a	00
b	01
c	10
d	11

symbol	codeword
a	0
b	10
c	110
d	111

The expected codeword length of the fixed-length code is 2, no matter what the probabilities of the source letters are. This is in contrast to the variable-length code. If the source letters are equally likely, then the expected codeword length of the variable-length code is $9/4 = 2.25$. However, if $p(a) = 1/2$ and $p(b) = p(c) = p(d) = 1/6$, then the expected length of the variable-length code is $11/6 = 1.8\bar{3}$.

Goals:

- Characterize unique decodability or self-punctuating codes.
- Characterize the minimum possible expected length. This result depends on the entropy of the source.
- Develop an algorithm for finding a variable-length code whose expected length is the minimum possible.

3.2. Codewords

Let D be a finite set containing D elements. Without loss of generality, take $D := \{0, 1, \dots, D-1\}$. The set D is called the **codeword alphabet**. Let D^* denote the set of all finite-length strings from D . For $\mathbf{d} \in D^*$, we let $\ell(\mathbf{d})$ denote the number of letters in $\mathbf{d}$. For example, if $D = 3$, $D = \{0, 1, 2\}$, and if $\mathbf{d} = 0020112$, then $\ell(\mathbf{d}) = 7$. Our **codewords** will be elements of D^* .

3.3. Codes

Let X be an arbitrary set of **source symbols**. Let $|X|$ denote the **cardinality** of X ; i.e., $|X|$ is the number of elements in X . A function or mapping $C: X \rightarrow D^*$ is called a **source code**. The **range** of C is the set $\{C(x) : x \in X\}$. In coding theory, the range of the function C is called the **codebook**. Typically, the codebook is always a proper subset of D^* .

In this course, we always work with finite or countably infinite sets X . In this case, we assume that we are given a **probability mass function** (pmf) p on X ; i.e., $p(x) \geq 0$ and $\sum_{x \in X} p(x) = 1$. When p is given, the **expected length** of the source code C is

$$L(C) := \sum_{x \in X} \ell(C(x))p(x).$$

In other words, if X is an X -valued **random variable** with pmf $p(x) = P(X = x)$, then

$$E[\ell(C(X))] = \sum_{x \in X} \ell(C(x))P(X = x) = \sum_{x \in X} \ell(C(x))p(x) = L(C).$$

Definition 3.1. A collection of codes is said to be **closed under codeword reassignment** if given any code in the collection, reordering its codewords yields another code in the collection.

Here are some collections of codes that are closed under codeword reassignment:

- Codes whose codewords have length at most n .
- Codes whose codewords have an even number of 1s.
- Prefix-free codes (defined later).

Here are some collections of codes that are *not* closed under codeword reassignment; here $X = \{a, b, c, d\}$:

- Codes with $\ell(C(a)) = 1$.
- Codes with $C(a)$ having an even number of 1s.
- Binary codes with $C(a)$ being the one's complement of $C(d)$.

Proposition 3.2. *Let $\mathcal{C}$ be a collection of codes that is closed under codeword reassignment. If $C \in \mathcal{C}$ has minimal expected length among all codes in $\mathcal{C}$, then C has the property that higher-probability codewords have smaller length; more precisely, if $p(u) > p(v)$, then $\ell(C(u)) \leq \ell(C(v))$.*

Proof. Suppose $C \in \mathcal{C}$ has minimal expected length; i.e., $L(C) \leq L(C')$ for all $C' \in \mathcal{C}$. Suppose also that $p(u) > p(v)$. Construct a new code C' by interchanging

the codewords assigned to u and v and leaving all other codewords the same. Thus,

$$\begin{aligned} L(C') &= \sum_{x \neq u, v} \ell(C'(x))p(x) + \ell(C'(u))p(u) + \ell(C'(v))p(v) \\ &= \sum_{x \neq u, v} \ell(C(x))p(x) + \ell(C(v))p(u) + \ell(C(u))p(v). \end{aligned}$$

Then

$$\begin{aligned} 0 \leq L(C') - L(C) &= \ell(C(v))[p(u) - p(v)] + \ell(C(u))[p(v) - p(u)] \\ &= [p(u) - p(v)][\ell(C(v)) - \ell(C(u))], \end{aligned}$$

where the inequality follows because $C' \in \mathcal{C}$ (on account of the closed reassignment property of $\mathcal{C}$) and because C has minimal expected length. Since $p(u) - p(v) > 0$, we can divide by this quantity and obtain the desired result. $\square$

3.4. Decodability

Good codes have small values of their expected length. But now suppose that we wish to compress a sequence of source symbols. If the codewords are separated by commas, there is no problem; however, this implies that the commas must also be stored!

Definition 3.3. A code $C: X \rightarrow D^*$ is **one-to-one** if $C(x) = C(y) \Rightarrow x = y$ for all $x, y \in X$.

Definition 3.4 (Extension of a Code). Given a code $C: X \rightarrow D^*$, we define its **extension** $C^*: X^* \rightarrow D^*$ by

$$C^*(x_1 x_2 \cdots x_n) := C(x_1)C(x_2) \cdots C(x_n).$$

In other words, for $x_1, \dots, x_n \in X$, if $\mathbf{x} := x_1 x_2 \cdots x_n \in X^*$ is the **concatenation** of $x_1, \dots, x_n$, then $C^*(\mathbf{x})$ is the concatenation of $C(x_1), \dots, C(x_n)$.

Definition 3.5 (Unique Decodability (Self Punctuating)). A code C is **uniquely decodable** or **self punctuating** if its extension C^* is one-to-one. In other words, C is uniquely decodable if given any sequence of source symbols, say $x_1, \dots, x_n$, we can recover them using only the concatenation of their corresponding codewords.

For a general uniquely decodable code, we need to examine $C(x_1) \cdots C(x_n)$ before we can determine x_1 .

Definition 3.6 (Prefix-Free Code (Instantaneous Code)). A code is said to be a **prefix-free code** or an **instantaneous code** if no codeword is a prefix of any other codeword.

Example 3.7 (A Prefix-Free Code [3]). With $X = \{a, b, c, d\}$ and $D = \{0, 1\}$, consider the prefix-free code

x	$C(x)$
a	0
b	10
c	110
d	111

If $C^*(\mathbf{x}) = 01011111010$, we see that $C^*(\mathbf{x}) = 0, 10, 111, 110, 10$ and so $x_1 = a, x_2 = b, x_3 = d, x_4 = c$, and $x_5 = b$. Putting this all together, we have $\mathbf{x} = abdc b$. Since we could figure out where to put the commas, the code is self punctuating.

Example 3.8 (Uniquely Decodable but not Prefix-Free [3]). With $X = \{a, b, c, d\}$ and $D = \{0, 1\}$, consider the code

x	$C(x)$
a	10
b	00
c	11
d	110

If $C^*(\mathbf{x}) = 10001101100$, we again decode $\mathbf{x} = abdc b$. However, we cannot decide that $x_3 = d$ until we see the first bit of $C(x_4) = 11$. Hence, this code is uniquely decodable, but it is not a prefix-free code.

Example 3.9 (One-to-One but not Uniquely Decodable [3]). Again suppose that $X = \{a, b, c, d\}$ and $D = \{0, 1\}$, consider the code

x	$C(x)$
a	0
b	010
c	01
d	10

With $C^*(\mathbf{x}) = 010$, we cannot distinguish among the choices $\mathbf{x} = b$, $\mathbf{x} = ad$, and $\mathbf{x} = ca$ since in each case $C^*(\mathbf{x}) = 010$. In other words, we need to put commas between the values of $C(x_i)$.

Proposition 3.10. *A nontrivial prefix-free code of minimal expected length must have at least two codewords of maximal length.*

Proof. Suppose otherwise that for some $u \in X$ with $p(u) > 0$, $\ell(C(u)) > \ell(C(x))$ for all $x \neq u$. Let $C(v)$ be a codeword whose length is the next longest after $C(u)$, say $C(v) = d_1 \cdots d_n$ and $C(u) = d'_1 \cdots d'_n d'_{n+1} \cdots d'_{n+k}$. If we define a new code C' by putting $C'(u) := d'_1 \cdots d'_n$ and $C'(x) := C(x)$ for $x \neq u$, then $L(C') < L(C)$, and C' is still a prefix-free code. This contradicts the assumption that the prefix-free code C has minimal expected length among all prefix-free codes. $\square$

3.5. Prefix-Free Codes and the Kraft Inequality

It is sometimes convenient to view codewords as being located at nodes in a tree. Let C be a code, and put $\ell_{\max} := \max_{x \in X} \ell(C(x))$. Construct a D -ary tree of depth $\ell_{\max}$ as shown in Figure 3.1 for $D = 2$ and $\ell_{\max} = 3$. We can then draw circles around

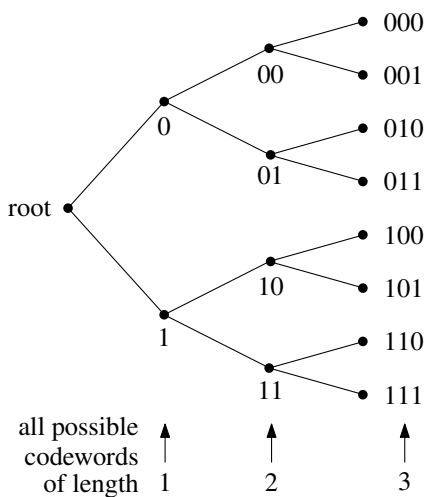


Figure 3.1. Binary tree for codewords of maximum length 3.

the nodes that correspond to codewords. For example, Figure 3.2 illustrates the case in which 0 and 10 are codewords. Notice that the 6 codewords connected to the right of 0 all contain 0 as a prefix. Similarly, the 2 codewords connected to the right of 10 contain 10 as a prefix. To add another codeword and maintain the prefix-free property, we must choose from 11, 110, or 111. If we choose 11, then there are no more codewords that can be added without violating the prefix-free property. On the

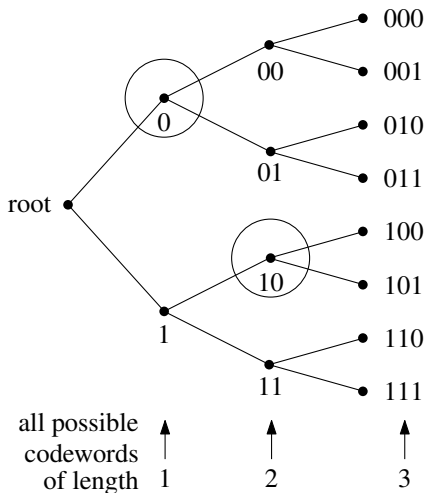


Figure 3.2. Codewords 0 and 10 are circled.

other hand, if we instead add 110 as the third codeword, we could still add 111 as a fourth codeword and maintain the prefix-free property.

Theorem 3.11 (Kraft Inequality, 1949). *Let X be a finite source alphabet of size m , and consider a code alphabet of size D . Then every prefix-free code C satisfies*

$$\sum_{x \in X} D^{-\ell(C(x))} \leq 1. \quad (3.1)$$

Conversely, if $\ell_1, \dots, \ell_m$ are positive integers satisfying $\sum_{i=1}^m D^{-\ell_i} \leq 1$, then there is a prefix-free code C whose codewords $\mathbf{d}_1, \dots, \mathbf{d}_m$ satisfy $\ell(\mathbf{d}_i) = \ell_i$. Furthermore, in the binary case, if (3.1) is strict, then C can be modified to obtain a new code that satisfies (3.1) with equality and whose expected length is not greater than the original.

Proof. Suppose C is a prefix-free code. Put $\ell_{\max} := \max_{x \in X} \ell(C(x))$. Construct the D -ary tree of potential codewords of maximum length $\ell_{\max}$ as shown in Figure 3.1 for $D = 2$ and $\ell_{\max} = 3$. If a codeword $C(x)$ has length $\ell(C(x))$, then there are $D^{\ell_{\max} - \ell(C(x))}$ words in the tree having length $\ell_{\max}$ that cannot be used as codewords for other source letters (do not worry about words in the tree having length less than $\ell_{\max}$ that cannot be used). Since there are $D^{\ell_{\max}}$ words in the tree having length $\ell_{\max}$,

and since we cannot eliminate more length- $\ell_{\max}$ words than there are, we must have

$$\sum_{x \in X} D^{\ell_{\max} - \ell(C(x))} \leq D^{\ell_{\max}},$$

which implies (3.1).

Now suppose codeword lengths ℓ_i are given with $\sum_{i=1}^m D^{-\ell_i} \leq 1$. Without loss of generality, assume $\ell_1 \leq \dots \leq \ell_m$. Then $\sum_{i=1}^m D^{\ell_m - \ell_i} \leq D^{\ell_m}$. Consider a D -ary tree of potential codewords of maximum length ℓ_m . Pick any word of length ℓ_1 . Then all words in the tree attached to the right of the codeword are not available to assign to other source letters. However,

$$D^{\ell_m} - D^{\ell_m - \ell_1} \geq \sum_{i=2}^m D^{\ell_m - \ell_i} \geq 1.$$

Hence, there are words in the tree having length ℓ_m that are available. Pick one and trace a path back to the root of the tree. This shows that there is a codeword of length ℓ_2 available. Continuing in this way, we obtain prefix-free codewords $\mathbf{d}_1, \dots, \mathbf{d}_m$ that satisfy $\ell(\mathbf{d}_i) = \ell_i$.

Suppose (3.1) is strict. Then

$$D^{\ell_m} - \sum_{i=1}^m D^{\ell_m - \ell_i} > 0.$$

This means that there is at least one available word of length $\ell_{\max}$ in the tree, call that word $\mathbf{d}$. If the sibling of $\mathbf{d}$ is a codeword $\mathbf{d}_i$ with length $\ell_i = \ell_{\max}$, then the source letter it corresponds to can be reassigned to the word of the parent node (remember, there is only one sibling by the binary assumption). This change decreases the number of available length- $\ell_{\max}$ words in the tree by one and reduces the value of ℓ_i by one (hence, this change cannot increase the expected length of the code). If the sibling of $\mathbf{d}$ is also available, then the word of their parent can be used as the codeword of some source letter that was originally assigned a length- $\ell_{\max}$ codeword (there is at least one such source letter by the definition of $\ell_{\max}$). In this case, the use of the parent reduces the number of available length- $\ell_{\max}$ words by two, but the reassignment increases it by one, resulting in a net reduction by one. In either case, one of the ℓ_i has been reduced by one. Call the new code C' . Compute its $\ell_{\max}$ and repeat the procedure. Continue until the (3.1) holds with equality. $\square$

Remark. The Kraft inequality is strict for a code C if and only if we could add another codeword of length at most $\ell_{\max}$ and maintain the prefix-free property.

3.6. Unique Decodability and McMillan's Theorem

Theorem 3.12 (Brockway McMillan, 1956). *Let $|\mathbf{X}| < \infty$. If $C: \mathbf{X} \rightarrow \mathbf{D}^*$ is uniquely decodable, then the Kraft inequality $\sum_{x \in \mathbf{X}} D^{-\ell(C(x))} \leq 1$ holds.*

Proof. (Karush, 1961). Let n be an arbitrary positive integer, and consider the identity

$$\begin{aligned} \left(\sum_{x \in \mathbf{X}} D^{-\ell(C(x))} \right)^n &= \prod_{k=1}^n \left(\sum_{x_k \in \mathbf{X}} D^{-\ell(C(x_k))} \right) \\ &= \sum_{x_1 \in \mathbf{X}} \sum_{x_2 \in \mathbf{X}} \cdots \sum_{x_n \in \mathbf{X}} D^{-[\ell(C(x_1)) + \cdots + \ell(C(x_n))]} . \end{aligned}$$

Since $\ell(C(x_1)) + \cdots + \ell(C(x_n)) = \ell(C^*(\mathbf{x}))$, where $\mathbf{x} = x_1 \cdots x_n$, we can write

$$\left(\sum_{x \in \mathbf{X}} D^{-\ell(C(x))} \right)^n = \sum_{\mathbf{x} \in \mathbf{X}^n} D^{-\ell(C^*(\mathbf{x}))} .$$

For $i \geq 1$, let $B_i := \{\mathbf{x} \in \mathbf{X}^n : \ell(C^*(\mathbf{x})) = i\}$. Put $I := \{i : B_i \neq \emptyset\}$. Note that if $\ell_{\max} := \max_{x \in \mathbf{X}} \ell(C(x))$, then $i \in I \Rightarrow i \leq n\ell_{\max}$ since $\ell(C^*(\mathbf{x})) = \sum_{k=1}^n \ell(C(x_k)) \leq \sum_{k=1}^n \ell_{\max} = n\ell_{\max}$. Next, observe that the B_i are disjoint. Hence,

$$\sum_{\mathbf{x} \in \mathbf{X}^n} D^{-\ell(C^*(\mathbf{x}))} = \sum_{i \in I} \sum_{\mathbf{x} \in B_i} D^{-i} .$$

As $\mathbf{x}$ runs over B_i , the codewords $C^*(\mathbf{x})$ are all different because the code is uniquely decodable. Furthermore, each such codeword has length i . Since there are at most D^i different codewords of length i , there can be at most D^i different source words $\mathbf{x} \in B_i$. Thus, $|B_i| \leq D^i$ and we can write

$$\sum_{i \in I} \sum_{\mathbf{x} \in B_i} D^{-i} = \sum_{i \in I} D^{-i} |B_i| \leq \sum_{i \in I} 1 = |I| \leq n\ell_{\max} .$$

So

$$\left(\sum_{x \in \mathbf{X}} D^{-\ell(C(x))} \right)^n \leq n\ell_{\max} ,$$

or

$$\sum_{x \in \mathbf{X}} D^{-\ell(C(x))} \leq (n\ell_{\max})^{1/n} \rightarrow 1, \quad \text{as } n \rightarrow \infty .$$

□

Remark. McMillan's Theorem shows that a uniquely decodable code satisfies the Kraft inequality. Hence, by Kraft's Theorem, we can always find a prefix code with the same code lengths as the uniquely decodable code. This prefix code has the same expected length as the uniquely decodable code, but the prefix code incurs less delay and is easier to decode.

3.7. Properties of Lengths that Satisfy the Kraft Inequality

Proposition 3.13. *Consider a uniquely decodable code with m codewords. If any codeword has length m or more, then the Kraft inequality is strict.*

Proof. Denote the lengths of the codewords by $\ell_1 \leq \dots \leq \ell_m$. We begin by considering the subproblem

$$\max_{1 \leq \ell_1 \leq \dots \leq \ell_{m-1}} \sum_{i=1}^{m-1} \frac{1}{D^{\ell_i}} \quad \text{subject to} \quad \sum_{i=1}^{m-1} \frac{1}{D^{\ell_i}} < 1.$$

We first want as many of the ℓ_i to be 1 as possible without violating the constraint. Since

$$\underbrace{\frac{1}{D} + \dots + \frac{1}{D}}_{D \text{ times}} = 1,$$

at most $D - 1$ of the ℓ_i can be 1. Similarly, since

$$\underbrace{\frac{1}{D} + \dots + \frac{1}{D}}_{D-1 \text{ times}} + \underbrace{\frac{1}{D^2} + \dots + \frac{1}{D^2}}_{D \text{ times}} = \frac{D-1}{D} + \frac{D}{D^2} = 1,$$

we conclude that at most $D - 1$ of the ℓ_i can be 2. The general approach is to write $m - 1 = q(D - 1) + r$, where q and r are integers with $0 \leq r < D - 1$. Then we want the first $D - 1$ values of $\ell_i = 1$, the next $D - 1$ values of $\ell_i = 2$, and so on, with $D - 1$ value of $\ell_i = q$, and only r of the $\ell_i = q + 1$. With the ℓ_i chosen this way,^a

$$\sum_{i=1}^{m-1} \frac{1}{D^{\ell_i}} = (D-1) \sum_{i=1}^q \frac{1}{D^i} + \frac{r}{D^{q+1}} = (D-1) \frac{1 - (1/D)^q}{D-1} + \frac{r}{D^{q+1}} = 1 - \frac{1}{D^q} + \frac{r}{D^{q+1}}.$$

^a Recall the **finite geometric series** $\sum_{k=0}^{n-1} z^k = \frac{1 - z^n}{1 - z}$, $z \neq 1$.

Now observe that since $D \geq 2$ and $r \geq 0$,

$$m - 1 = q(D - 1) + r \geq q.$$

Hence, if $\ell_m \geq m$, then $\ell_m \geq q + 1$. This further implies $1/D^{\ell_m} \leq 1/D^{q+1}$, and we have

$$\sum_{i=1}^m \frac{1}{D^{\ell_i}} \leq \left[1 - \frac{1}{D^q} + \frac{r}{D^{q+1}} \right] + \frac{1}{D^{q+1}} = 1 - \frac{1}{D^q} + \frac{r+1}{D^{q+1}} < 1 - \frac{1}{D^q} + \frac{D}{D^{q+1}} = 1.$$

Thus, if $\ell_m \geq m$, the Kraft inequality is strict. □

Remark. It is sometimes useful to restate the proposition in the contrapositive; i.e., if the Kraft inequality holds with equality, then all codewords lengths are upper bounded by $m - 1$. In particular, if the Kraft inequality holds with equality, then there are finitely many prefix-free codes with m codewords.

3.8. Entropy

Definition 3.14 (Entropy). If p is a probability mass function on X , the **entropy** of p is

$$H(p) := \sum_x p(x) \log \frac{1}{p(x)} = - \sum_x p(x) \log p(x) \geq 0,$$

where the sum is understood to be over those x with $p(x) > 0$. Equivalently, we can take the sum over all $x \in X$ if we put $0 \log 0 := \lim_{t \downarrow 0} t \log t = 0$. The base of the logarithm can be any real number greater than one. If the base is 2, the entropy is measured in **bits**; if the base is e , the entropy is measured in **nats**. If X is an X -valued random variable with pmf $P_X(x) := P(X = x)$, we almost always use the shorthand notation $H(X)$ instead of $H(P_X)$.

You will show in the problems that $H(X) \geq 0$, with equality if and only if X is a constant random variable. You will also show that if X is a finite set, then $H(X) \leq \log |X|$, with equality if and only if $p(x) = 1/|X|$ for all x ; i.e., X is uniformly distributed on X . These observations suggest that **entropy is a measure of uncertainty**. Suppose we know the distribution of X . If $P(X = x_0) = 1$ and $P(X = x) = 0$ for $x \neq x_0$, then there is *no* uncertainty about the value of X because X is a constant random variable with value x_0 . If we know that X is uniformly distributed on X , then all values of x are equally likely to occur; i.e., we have *maximum* uncertainty about which value of X will occur.

In order to simplify derivations of properties of entropy, it is convenient to write entropy as an expectation. Recall that for a real-valued function $g(x)$,

$$\mathbb{E}[g(X)] = \sum_x g(x)P_X(x).$$

The special case $g(x) = -\log P_X(x)$ yields

$$\mathbb{E}[-\log P_X(X)] = \sum_x [-\log P_X(x)]P_X(x) = H(X).$$

Thus,

$$H(X) = \mathbb{E}[-\log P_X(X)].$$

Example 3.15. For a pair of random variables X and Y with joint pmf $P_{XY}(x, y)$, their **joint entropy** is

$$H(XY) = \mathbb{E}[-\log P_{XY}(X, Y)] = \sum_{x,y} P_{XY}(x, y) \log \frac{1}{P_{XY}(x, y)}.$$

Similarly, the **conditional entropy**

$$H(X|Y) = H(P_{X|Y}) = \mathbb{E}[-\log P_{X|Y}(X|Y)] = \sum_{x,y} P_{XY}(x, y) \log \frac{1}{P_{X|Y}(x|y)}.$$

Using the identity $P_{XY}(x, y) = P_{X|Y}(x|y)P_Y(y)$, we see that

$$\begin{aligned} H(XY) &= \mathbb{E}[-\log P_{X|Y}(X|Y)P_Y(Y)] = \mathbb{E}[-\log P_{X|Y}(X|Y)] + \mathbb{E}[-\log P_Y(Y)] \\ &= H(X|Y) + H(Y). \end{aligned} \quad (3.2)$$

The reader should check that we can also write

$$H(XY) = H(Y|X) + H(X) \quad (3.3)$$

and

$$H(XYZ) = H(X|YZ) + H(Y|Z) + H(Z), \quad (3.4)$$

where $H(X|YZ) = \mathbb{E}[-\log P_{X|YZ}(X|Y, Z)]$. Formulas (3.2)–(3.4) are instances of the **chain rule** for entropy. In particular, if X , Y , and Z are independent random variables, then $P_{X|YZ}(x|y, z) = P_X(x)$ and $P_{Y|Z}(y|z) = P_Y(y)$. Thus,

$$H(XYZ) = H(X) + H(Y) + H(Z), \quad \text{if } X, Y, \text{ and } Z \text{ are independent.}$$

In other words, the entropy of a block of independent random variables is the sum of the individual entropies.

Definition 3.16. An independent, identically distributed (i.i.d.) sequence of X -valued random variables $\{X_n\}_{n=1}^{\infty}$ is called a **discrete memoryless source (DMS)**. If the X_n have common pmf $p(x) = P(X_n = x)$, then $H(p) = H(X_1)$ is called the entropy of the DMS.

Theorem 3.17 (Log Inequality).

$$\log t \leq (\log e)(t - 1), \quad t > 0,$$

with equality if and only if $t = 1$.

Proof. First note that since $\log t = (\log e) \ln t$, it suffices to prove the result for the natural logarithm. The log inequality is so fundamental to information theory that we give three proofs.

For the first proof, we use the fact that $1/t$ is strictly decreasing as follows (see Figure 3.3). For $t > 1$,

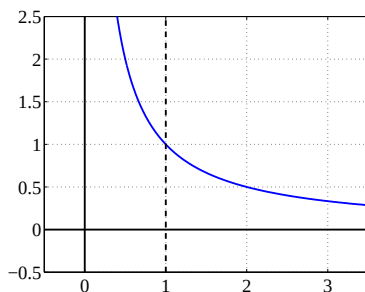


Figure 3.3. Plot of $1/t$ vs. t .

$$\ln t = \int_1^t \frac{d\tau}{\tau} < \int_1^t 1 \, d\tau = t - 1,$$

and for $0 < t < 1$,

$$\ln t = \int_1^t \frac{d\tau}{\tau} = - \int_t^1 \frac{d\tau}{\tau} < - \int_t^1 1 \, d\tau = t - 1.$$

For the second proof, we apply the **mean-value theorem** [19, p. 108, Th. 5.10]

$$f(t) = f(\tau) + f'(s)(t - \tau)$$

to $f(t) := \ln t$. In this formula, s lies strictly between t and τ . Hence, because the derivative $f'(s) = 1/s$ is strictly decreasing,^b

$$f(t) < f(\tau) + f'(\tau)(t - \tau), \quad t \neq \tau.$$

The reader should check the two cases $\tau < s < t$ and $t < s < \tau$. Now take $\tau = 1$ and $t \neq 1$ to get $\ln t < 0 + 1(t - 1) = t - 1$.

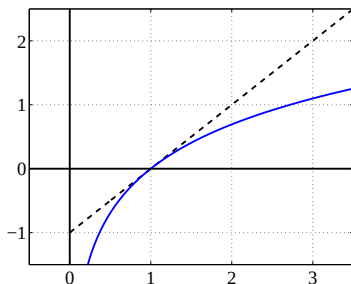


Figure 3.4. Graph of $\ln t$ and its tangent at $t = 1$.

For the third proof, we expand the function $\ln t$ about the point $t = 1$ using a Taylor series with remainder [19, pp. 110–111]. Thus,

$$\ln t = (t - 1) - \frac{(t - 1)^2}{2\tau^2} \leq t - 1,$$

where τ is strictly between 1 and t . Note that the inequality is strict for $t \neq 1$. □

Remark. Although the foregoing three proofs seem very different, at a certain level they are all nearly the same. The first two proofs use the fact that the derivative is strictly decreasing. The third proof uses the fact that the second derivative is strictly negative, which implies that the first derivative is strictly decreasing.

3.9. Expected Lengths of Codes

Theorem 3.18. *Let $|\mathcal{X}| < \infty$. If $C : \mathcal{X} \rightarrow \mathcal{D}^*$ is a uniquely decodable code, and if X is an $\mathcal{X}$ -valued random variable, then*

$$\mathbb{E}[\ell(C(X))] \geq \frac{H(X)}{\log D}.$$

^bThis inequality says that the curve f lies below that tangent line to the curve at the point $(\tau, f(\tau))$. This is illustrated in Figure 3.4 with $f(t) = \ln t$ and $\tau = 1$. The inequality implies that f is concave, but we do not need this fact here.

Proof. We show that

$$H(X) - (\log D)E[\ell(C(X))] \leq 0.$$

Let p denote the pmf of X , and write

$$\begin{aligned} H(X) - (\log D)E[\ell(C(X))] &= E[-\log p(X)] + E[\log D^{-\ell(C(X))}] \\ &= E\left[\log \frac{D^{-\ell(C(X))}}{p(X)}\right] \\ &= (\log e) \sum_x p(x) \ln\left(\frac{D^{-\ell(C(x))}}{p(x)}\right) \\ &\leq (\log e) \sum_x p(x) \left(\frac{D^{-\ell(C(x))}}{p(x)} - 1\right) \end{aligned} \quad (3.5)$$

$$\begin{aligned} &= (\log e) \left(\sum_x D^{-\ell(C(x))} - \sum_x p(x) \right) \\ &= (\log e) \left(\sum_x D^{-\ell(C(x))} - 1 \right) \\ &\leq (\log e)(1 - 1) \\ &= 0, \end{aligned} \quad (3.6)$$

where we used the log inequality in (3.5) and the Kraft inequality (McMillan's Theorem) in (3.6). $\square$

This proves that there is a lower bound on the expected length of a uniquely decodable source code. In other words, it is impossible to have a uniquely decodable source code with an expected length lower than $H(X)/\log D$. The next theorem shows that we can *always* achieve an expected length that is within one of this lower bound.

Theorem 3.19 (Shannon Codes). *Let $|X| < \infty$, and let X be an X -valued random variable. Then there exists a prefix code $C : X \rightarrow D^*$ such that*

$$\frac{H(X)}{\log D} \leq E[\ell(C(X))] < \frac{H(X)}{\log D} + 1.$$

Proof. Let p denote the pmf of X . The lower bound was shown in the previous theorem. For the upper bound, observe that equality in (3.5) holds if and only if

$$\frac{D^{-\ell(C(x))}}{p(x)} = 1$$

for each x with $p(x) > 0$. In other words, we would like to have $\ell(C(x)) \approx \log_D \frac{1}{p(x)}$.

Since $|\mathbf{X}| < \infty$, let $\{x_1, \dots, x_m\}$ be an enumeration of all the points in $\mathbf{X}$ which have a positive probability. Then $\ell_i := \lceil \log_D \frac{1}{p(x_i)} \rceil$ satisfies^c

$$D^{-\ell_i} \leq p(x_i) < D^{-\ell_i+1}. \quad (3.7)$$

This leads to two results. First,

$$\sum_{i=1}^m D^{-\ell_i} \leq \sum_{i=1}^m p(x_i) = 1.$$

Hence, by Kraft's Theorem, there exists a prefix code with code lengths $\{\ell_i\}_{i=1}^m$. Second,

$$\log p(x_i) < (-\ell_i + 1)(\log D).$$

So

$$\ell_i < \frac{-\log p(x_i)}{\log D} + 1,$$

which implies

$$\sum_{i=1}^m \ell_i p(x_i) < \frac{-1}{\log D} \sum_{i=1}^m p(x_i) \log p(x_i) + \sum_{i=1}^m p(x_i),$$

or

$$\mathbb{E}[\ell(C(X))] < \frac{H(X)}{\log D} + 1. \quad \square$$

Although Shannon codes are good, in some cases better codes can be found.

Example 3.20 (A Shannon Code That Can Be Improved [3]). Suppose that $\mathbf{X} = \{\mathbf{a}, \mathbf{b}\}$ with $P(X = \mathbf{a}) = .9999$ and $P(X = \mathbf{b}) = .0001$. For a code alphabet of size $D = 2$, the Shannon codeword lengths are $\ell_1 = 1$ and $\ell_2 = 14$, giving an expected length of $\mathbb{E}[\ell(C(X))] = 1.0013$. But the code $C(\mathbf{a}) = 0$ and $C(\mathbf{b}) = 1$ has lengths $\ell_1 = 1$ and $\ell_2 = 1$, and its expected length is $\mathbb{E}[\ell(C(X))] = 1$.

^c Observe that $\ell_i := \lceil \log_D \frac{1}{p(x_i)} \rceil$ satisfies $\ell_i \geq \log_D \frac{1}{p(x_i)} > \ell_i - 1$. Multiplying this by minus one yields $-\ell_i \leq \log_D p(x_i) < -\ell_i + 1$, and (3.7) follows.

3.9.1. Shannon Codes under Mismatch [3]

Theorem 3.19 takes $\ell_i = \lceil \log_D \frac{1}{p(x_i)} \rceil$, where p is the pmf of X . Suppose we have $\tilde{\ell}_i := \lceil \log_D \frac{1}{q(x_i)} \rceil$ instead, where q is the “wrong” pmf, perhaps obtained by a statistical estimation procedure. Let us write

$$\begin{aligned} \sum_{i=1}^m \tilde{\ell}_i p(x_i) &\approx \sum_{i=1}^m p(x_i) \log_D \frac{1}{q(x_i)} \\ &= \frac{1}{\log D} \sum_{i=1}^m p(x_i) \log \frac{1}{q(x_i)} \\ &= \frac{1}{\log D} \sum_{i=1}^m p(x_i) \left[\log \frac{p(x_i)}{q(x_i)} + \log \frac{1}{p(x_i)} \right] \\ &= \frac{D(p\|q) + H(p)}{\log D}, \end{aligned}$$

where^d

$$D(p\|q) := \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(X)}{q(X)} \right]$$

is the **Kullback–Leibler informational divergence** or **relative entropy**. Since

$$\frac{1}{\log D} \log \frac{1}{q(x_i)} \leq \tilde{\ell}_i < \frac{1}{\log D} \log q(x_i) + 1,$$

we see that

$$\frac{D(p\|q) + H(p)}{\log D} \leq \sum_{i=1}^m \tilde{\ell}_i p(x_i) < \frac{D(p\|q) + H(p)}{\log D} + 1.$$

Thus, when the “wrong” lengths are used, the expected length of the code, where the expectation is over the correct pmf, satisfies the above inequality. Since $D(p\|p) = 0$, this recovers the result of Theorem 3.19.

3.10. Coding Sequences of Source Letters

3.10.1. A Simple Method

Suppose we have a uniquely decodable code $C: X \rightarrow D^*$ that satisfies

$$\frac{H(X)}{\log D} \leq \mathbb{E}[\ell(C(X))] < \frac{H(X)}{\log D} + 1.$$

^dThe notation $\mathbb{E}_p[g(X)]$ means that the expectation should be computed using the pmf p for X .

How might we encode $X_1, \dots, X_n$ if they are i.i.d. with the same pmf as X ? Consider the code $\tilde{C}_n(X_1, \dots, X_n) := C(X_1) \cdots C(X_n)$. Then $E[\ell(\tilde{C}_n(X_1, \dots, X_n))] = E[\ell(C(X_1)) + \cdots + \ell(C(X_n))] = nE[\ell(C(X_1))] = nE[\ell(C(X))]$. Substituting this in the middle of the above inequality, we see that

$$\frac{H(X)}{\log D} \leq \frac{E[\ell(\tilde{C}_n(X_1, \dots, X_n))]}{n} < \frac{H(X)}{\log D} + 1,$$

where the middle quotient is the **expected number of code letters per source letter**.

3.10.2. A Better Method

Theorem 3.21 (Variable-Length Source Coding Theorem). *Let X be a finite set. If $X_1, X_2, \dots$ is an X -valued DMS with common pmf p , then for every blocklength $n \geq 1$, there is a prefix code $C_n : X^n \rightarrow D^*$ such that*

$$\frac{H(p)}{\log D} \leq \frac{E[\ell(C_n(X_1, \dots, X_n))]}{n} < \frac{H(p)}{\log D} + \frac{1}{n}.$$

Proof. Applying Theorem 3.19 to $Y = (X_1, \dots, X_n)$, we have

$$\frac{H(Y)}{\log D} \leq E[\ell(C_n(Y))] < \frac{H(Y)}{\log D} + 1,$$

or, somewhat more explicitly,

$$\frac{H(Y)}{\log D} \leq E[\ell(C_n(X_1, \dots, X_n))] < \frac{H(Y)}{\log D} + 1,$$

Since the X_n are i.i.d., we know from Example 3.15 that

$$H(Y) = H(X_1 \cdots X_n) = H(X_1) + \cdots + H(X_n) = nH(p).$$

Making this substitution and dividing by n completes the proof. □

3.11. Optimal Variable-Length Prefix Codes (Huffman, 1952)

Example 3.22 (Cover and Thomas [3]). Consider the source alphabet with probabilities given in table at the left in Figure 3.5. The first step in constructing a binary Huffman code is to join the two source letters with the lowest probabilities as shown

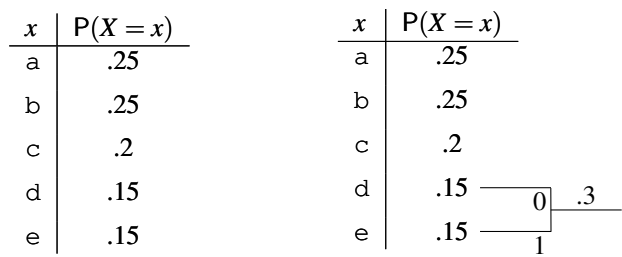


Figure 3.5. Probabilities for Example 3.22 (left). First step of Huffman code construction (right).

at the right in Figure 3.5. We label the upper symbol with a 0 and the lower symbol with a 1. We also add the two probabilities and put the result on the outgoing edge. Although we started with five source letters, we now think of having only four letters, where the fourth letter has probability .3. Of these four letters, we join the two with the lowest probabilities as shown at the left in Figure 3.6. We now think of having

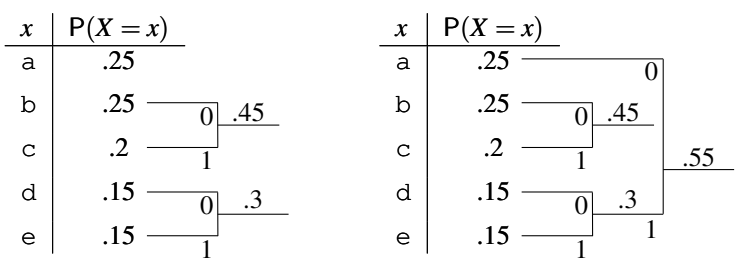


Figure 3.6. Huffman code construction step 2 (left) step 3 (right).

three source letters with probabilities .25, .45, and .3. Again we join the two with the lowest probabilities as shown at the right in Figure 3.6. It remains to join the last two probabilities as shown in Figure 3.7. When the probabilities sum to 1.0, we can then find the codewords. For each symbol, follow the tree from left to right reading off the labeled bits. For instance, the symbol d reads off 011. So the codeword is made up of these bits in reverse order to get 110. To conclude, we note that $E[\ell(C(X))]$ = 2.3 bits.



Example 3.23 (Cover and Thomas [3]). The construction of a ternary Huffman code is similar, except that at each stage, we join the three symbols of lowest probability as shown in Figure 3.8. In this case, $E[\ell(C(X))]$ = 1.5 ternary digits.



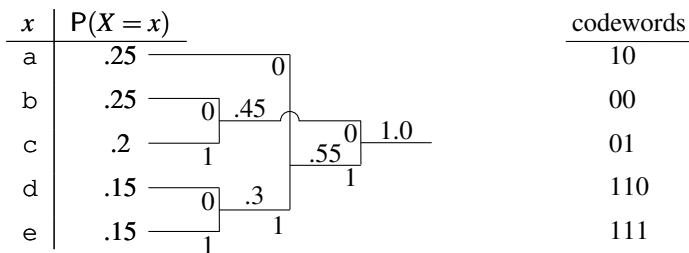


Figure 3.7. Fourth step of Huffman code construction.

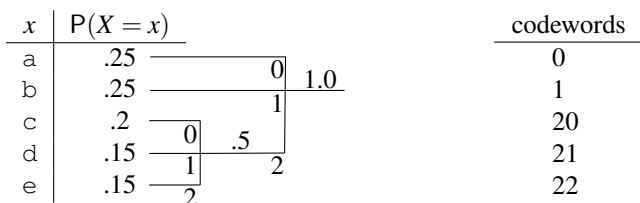


Figure 3.8. A ternary Huffman code.

Example 3.24 (Cover and Thomas [3]). For nonbinary Huffman codes, the number of source letters must be such that we can always take groups of three. In Figure 3.9 you can see that we have added a dummy symbol with zero probability. We find that $E[\ell(C(X))] = 1.7$ ternary digits.

Here is the general rule for adding dummy symbols. For $D \geq 3$, we may have to add dummy symbols of probability zero to the source alphabet. Each reduction of the Huffman procedure reduces the number of symbols by $D - 1$ until there are D symbols left. Thus,

$$|X| + \# \text{ dummy symbols} = 1 + k(D - 1),$$

where k is the number of reductions. For example, if $|X| = 5$ and $D = 3$, then we must solve

$$5 + ? = 1 + k(2), \text{ which implies } ? = 0,$$

while if $|X| = 6$ and $D = 3$, then we must solve

$$6 + ? = 1 + k(2), \text{ which implies } ? = 1.$$

x	$P(X=x)$		codewords
a	.25	0	0
b	.25		1
c	.2	0	20
d	.1		21
e	.1	0	220
f	.1		221
dummy	0	2	222

Figure 3.9. Another ternary Huffman code.

Problems

- Read the first two paragraphs about Morse code in Wikipedia.
 - When decoding Morse code, how do you know when the code for a letter is finished?
 - Is Morse code self-punctuating? Why or why not?
- Consider the variable-length code

x	$C(x)$
a	0
b	10
c	01
d	11

- Is it uniquely decodable?
 - Find $L(C)$ if all source symbols are equally likely.
- Show that

$$\frac{\ln \lfloor e^{nR} \rfloor}{n} \rightarrow R \text{ as } n \rightarrow \infty.$$

Hints: (i) Replace n by the continuous variable t . (ii) The inequality $x - 1 \leq \lfloor x \rfloor \leq x$ may be helpful.

- Show that $H(XY) = H(Y|X) + H(X)$.
- Recall that the entropy of a discrete random variable X is

$$H(X) = \sum_{x:p(x)>0} p(x) \log \frac{1}{p(x)} = - \sum_{x:p(x)>0} p(x) \log p(x),$$

where $p(x) := P(X = x)$ is the probability mass function of X . Show that $H(X) \geq 0$.

6. Let X be a finite set, and let X be an X -valued random variable with probability mass function p .

(a) Show that $H(p) \leq \log |X|$. *Hint:* Show that $H(p) - \log |X| \leq 0$.

(b) Show that $H(p) = \log |X|$ if and only if $p(x) = 1/|X|$ for all x .

(c) How that $H(p) = 0$ if and only if $p(x_0) = 1$ for some $x_0 \in X$, and $p(x) = 0$ for $x \neq x_0$.

7. Given nonnegative numbers a_i and b_i , put $a := \sum_i a_i$ and $b := \sum_i b_i$.

(a) Assuming $b < \infty$, show that

$$\sum_{i: 0 < a_i < b_i} a_i \log \frac{b_i}{a_i} < \infty.$$

(b) If $0 < a < \infty$ and $b < \infty$, derive the **log-sum inequality**

$$\sum_i a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b},$$

where the sum is understood as being over those i with $a_i > 0$.^e *Hint:* First note that if the left-hand side is infinite, there is nothing to prove. One way this can happen is if for some i with $a_i > 0$, we have $b_i = 0$. In this case, we take $a_i \log(a_i/0) = a_i \log \infty = a_i \cdot \infty = \infty$ since $a_i > 0$. Thus, the left-hand side is infinite and there is nothing to prove. The right-hand side is finite unless all the b_i are zero. But in this case, the inequality reduces to $\infty \geq \infty$, which is true. So, in your derivation, you may assume that the left-hand side is finite and that $b_i > 0$ for all i with $a_i > 0$.

(c) Show that if the two sides of the log-sum inequality are finite and equal, then $ba_i = ab_i$ for all i .

Remark. By the log inequality, $a \log(a/b) \geq a - b$. Hence,

$$\sum_i a_i \left[\log \frac{a_i}{b_i} - a_i + b_i \right] \geq 0.$$

^e The sum is more correctly understood as

$$\left(\sum_{i: a_i > 0 \text{ and } a_i \geq b_i} a_i \log \frac{a_i}{b_i} \right) - \left(\sum_{i: 0 < a_i < b_i} a_i \log \frac{b_i}{a_i} \right).$$

Since this second sum is finite by part (a), the left-hand side of the log-sum inequality is well defined, though it may be infinite.

This inequality often appears in the study of **nonnegative matrix factorization** [5], [13, eq. (3)], [14, eq. (2)].

8. **Kullback–Leibler Informational Divergence.** The Kullback–Leibler informational divergence or **relative entropy** of two pmfs p and q is

$$D(p\|q) := \sum_x p(x) \log \frac{p(x)}{q(x)},$$

where the sum is understood to be over those x with $p(x) > 0$. Use the **log-sum inequality** to show that $D(p\|q) \geq 0$, with equality if and only if $p(x) = q(x)$ for all x .

9. Show that if p is a pmf on a finite set X , then

$$H(p) \leq \log |X| \leq \frac{1}{|X|} \sum_x \log \frac{1}{p(x)}.$$

Hint: Let u denote the uniform pmf on X , and consider the quantities $D(p\|u)$ and $D(u\|p)$.

Remark. Both the left and right inequalities are achieved by $p = u$.

10. **The Binary Entropy Function.** If X has two points, say $X = \{a, b\}$, and if $p(a) = \theta$, then $p(b) = 1 - \theta$. Hence,

$$H(p) = \theta \log \frac{1}{\theta} + (1 - \theta) \log \frac{1}{1 - \theta} =: h(\theta),$$

where h is called the **binary entropy function**. It is understood that $h(0) = h(1) = 1$. We can also write $h(\theta) = -[\theta \log \theta + (1 - \theta) \log(1 - \theta)]$.

- For $0 < \theta < 1$, show that $h'(\theta) = \log(1 - \theta) - \log \theta$.
- For $0 < \theta < 1$, show that $h''(\theta) = (-\log e)/[\theta(1 - \theta)]$.
- Show that h has a maximum at $\theta = 1/2$.
- MATLAB.** The MATLAB function `binaryentropy` is given below. Run the commands

```
t = linspace(0,1,200);
y = binaryentropy(t);
plot(t,y)
```

to plot the binary entropy function:

```
function y = binaryentropy(p)
y = zeros(size(p));
i = find(0<p & p<1);
y(i) = -(p(i).*log2(p(i)) + (1-p(i)).*log2(1-p(i)));
```

11. **Entropy of the Geometric Distribution.** Let g_θ denote the geometric distribution on the positive integers; i.e., $g_\theta(k) = (1 - \theta)\theta^{k-1}$ for $k = 1, 2, \dots$, where $0 \leq \theta < 1$.

(a) Show that $H(g_\theta) = h(\theta)/(1 - \theta)$. *Hint:* The mean of the geometric distribution is $1/(1 - \theta)$; i.e., $\sum_{k=1}^{\infty} k(1 - \theta)\theta^{k-1} = 1/(1 - \theta)$.

(b) Show that $\lim_{\theta \rightarrow 1} H(g_\theta) = \infty$.

(c) **MATLAB.** Plot $H(g_\theta)$ as a function of θ .

(d) Put $f(\theta) := h(\theta)/(1 - \theta)$. Show that $f'(\theta) = (-\log \theta)/(1 - \theta)^2$.

12. **A Distribution with Infinite Entropy.**^f

(a) Use an integral comparison test to show that

$$\sum_{k=2}^{\infty} \frac{1}{k(\ln k)^2} < \infty.$$

Hint: Evaluate $(d/dt)(-1/\ln t)$.

(b) Use an integral comparison test to show that

$$\sum_{k=2}^{\infty} \frac{1}{k \ln k} = \infty.$$

Hint: Evaluate $(d/dt) \ln(\ln t)$.

(c) On account of part (a), there is a finite, positive constant C such that $p(k) := C/(k \ln^2 k)$ is a pmf on $\{2, 3, \dots\}$. Show that $H(p) = \infty$.

13. **A Sufficient Condition for Finite Entropy.** Show that if p is a distribution on the positive integers with $\sum_{k=1}^{\infty} k^r p(k) < \infty$ for some positive r , then $H(p) < \infty$. *Hints:* In the sum for $H(p)$, break the sum into two parts: $\{k \geq 1 : p(k) \leq e^{-k^r}\}$ and $\{k \geq 1 : p(k) > e^{-k^r}\}$. Use the fact that for $0 \leq t \leq e^{-1}$, the function $f(t) = -t \log t$ is increasing. The other fact that you will need is that $p(k) > e^{-k^r}$ is equivalent to $-\log p(k) \leq k^r (\log e)$.

14. Find a binary Huffman code and a ternary Huffman code for the seven-letter alphabet $X := \{a, b, c, d, e, f, g\}$ if

x	$P(X = x)$
a	.49
b	.26
c	.12
d	.04
e	.04
f	.03
g	.02

^fThis construction was inspired by Baer [1].

In each case, compute the expected length of the code.

15. Consider the first Huffman code example. Draw a binary tree as follows. From the root, draw one edge and label its end with the source letters whose first bit is a one. Draw the other edge and label its end with the other source letters. From each group of source letters, draw two edges. At then ends, put the source letters whose second bit is a one/zero. Continue until each source letter is at the end of its own edge.

Remark. The point of this problem is that a prefix code defines a way of successively partitioning a set until each partition set contains only one letter. It is a way of playing “20 Questions.” If you know the item probabilities, then the best way to play is to ask questions according to the Huffman code.

16. **MATLAB.** The MATLAB function `entropy2` is given below. Put it in an M-file and use the commands

```
p = [ 1/6 1/3 1/2 ];
H = entropy2(p)
```

to compute the entropy of $p = (1/6, 1/3, 1/2)$. How does your answer compare with $\log_2(3)$?

```
function H = entropy2(pin,varargin);
% H = entropy2(p,q) = -sum_{i:q(i)>0} q(i)*log2(p(i))
% If q=p, use H = entropy2(p).
% p and q can be matrices; e.g.,
%
% -sum_{i,j} q(i,j)*log p(j|i)
%
% where q(i,j) may be the joint pmf.
%
% You can also compute the relative entropy or
% informational divergence D(q||p) = entropy2(p./q,q)
if nargin > 1
    qin = varargin{1};
else
    qin = pin;
end
lt = prod(size(pin));
p = reshape(pin,lt,1); % convert pin to column vector
q = reshape(qin,1,lt); % convert qin to row vector
i = find(q>0);
H = - q(i) * log2( p(i) );
```

17. **MATLAB.** The MATLAB function `relent` is given below. Put it in an M-file and use the commands

```
p = [ 1/6 1/3 1/2 ];  
q = [ 7/10 1/5 1/10];  
D = relent(p,q)
```

to compute $D(p\|q)$, where $p = (1/6, 1/3, 1/2)$ and $q = (7/10, 1/5, 1/10)$.

```
function D = relent(p,q)  
% Compute the relative entropy or Kullback-Leibler  
% informational divergence sum_i p(i) log2(p(i)/q(i)).  
D = entropy2(q./p,p);
```

CHAPTER 4

Data Compression Using Fixed-Length Codes

4.1. Introduction

Example 4.1. Let X_1, X_2, X_3 be i.i.d. with $P(X_i = 1) = p$ where, $0 \leq p \leq 1$. Consider the compression scheme $(x_1, x_2, x_3) \rightarrow (y_1, y_2)$ defined by the table

$x_1 x_2 x_3$	$y_1 y_2$
0 0 0	0 0
0 0 1	0 1
0 1 0	1 0
1 0 0	1 1
0 1 1	0 0
1 0 1	0 0
1 1 0	0 0
1 1 1	0 0

We use the decompression scheme $(y_1, y_2) \rightarrow (\hat{x}_1, \hat{x}_2, \hat{x}_3)$ defined by the table

$y_1 y_2$	$\hat{x}_1 \hat{x}_2 \hat{x}_3$
0 0	0 0 0
0 1	0 0 1
1 0	0 1 0
1 1	1 0 0

At first glance, this may seem like a terrible way to compress data. After all, half of the input sequences cannot be correctly recovered. However, notice that the probability of error is $p^3 + 3p^2(1-p)$. If $p = 1/1000$, then the error probability is 2.998×10^{-6} , or about 3×10^{-6} . If $p = 0$, then the system never makes an error.

4.2. Achievable Rates for Source Codes

Recall that for variable-length codes, the goal was to minimize the expected code-length, $E[\ell(C_n(X_1, \dots, X_n))]$. For fixed-length codes we take a different point of view. Suppose $|X| < \infty$. Then there are $|X|^n = \exp(n \log |X|)$ different possible realizations

of the random source sequence $(X_1, \dots, X_n)$. Our goal here is to find a subset of X^n that has small cardinality and large probability. Somewhat more precisely, we want to find a set $A_n \subset X^n$ with

$$|A_n| < \exp(n \log |X|) \quad \text{and} \quad P((X_1, \dots, X_n) \in A_n) \approx 1.$$

Our results will be *asymptotic* in that we obtain both conditions together only when the blocklength n is sufficiently large.

Definition 4.2. A real number $R \geq 0$ is an **achievable rate** for a source $X_1, X_2, \dots$, if $\forall \lambda > 0, \forall \Delta R > 0, \exists n_0$ such that $\forall n \geq n_0, \exists A_n \subset X^n$ with

$$\frac{\log |A_n|}{n} < R + \Delta R \quad \text{and} \quad P((X_1, \dots, X_n) \notin A_n) < \lambda.$$

The left-hand inequality is equivalent to $|A_n| < \exp(n[R + \Delta R])$.

Implications of the Definition

The first thing to point out is that $R = \log |X|$ is always an achievable rate. To see this, take $A_n = X^n$ and write

$$\frac{\log |A_n|}{n} = \frac{\log |X^n|}{n} = \frac{n \log |X|}{n} = \log |X| < \log |X| + \Delta R,$$

and

$$P((X_1, \dots, X_n) \notin A_n) = P((X_1, \dots, X_n) \notin X^n) = 0 < \lambda.$$

Thus, there is no question that there *are* achievable rates. We will show later that if R is an achievable rate, then any $R' > R$ is also achievable. Hence, any $R \geq \log |X|$ is an achievable rate. In the next section, we investigate under what conditions there are any numbers less than $\log |X|$ that are achievable.

4.3. Shannon's Source Coding Theorem

Theorem 4.3 (Shannon's Source Coding Theorem). *For a DMS with pmf p , we have the following.*

- (i) (Forward part): Every rate $R \geq H(p)$ is achievable
(ii) (Converse part): Every rate $0 \leq R < H(p)$ is not achievable.

Proof. (Forward part). We show that every $R > H(p)$ is achievable. Since any limit of achievable rates is achievable (Lemma 4.5 below), it will follow that $H(p) = \lim_{k \rightarrow \infty} H(p) + 1/k$ is also achievable. For $\Delta R > 0$, $\lambda > 0$, and $n = 1, 2, \dots$, consider the set

$$A_n := \left\{ (x_1, \dots, x_n) \in \mathcal{X}^n : \left| \frac{1}{n} \sum_{k=1}^n \log \frac{1}{p(x_k)} - H(p) \right| \leq \Delta R \right\}.$$

By the **weak law of large numbers (WLLN)** [7, p. 576],^a

$$\mathbb{P}((X_1, \dots, X_n) \notin A_n) = \mathbb{P}\left(\left| \frac{1}{n} \sum_{k=1}^n \log \frac{1}{p(X_k)} - H(p) \right| > \Delta R\right) \rightarrow 0.$$

Hence, there is an n_0 such that for all $n \geq n_0$,

$$\mathbb{P}((X_1, \dots, X_n) \notin A_n) < \lambda.$$

Using the notation $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{X} = (X_1, \dots, X_n)$, observe that

$$\begin{aligned} \mathbf{x} \in A_n &\Leftrightarrow -\Delta R \leq \frac{1}{n} \sum_{k=1}^n \log \frac{1}{p(x_k)} - H(p) \leq \Delta R \\ &\Leftrightarrow H(p) - \Delta R \leq \frac{1}{n} \log \frac{1}{p(x_1) \cdots p(x_n)} \leq H(p) + \Delta R \\ &\Leftrightarrow \exp(-n[H(p) - \Delta R]) \geq \mathbb{P}(\mathbf{X} = \mathbf{x}) \geq \exp(-n[H(p) + \Delta R]). \end{aligned}$$

Hence,

$$\begin{aligned} 1 \geq \mathbb{P}(\mathbf{X} \in A_n) &= \sum_{\mathbf{x} \in A_n} \mathbb{P}(\mathbf{X} = \mathbf{x}) \\ &\geq \sum_{\mathbf{x} \in A_n} \exp(-n[H(p) + \Delta R]) \\ &= |A_n| \exp(-n[H(p) + \Delta R]). \end{aligned}$$

^aThe WLLN says that if $Y_1, Y_2, \dots$ are i.i.d. with finite mean $\mathbb{E}[Y_1]$, then for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{k=1}^n Y_k - \mathbb{E}[Y_1] \right| > \varepsilon\right) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Rearranging this yields $\exp(n[H(p) + \Delta R]) \geq |A_n|$, or

$$\frac{\log |A_n|}{n} \leq H(p) + \Delta R < R + \Delta R.$$

Thus, $R > H(p)$ is an achievable rate.

(Converse part). Suppose $0 \leq R < H(p)$ and $0 < \lambda < 1$. To obtain a contradiction, suppose that R is achievable. Since $0 \leq R < H(p)$, we can choose $\Delta R > 0$ so small that $R + \Delta R < H(p) - 2\Delta R$; i.e., $0 < \Delta R < (H(p) - R)/3$. Furthermore, $\exists n_0$ such that $\forall n \geq n_0$, $\exists G_n \subset X^n$ with

$$|G_n| < \exp(n[R + \Delta R]) \leq \exp(n[H(p) - 2\Delta R]) \quad \text{and} \quad P(\mathbf{X} \notin G_n) < \lambda. \quad (4.1)$$

We may further assume n is so large that

$$\exp(-n\Delta R) < (1 - \lambda)/2 \quad \text{and} \quad P(\mathbf{X} \notin A_n) < (1 - \lambda)/2,$$

where A_n is defined as before. Of course, we also have $P(\mathbf{X} = \mathbf{x}) \leq \exp(-n[H(p) - \Delta R])$ for $\mathbf{x} \in A_n$. Next write

$$\begin{aligned} P(\mathbf{X} \in G_n) &= P(\mathbf{X} \in [G_n \cap A_n]) + P(\mathbf{X} \in [G_n \cap A_n^c]) \\ &\leq P(\mathbf{X} \in [G_n \cap A_n]) + P(\mathbf{X} \in A_n^c) \\ &\leq P(\mathbf{X} \in [G_n \cap A_n]) + (1 - \lambda)/2. \end{aligned}$$

We also have

$$\begin{aligned} P(\mathbf{X} \in [G_n \cap A_n]) &= \sum_{\mathbf{x} \in G_n \cap A_n} P(\mathbf{X} = \mathbf{x}) \\ &\leq \sum_{\mathbf{x} \in G_n \cap A_n} \exp(-n[H(p) - \Delta R]) \\ &\leq \sum_{\mathbf{x} \in G_n} \exp(-n[H(p) - \Delta R]) \\ &= |G_n| \exp[-n(H(p) - \Delta R)] \\ &\leq \exp(n[H(p) - 2\Delta R]) \exp(-n[H(p) - \Delta R]) \\ &= \exp(-n\Delta R) \\ &< (1 - \lambda)/2. \end{aligned}$$

Hence, for large n , we have $P(\mathbf{X} \in G_n) \leq 1 - \lambda$, which contradicts the right-hand side of (4.1). $\square$

4.3.1. The Asymptotic Equipartition Property and Typical Sequences

The key ideas in the proof of the forward part of Shannon's Source Coding Theorem are summarized in the following result, except for Property (iv).

Theorem 4.4 (Asymptotic Equipartition Property (AEP)). *Using the notation $p^n(\mathbf{x}) := p(x_1) \cdots p(x_n)$, the set A_n has the following properties:*

- (i) For all n , $|A_n| \leq \exp(n[H(p) + \Delta R])$.
- (ii) For all n and $\mathbf{x} \in A_n$, $\exp(-n[H(p) + \Delta R]) \leq p^n(\mathbf{x}) \leq \exp(-n[H(p) - \Delta R])$.
- (iii) For all sufficiently large n , $p^n(A_n) \approx 1$.
- (iv) Given $0 < \lambda < 1$, for all sufficiently large n , $|A_n| \geq (1 - \lambda) \exp(n[H(p) - \Delta R])$.

Proof. Only the fourth property needs justification. Given $0 < \lambda < 1$, by (iii), we have $p^n(A_n) > 1 - \lambda$ for all sufficiently large n . Hence,

$$\begin{aligned} 1 - \lambda &\leq p^n(A_n) = \sum_{\mathbf{x} \in A_n} p^n(\mathbf{x}) \\ &\leq \sum_{\mathbf{x} \in A_n} \exp(-n[H(p) - \Delta R]), \quad \text{by (ii),} \\ &= |A_n| \exp(-n[H(p) - \Delta R]). \end{aligned} \quad \square$$

Discussion

Property (iii) of the AEP says that a random source block $(X_1, \dots, X_n)$ lies in A_n with probability close to one; for this reason, the sequences $\mathbf{x} \in A_n$ are called **typical**. Property (ii) says that for $\mathbf{x} \in A_n$, $P(\mathbf{X} = \mathbf{x}) \approx \exp[-nH(p)]$. Properties (i) and (iv) together say that $|A_n| \approx \exp[nH(p)]$. Hence, the sequences $\mathbf{x} \in A_n$ are nearly uniformly distributed.

4.3.2. Interpretation of Fixed-Length Codes for Data Compression

Suppose $\mathcal{X} = \{0, 1\}$ and that $\log$ and $\exp$ are to the base 2. Then there are 2^n possibilities for the source sequence $(X_1, \dots, X_n)$. Suppose $|A_n| = 2^{nR}$, where $R \leq \log |\mathcal{X}| = \log_2 2 = 1$. If $R = 3/4$ and $n = 16$, $|A_n| = 2^{16(3/4)} = 2^{12}$. Thus, if $(X_1, \dots, X_n) \in A_n$, we can represent this 16-bit source sequence by a 12-bit codeword. Of course, there is a positive probability (hopefully small) that $(X_1, \dots, X_n) \notin A_n$. In this case there is an error. The Source Coding Theorem tells for what rates R the probability of an error can be made small as $n \rightarrow \infty$, namely, rates greater than $\text{Fix } 0 < \lambda < 1$ or equal to $H(p)$. Thus, the minimum ratio of codeword length to number

of input symbols is $nR/n = R \geq H(p)$. Compare this to the Variable-Length Source Coding Theorem 3.21, which says, with $D = 2$,

$$H(p) \leq \frac{E[\ell(C_n(X_1, \dots, X_n))]}{n} < H(p) + \frac{1}{n}.$$

Thus, the *expected* length for variable-length codes (per source letter) can also be made close to $H(p)$. However, $\ell(C_n(X_1, \dots, X_n))/n$ is a random variable, and with positive probability could be much greater than $H(p)$.

To summarize the down side differences:

- Fixed length codes: occasional errors.
- Variable length codes: occasional long delays.

4.4. Properties of Achievable Rates

Let $\mathcal{R} := \{R : R \text{ is achievable}\}$.

Lemma 4.5. $\mathcal{R}$ is a closed set.

Proof. We must show that if $R_k \in \mathcal{R}$ is a sequence of achievable rates and if $R_k \rightarrow R$, then $R \in \mathcal{R}$; i.e., a limit of achievable rates is also achievable. To show $R = \lim_{k \rightarrow \infty} R_k$ is achievable, let $\Delta R > 0$ and $\lambda > 0$ be given. Let $k \geq K$ imply $|R_k - R| < \Delta R/2$. Note that $|R_k - R| < \Delta R/2 \Rightarrow R_k < R + \Delta R/2$. Since R_k is achievable, $\exists n_0 \forall n \geq n_0, \exists A_n \subset X^n$ with

$$\frac{\log |A_n|}{n} < R_k + \frac{\Delta R}{2} \quad \text{and} \quad P((X_1, \dots, X_n) \notin A_n) < \lambda,$$

Which implies

$$\frac{\log |A_n|}{n} < \left(R + \frac{\Delta R}{2}\right) + \frac{\Delta R}{2} = R + \Delta R \quad \text{and} \quad P((X_1, \dots, X_n) \notin A_n) < \lambda.$$

Hence R is achievable. □

Definition 4.6. A subset C of a vector space V is **convex** if $\forall x, y \in C, \forall 0 \leq \alpha \leq 1, \alpha x + (1 - \alpha)y \in C$. In other words C is convex if $\forall x, y \in C, C$ contains the line joining x and y . Examples of convex and nonconvex sets are shown in Figure 4.1.

Lemma 4.7. $\mathcal{R}$ is convex.

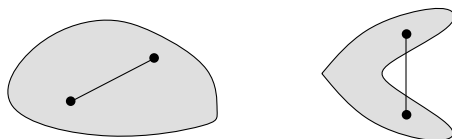


Figure 4.1. A convex set (left) and a nonconvex set (right).

Proof. We prove a stronger result: If $R \in \mathcal{R}$ and $R' > R$, then $R' \in \mathcal{R}$. In other words, if R is achievable, all rates greater than R are also achievable. Let R be achievable, and let $R' > R$. Since R is achievable, given any $\lambda > 0$ and $\Delta R > 0$, $\exists n_0 \forall n \geq n_0, \exists A_n \subset X^n$ with

$$\frac{\log |A_n|}{n} < R + \Delta R < R' + \Delta R \quad \text{and} \quad P((X_1, \dots, X_n) \notin A_n) < \lambda.$$

Hence, R' is achievable. □

Corollary 4.8. $\mathcal{R}$ is closed and convex.

For an i.i.d. source with common pmf p , Shannon's Source Coding Theorem shows that $\mathcal{R} = [H(p), \infty)$, which is closed and convex.

Problems

- The Set of Distributions is Convex.** Let X be a finite or countably infinite set. We denote by $\mathcal{D}(X)$ the set of all pmfs on X . In other words, $\mathcal{D}(X)$ is the set of all *nonnegative* functions p on X with the property that $\sum_x p(x) = 1$. Show that $\mathcal{D}(X)$ is convex; i.e., show that if $p, q \in \mathcal{D}(X)$ and if $0 \leq \lambda \leq 1$, then $r := \lambda p + (1 - \lambda)q$ also belongs to $\mathcal{D}(X)$.
- The Unit Simplex.** If $|X| = n$, we can identify $\mathcal{D}(X)$ with the subset of vectors in $\mathbb{R}^n$ having nonnegative components that sum to one; i.e.,

$$\Delta^{n-1} := \left\{ p = (p_1, \dots, p_n) \in \mathbb{R}^n : p_k \geq 0 \text{ and } \sum_{k=1}^n p_k = 1 \right\}.$$

The set Δ^{n-1} is called the **unit simplex** in $\mathbb{R}^n$. Sketch the unit simplex in $\mathbb{R}^2$ and in $\mathbb{R}^3$.

- Concavity of Entropy.** Let

$$f(t) := \begin{cases} -t \log t, & 0 < t \leq 1, \\ 0, & t = 0. \end{cases}$$

Then if p is a pmf on X , we can write $H(p) = \sum_x f(p(x))$. Use the fact that f is a concave function of $t \in [0, 1]$ to show that H is a concave function of $p \in \mathcal{D}(X)$; i.e., show that for $0 \leq \lambda \leq 1$,

$$H(\lambda p + (1 - \lambda)q) \geq \lambda H(p) + (1 - \lambda)H(q).$$

4. Let p be a pmf on X , and let $W(y|x)$ be a **transition probability** from X into Y . In other words, for each fixed $x \in X$, $W(y|x)$ is a pmf in the variable $y \in Y$. Define a marginal pmf on Y by

$$(pW)(y) := \sum_x p(x)W(y|x).$$

Show that $g(p) := H(pW)$ is a concave function of p .

5. Let X be a countably infinite set. We saw in Problem 3.12 that some pmfs have infinite entropy. Let $\widehat{\mathcal{D}}(X)$ denote the set of all pmfs on X that have finite entropy. Show that $\widehat{\mathcal{D}}(X)$ is a convex set.

6. **Distance Properties of Distributions.** We can define a notion of **distance** between pmfs $p, q \in \mathcal{D}(X)$. We put^b

$$d(p, q) := \sum_x |p(x) - q(x)|.$$

It is obvious that $d(p, q) = d(q, p)$ and that $d(p, q) = 0$ if and only if $p(x) = q(x)$ for all x .

- (a) Show that d satisfies the **triangle inequality**; i.e., if $p, q, r \in \mathcal{D}(X)$, then

$$d(p, q) \leq d(p, r) + d(r, q).$$

- (b) Show that $d(p, q) \leq 2$.

- (c) If $r = \lambda p + (1 - \lambda)q$, show that

$$d(r, p) \leq (1 - \lambda)d(p, q) \leq 2(1 - \lambda) \quad \text{and} \quad d(r, q) \leq \lambda d(p, q) \leq 2\lambda.$$

7. (a) Show that $d(p_n, p) \rightarrow 0$ implies $p_n(x) \rightarrow p(x)$ for each $x \in X$.
 (b) Let $p_n \in \mathcal{D}(X)$, and suppose that for each $x \in X$, the sequence of numbers $p_n(x)$ converges to a limit, which we denote by $p(x)$. If X is a finite set, show that $p \in \mathcal{D}(X)$ and that $d(p_n, p) \rightarrow 0$.

^b Sometimes d is called the **variational distance**. When $|X| = n$, we can regard $\mathcal{D}(X)$ as a subset of $\mathbb{R}^n$ and write $d(p, q) = \|p - q\|_1$, where $\|\cdot\|_1$ denotes the 1-norm on $\mathbb{R}^n$. Similarly, when X is countably infinite, we can regard $\mathcal{D}(X)$ as a subset of ℓ^1 (the set of all absolutely summable sequences) and write $d(p, q) = \|p - q\|_1$, where now $\|\cdot\|_1$ is the 1-norm on ℓ^1 .

- (c) Show that in part (b), the assumption $|\mathcal{X}| < \infty$ is essential. *Hint:* Give an example of p_n on a countably infinite set $\mathcal{X}$ for which $\lim_{n \rightarrow \infty} p_n(x) = 0$ for all x .
- (d) Let p_n be any sequence in $\mathcal{D}(\mathcal{X})$, where $\mathcal{X}$ is a finite set. Show that there is a subsequence p_{n_k} that converges to a limit $p \in \mathcal{D}(\mathcal{X})$. *Hints:* Assume $|\mathcal{X}| = 2$ and apply the Bolzano–Weierstrass Theorem.^c

8. Continuity of Entropy. Let f be as in Problem 4.3. Observe that from Problem 1.2, f is a continuous function of $t \in [0, 1]$. When $\mathcal{X}$ is a finite set, use the continuity of f on $[0, 1]$ to show that H is a continuous function on $\mathcal{D}(\mathcal{X})$. In other words, given $p \in \mathcal{D}(\mathcal{X})$ and $\varepsilon > 0$, show that there is a $\delta > 0$ such that if $q \in \mathcal{D}(\mathcal{X})$ and $d(p, q) < \delta$, then $|H(p) - H(q)| < \varepsilon$.

9. The notion of distance between pmfs on $\mathcal{X}$ can be extended to pmfs on $\mathcal{X} \times \mathcal{Y}$ by

$$d(r, r') = \sum_{x,y} |r(x, y) - r'(x, y)|,$$

where $r, r' \in \mathcal{D}(\mathcal{X} \times \mathcal{Y})$. If we define the marginal pmfs

$$p(x) := \sum_y r(x, y) \quad \text{and} \quad p'(x) := \sum_y r'(x, y),$$

show that

$$d(p, p') \leq d(r, r').$$

10. Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets, and let $P_{XY}(x, y)$ be a joint pmf on $\mathcal{X} \times \mathcal{Y}$. Show that $H(X|Y) = H(P_{X|Y})$ is a continuous function of P_{XY} . More specifically, given $\varepsilon > 0$, show that there is a $\delta > 0$ such that

$$d(P_{XY}, Q_{XY}) < \delta \Rightarrow |H(P_{X|Y}) - H(Q_{X|Y})| < \varepsilon.$$

Hint: Recall that $H(X|Y) = H(XY) - H(Y)$.

11. Let $P_{XY}(x, y)$ and $Q_{XY}(x, y)$ be two joint pmfs on $\mathcal{X} \times \mathcal{Y}$. Determine whether or not $D(P_X \| Q_X) \leq D(P_{XY} \| Q_{XY})$.

^cThe Bolzano–Weierstrass Theorem says that if t_n is a bounded sequence of real numbers, then there is a converging subsequence t_{n_k} . Note that the definition of a subsequence requires that $n_k \rightarrow \infty$ as $k \rightarrow \infty$.

CHAPTER 5

Introduction to Channel Coding

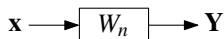
The notion of a channel as a transition probability (i.e., a conditional probability) is developed. This is followed by the general development of encoders and decoders.

5.1. Channels

Given a set $\mathbf{X}$ of **channel input symbols**, a set $\mathbf{Y}$ of **channel output symbols**, and a positive integer n , a **channel** W_n is a **transition probability** from $\mathbf{X}^n$ into $\mathbf{Y}^n$. In other words, if $\mathbf{X} = (X_1, \dots, X_n)$ is an $\mathbf{X}^n$ -valued random variable and $\mathbf{Y} = (Y_1, \dots, Y_n)$ is a $\mathbf{Y}^n$ -valued random variable, then

$$W_n(B|\mathbf{x}) = P(\mathbf{Y} \in B | \mathbf{X} = \mathbf{x}), \quad B \subset \mathbf{Y}^n,$$

is the conditional probability that the random sequence $\mathbf{Y}$ of channel output symbols lies in B given that the channel input sequence is $\mathbf{X} = \mathbf{x}$.



Discrete Channels

If $\mathbf{Y}$ is a finite set, then so is $\mathbf{Y}^n$. In this case, W_n is given by a conditional probability mass function, which we also denote by W_n , and we have

$$W_n(B|\mathbf{x}) = \sum_{\mathbf{y} \in B} W_n(\mathbf{y}|\mathbf{x}).$$

In many applications, W_n is not explicitly given, but is defined implicitly.

Example 5.1. Let $\mathbf{X} = \mathbf{Y}$ denote the integers $\{0, \dots, m-1\}$ under mod- m addition. Let the components of $\mathbf{Z} := (Z_1, \dots, Z_n)$ be i.i.d. $\mathbf{X}$ -valued random variables with common pmf $p_Z(z) = P(Z = z)$ for $z \in \mathbf{X}$. Let $\mathbf{X}$ be independent of $\mathbf{Z}$, and put $\mathbf{Y} := \mathbf{X} + \mathbf{Z}$, where the componentwise addition is mod- m . Then

$$\begin{aligned} W_n(\mathbf{y}|\mathbf{x}) &:= P(\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{x}) = P(\mathbf{X} + \mathbf{Z} = \mathbf{y} | \mathbf{X} = \mathbf{x}) \\ &= P(\mathbf{x} + \mathbf{Z} = \mathbf{y} | \mathbf{X} = \mathbf{x}), && \text{by substitution,} \\ &= P(\mathbf{Z} = \mathbf{y} - \mathbf{x}), && \text{by independence,} \\ &= \prod_{k=1}^n p_Z(y_k - x_k), && \text{by independence.} \end{aligned}$$

Example 5.2 (Discrete Memoryless Channel (DMC)). Let us temporarily write $X^k := (X_1, \dots, X_k)$ and similarly for x^k , Y^k , and y^k . Then $W_n(\mathbf{y}|\mathbf{x}) = W_n(y^n|x^n)$. Using the definition of conditional probability,

$$\begin{aligned} W_n(y^n|x^n) &= P(Y^n = y^n | X^n = x^n) \\ &= P(Y_n = y_n | Y^{n-1} = y^{n-1}, X^n = x^n) \\ &\quad \cdot \frac{P(X_n = x_n | Y^{n-1} = y^{n-1}, X^{n-1} = x^{n-1})}{P(X_n = x_n | X^{n-1} = x^{n-1})} W_{n-1}(y^{n-1}|x^{n-1}). \end{aligned}$$

Let us consider the fractional factor. If we assume that the numerator is equal to the denominator, then we are saying that the channel input at time n , X_n , does not depend on any past channel outputs $Y^{n-1} = (Y_1, \dots, Y_{n-1})$; in other words, there is no **feedback** of past channel output symbols to the encoder. Under this assumption,

$$W_n(y^n|x^n) = P(Y_n = y_n | Y^{n-1} = y^{n-1}, X^n = x^n) W_{n-1}(y^{n-1}|x^{n-1}).$$

If we next assume that $P(Y_n = y_n | Y^{n-1} = y^{n-1}, X^n = x^n)$ does not depend on y^{n-1} or x^{n-1} , then this probability is equal to $P(Y_n = y_n | X_n = x_n)$ [7, Example 12.1]. This assumption says that the channel is **memoryless**. If we apply the same assumptions to $W_{n-1}(y^{n-1}|x^{n-1})$ and so on, we arrive at

$$W_n(y^n|x^n) = \prod_{k=1}^n P(Y_k = y_k | X_k = x_k).$$

If the function of x and y defined by $P(Y_k = y | X_k = x)$ does not depend on k , then we say that the channel is **stationary**, and we put $W(y|x) := P(Y_k = y | X_k = x)$. With all of these assumptions,

$$W_n(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^n W(y_k|x_k) =: W^n(\mathbf{y}|\mathbf{x}).$$

This is the **discrete memoryless channel** (DMC). Hence, it is important to realize that the DMC terminology implies not only memorylessness, but also stationarity and no feedback. The channel of Example 5.1 is a DMC.

Example 5.3 (Binary Symmetric Channel (BSC)). This channel is obtained by taking $m = 2$ in Example 5.1 and putting $p_Z(1) = \varepsilon$. This results in the probability of flipping a bit being ε . For this reason, we say that this is a BSC with **crossover probability** ε . This is illustrated in Figure 5.1.

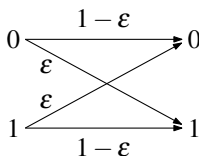


Figure 5.1. Binary symmetric channel with crossover probability ε .

Continuous Channels

If $\mathcal{Y} = \mathbb{R}$, then we take W_n to be given by a conditional probability density function, which we denote by w_n , and we have

$$W_n(B|\mathbf{x}) = \int_B w_n(\mathbf{y}|\mathbf{x}) d\mathbf{y}.$$

Here too, w_n is often given implicitly.

Example 5.4. Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}$. Let the components of $\mathbf{Z} := (Z_1, \dots, Z_n)$ be i.i.d. real-valued random variables with common probability density $p_Z(z)$. Let $\mathbf{X}$ be independent of $\mathbf{Z}$, and put $\mathbf{Y} := \mathbf{X} + \mathbf{Z}$. Then

$$\begin{aligned} P(\mathbf{Y} \leq \mathbf{y} | \mathbf{X} = \mathbf{x}) &= P(\mathbf{X} + \mathbf{Z} \leq \mathbf{y} | \mathbf{X} = \mathbf{x}) = P(\mathbf{Z} \leq \mathbf{y} - \mathbf{x}) = \prod_{k=1}^n P(Z_k \leq y_k - x_k) \\ &= \prod_{k=1}^n \int_{-\infty}^{y_k - x_k} p_Z(z_k) dz_k. \end{aligned}$$

Differentiating with respect to the y_k , we have

$$w_n(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^n p_Z(y_k - x_k).$$

This is an example of continuous memoryless channel.

For continuous memoryless channels, the most common choice for p_Z is a **Gaussian or normal density function**, which results in the **discrete-time Gaussian channel**. The standard Gaussian density is

$$\xi(z) := \frac{e^{-z^2/2}}{\sqrt{2\pi}}.$$

It has zero mean and unit variance. The Gaussian density with mean m and variance σ^2 is given by

$$p_Z(z) = \xi \left(\frac{z-m}{\sigma} \right) / \sigma.$$

In this case we write $p_Z \sim N(m, \sigma^2)$. Of course, $\xi \sim N(0, 1)$.

Waveform Channels

Example 5.5. Take $\mathsf{X} := \mathbb{R}$, and consider a unit-energy signaling waveform s on some time interval, say $[0, T]$. The transmitter applies a gain $x \in \mathbb{R}$ to the signaling waveform s so that the waveform $r(t) := xs(t) + N(t)$ arrives at the destination, where N is **additive white Gaussian noise (AWGN)**, which we take to be zero mean. The receiver front end applies r to a correlator (or sampled lowpass filter) to produce the effective channel output

$$Y := \int_0^T r(t)s(t) dt = x + Z.$$

where Z is a zero-mean Gaussian random variable. Since the AWGN in different time intervals is independent, signaling in the manner over n time intervals yields a memoryless channel like the one in Example 5.4.

Example 5.6. Take $\mathsf{X} := \{0, 1\}$, and consider two equal-energy waveforms s_0 and s_1 on the time interval $[0, T]$. When $x \in \mathsf{X} = \{0, 1\}$ is applied to the channel input, the waveform $r(t) := s_x(t) + N(t)$ arrives at the destination, where N is again zero-mean AWGN. This time the receiver front end produces the effective channel output^a

$$Y := \begin{cases} 1, & \text{if } \int_0^T r(t)s_1(t) dt \geq \int_0^T r(t)s_0(t) dt, \\ 0, & \text{otherwise.} \end{cases}$$

Again, since the AWGN in different time intervals is independent, signaling in this manner yields a DMC. In fact, it is a BSC.

5.2. Encoders and Decoders

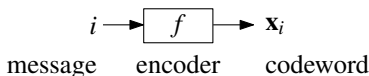
Encoders

A mapping $f: \{1, \dots, N\} \rightarrow \mathsf{X}^n$ is called an **encoder**. The integers $1, \dots, N$ are called **messages**.^b The corresponding **codewords** are denoted by $\mathbf{x}_i := f(i)$. Thus,

^aThe optimality of this procedure is usually established in a course on detection and estimation or in a course on digital communication systems.

^bOccasionally, we use $\{0, \dots, N-1\}$ for the set of messages.

each $\mathbf{x}_i \in \mathcal{X}^n$. The collection $(\mathbf{x}_1, \dots, \mathbf{x}_N)$ is the corresponding **codebook**. Knowing the codebook $(\mathbf{x}_1, \dots, \mathbf{x}_N)$ is the same as knowing the encoder f , and we sometimes just write $f = (\mathbf{x}_1, \dots, \mathbf{x}_N)$, which is an element of $(\mathcal{X}^n)^N$.



Code Rates

If there are $N = 2^k$ messages, then each message can be thought of as a k -bit word. Even for N not a power of 2, we think of each message as consisting of $\log_2 N$ bits. If we use natural logarithms, we say that each message consists of $\ln N$ **nats**. We often do not specify the base of the logarithm and just write $\log$. In this case, we use $\exp$ for the inverse of $\log$. Hence, if it is understood that the logarithm base is b , then $\exp(x) = b^x$. In particular, if the logarithm base is 2, then $\exp(x) = 2^x$, and if the logarithm base is e , then $\exp(x) = e^x$.

Given a code $f: \{1, \dots, N\} \rightarrow \mathcal{X}^n$, or equivalently a codebook $(\mathbf{x}_1, \dots, \mathbf{x}_N)$, the **rate** of the code is

$$R = \frac{\log N}{n}. \quad (5.1)$$

In this expression, n is the number of channel uses or channel symbols transmitted for each message. If the logarithm base is 2, then the numerator has units of bits, and the quotient R has units of **bits per channel use** or **bits per channel symbol**. If the channel transmits R_c channel symbols per second, then

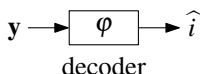
$$R_b := R_c \frac{\log_2 N}{n}$$

has units of bits per second. Combining this with (5.1) yields

$$R_b = R_c R.$$

Decoders

A mapping $\varphi: \mathcal{Y}^n \rightarrow \{1, \dots, N\}$ is called a **decoder**. If for $i = 1, \dots, N$ we put $D_i := \{\mathbf{y} : \varphi(\mathbf{y}) = i\}$, then the D_i are disjoint and their union is $\mathcal{Y}^n$. We call the D_i **decoding sets**.



5.3. Probabilistic Model

If we apply the output of the encoder to the channel, and we apply the output of the channel to the decoder, then we obtain the **channel coding system** shown in Figure 5.2. Notice that the input to the system is i and the output is $\hat{i}$. We hope that

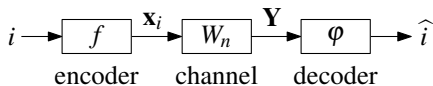


Figure 5.2. A channel coding system.

most of the time $\hat{i} = i$. To make this more precise, we assume that the input to the encoder is a $\{1, \dots, N\}$ -valued random variable M , and we focus on the probability of error $P(\varphi(\mathbf{Y}) \neq M)$. However, this problem is not well defined unless we specify the joint distribution of M and $\mathbf{Y}$. We take

$$P(M = i, \mathbf{Y} \in B) := \frac{1}{N} \cdot W_n(B|\mathbf{x}_i), \quad i = 1, \dots, N, B \subset \mathcal{Y}^n. \quad (5.2)$$

Notice that since $W_n(\mathcal{Y}^n|\mathbf{x}_i) = 1$, (5.2) implies that $P(M = i) = 1/N$; i.e., the random message to be sent is chosen uniformly from the message set $\{1, \dots, N\}$. It then follows that

$$P(\mathbf{Y} \in B|M = i) = \frac{P(M = i, \mathbf{Y} \in B)}{P(M = i)} = W_n(B|\mathbf{x}_i). \quad (5.3)$$

On account of the discussion following Definition 5.8 below, in the context of channel coding, we almost always assume M and $\mathbf{Y}$ have the joint distribution given by (5.2) where M is uniformly distributed.

5.4. Probability of Error and Achievable Rates

Example 5.7 (Repetition Codes on the BSC). Consider a BSC with crossover probability $\varepsilon = 0.1$. Suppose we have $N = 2$ messages, say $\{0, 1\}$. For $n = 1$, we might use codewords $\mathbf{x}_0 := 0$ and $\mathbf{x}_1 := 1$. In this case, the probability of error is just the crossover probability $\varepsilon = 0.1$. For $n = 3$, we might use codewords $\mathbf{x}_0 := (0, 0, 0)$ and $\mathbf{x}_1 := (1, 1, 1)$ with **majority rule decoding**; i.e., if the number of 1s in $\mathbf{y}$ is two or more, we decode $\hat{i} = 1$, and zero otherwise. With this decoding rule, an error occurs

if two or three of the transmitted bits are flipped by the channel.^c This probability is 0.028. If we proceed in a similar way with $n = 5$ bits, the error probability is 0.0086. This analysis suggests that if we fix $N = 2$ and let $n \rightarrow \infty$, we can drive the error probability to zero. The downside is that the code rate $(\log N)/n$ also goes to zero.

The preceding example suggests that to get an interesting problem, we should allow N to depend on n and then ask how the error probability behaves as a function of n . Ideally, we want the error probability to get smaller as n increases. We also want the code rate to stay away from zero.

Fix a *desired* code rate R . Suppose that given any error tolerance $\lambda > 0$, for any sufficiently large block length n , there are codewords $\mathbf{x}_1, \dots, \mathbf{x}_N$ each in $\mathcal{X}^n$ and a decoder $\varphi: \mathcal{Y}^n \rightarrow \{1, \dots, N\}$ such that the rate of *this* code, $(\log N)/n$, is not much less than R , and the error probability satisfies $P(\varphi(\mathbf{Y}) \neq M) < \lambda$. If this is the case, then we say that R is an **achievable rate**.

Since this definition is not quite precise enough to be useful, we will use Definition 5.8, which is given below, after we make a few preliminary observations. To begin, note that by the law of substitution,

$$\begin{aligned} P(\varphi(\mathbf{Y}) \neq M | M = i) &= P(\varphi(\mathbf{Y}) \neq i | M = i) \\ &= P(\mathbf{Y} \in \{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | M = i) \\ &= W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i), \quad \text{by (5.3).} \end{aligned} \quad (5.4)$$

Using the law of total probability,

$$\begin{aligned} P(\varphi(\mathbf{Y}) \neq M) &= \sum_{i=1}^N P(\varphi(\mathbf{Y}) \neq M | M = i) P(M = i) \\ &= \frac{1}{N} \sum_{i=1}^N P(\varphi(\mathbf{Y}) \neq M | M = i), \quad \text{since } M \text{ is uniform,} \\ &= \frac{1}{N} \sum_{i=1}^N W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i), \quad \text{by (5.4).} \end{aligned} \quad (5.5)$$

Definition 5.8. A real number $R \geq 0$ is an **achievable rate** for a channel (more precisely, a sequence of channels W_n), under the **average probability of error criterion**, if $\forall \lambda > 0, \forall \Delta R > 0, \exists n_0$ such that $\forall n \geq n_0, \exists$ codewords $\mathbf{x}_1, \dots, \mathbf{x}_N$ each in

^c Let S_n denote the number of flips in n bits. An error occurs if $S_n \geq n/2$. Hence, the error probability is $1 - P(S_n < n/2)$. Since S_n is a binomial(n, ϵ) random variable, the error probability is given by the MATLAB command `1 - binocdf(floor(n/2), n, epsilon)` when n is odd.

X^n and there exists a decoder $\varphi: Y^n \rightarrow \{1, \dots, N\}$ with

$$\frac{\log N}{n} > R - \Delta R \quad \text{and} \quad \frac{1}{N} \sum_{i=1}^N W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) < \lambda.$$

Although the reader may have reservations about restricting M to be uniformly distributed, in many cases, we will be able to obtain a bound on (5.4) that does not depend on i , say^d

$$W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) < \lambda, \quad i = 1, \dots, N. \quad (5.6)$$

If we can establish such a bound, then for any probability mass function $q(i)$ on $\{1, \dots, N\}$, if we replace (5.2) with

$$P(M = i, \mathbf{Y} \in B) := q(i) W_n(B | \mathbf{x}_i), \quad i = 1, \dots, N, B \subset Y^n,$$

then (5.3) still holds, and it follows that

$$\begin{aligned} P(\varphi(\mathbf{Y}) \neq M) &= \sum_{i=1}^N P(\varphi(\mathbf{Y}) \neq M | M = i) P(M = i) \\ &= \sum_{i=1}^N P(\varphi(\mathbf{Y}) \neq M | M = i) q(i) \\ &= \sum_{i=1}^N W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) q(i), \quad \text{by (5.4),} \\ &< \sum_{i=1}^N \lambda q(i), \quad \text{by (5.6),} \\ &= \lambda. \end{aligned} \quad (5.7)$$

Definition 5.9. A real number $R \geq 0$ is an **achievable rate** for a channel (more precisely, a sequence of channels W_n), under the **maximum probability of error criterion**, if $\forall \lambda > 0, \forall \Delta R > 0, \exists n_0$ such that $\forall n \geq n_0, \exists$ codewords $\mathbf{x}_1, \dots, \mathbf{x}_N$ each in X^n and there exists a decoder $\varphi: Y^n \rightarrow \{1, \dots, N\}$ with

$$\frac{\log N}{n} > R - \Delta R \quad \text{and} \quad \max_{1 \leq i \leq N} W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) < \lambda.$$

^dThe condition (5.6) can also be expressed as $\max_{1 \leq i \leq N} W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) < \lambda$.

To say that a rate R is achievable according to the above definition means that given any error tolerance $\lambda > 0$, if we take any sufficiently large block length n , then there are N codewords in $\mathcal{X}^n$ and a decoder ϕ such that the code rate $(\log N)/n$ is not much less than R and $\max_{1 \leq i \leq N} \mathbb{P}(\phi(\mathbf{Y}) \neq M | M = i) < \lambda$ (cf. (5.4)). Of course, as shown by (5.7), no matter what distribution q we use for M , we will have $\mathbb{P}(\phi(\mathbf{Y}) \neq M) < \lambda$ as well. Since this holds in particular for q being the uniform distribution, it is clear that any rate that is achievable under the maximum probability of error criterion is achievable under the average probability of error criterion.

5.5. Capacity Regions

The **capacity region** is the set of achievable rates for a given channel under a given probability of error criterion. A **weak converse** theorem says that a capacity region is contained in a *specified* bigger set.^e A **forward theorem** says that a capacity region contains a *specified* smaller subset. If the bigger set and the smaller set are the same, then we say that the capacity region has been characterized. A central goal of information theory is to characterize the capacity regions of various channels under various assumptions, and to obtain as simple a characterization as possible.

Our focus will be on the capacity regions

$$\mathcal{C}_a := \{R \geq 0 : R \text{ is achievable under the average probability of error criterion}\},$$

and

$$\mathcal{C}_m := \{R \geq 0 : R \text{ is achievable under the maximum probability of error criterion}\}.$$

It is easy to check that $R = 0$ is achievable under either criterion. It is also easy to see that if $R > 0$ is achievable, then any $0 \leq R' < R$ is also achievable. Hence, each capacity region must be an interval (unless the region consist only of the origin). It is also easy to show that any limit of achievable rates is achievable (Problem 5.3). This implies that each capacity region must be a closed interval.^f The right-hand endpoint of this interval, which is the maximum achievable rate, is called the **capacity** of the channel (more precisely, the sequence of channels W_n) under the corresponding probability of error criterion.

Theorem 5.10. $\mathcal{C}_m = \mathcal{C}_a$.

^e There is a notion of a **strong converse**, which is discussed briefly in a later chapter.

^f A singleton set such as $\{0\} = [0, 0]$ is considered a (degenerate) interval.

Proof. The fact that $\mathcal{C}_m \subset \mathcal{C}_a$ follows from the paragraph below Definition 5.9. It is the reverse inclusion that requires some work. This work includes two key observations. The first one is that if a numerical average is less than λ , then at least half the terms must individually be less than 2λ . The second observation provides a simple method to construct a new code from an existing one by throwing away codewords. These two observations are combined in Section 5.5.3, where the proof that $\mathcal{C}_a \subset \mathcal{C}_m$ is completed. $\square$

5.5.1. Small Averages

Suppose $\theta_1, \dots, \theta_N$ are nonnegative numbers whose numerical average is less than λ . Then more than half of the θ_i are individually less than 2λ . To establish this claim, we first express it in symbolic terms. Suppose that

$$\frac{1}{N} \sum_{i=1}^N \theta_i < \lambda,$$

where $0 < \lambda < \infty$. Put $G := \{i : \theta_i < 2\lambda\}$. We show that $|G| > N/2$. Write

$$\lambda > \frac{1}{N} \sum_{i=1}^N \theta_i = \frac{1}{N} \left(\sum_{i \in G^c} \theta_i + \sum_{i \in G} \theta_i \right) \geq \frac{1}{N} \sum_{i \in G^c} \theta_i \geq \frac{1}{N} \sum_{i \in G^c} 2\lambda = \frac{2\lambda}{N} |G^c|.$$

It follows that $|G^c| < N/2$, and then $|G| = N - |G^c| > N - N/2 = N/2$.

5.5.2. New Codes from Old Codes by Throwing Away Codewords

Let $f: \{1, \dots, N\} \rightarrow \mathcal{X}^n$ be any encoder, and let $\varphi: \mathcal{Y}^n \rightarrow \{1, \dots, N\}$ be any decoder. Then

$$D_i := \{\mathbf{y} : \varphi(\mathbf{y}) = i\}$$

contains exactly those outputs $\mathbf{y}$ that are decoded to message i .

If G is any subset of $\{1, \dots, N\}$, we define the **modified encoder** $f_G: G \rightarrow \mathcal{X}^n$ by $f_G(i) := f(i) = \mathbf{x}_i$, $i \in G$, and we define the **modified decoder** $\varphi_G: \mathcal{Y}^n \rightarrow G$ by

$$\varphi_G(\mathbf{y}) := \begin{cases} \varphi(\mathbf{y}), & \mathbf{y} \in \bigcup_{i \in G} D_i, \\ i_0, & \text{otherwise,} \end{cases}$$

where i_0 can be any fixed element of G . Hence, for $i \in G$ with $i \neq i_0$, $\varphi_G(\mathbf{y}) = i \Leftrightarrow \mathbf{y} \in D_i$, and we can write

$$\{\mathbf{y} : \varphi_G(\mathbf{y}) \neq i\} = \{\mathbf{y} : \varphi(\mathbf{y}) \neq i\}, \quad \text{for } i \in G, i \neq i_0. \quad (5.8)$$

Furthermore, if we put

$$H := \left(\bigcup_{i \in G} D_i \right)^c,$$

then $\varphi_G(\mathbf{y}) = i_0 \Leftrightarrow \mathbf{y} \in D_{i_0} \cup H$. We can therefore write

$$\{\mathbf{y} : \varphi_G(\mathbf{y}) \neq i\} \subset D_i^c = \{\mathbf{y} : \varphi(\mathbf{y}) \neq i\}, \quad \underline{\text{for all } i \in G}. \quad (5.9)$$

Suppose that instead of (5.6), we have only

$$W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) < \lambda, \quad i \in G,$$

where G is a proper subset of $\{1, \dots, N\}$. Let f_G and φ_G be the encoder and decoder just described. Then by (5.9), we can write

$$W_n(\{\mathbf{y} : \varphi_G(\mathbf{y}) \neq i\} | \mathbf{x}_i) \leq W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) < \lambda, \quad i \in G. \quad (5.10)$$

5.5.3. Proof that $\mathcal{C}_a \subset \mathcal{C}_m$

Suppose R is achievable under the average probability of error criterion. Then for given $\lambda > 0$ and $\Delta R > 0$, for all sufficiently large n , there exist codewords $\mathbf{x}_1, \dots, \mathbf{x}_N$ each in X^n and there exists a decoder $\varphi: Y^n \rightarrow \{1, \dots, N\}$ with

$$\frac{\log N}{n} > R - \Delta R/2 \quad \text{and} \quad \frac{1}{N} \sum_{i=1}^N W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) < \lambda/2.$$

By our observation about small averages in Section 5.5.1, there must be a subset $G \subset \{1, \dots, N\}$ with $|G| > N/2$ and

$$W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) < 2(\lambda/2) = \lambda, \quad i \in G.$$

By (5.10),

$$W_n(\{\mathbf{y} : \varphi_G(\mathbf{y}) \neq i\} | \mathbf{x}_i) < \lambda, \quad i \in G.$$

Since there is no harm in assuming n is so large that $(\log 2)/n \leq \Delta R/2$, we have

$$\begin{aligned} \frac{\log |G|}{n} &> \frac{\log(N/2)}{n} = \frac{\log N}{n} - \frac{\log 2}{n} \\ &> R - \Delta R/2 - \frac{\log 2}{n} \\ &\geq R - \Delta R/2 - \Delta R/2 \\ &= R - \Delta R. \end{aligned}$$

We thus have an encoder-decoder pair (f_G, φ_G) with rate greater than $R - \Delta R$ and with maximum error probability less than λ ; i.e., R is achievable under the maximum probability of error criterion. $\square$

Problems

1. Using the notation of Example 5.2, suppose that

$$P(Y^n = y^n | X^n = x^n) = W^n(y^n | x^n) := W(y_n | x_n) \cdots W(y_1 | x_1)$$

for some transition probability $W(\cdot | \cdot)$. Show that

$$P(Y_n = y_n | Y^{n-1} = y^{n-1}, X^n = x^n) = W(y_n | x_n)$$

and that

$$P(X_n = x_n | Y^{n-1} = y^{n-1}, X^{n-1} = x^{n-1}) = P(X_n = x_n | X^{n-1} = x^{n-1}).$$

2. **MATLAB.** When X and Y are finite sets, we can identify X with the positive integers $\{1, \dots, |X|\}$, and we can identify Y with the positive integers $\{1, \dots, |Y|\}$. Consider a joint pmf of the form $P_{XY}(i, j) = p(i)W(j|i)$. We regard the input pmf $p = [p(1), \dots, p(|X|)]$ as a row vector. Since $P_Y(j) = \sum_i P_{XY}(i, j) = \sum_i p(i)W(j|i)$, we can also regard P_Y as a row vector. In fact, if we let W be the matrix whose ij component is $W(j|i)$, then in MATLAB, the expression $PY = p * W$ computes the row vector of output probabilities P_Y . The commands

```
[m,n] = size(W);           % m = # inputs, n = # outputs
Pmat = repmat(p', 1, n);
PXY = Pmat.*W;
```

compute the matrix whose ij component is $P_{XY}(i, j) = p(i)W(j|i)$. The commands

```
PY = p*W
PYmat = repmat(PY, m, 1);
PXgY = PXY./PYmat;
```

compute the matrix whose ij component is $P_{X|Y}(i|j) = P_{XY}(i, j)/P_Y(j)$. Run the script

```
W = [ 1/2 1/2 0; 1/4 0 3/4 ; 1/2 1/2 0 ];
p = [ 5/12 1/3 1/4]

[m,n] = size(W);
Pmat = repmat(p', 1, n);
PXY = Pmat.*W;

PY = p*W
PYmat = repmat(PY, m, 1);
PXgY = PXY./PYmat;
```

```

HX = entropy2(p)
HY = entropy2(PY)
HXY = entropy2(PXY)
HYgX = entropy2(W, PXY)
HXgY = entropy2(PXgY, PXY)

```

to compute $H(X)$, $H(Y)$, $H(XY)$, $H(Y|X)$, and $H(X|Y)$.

3. **Limits of Achievable Rates Are Achievable.** Let R_k be a sequence of achievable rates under the average probability of error criterion. If $R_k \rightarrow R$, show that R is also an achievable rate under the average probability of error criterion.
4. The average on a class midterm is less than 40 on a 100-point test. Consider the following statement:

Fewer than half the students got at least 80 points.

Determine whether the statement is true or false. **Justify your answer.**

CHAPTER 6

Converse Theorems for the DMC

As shown in Chapter 5, $\mathcal{C}_m = \mathcal{C}_a$. In this chapter, we show that for the DMC with transition probability W ,

$$\mathcal{C}_a \subset [0, \sup_p I(p \times W)],$$

where p runs over all pmfs on the set of channel input symbols X , and $I(p \times W)$ is the mutual information defined below. We can express the subset relation in words by saying that if R is achievable under the average probability of error criterion for a DMC W , then $R \leq \sup_p I(p \times W)$. This result is known as the **weak converse**. At the end of the chapter, we discuss the **strong converse** for the DMC.

6.1. Preliminaries

6.1.1. Notation

If p and q are pmfs on X and Y , respectively, then $p \times q$ is the joint pmf on $\mathsf{X} \times \mathsf{Y}$ defined by

$$(p \times q)(x, y) := p(x)q(y).$$

Thus, $p \times q$ represents the joint pmf of a pair of independent random variables. More generally, if p is a pmf on X and $W(y|x)$ is a transition probability from X into Y , then $p \times W$ is the joint pmf on $\mathsf{X} \times \mathsf{Y}$ defined by

$$(p \times W)(x, y) := p(x)W(y|x).$$

In this case, $p \times W$ represents the joint pmf of a pair of dependent random variables. Of course the marginal on X is just p , while the marginal on Y is

$$(pW)(y) := \sum_x p(x)W(y|x).$$

6.1.2. Mutual Information

If $P_{XY}(x, y)$ is any joint pmf, we define the (average) mutual information

$$I(P_{XY}) := \mathbb{E} \left[\log \frac{P_{XY}(X, Y)}{P_X(X)P_Y(Y)} \right] = D(P_{XY} \| P_X \times P_Y),$$

where D is the **Kullback–Leibler informational divergence** or **relative entropy**. We often write $I(X \wedge Y)$ instead of $I(P_{XY})$. Some authors write $I(X; Y)$ instead of $I(X \wedge Y)$.

Several properties of mutual information (and entropy) are readily apparent. First, because mutual information can be expressed as a divergence, we see that $I(X \wedge Y) \geq 0$, with equality if and only if X and Y are independent. Second, since

$$\begin{aligned} 0 \leq I(X \wedge Y) &= \mathbb{E} \left[\log \frac{P_{XY}(X, Y)}{P_X(X)P_Y(Y)} \right] \\ &= \mathbb{E}[-\log P_X(X)] + \mathbb{E}[-\log P_Y(Y)] - \mathbb{E}[-\log P_{XY}(X, Y)], \end{aligned}$$

we see that

$$H(XY) \leq H(X) + H(Y), \quad (6.1)$$

with equality if and only if X and Y are independent. Third,

$$I(X \wedge Y) = H(Y) - H(Y|X) = H(X) - H(X|Y),$$

from which it follows that

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

Since entropy is a measure of uncertainty, this last property is summarized by saying, “conditioning reduces uncertainty.”

6.1.3. Markov Chains

When X , Y , and Z are random variables that model the signals in a cascade of systems as in Figure 6.1 the joint distribution of X , Y , and Z usually implies that

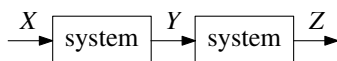


Figure 6.1. A cascade of systems.

$P(Z = z|Y = y, X = x)$ does not depend on x . We use the notation $X \rightarrow Y \rightarrow Z$ as shorthand to indicate that $P(Z = z|Y = y, X = x)$ does not depend on x . The notation $X \rightarrow Y \rightarrow Z$ is read as “ X , Y , and Z form a **Markov chain**,” although in writing we frequently use the redundant expression, “ $X \rightarrow Y \rightarrow Z$ is a Markov chain.” As shown in the problems, the Markov-chain condition is equivalent to having X and Z conditionally independent given Y .

Example 6.1. Let X and Y have any joint pmf, and suppose $Z = \varphi(Y)$ for some deterministic function φ . In this case we show that $X \rightarrow Y \rightarrow Z$ is a Markov chain. To begin, write

$$\begin{aligned} P(Z = z|Y = y, X = x) &= P(\varphi(Y) = z|Y = y, X = x) = P(\varphi(y) = z|Y = y, X = x) \\ &= P(\varphi(y) = z), \end{aligned}$$

where we have used the law of substitution and the fact that a deterministic event is independent of everything. Since the probability of a deterministic condition is one if the condition is true and zero otherwise,

$$P(Z = z|Y = y, X = x) = \begin{cases} 1, & \text{if } \varphi(y) = z, \\ 0, & \text{otherwise,} \end{cases}$$

which does not depend on x .

Example 6.2. Suppose M and $\mathbf{Y}$ are random variables with $P(\mathbf{Y} \in B, M = i) = W_n(B|\mathbf{x}_i)q(i)$, where q is a pmf on $\{1, \dots, N\}$, and $\mathbf{x}_1, \dots, \mathbf{x}_N$ are given. Put $f(i) := \mathbf{x}_i$, and define a new random variable $\mathbf{X} := f(M) = \mathbf{x}_M$. We show that $M \rightarrow \mathbf{X} \rightarrow \mathbf{Y}$ is a Markov chain. To begin, write

$$\begin{aligned} P(\mathbf{Y} \in B, \mathbf{X} = \mathbf{x}, M = i) &= P(\mathbf{Y} \in B, f(M) = \mathbf{x}, M = i) \\ &= P(\mathbf{Y} \in B, f(i) = \mathbf{x}, M = i) \\ &= \begin{cases} P(\mathbf{Y} \in B, M = i), & \text{if } \mathbf{x} = \mathbf{x}_i, \\ 0, & \text{otherwise,} \end{cases} \\ &= W_n(B|\mathbf{x})\mathbf{1}_{\{\mathbf{x}_i\}}(\mathbf{x})q(i). \end{aligned} \tag{6.2}$$

Taking $B = \mathbf{Y}^n$ shows that

$$P(\mathbf{X} = \mathbf{x}, M = i) = \mathbf{1}_{\{\mathbf{x}_i\}}(\mathbf{x})q(i).$$

Thus,^a

$$P(\mathbf{Y} \in B|\mathbf{X} = \mathbf{x}, M = i) = \frac{P(\mathbf{Y} \in B, \mathbf{X} = \mathbf{x}, M = i)}{P(\mathbf{X} = \mathbf{x}, M = i)} = \frac{W_n(B|\mathbf{x})\mathbf{1}_{\{\mathbf{x}_i\}}(\mathbf{x})q(i)}{\mathbf{1}_{\{\mathbf{x}_i\}}(\mathbf{x})q(i)} = W_n(B|\mathbf{x}),$$

^aThe value assigned to $P(\mathbf{Y} \in B|\mathbf{X} = \mathbf{x}, M = i)$ when the denominator is zero can be any probability measure on B ; what is important about conditional probability is that its product with the denominator always equal the joint probability; i.e., we need

$$P(\mathbf{Y} \in B|\mathbf{X} = \mathbf{x}, M = i)P(\mathbf{X} = \mathbf{x}, M = i) = P(\mathbf{Y} \in B, \mathbf{X} = \mathbf{x}, M = i).$$

This clearly holds for all $\mathbf{x}$ — check the two cases $\mathbf{x} = \mathbf{x}_i$ and $\mathbf{x} \neq \mathbf{x}_i$.

which does not depend on i . You are asked to show in the problems that for any deterministic function $\varphi(\mathbf{y})$,

$M \rightarrow \mathbf{X} \rightarrow \mathbf{Y}$ is a Markov chain $\Rightarrow M \rightarrow \mathbf{X} \rightarrow \varphi(\mathbf{Y})$ is a Markov chain.

6.1.4. Data Processing Inequalities

Data processing inequalities say that if $X \rightarrow Y \rightarrow Z$ is a Markov chain, then “the mutual information between close things is greater than or equal to the mutual information between far things.” More precisely, it is shown in the problems that

$$X \rightarrow Y \rightarrow Z \Rightarrow I(X \wedge Z) \leq \begin{cases} I(X \wedge Y) \\ I(Y \wedge Z). \end{cases}$$

Lemma 6.3 (Data Processing). *Let M , $\mathbf{X}$, and $\mathbf{Y}$ be as in Example 6.2. Then*

$$I(M \wedge \varphi(\mathbf{Y})) \leq I(\mathbf{X} \wedge \varphi(\mathbf{Y})) \leq I(\mathbf{X} \wedge \mathbf{Y}).$$

Proof. The left-hand inequality follows because we noted in Example 6.2 that $M \rightarrow \mathbf{X} \rightarrow \varphi(\mathbf{Y})$ is a Markov chain. The right-hand inequality follows because we showed in Example 6.1 that $\mathbf{X} \rightarrow \mathbf{Y} \rightarrow \varphi(\mathbf{Y})$ is a Markov chain. $\square$

Lemma 6.4. *If $\mathbf{X}$ and $\mathbf{Y}$ are as in Example 6.2 and $W_n = W^n$ is a DMC, then*

$$H(\mathbf{Y}|\mathbf{X}) = \sum_{k=1}^n H(Y_k|X_k) \quad \text{and} \quad I(\mathbf{X} \wedge \mathbf{Y}) \leq \sum_{k=1}^n I(X_k \wedge Y_k).$$

Proof. First, it follows from (6.2), that $P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}) = W^n(\mathbf{y}|\mathbf{x})$. From this it follows that $P(Y_k = y_k|\mathbf{X} = \mathbf{x}) = W(y_k|x_k)$, which implies (cf. Problem 6.6) $P(Y_k = y_k|X_k = x_k) = W(y_k|x_k)$. Hence,

$$H(\mathbf{Y}|\mathbf{X}) = \mathbb{E}[-\log W^n(\mathbf{Y}|\mathbf{X})] = \sum_{k=1}^n \mathbb{E}[-\log W(Y_k|X_k)] = \sum_{k=1}^n H(Y_k|X_k).$$

To derive the inequality, write

$$I(\mathbf{X} \wedge \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X})$$

$$\begin{aligned}
&= H(\mathbf{Y}) - \sum_{k=1}^n H(Y_k | X_k) \\
&\leq \sum_{k=1}^n H(Y_k) - \sum_{k=1}^n H(Y_k | X_k), \quad \text{by (6.1),} \\
&= \sum_{k=1}^n I(X_k \wedge Y_k).
\end{aligned}$$

□

6.1.5. Fano's Inequality

Theorem 6.5 (Fano's Inequality). *Let M and V be two random variables taking values in $\{1, \dots, N\}$. Then*

$$H(M|V) \leq P(M \neq V) \log(N-1) + h(P(M \neq V)),$$

where h is the binary entropy function (Problem 3.10).

Corollary 6.6. $H(M) \leq P(M \neq V) \log N + \log 2 + I(M \wedge V)$.

Proof of Corollary. Since h is the entropy of a pmf on two points, $h(P(M \neq V)) \leq \log 2$. Since $N-1 < N$, Fano's inequality implies

$$0 \leq P(M \neq v) \log N + \log 2 - H(M|V).$$

Now add $H(M)$ to both sides to obtain the result, since $H(M) - H(M|V) = I(M \wedge V)$. □

Proof of Fano's Inequality. Consider the random variable $Z := \mathbf{1}_{\{M \neq V\}}$. Using the law of total probability,^b

$$\begin{aligned}
H(M|V) &:= \mathbb{E} \left[\log \frac{1}{P(M|V)} \right] = \mathbb{E} \left[\mathbb{E} \left[\log \frac{1}{P(M|V)} \middle| Z \right] \right] \\
&= \mathbb{E} \left[\log \frac{1}{P(M|V)} \middle| Z = 1 \right] P(M \neq V) + \mathbb{E} \left[\log \frac{1}{P(M|V)} \middle| Z = 0 \right] P(M = V).
\end{aligned}$$

Next, by Jensen's inequality (Problem 1.1), the conditional expectations satisfy

$$\mathbb{E} \left[\log \frac{1}{P(M|V)} \middle| Z = 1 \right] \leq \log \mathbb{E} \left[\frac{1}{P(M|V)} \middle| Z = 1 \right]$$

^b We write $P(m|v)$ for $P_{M|V}(m|v)$ and $P(v)$ for $P_V(v)$ to simplify the notation.

and

$$\mathbb{E} \left[\log \frac{1}{P(M|V)} \middle| Z=0 \right] \leq \log \mathbb{E} \left[\frac{1}{P(M|V)} \middle| Z=0 \right].$$

It remains to evaluate the two conditional expectations on the right. Observe that

$$P(M=m, V=v | M \neq V) = \frac{P(M=m, V=v)}{P(M \neq V)} \mathbf{1}_{\{m\}^c}(v) = \frac{P(m|v)P(v)}{P(M \neq V)} \mathbf{1}_{\{m\}^c}(v)$$

and

$$P(M=m, V=v | M=V) = \frac{P(M=m, V=v)}{P(M=V)} \mathbf{1}_{\{m\}}(v) = \frac{P(m|v)P(v)}{P(M=V)} \mathbf{1}_{\{m\}}(v).$$

Hence,

$$\begin{aligned} \mathbb{E} \left[\frac{1}{P(M|V)} \middle| Z=1 \right] &= \sum_v \sum_m \frac{1}{P(m|v)} \frac{P(m|v)P(v)}{P(M \neq V)} \mathbf{1}_{\{m\}^c}(v) \\ &= \frac{1}{P(M \neq V)} \sum_v P(v) \sum_m \mathbf{1}_{\{m\}^c}(v) = \frac{N-1}{P(M \neq V)}. \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{E} \left[\frac{1}{P(M|V)} \middle| Z=0 \right] &= \sum_v \sum_m \frac{1}{P(m|v)} \frac{P(m|v)P(v)}{P(M=V)} \mathbf{1}_{\{m\}}(v) \\ &= \frac{1}{P(M=V)} \sum_v P(v) \sum_m \mathbf{1}_{\{m\}}(v) = \frac{1}{P(M=V)}. \end{aligned}$$

Putting this all together yields

$$\begin{aligned} H(M|V) &\leq P(M \neq V) \log \frac{N-1}{P(M \neq V)} + P(M=V) \log \frac{1}{P(M=V)} \\ &= P(M \neq V) \log(N-1) + h(P(M \neq V)). \end{aligned}$$

□

6.2. The Weak Converse

Theorem 6.7 (Weak Converse for the DMC). *If R is an achievable rate for a DMC W under the average probability of error criterion, then $R \leq \sup_p I(p \times W)$.*

Proof. Fix any $0 < \lambda < 1$ and any $\Delta R > 0$. Since R is achievable, for all sufficiently large n , there exist codewords $\mathbf{x}_1, \dots, \mathbf{x}_N$ and there exists a decoder φ with

$$\frac{\log N}{n} > R - \Delta R \quad \text{and} \quad P(\varphi(\mathbf{Y}) \neq M) < \lambda,$$

where $M, \mathbf{X}, \mathbf{Y}$ have joint pmf (cf. Example 6.2)

$$P(M = i, \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = \frac{1}{N} \mathbf{1}_{\{\mathbf{x}_i\}}(\mathbf{x}) W^n(\mathbf{y}|\mathbf{x}).$$

Since M is uniformly distributed, $H(M) = \log N$. By the corollary to Fano's inequality,

$$\log N = H(M) \leq \lambda \log N + \log 2 + I(M \wedge \varphi(\mathbf{Y})).$$

Rearrange this as

$$(1 - \lambda) \log N \leq I(M \wedge \varphi(\mathbf{Y})) + \log 2.$$

Since $(\log N)/n > R - \Delta R$, we have $\log N > n(R - \Delta R)$, and so

$$\begin{aligned} (1 - \lambda)n(R - \Delta R) &\leq I(\mathbf{X} \wedge \mathbf{Y}) + \log 2, && \text{by the Data Processing Lemma,} \\ &\leq \sum_{k=1}^n I(X_k \wedge Y_k) + \log 2, && \text{by Lemma 6.4,} \\ &\leq n \cdot \sup_p I(p \times W) + \log 2. \end{aligned}$$

Hence,

$$R \leq \frac{\sup_p I(p \times W)}{1 - \lambda} + \frac{\log 2}{n(1 - \lambda)} + \Delta R.$$

Since $\lambda > 0$ and $\Delta R > 0$ can be arbitrarily small, and n can be arbitrarily large, we must have $R \leq \sup_p I(p \times W)$. $\square$

6.3. The Strong Converse

Definition 6.8. Given $\lambda > 0$, a real number $R \geq 0$ is **λ -achievable** for a channel (more precisely, a sequence of channels W_n), under the average probability of error criterion, if $\forall \Delta R > 0$, $\exists n_0$ such that $\forall n \geq n_0$, $\exists$ codewords $\mathbf{x}_1, \dots, \mathbf{x}_N$ each in $\mathcal{X}^n$ and there exists a decoder $\varphi: \mathcal{Y}^n \rightarrow \{1, \dots, N\}$ with

$$\frac{\log N}{n} > R - \Delta R \quad \text{and} \quad P(\varphi(\mathbf{Y}) \neq M) < \lambda,$$

where $P(M = i, \mathbf{Y} = \mathbf{y}) = W^n(\mathbf{y}|\mathbf{x}_i)/N$.

A rate R is achievable $\Leftrightarrow \forall \lambda > 0$, R is λ -achievable. In other words, if we let $\mathcal{C}_{a,\lambda}$ denote the set of λ -achievable rates, then

$$\mathcal{C}_a = \bigcap_{\lambda > 0} \mathcal{C}_{a,\lambda}.$$

Since $0 < \lambda_1 < \lambda_2$ implies $\mathcal{C}_{a,\lambda_1} \subset \mathcal{C}_{a,\lambda_2}$,

$$\bigcap_{0 < \lambda < 1} \mathcal{C}_{a,\lambda} \subset \bigcap_{\lambda > 0} \mathcal{C}_{a,\lambda}.$$

Since the reverse inclusion is trivially true, we have

$$\mathcal{C}_a = \bigcap_{0 < \lambda < 1} \mathcal{C}_{a,\lambda}.$$

We next observe that for fixed $\lambda > 0$, $\mathcal{C}_a \subset \mathcal{C}_{a,\lambda}$. A strong converse says that for $0 < \lambda < 1$, $\mathcal{C}_{a,\lambda} \subset \mathcal{C}_a$. If a strong converse holds, then $\mathcal{C}_{a,\lambda} = \mathcal{C}_a$ for $0 < \lambda < 1$.

Since $\mathcal{C}_{a,\lambda} \subset \mathcal{C}_a$ is equivalent to $\mathcal{C}_a^c \subset \mathcal{C}_{a,\lambda}^c$, we can say the following: If a rate R is such that you cannot drive the error probability to zero, then asymptotically, you cannot drive the error probability to any number less than λ .

Theorem 6.9 (Strong Converse for the DMC). *For a DMC W and $0 < \lambda < 1$, $\mathcal{C}_{a,\lambda} \subset \mathcal{C}_a$.*

Proof. We prove the equivalent formula $\mathcal{C}_a^c \subset \mathcal{C}_{a,\lambda}^c$. Put $C := \sup_p I(p \times W)$. The weak converse shows that $\mathcal{C}_a \subset [0, C]$. In a later chapter, we show that $\mathcal{C}_a \supset [0, C]$ (the **forward theorem**). Thus, $\mathcal{C}_a = [0, C]$. It now follows that $R \in \mathcal{C}_a^c$ implies $R > C$. Choose any $0 < \Delta R < (R - C)/2$. Suppose (f, ϕ) is an encoder-decoder pair with

$$\frac{\log(N)}{n} > R - \Delta R > C \quad \text{and} \quad P(\phi(\mathbf{Y}) \neq M) < \lambda.$$

It can be proved that^c

$$P(\phi(\mathbf{Y}) \neq M) \geq 1 - \frac{4A}{n(R - C)^2} - \exp \left[-n \left(\frac{R - C}{2} - \Delta R \right) \right] \rightarrow 1,$$

where $0 < A < \infty$ depends on W but not on n or N . The conditions $P(\phi(\mathbf{Y}) \neq M) < \lambda$ and $P(\phi(\mathbf{Y}) \neq M) \rightarrow 1$ are contradictory. $\square$

^c See, e.g., Gallager [6, pp. 173–175] for a proof of this result, which is due to Wolfowitz (1957).

Problems

- Show that if X and Y have the same pmf, then $H(Y|X) = H(X|Y)$.
 - MATLAB.** If $P_X = P_Y$, is it true that $P_{X|Y} = P_{Y|X}$? Explore this question as follows. Recall Problem 5.2. In the script of that problem, $P_Y = P_X$. Run the script and print out $P_{X|Y}$ and W . Are they the same?
 - MATLAB.** Compute $\text{relent}(P_{XY}, p' * P_Y)$ and compare it with the mutual information $H(Y) - H(Y|X)$.
- Chain Rule for Mutual Information.** Recall that the mutual information between X and Y is

$$I(X \wedge Y) := \mathbb{E} \left[\log \frac{P_{XY}(X, Y)}{P_X(X)P_Y(Y)} \right].$$

Similarly, the mutual information between X and YZ is

$$I(X \wedge YZ) := \mathbb{E} \left[\log \frac{P_{XYZ}(X, Y, Z)}{P_X(X)P_{YZ}(Y, Z)} \right].$$

The conditional mutual information is

$$I(X \wedge Z|Y) := \mathbb{E} \left[\log \frac{P_{XZ|Y}(X, Z|Y)}{P_{X|Y}(X|Y)P_{Z|Y}(Z|Y)} \right].$$

Derive the chain rules for mutual information,

$$I(X \wedge YZ) = I(X \wedge Y) + I(X \wedge Z|Y)$$

and

$$I(X \wedge YZ) = I(X \wedge Z) + I(X \wedge Y|Z).$$

- Show that conditional mutual information is nonnegative. In other words, show that $I(X \wedge Y|Z) \geq 0$, with equality if and only if X and Y are conditionally independent given Z . *Hint:* Explain why

$$I(X \wedge Y|Z = z) := \mathbb{E} \left[\log \frac{P_{XY|Z}(X, Y|z)}{P_{X|Z}(X|z)P_{Y|Z}(Y|z)} \middle| Z = z \right] \geq 0$$

for all z , and then use the laws of total probability and substitution.

- Show that

$$M \rightarrow \mathbf{X} \rightarrow \mathbf{Y} \text{ is a Markov chain} \Rightarrow M \rightarrow \mathbf{X} \rightarrow \varphi(\mathbf{Y}) \text{ is a Markov chain.}$$

5. **Data Processing Inequality.** If X and Z are conditionally independent given Y , show that $I(X \wedge Y) \geq I(X \wedge Z)$. *Hint:* Use the two chain rules for mutual information.
6. Show that if $P_{Z|YX}(z|y, x)$ does not depend on x , then X and Z are conditionally independent given Y . *Hint:* Since

$$P_{XZ|Y}(x, z|y) = \frac{P_{XYZ}(x, y, z)}{P_Y(y)} = \frac{P_{Z|YX}(z|y, x)P_{YX}(y, x)}{P_Y(y)} = P_{Z|YX}(z|y, x)P_{X|Y}(x|y),$$

it suffices to show that $P_{Z|YX}(z|y, x) = P_{Z|Y}(z|y)$.

7. **Data Processing Inequality Again.** Suppose X and Z are conditionally independent given Y . Show that $I(Y \wedge Z) \geq I(X \wedge Z)$. *Hint:* Expand $I(XY \wedge Z)$ in two different ways.
8. Put^d

$$H(Y|X = x) := \mathbb{E}[-\log P_{Y|X}(Y|X)|X = x] = \sum_y P_{Y|X}(y|x) \log \frac{1}{P_{Y|X}(y|x)}.$$

Use the law of total probability to show that

$$H(Y|X) = \sum_x H(Y|X = x)P(X = x).$$

9. In this problem you will derive three intuitive results. Part (c) says there is less uncertainty in $\varphi(Y)$ than in Y . Part (e) says that reducing conditioning increases uncertainty. Part (g) says that there is more mutual information between X and Y than between X and $\varphi(Y)$. Part (h) says that the uncertainty in X given $\varphi(Y)$ is more than the uncertainty of X given Y .

- (a) How that $H(\varphi(Y)|Y) = 0$. *Hint:* Conditioned on $Y = y$, $\varphi(Y)$ is a constant random variable.
- (b) Show that $I(Y \wedge \varphi(Y)) = H(\varphi(Y))$.
- (c) Show that $H(\varphi(Y)) \leq H(Y)$.
- (d) Show that $I(Z \wedge X|Y) = H(Z|Y) - H(Z|XY)$.
- (e) Show that $H(Z|XY) \leq H(Z|Y)$.
- (f) Show that $I(X \wedge \varphi(Y)|Y) = 0$.
- (g) Show that $I(X \wedge Y) \geq I(X \wedge \varphi(Y))$.

^dThe second equality follows by the law of substitution and evaluation of the resulting conditional expectation.

(h) Show that $H(X|\varphi(Y)) \geq H(X|Y)$.

10. Show that $I(p \times W)$ is a concave function of p . *Hint:* Use the fact that

$$\begin{aligned} I(X \wedge Y) &= H(Y) - H(Y|X) \\ &= H(pW) - \sum_x H(Y|X=x)p(x) \\ &= H(pW) - \sum_x H(W(\cdot|x))p(x). \end{aligned}$$

Problem 4.4 may also be helpful.

11. **Continuity of Mutual Information.** Let X and Y be finite sets, and let $P_{XY}(x,y)$ be a joint pmf on $X \times Y$. Show that $I(X \wedge Y) = I(P_{XY})$ is a continuous function of P_{XY} . More specifically, given $\varepsilon > 0$, show that there is a $\delta > 0$ such that

$$d(P_{XY}, Q_{XY}) < \delta \Rightarrow |I(P_{XY}) - I(Q_{XY})| < \varepsilon.$$

CHAPTER 7

General Techniques for Studying Channel Codes

7.1. More about Decoders

Given codewords $\mathbf{x}_1, \dots, \mathbf{x}_N$, each in X^n , decoder design usually starts with a “bright idea” in the form of subsets of Y^n , say $B_1, \dots, B_N$, that depend on the codewords. If $\mathbf{y} \in B_i$, we would like to put $\varphi(\mathbf{y}) := i$. However, we need to say more in order to guarantee that φ is well defined.

Example 7.1 (Maximum-Likelihood Decoding Sets). The sets

$$B_i := \{\mathbf{y} : W_n(\mathbf{y}|\mathbf{x}_i) \geq W_n(\mathbf{y}|\mathbf{x}_j) \text{ for all } j\},$$

are used for **maximum-likelihood (ML) decoding**. Later we need the fact that $\bigcup_{i=1}^N B_i = Y^n$. This can be seen examining the numbers $W_n(\mathbf{y}|\mathbf{x}_1), \dots, W_n(\mathbf{y}|\mathbf{x}_N)$. Find the largest; if there is a tie, it just means that $\mathbf{y}$ belongs to more than one of the B_i .

Although maximum likelihood is a good idea, it has a problem as far as using it as a decoder. The most obvious problem is that the B_i may not be disjoint. For example, if $\mathbf{y} \in B_i$ and if $W_n(\mathbf{y}|\mathbf{x}_i) = W_n(\mathbf{y}|\mathbf{x}_{i'})$ for some $\mathbf{x}_i \neq \mathbf{x}_{i'}$, then $\mathbf{y} \in B_{i'}$ as well. In this case, how should we define $\varphi(\mathbf{y})$? Should we put $\varphi(\mathbf{y}) = i$ or should we put $\varphi(\mathbf{y}) = i'$?

Suppose we had taken

$$B_i := \{\mathbf{y} : W_n(\mathbf{y}|\mathbf{x}_i) > W_n(\mathbf{y}|\mathbf{x}_j) \text{ for all } j \neq i\}.$$

In this case, the B_i would be disjoint. However, as noted in the previous paragraph, it may happen that $W_n(\mathbf{y}|\mathbf{x}_i) = W_n(\mathbf{y}|\mathbf{x}_{i'})$ for some $\mathbf{x}_i \neq \mathbf{x}_{i'}$. In this case $\mathbf{y}$ does not belong to any B_i ; how should we define $\varphi(\mathbf{y})$?

The next section addresses both situations; i.e., what to do if the B_i are not disjoint and what to do if $\bigcup_{i=1}^N B_i \neq Y^n$.

7.1.1. Decoders from Subsets

Given arbitrary subsets $B_1, \dots, B_N$ of Y^n , if $\mathbf{y}$ belongs to more than one B_i , announce the smallest corresponding value of i . If $\mathbf{y}$ does not belong to any B_i , announce some fixed message, say N .

Notation

In order to work with this definition mathematically, it is convenient to introduce the following notation. Let

$$F_1 := B_1$$

$$F_i := B_i \cap B_{i-1}^c \cap \cdots \cap B_1^c, \quad i = 2, \dots, N.$$

Note that $F_i \cap F_j = \emptyset$ for $i \neq j$. Next, put

$$F := \left(\bigcup_{i=1}^N F_i \right)^c = F_1^c \cap \cdots \cap F_N^c.$$

Then $F \cap F_i = \emptyset$ for $i = 1, \dots, N$, and

$$F_1 \cup \cdots \cup F_N \cup F = \mathcal{Y}^n.$$

If we put

$$\varphi(\mathbf{y}) := \sum_{i=1}^N i \mathbf{1}_{F_i}(\mathbf{y}) + N \mathbf{1}_F(\mathbf{y}),$$

then

$$\begin{aligned} \varphi(\mathbf{y}) = i &\Leftrightarrow \mathbf{y} \in F_i, & i = 1, \dots, N-1, \\ \varphi(\mathbf{y}) = N &\Leftrightarrow \mathbf{y} \in F_N \cup F. \end{aligned}$$

7.1.2. Decoder Properties

Now that we have φ expressed in terms of the foregoing notation, we can make some important observations that will enable us to derive bounds on the quantity $W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i)$ in (5.5) and (5.6).

We begin with

$$\varphi(\mathbf{y}) \neq i \Leftrightarrow \mathbf{y} \in F_i^c, \quad i = 1, \dots, N-1. \quad (7.1)$$

Next, since $\varphi(\mathbf{y}) = N \Leftrightarrow \mathbf{y} \in F_N \cup F$, we have

$$\varphi(\mathbf{y}) \neq N \Leftrightarrow \mathbf{y} \in F_N^c \cap F^c,$$

where it is important to note that

$$F_N^c \cap F^c \subset F_N^c.$$

Hence,

$$\varphi(\mathbf{y}) \neq i \Rightarrow \mathbf{y} \in F_i^c, \quad i = 1, \dots, N. \quad (7.2)$$

To conclude, observe that

$$F_i^c = \begin{cases} B_1^c, & i = 1, \\ B_i^c \cup \left(\bigcup_{j < i} B_j \right), & i = 2, \dots, N. \end{cases} \quad (7.3)$$

We can now write

$$F_i^c \subset B_i^c \cup \left(\bigcup_{j \neq i} B_j \right), \quad i = 1, \dots, N. \quad (7.4)$$

From (7.2) and (7.4), we have

$$\boxed{\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} \subset B_i^c \cup \left(\bigcup_{j \neq i} B_j \right), \quad \underline{\underline{i = 1, \dots, N}},} \quad (7.5)$$

and from (7.1) and (7.3) we have

$$\boxed{\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} = F_i^c \supset B_i^c, \quad \underline{\underline{i = 1, \dots, N-1}}.} \quad (7.6)$$

There are two special cases to be mentioned. First, if the B_i are disjoint, then $F_i = B_i$, and we have from (7.2) that

$$\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} \subset B_i^c, \quad i = 1, \dots, N.$$

Second, if the union of the B_i is equal to $\mathcal{Y}^n$ (e.g., ML decoding), then

$$B_i^c \subset \bigcup_{j \neq i} B_j.$$

It then follows from (7.5) that

$$\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} \subset \bigcup_{j \neq i} B_j, \quad i = 1, \dots, N. \quad (7.7)$$

7.2. The Random Coding Argument

Let g be a nonnegative function defined on some set $\mathcal{Z}$, and let Z be a $\mathcal{Z}$ -valued random variable such that $\mathbb{E}[g(Z)] < \lambda$. Then there is at least one $z \in \mathcal{Z}$ with $g(z) < \lambda$. To see this, suppose otherwise that $g(z) \geq \lambda$ for all z . Then we would have $\mathbb{E}[g(Z)] \geq \lambda$, which contradicts the hypothesis that $\mathbb{E}[g(Z)] < \lambda$.

We employ the **random coding argument** as follows. Write the error probability in (5.5) as a function of the codewords,^a

$$e_n(\mathbf{x}_1, \dots, \mathbf{x}_N) := \frac{1}{N} \sum_{i=1}^N W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i).$$

If we can find a *random* codebook $(\mathbf{X}_1, \dots, \mathbf{X}_N)$ such that $E[e_n(\mathbf{X}_1, \dots, \mathbf{X}_N)] < \lambda$, then there must be at least one codebook $(\mathbf{x}_1, \dots, \mathbf{x}_N) \in (\mathcal{X}^n)^N$ with $e_n(\mathbf{x}_1, \dots, \mathbf{x}_N) < \lambda$.

For later use, note that

$$\begin{aligned} E[e_n(\mathbf{X}_1, \dots, \mathbf{X}_N)] &= E\left[\frac{1}{N} \sum_{i=1}^N W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)\right] \\ &= \frac{1}{N} \sum_{i=1}^N E[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)]. \end{aligned} \quad (7.8)$$

Our approach is to obtain a bound on $E[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)]$ that does *not* depend on i . To this end, we first compute the conditional expectation

$$E[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}].$$

We then have by the law of total probability for expectation that

$$E[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)] = \sum_{\mathbf{x}} E[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}] P(\mathbf{X}_i = \mathbf{x}). \quad (7.9)$$

Problems

1. Consider the following subsets,

$$B_1 := \{1101, 0111, 1001\}$$

$$B_2 := \{0101, 1101, 0001\}$$

$$B_3 := \{1110, 1001, 0101, 1000\}.$$

Your friend suggests defining a decoder φ by the following rule: If $\mathbf{y} = b_1 b_2 b_3 b_4$, put $\varphi(\mathbf{y}) = i$ if $\mathbf{y} \in B_i$. If this is a well-defined decoding rule, prove it. If the rule is not well defined, explain why and modify the rule to make it well defined.

^aIn general, the decoder φ is defined in terms of the codewords. However, to keep the notation from getting out of hand, we just write $\varphi(\mathbf{y})$ instead of the more explicit $\varphi_{\mathbf{x}_1, \dots, \mathbf{x}_N}(\mathbf{y})$.

CHAPTER 8

Forward Theorem for the DMC

8.1. Joint Typicality

The success of the asymptotic equipartition property (AEP) (Theorem 4.4) in summarizing the key concepts used to prove Shannon's Source Coding Theorem suggests the following extension.

When $P_{XY}(x, y)$ is any joint probability mass function on $\mathsf{X} \times \mathsf{Y}$, we use the notation

$$P_X(x) := \sum_y P_{XY}(x, y) \quad \text{and} \quad P_Y(y) := \sum_x P_{XY}(x, y)$$

for the marginal probability mass functions on X and Y , respectively. Next, for $\mathbf{x} = (x_1, \dots, x_n) \in \mathsf{X}^n$ and $\mathbf{y} = (y_1, \dots, y_n) \in \mathsf{Y}^n$, we put

$$P_{XY}^n(\mathbf{x}, \mathbf{y}) := \prod_{k=1}^n P_{XY}(x_k, y_k).$$

The reader should verify that the marginals on X^n and Y^n satisfy

$$\sum_{\mathbf{y}} P_{XY}^n(\mathbf{x}, \mathbf{y}) = \prod_{k=1}^n P_X(x_k) =: P_X^n(\mathbf{x}) \quad \text{and} \quad \sum_{\mathbf{x}} P_{XY}^n(\mathbf{x}, \mathbf{y}) = \prod_{k=1}^n P_Y(y_k) =: P_Y^n(\mathbf{y}).$$

Given $\varepsilon > 0$, we say that $\mathbf{x}$ and $\mathbf{y}$ are **jointly typical** if pair $(\mathbf{x}, \mathbf{y})$ belongs to the set

$$A_n := \left\{ (\mathbf{x}, \mathbf{y}) \in \mathsf{X}^n \times \mathsf{Y}^n : \begin{aligned} & \left| \frac{1}{n} \log \frac{1}{P_X^n(\mathbf{x})} - H(P_X) \right| \leq \varepsilon, \\ & \left| \frac{1}{n} \log \frac{1}{P_Y^n(\mathbf{y})} - H(P_Y) \right| \leq \varepsilon, \\ & \text{and } \left| \frac{1}{n} \log \frac{1}{P_{XY}^n(\mathbf{x}, \mathbf{y})} - H(P_{XY}) \right| \leq \varepsilon \end{aligned} \right\}.$$

By writing $|t| \leq \varepsilon$ as $-\varepsilon \leq t \leq \varepsilon$, it is easy to show that

$$\begin{aligned} A_n = \left\{ (\mathbf{x}, \mathbf{y}) \in \mathsf{X}^n \times \mathsf{Y}^n : \right. & \exp(-n[H(P_X) + \varepsilon]) \leq P_X^n(\mathbf{x}) \leq \exp(-n[H(P_X) - \varepsilon]), \\ & \exp(-n[H(P_Y) + \varepsilon]) \leq P_Y^n(\mathbf{y}) \leq \exp(-n[H(P_Y) - \varepsilon]), \\ & \left. \text{and } \exp(-n[H(P_{XY}) + \varepsilon]) \leq P_{XY}^n(\mathbf{x}, \mathbf{y}) \leq \exp(-n[H(P_{XY}) - \varepsilon]) \right\}. \end{aligned}$$

We thus have upper and lower bounds on the marginal probabilities $P_X^n(\mathbf{x})$ and $P_Y^n(\mathbf{y})$ and on the joint probabilities $P_{XY}^n(\mathbf{x}, \mathbf{y})$.

Theorem 8.1 (Joint Asymptotic Equipartition Property (Joint AEP)). *Let A_n be defined as in Section 8.1, where $0 < \varepsilon < 1$ and $P_{XY}(x, y)$ have been given. The set of jointly typical pairs A_n has small cardinality in the sense that*

$$|A_n| \leq \exp(n[H(P_{XY}) + \varepsilon]).$$

For large n , the set A_n has high P_{XY}^n probability in the sense that

$$\lim_{n \rightarrow \infty} P_{XY}^n(A_n) = 1.$$

It then follows that for large n ,

$$|A_n| \geq (1 - \varepsilon) \exp(n[H(P_{XY}) - \varepsilon]).$$

If $Q_n(\mathbf{x}, \mathbf{y}) := P_X^n(\mathbf{x})P_Y^n(\mathbf{y})$, then

$$Q_n(A_n) \leq \exp(-n[I(P_{XY}) - 3\varepsilon]),$$

and for sufficiently large n ,

$$Q_n(A_n) \geq (1 - \varepsilon) \exp(-n[I(P_{XY}) + 3\varepsilon]).$$

Proof. Because we have a lower bound on $P_{XY}(\mathbf{x}, \mathbf{y})$ for $(\mathbf{x}, \mathbf{y}) \in A_n$, we can write

$$\begin{aligned} 1 &= \sum_{(\mathbf{x}, \mathbf{y})} P_{XY}^n(\mathbf{x}, \mathbf{y}) \geq \sum_{(\mathbf{x}, \mathbf{y}) \in A_n} P_{XY}^n(\mathbf{x}, \mathbf{y}) \geq \sum_{(\mathbf{x}, \mathbf{y}) \in A_n} \exp(-n[H(P_{XY}) + \varepsilon]) \\ &= |A_n| \exp(-n[H(P_{XY}) + \varepsilon]). \end{aligned}$$

Rearranging yields the upper bound on $|A_n|$.

To prove the limit result, it suffices to prove that $1 - P_{XY}^n(A_n) = P_{XY}^n(A_n^c) \rightarrow 0$. Let $(X_1, Y_1), (X_2, Y_2), \dots$ be i.i.d. pairs with common probability mass function P_{XY} . Put $\mathbf{X} := (X_1, \dots, X_n)$ and $\mathbf{Y} := (Y_1, \dots, Y_n)$. Then

$$\begin{aligned} \mathbb{P}((\mathbf{X}, \mathbf{Y}) \notin A_n) &= \mathbb{P}\left(\left|\frac{1}{n} \sum_{k=1}^n \log \frac{1}{P_X(X_k)} - H(P_X)\right| > \varepsilon, \right. \\ &\quad \text{or } \left|\frac{1}{n} \sum_{k=1}^n \log \frac{1}{P_Y(Y_k)} - H(P_Y)\right| > \varepsilon, \\ &\quad \left. \text{or } \left|\frac{1}{n} \sum_{k=1}^n \log \frac{1}{P_{XY}(X_k, Y_k)} - H(P_{XY})\right| > \varepsilon\right) \end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^n \log \frac{1}{P_X(X_k)} - H(P_X)\right| > \varepsilon\right) \\
&\quad + \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^n \log \frac{1}{P_Y(Y_k)} - H(P_Y)\right| > \varepsilon\right) \\
&\quad + \mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^n \log \frac{1}{P_{XY}(X_k, Y_k)} - H(P_{XY})\right| > \varepsilon\right).
\end{aligned}$$

In each probability on the right, the terms in the sums are i.i.d., and the expectation of each logarithm is the corresponding entropy. Hence, by the weak law of large numbers, each probability tends to zero as $n \rightarrow \infty$.

Since $P_{XY}^n(A_n) \rightarrow 1$, for large n , we have $P_{XY}^n(A_n) \geq 1 - \varepsilon$. Hence,

$$\begin{aligned}
1 - \varepsilon &\leq P_{XY}^n(A_n) \\
&= \sum_{(\mathbf{x}, \mathbf{y}) \in A_n} P_{XY}^n(\mathbf{x}, \mathbf{y}) \\
&\leq \sum_{(\mathbf{x}, \mathbf{y}) \in A_n} \exp(-n[H(P_{XY}) - \varepsilon]) \\
&= |A_n| \exp(-n[H(P_{XY}) - \varepsilon]).
\end{aligned}$$

Rearranging yields the lower bound on $|A_n|$.

To establish the last two inequalities, write

$$\begin{aligned}
Q_n(A_n) &= \sum_{(\mathbf{x}, \mathbf{y}) \in A_n} P_X^n(\mathbf{x}) P_Y^n(\mathbf{y}) \\
&\leq \sum_{(\mathbf{x}, \mathbf{y}) \in A_n} \exp(-n[H(P_X) - \varepsilon]) \exp(-n[H(P_Y) - \varepsilon]) \\
&= |A_n| \exp(-n[H(P_X) - \varepsilon]) \exp(-n[H(P_Y) - \varepsilon]) \\
&\leq \exp(n[H(P_{XY}) + \varepsilon]) \exp(-n[H(P_X) - \varepsilon]) \exp(-n[H(P_Y) - \varepsilon]) \\
&= \exp(-n[I(P_{XY}) - 3\varepsilon]),
\end{aligned}$$

and

$$\begin{aligned}
Q_n(A_n) &= \sum_{(\mathbf{x}, \mathbf{y}) \in A_n} P_X^n(\mathbf{x}) P_Y^n(\mathbf{y}) \\
&\geq \sum_{(\mathbf{x}, \mathbf{y}) \in A_n} \exp(-n[H(P_X) + \varepsilon]) \exp(-n[H(P_Y) + \varepsilon]) \\
&= |A_n| \exp(-n[H(P_X) + \varepsilon]) \exp(-n[H(P_Y) + \varepsilon]) \\
&\geq (1 - \varepsilon) \exp(n[H(P_{XY}) - \varepsilon]) \exp(-n[H(P_X) + \varepsilon]) \exp(-n[H(P_Y) + \varepsilon]) \\
&= (1 - \varepsilon) \exp(-n[I(P_{XY}) + 3\varepsilon]).
\end{aligned}$$

□

8.2. The Typicality Decoder for the DMC

Recall that in ML decoding, the sets B_i were defined only in terms of the channel transition probability W_n . In **typicality decoding**, the sets B_i depend not only on W_n but also on a probability mass function p on $\mathcal{X}$. In the next section, we use p to construct a random codebook, but for now, p is just a parameter used to construct the B_i .

For a DMC, recall that $W_n(\mathbf{y}|\mathbf{x}) = W^n(\mathbf{y}|\mathbf{x}) := \prod_{k=1}^n W(y_k|x_k)$. The typicality decoder is defined in terms of the sets^a

$$B_i := \{\mathbf{y} : (\mathbf{x}_i, \mathbf{y}) \in A_n\},$$

where now we take $P_{XY}(x, y) := p(x)W(y|x)$ in the definition of A_n (Section 8.1). In other words, the decoder decides message i was sent if $\mathbf{x}_i$ and $\mathbf{y}$ are jointly typical. If $\mathbf{y}$ is jointly typical with more than one codeword, we decide in favor of the smallest value of i . If there is no codeword with which $\mathbf{y}$ is jointly typical, then we announce message N .

8.3. Forward Theorem for the DMC

For a DMC W , put

$$C := \sup_p I(p \times W).$$

Theorem 8.2 (Forward Theorem for the DMC). *For a DMC W , if $R \leq C$, then R is achievable under the average probability of error criterion; i.e., $[0, C] \subset \mathcal{C}_a$.*

Proof. Since limits of achievable rates are achievable, it suffices to show that for fixed p , $I(p \times W)$ is an achievable rate. Again, since limits of achievable rates are achievable, it suffices to show that any $0 < R < I(p \times W)$ is an achievable rate. Fix any $0 < \lambda < 1$ and any $\Delta R > 0$. Choose any

$$0 < \varepsilon < \frac{I(p \times W) - R}{3}.$$

^aWhen we use a random codebook, $B_i = \{\mathbf{y} : (\mathbf{X}_i, \mathbf{y}) \in A_n\}$ depends on the random codeword $\mathbf{X}_i$.

Put $P_{XY}(x, y) := p(x)W(y|x)$, and for $n \geq 1$, define the jointly typical sets A_n as in Section 8.1. Put $N := \lfloor \exp(nR) \rfloor$. For all sufficiently large n , the following three conditions all hold:

$$\exp(-n[I(p \times W) - R - 3\epsilon]) < \lambda/2,$$

$$\frac{\log N}{n} > R - \Delta R, \quad \text{and} \quad P_{XY}^n(A_n^c) < \lambda/2.$$

The first inequality holds because $I(p \times W) - R - 3\epsilon > 0$; the second one holds by Problem 3.3; the third one holds by the joint AEP.

It remains to prove the existence of a codebook and decoder with $P(\varphi(\mathbf{Y}) \neq M) < \lambda$. First note that for any codebook and decoder, we have by (7.5) that

$$W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) \leq W_n(B_i^c | \mathbf{x}_i) + W_n\left(\bigcup_{j \neq i} B_j \middle| \mathbf{x}_i\right).$$

We next apply the random coding argument of Section 7.2. Consider random code-words $\mathbf{X}_1, \dots, \mathbf{X}_N$ chosen i.i.d. with $P(\mathbf{X}_i = \mathbf{x}) := p^n(\mathbf{x})$. Using the above right-hand side in (7.9), we get

$$\mathbb{E}[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)] \leq T_1 + T_2, \quad (8.1)$$

where

$$T_1 := \sum_{\mathbf{x}} \mathbb{E}[W_n(B_i^c | \mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}] p^n(\mathbf{x}),$$

and

$$T_2 := \sum_{\mathbf{x}} \mathbb{E}\left[W_n\left(\bigcup_{j \neq i} B_j \middle| \mathbf{X}_i\right) \middle| \mathbf{X}_i = \mathbf{x}\right] p^n(\mathbf{x}).$$

Below we show that if we use a typicality decoder with

$$B_i := \{\mathbf{y} : (\mathbf{x}_i, \mathbf{y}) \in A_n\},$$

then T_1 and T_2 are each upper bounded by $\lambda/2$. Then (8.1), and therefore (7.9), will be upper bounded by λ , which implies (7.8) is also upper bounded by λ . Hence, as argued in Section 7.2 there will exist an encoder-decoder pair with the required rate and with $P(\varphi(\mathbf{Y}) \neq M) < \lambda$; i.e., R will be achievable under the average probability of error criterion.

To bound T_1 , write

$$T_1 = \sum_{\mathbf{x}} \mathbb{E}\left[\sum_{\mathbf{y} : (\mathbf{X}_i, \mathbf{y}) \notin A_n} W_n(\mathbf{y} | \mathbf{X}_i) \middle| \mathbf{X}_i = \mathbf{x}\right] p^n(\mathbf{x})$$

$$\begin{aligned}
&= \sum_{\mathbf{x}} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{x}, \mathbf{y}) \notin A_n} W_n(\mathbf{y}|\mathbf{x}) \middle| \mathbf{X}_i = \mathbf{x} \right] p^n(\mathbf{x}) \\
&= \sum_{\mathbf{x}} \sum_{\mathbf{y}: (\mathbf{x}, \mathbf{y}) \notin A_n} W_n(\mathbf{y}|\mathbf{x}) p^n(\mathbf{x}).
\end{aligned}$$

Since we are working with a DMC, $W_n(\mathbf{y}|\mathbf{x}) = W^n(\mathbf{y}|\mathbf{x})$. Hence, $W_n(\mathbf{y}|\mathbf{x})p^n(\mathbf{x}) = W^n(\mathbf{y}|\mathbf{x})p^n(\mathbf{x}) = P_{XY}^n(\mathbf{x}, \mathbf{y})$. In other words, $T_1 = P_{XY}^n(A_n^c) < \lambda/2$.

To bound T_2 , write

$$\begin{aligned}
T_2 &\leq \sum_{j \neq i} \sum_{\mathbf{x}} \mathbb{E} [W_n(B_j|\mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}] p^n(\mathbf{x}), \quad \text{by the union bound,} \\
&= \sum_{j \neq i} \sum_{\mathbf{x}} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{X}_j, \mathbf{y}) \in A_n} W_n(\mathbf{y}|\mathbf{X}_i) \middle| \mathbf{X}_i = \mathbf{x} \right] p^n(\mathbf{x}) \\
&= \sum_{j \neq i} \sum_{\mathbf{x}} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{X}_j, \mathbf{y}) \in A_n} W_n(\mathbf{y}|\mathbf{x}) \middle| \mathbf{X}_i = \mathbf{x} \right] p^n(\mathbf{x}), \quad \text{by the law of subst.,} \\
&= \sum_{j \neq i} \sum_{\mathbf{x}} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{X}_j, \mathbf{y}) \in A_n} W_n(\mathbf{y}|\mathbf{x}) \right] p^n(\mathbf{x}), \quad \text{by independence,} \\
&= \sum_{j \neq i} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{X}_j, \mathbf{y}) \in A_n} \sum_{\mathbf{x}} W^n(\mathbf{y}|\mathbf{x}) p^n(\mathbf{x}) \right] \\
&= \sum_{j \neq i} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{X}_j, \mathbf{y}) \in A_n} P_Y^n(\mathbf{y}) \right] \\
&= \sum_{j \neq i} \sum_{\mathbf{x}'} p^n(\mathbf{x}') \sum_{\mathbf{y}: (\mathbf{x}', \mathbf{y}) \in A_n} P_Y^n(\mathbf{y}) \\
&= (N-1) Q_n(A_n) \\
&\leq (N-1) \exp(-n[I(P_{XY}) - 3\epsilon]), \quad \text{by the joint AEP,} \\
&\leq N \exp(-n[I(P_{XY}) - 3\epsilon]) \\
&\leq \exp(nR) \exp(-n[I(P_{XY}) - 3\epsilon]), \quad \text{since } N := \lfloor \exp(nR) \rfloor \leq \exp(nR), \\
&= \exp(-n[I(p \times W) - R - 3\epsilon]), \quad \text{since } P_{XY}(x, y) := p(x)W(y|x), \\
&< \lambda/2.
\end{aligned}$$

□

8.4. Generalization to Countably Infinite Sets

There is a difficulty in extending our proof of the Forward Theorem to the case when X or Y is a countably infinite set such as the positive integers. Recall the jointly typical sets A_n in Section 8.1. If X or Y is countably infinite, it may happen

that $H(P_X)$ or $H(P_Y)$ is infinite (e.g., Problem 3.12). In this case, the definition of A_n does not make sense. In some cases, this difficulty can be circumvented by carrying out the proof only for the p that maximizes $I(p \times W)$ if it can be shown that for *this* p , $H(p)$, $H(pW)$, and $H(p \times W)$ are all finite.

CHAPTER 9

Evaluation of the DMC Capacity

9.1. Weakly Symmetric Channels

Let W be a transition probability from a finite set X into a finite set Y . Recall that $C = \sup_p I(p \times W)$. Now, if $P_{XY}(x, y) := p(x)W(y|x)$,

$$\begin{aligned}
 I(p \times W) &= I(X \wedge Y) = H(Y) - H(Y|X) \\
 &= H(Y) - \sum_{x \in X} p(x)H(Y|X=x) \\
 &= H(Y) - \sum_{x \in X} p(x)H(W(\cdot|x)) \\
 &\leq \log |Y| - \sum_{x \in X} p(x)H(W(\cdot|x)),
 \end{aligned}$$

where equality holds $\Leftrightarrow P_Y(y) = \sum_{x \in X} p(x)W(y|x)$ is the uniform pmf on Y .

The foregoing observation suggests that to find channels whose capacity is easy to evaluate, we look for channels with the properties that

- (i) $H(W(\cdot|x))$ does not depend on x ; in this case, $H(Y|X)$ does not depend on p .
- (ii) We can find an input pmf p such that the output pmf pW is the uniform pmf on Y ; i.e., $(pW)(y) = 1/|Y|$.

For such channels, the maximum mutual information is $\log |Y| - H(W(\cdot|x))$, which does not depend on x .

Definition 9.1. A DMC W is **symmetric** if

- (i) For fixed $x, x' \in X$, the numbers $\{W(\cdot|x)\}$ are a permutation of $\{W(\cdot|x')\}$.
- (ii) For fixed $y, y' \in Y$, the numbers $\{W(y|\cdot)\}$ are a permutation of $\{W(y'|\cdot)\}$.

If we arrange the entries of W in a matrix, where x is the row index and y is the column index, then the first condition says that the rows are permutations of each other, and the second condition says that the columns are permutations of each other.

The main implication of the symmetry conditions stems from the fact that when you add up a list of numbers, the *order* in which you add does not change the sum.

More specifically, the first condition of a symmetric channel implies that

$$H(W(\cdot|x)) = \sum_y W(y|x) \log \frac{1}{W(y|x)}$$

does not depend on x . The second condition implies that if p is the uniform pmf on $\mathbf{X}$; i.e., $p(x) = 1/|\mathbf{X}|$, then

$$(pW)(y) = \frac{1}{|\mathbf{X}|} \sum_x W(y|x)$$

does not depend on y . Since the $(pW)(y)$ are nonnegative and sum to one, we must have $(pW)(y) = 1/|\mathbf{Y}|$ as we want.

Example 9.2. The mod- D adder channels of Example 5.1, of which the BSC of Example 5.3 (Figure 9.1) is a special case, are symmetric channels. Since $W(y|x) =$

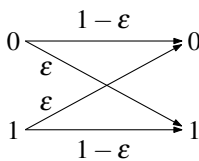


Figure 9.1. Binary symmetric channel with crossover probability ε .

$p_Z(y-x)$, for fixed x , as y runs over $\{0, \dots, D-1\}$, so does the mod- D difference $y-x$. This also happens if we fix y and let x run over $\{0, \dots, D-1\}$. Since $H(W(\cdot|x)) = H(p_Z(\cdot-x)) = H(p_Z)$, the capacity of the mod- D adder channel is $C = \log D - H(p_Z)$. In particular, the BSC with crossover probability $p_Z(1) = \varepsilon$ has capacity

$$C = 1 - h(\varepsilon) \text{ bits per channel use,}$$

where h is the binary entropy function and base-2 logarithms are used.

A little thought shows that the second condition in the definition of a symmetric channel is stronger than we need. This suggests the following definition.

Definition 9.3. A DMC W is **weakly symmetric** if

- (i) For fixed $x, x' \in \mathbf{X}$, the numbers $\{W(\cdot|x)\}$ are a permutation of $\{W(\cdot|x')\}$.
- (ii) $\sum_x W(y|x)$ does not depend on y .

If we think of W as a matrix, then the second condition says that the column sums are all the same.

Example 9.4 (Cover and Thomas [3]). Let $X = \{0, 1\}$, $Y = \{0, 1, 2\}$,

$$W(y|0) = \left(\frac{1}{3}, \frac{1}{6}, \frac{1}{2}\right)$$

$$W(y|1) = \left(\frac{1}{3}, \frac{1}{2}, \frac{1}{6}\right).$$

Then W is weakly symmetric, but not symmetric.

Example 9.5 (Binary Erasure Channel (BEC)). The **binary erasure channel (BEC)** is shown in Figure 9.2. It is defined as follows. Let $Y = \{0, 1, E\}$,

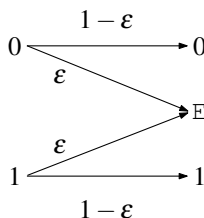


Figure 9.2. Binary erasure channel with erasure probability ε .

$$W(\cdot|0) = (1 - \varepsilon, \varepsilon, 0)$$

$$W(\cdot|1) = (0, \varepsilon, 1 - \varepsilon).$$

Although the rows are permutations of each other, the column sums are not constant; hence this channel is not weakly symmetric. However, since its rows are permutations of each other $H(W(\cdot|x)) = h(\varepsilon)$ does *not* depend on x . Let us now examine pW . Letting $\beta := p(1)$, we have

$$(pW)(0) = (1 - \varepsilon)(1 - \beta)$$

$$(pW)(E) = \varepsilon$$

$$(pW)(1) = (1 - \varepsilon)\beta.$$

Hence,

$$\begin{aligned} H(pW) &= (1 - \beta)(1 - \varepsilon) \log \frac{1}{(1 - \beta)(1 - \varepsilon)} + \varepsilon \log \frac{1}{\varepsilon} + \beta(1 - \varepsilon) \log \frac{1}{\beta(1 - \varepsilon)} \\ &= (1 - \varepsilon)h(\beta) + h(\varepsilon). \end{aligned}$$

Then

$$I(p \times W) = [(1 - \varepsilon)h(\beta) + h(\varepsilon)] - h(\varepsilon) = (1 - \varepsilon)h(\beta),$$

Maximizing over p means maximizing over $0 \leq \beta \leq 1$. Since $\beta = 1/2$ maximizes the binary entropy function with value $\log 2$, we have

$$C = \sup_{0 \leq \beta \leq 1} (1 - \varepsilon)h(\beta) = (1 - \varepsilon) \log 2.$$

Using base-2 logarithms, we see that the capacity of the BEC is $1 - \varepsilon$ bits per channel use. In other words, on average, a fraction of ε bits are erased over many uses of the channel.

Problems

1. Let $X = Y = \{1, 2, 3\}$, and let

$$W(\cdot|1) = (\tfrac{2}{3}, \tfrac{1}{3}, 0),$$

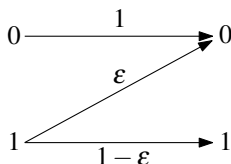
$$W(\cdot|2) = (\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}),$$

$$W(\cdot|3) = (0, \tfrac{1}{3}, \tfrac{2}{3}).$$

The goal of this problem is to find the capacity of the DMC with transition probability W .

- (a) Let $P_{X,Y}(x,y) := p(x)W(y|x)$. If $p(2)$ is fixed, show that $H(Y)$ is maximized if $p(1) = p(3)$, and this maximum value does not depend on $p(2)$.
 (b) Find $\sup_p I(p \times W)$. What is the maximizing distribution p ?

2. Consider the discrete memoryless Z channel:



Letting $p = P_X(1)$, find p to maximize $I(X \wedge Y)$.

3. Let X and Y be the binary-valued input and output of a DMC defined as follows. Let X , N_1 , and N_2 be independent $\{0, 1\}$ -valued random variables, and put $Y := X \oplus N_1 \oplus N_2$, where $\oplus$ denotes mod-2 addition (exclusive or):

$$0 \oplus 0 = 0, \quad 1 \oplus 0 = 1, \quad 0 \oplus 1 = 1, \quad \text{and} \quad 1 \oplus 1 = 0.$$

If $P(N_1 = 1) = P(N_2 = 1) = \varepsilon$, and if $W(y|x) := P(Y = y|X = x)$, find the capacity of the DMC W .

4. Let W_1 and W_2 be two discrete memoryless channels with capacities

$$C_1 := \sup_q I(X_1 \wedge Y_1) \quad \text{and} \quad C_2 := \sup_q I(X_2 \wedge Y_2),$$

where $P_{X_1 Y_1}(x, y) := q(x)W_1(y|x)$ and $P_{X_2 Y_2}(x, y) := q(x)W_2(y|x)$. If we define the new DMC

$$W(y_1, y_2 | x_1, x_2) := W_1(y_1 | x_1)W_2(y_2 | x_2),$$

then W can be considered as modeling **parallel** DMCs that are independent. Show that the capacity of W is $C_1 + C_2$; i.e., show that

$$\sup_p I(X_1 X_2 \wedge Y_1 Y_2) = C_1 + C_2,$$

where

$$P_{X_1 X_2 Y_1 Y_2}(x_1, x_2, y_1, y_2) = p(x_1, x_2)W_1(y_1 | x_1)W_2(y_2 | x_2).$$

5. The **cascade** of DMCs $W(y|x)$ and $V(z|y)$ is

$$U(z|x) := \sum_y V(z|y)W(y|x).$$

Given an arbitrary input pmf $p(x)$, define the joint pmf

$$P_{XYZ}(x, y, z) := p(x)W(y|x)V(z|y).$$

Show that

$$\sup_p I(X \wedge Z) \leq \min(C_W, C_V),$$

where

$$C_W := \sup_p I(p \wedge W), \quad \text{and} \quad C_V := \sup_q I(q \times V).$$

CHAPTER 10

Codeword Constraints

If a is nonnegative “cost” function defined on X , we may require that all codewords satisfy

$$a_n(\mathbf{x}) := \frac{1}{n} \sum_{k=1}^n a(x_k) \leq A$$

for some constant A . In other words, the total cost to transmit a codeword cannot exceed nA . When $\mathsf{X} = \mathbb{R}$, we usually take $a(x) = x^2$ as a measure of the energy required to send the symbol x , and $a_n(\mathbf{x}) = \|\mathbf{x}\|^2/n$ is the average power used to transmit the codeword.

If $A_{\max} := \sup_{x \in \mathsf{X}} a(x)$ is finite, and if $A \geq A_{\max}$, then *all* codewords satisfy $a_n(\mathbf{x}) \leq A$, and the constraint is said to be **inactive**. If $A < A_{\min} := \inf_{x \in \mathsf{X}} a(x)$, then *no* codewords satisfy $a_n(\mathbf{x}) \leq A$.

Definition 10.1. A real number $R \geq 0$ is an **achievable rate** for a channel (more precisely, a sequence of channels W_n), under the average probability of error criterion and codeword constraint A , if $\forall \lambda > 0, \forall \Delta R > 0, \exists n_0$ such that $\forall n \geq n_0, \exists$ codewords $\mathbf{x}_1, \dots, \mathbf{x}_N$ each in X^n and satisfying $a_n(\mathbf{x}_i) \leq A$, and there exists a decoder $\varphi: \mathsf{Y}^n \rightarrow \{1, \dots, N\}$ with

$$\frac{\log N}{n} > R - \Delta R \quad \text{and} \quad \frac{1}{N} \sum_{i=1}^N W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) < \lambda.$$

The definition of **achievable rate** under the maximum probability of error and codeword constraint A is similar.

10.1. Capacity Regions

Let $\mathcal{D}_A(\mathsf{X})$ denote the set of pmfs p on X with the property that $\sum_x a(x)p(x) \leq A$. We put

$$C(A) := \sup_{p \in \mathcal{D}_A(\mathsf{X})} I(p \times W).$$

If X is a random variable with pmf p , then $E[a(X)] = \sum_x a(x)p(x)$. Hence, we can also write

$$C(A) = \sup\{I(X \wedge Y) : E[a(X)] \leq A \text{ and } P_{Y|X}(y|x) = W(y|x)\}.$$

We denote by $\mathcal{C}_a(A)$ the set of all rates that are achievable under Definition 10.1. We denote by $\mathcal{C}_m(A)$ the set of achievable rates under the maximum probability of error and codeword constraint A . In this chapter we prove the weak-converse result that $\mathcal{C}_a(A) \subset [0, C(A)]$ and the forward result $[0, C(A)] \subset \mathcal{C}_m(A)$. Since we trivially have $\mathcal{C}_m(A) \subset \mathcal{C}_a(A)$, it follows that

$$\mathcal{C}_m(A) = \mathcal{C}_a(A) = [0, C(A)].$$

10.1.1. Capacity-Cost Function

We call $C(A)$ the **capacity-cost function**. As mentioned in the last footnote, for $A < A_{\min}$, no codeword can satisfy $a_n(\mathbf{x}) \leq A$. Hence, for such A there are no achievable rates, and the capacity region is empty. Although we may put $C(A) := -\infty$ when the capacity region is empty, we usually just restrict attention to $A \geq A_{\min}$. However, if X is an infinite set, it may happen that there is no $x \in X$ that achieves the infimum $A_{\min} := \inf_{x \in X} a(x)$. In this case, no codeword can satisfy $a_n(\mathbf{x}) \leq A_{\min}$ and the capacity region is empty. In these cases, it is understood that we would restrict attention to $A > A_{\min}$. Capacity-cost curves are always nondecreasing as shown in Figure 10.1. This can be established mathematically as follows. From

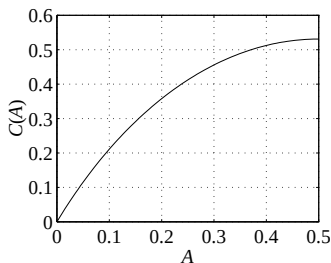


Figure 10.1. A typical capacity-cost function $C(A)$.

the definition of achievable rate, it is clear that $A_1 \leq A_2$ implies $\mathcal{C}(A_1) \subset \mathcal{C}(A_2)$, which then implies $C(A_1) \leq C(A_2)$. Hence, $C(A)$ is nondecreasing. We also point out that if $A_{\max} := \sup_{x \in X} a(x)$ is finite, then $C(A)$ becomes constant for large A ; in particular, for $A \geq A_{\max}$, $C(A) = C(A_{\max})$. This follows because, as mentioned in the last footnote, if $A \geq A_{\max}$, then all codewords satisfy $a_n(\mathbf{x}) \leq A$.

10.2. Weak Converse

Theorem 10.2 (Weak Converse for the DMC under a Codeword Constraint). *If R is an achievable rate for a DMC W under the average probability of error criterion and codeword constraint A , then $R \leq C(A)$.*

Proof. Fix any $0 < \lambda < 1$ and any $\Delta R > 0$. Since R is achievable, for all sufficiently large n , there exist codewords $\mathbf{x}_1, \dots, \mathbf{x}_N$ with $a_n(\mathbf{x}_i) \leq A$ and there exists a decoder φ with

$$\frac{\log N}{n} > R - \Delta R \quad \text{and} \quad P(\varphi(\mathbf{Y}) \neq M) < \lambda,$$

where $M, \mathbf{X}, \mathbf{Y}$ have joint pmf (cf. Example 6.2)

$$P(M = i, \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = \frac{1}{N} I_{\{i\}}(\mathbf{x}) W^n(\mathbf{y}|\mathbf{x}). \quad (10.1)$$

Since M is uniformly distributed, $H(M) = \log N$. By the corollary to Fano's inequality,

$$\log N = H(M) \leq \lambda \log N + \log 2 + I(M \wedge \varphi(\mathbf{Y})).$$

Rearrange this as

$$(1 - \lambda) \log N \leq I(M \wedge \varphi(\mathbf{Y})) + \log 2.$$

Since $(\log N)/n > R - \Delta R$, we have $\log N > n(R - \Delta R)$, and so

$$\begin{aligned} (1 - \lambda)n(R - \Delta R) &\leq I(\mathbf{X} \wedge \mathbf{Y}) + \log 2, && \text{by the Data Processing Lemma,} \\ &\leq \sum_{k=1}^n I(X_k \wedge Y_k) + \log 2, && \text{by Lemma 6.4.} \end{aligned}$$

Rewrite this as

$$(1 - \lambda)(R - \Delta R) \leq \frac{1}{n} \sum_{k=1}^n I(X_k \wedge Y_k) + \frac{\log 2}{n}.$$

To see what is really happening in this average of mutual informations, consider the random variables $\widehat{V}$, $\widehat{X}$, and $\widehat{Y}$ with joint pmf

$$P_{\widehat{V}\widehat{X}\widehat{Y}}(k, x, y) := \frac{1}{n} p_k(x) W(y|x), \quad y \in \mathcal{Y}, x \in \mathcal{X}, k \in \{1, \dots, n\}. \quad (10.2)$$

where $p_k(x) := P(X_k = x)$ is induced by the joint pmf in (10.1). Then

$$\begin{aligned}
 I(\widehat{X} \wedge \widehat{Y} | \widehat{V}) &= \sum_{k=1}^n I(\widehat{X} \wedge \widehat{Y} | \widehat{V} = k) P(\widehat{V} = k) \\
 &= \frac{1}{n} \sum_{k=1}^n \sum_{x,y} P_{\widehat{X}\widehat{Y}|\widehat{V}}(x,y|k) \log \frac{P_{\widehat{X}\widehat{Y}|\widehat{V}}(x,y|k)}{P_{\widehat{X}|\widehat{V}}(x|k) P_{\widehat{Y}|\widehat{V}}(y|k)} \\
 &= \frac{1}{n} \sum_{k=1}^n \sum_{x,y} p_k(x) W(y|x) \log \frac{W(y|x)}{(p_k W)(y)} \\
 &= \frac{1}{n} \sum_{k=1}^n I(X_k \wedge Y_k).
 \end{aligned}$$

We can now write

$$\begin{aligned}
 R &\leq \frac{I(\widehat{X} \wedge \widehat{Y} | \widehat{V})}{1 - \lambda} + \frac{\log 2}{n(1 - \lambda)} + \Delta R \\
 &\leq \frac{I(\widehat{X} \wedge \widehat{Y})}{1 - \lambda} + \frac{\log 2}{n(1 - \lambda)} + \Delta R, \quad \text{by Problem 10.3,} \\
 &\leq \frac{C(A)}{1 - \lambda} + \frac{\log 2}{n(1 - \lambda)} + \Delta R,
 \end{aligned}$$

since $P_{\widehat{Y}|\widehat{X}}(y|x) = W(y|x)$, and $E[a(\widehat{X})] \leq A$ by Problem 10.2.

Since $\lambda > 0$ and $\Delta R > 0$ can be arbitrarily small, and n can be arbitrarily large, we must have $R \leq C(A)$. $\square$

10.3. A Forward Theorem

Theorem 10.3 (Forward Theorem for the DMC under a Codeword Constraint). *For a DMC W , if $R \leq C(A)$, then R is achievable under the maximum probability of error criterion and codeword constraint A ; i.e., $[0, C(A)] \subset \mathcal{C}_m(A)$.*

Proof. Since limits of achievable rates are achievable, it suffices to show that for fixed p with $\sum_x a(x)p(x) \leq A$, $I(p \times W)$ is an achievable rate. Again, since limits of

achievable rates are achievable, it suffices to show that any $0 < R < I(p \times W)$ is an achievable rate. Fix any $0 < \lambda' < 1$ and any $\Delta R > 0$. Put $\lambda := \lambda'/2$. Choose any

$$0 < \varepsilon < \frac{I(p \times W) - R}{3}.$$

Put $P_{XY}(x, y) := p(x)W(y|x)$, and for $n \geq 1$, define the jointly typical sets A_n as in Section 8.1. Put $N := \lfloor \exp(nR) \rfloor$. For all sufficiently large n , the following three conditions all hold:

$$\exp(-n[I(p \times W) - R - 3\varepsilon]) < \lambda/2,$$

$$\frac{\log N}{n} > R - \Delta R/2, \quad \text{and} \quad P_{XY}^n(A_n^c) < \lambda/4.$$

The first inequality holds because $I(p \times W) - R - 3\varepsilon > 0$; the second one holds by Problem 3.3; the third one holds by the joint AEP.

We need that two other conditions hold for all sufficiently large n . The first is that $(\log 4)/n < \Delta R/2$. The second is that if $\mathbf{X} = (X_1, \dots, X_n)$ with the X_k being i.i.d. with pmf p such that $\mathbb{E}[a(X_k)] = \sum_x a(x)p(x) \leq A$, then

$$\mathbb{P}(a_n(\mathbf{X}) > A) = \mathbb{P}\left(\frac{1}{n} \sum_{k=1}^n a(X_k) > A\right) < \lambda/4.$$

This can be done by the weak law of large numbers.

It remains to prove the existence of a codebook with $a_n(\mathbf{x}_i) \leq A$ and decoder with $\mathbb{P}(\varphi(\mathbf{Y}) \neq M) < \lambda$. First note that for any codebook and decoder, we have by (7.5) that

$$W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) \leq W_n(B_i^c | \mathbf{x}_i) + W_n\left(\bigcup_{j \neq i} B_j \middle| \mathbf{x}_i\right).$$

We next apply the random coding argument of Section 7.2. Consider random code-words $\mathbf{X}_1, \dots, \mathbf{X}_N$ chosen i.i.d. with $\mathbb{P}(\mathbf{X}_i = \mathbf{x}) := p^n(\mathbf{x})$. Using the above right-hand side in (7.9), we get

$$\mathbb{E}[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)] \leq T_1 + T_2, \quad (10.3)$$

where

$$T_1 := \sum_{\mathbf{x}} \mathbb{E}[W_n(B_i^c | \mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}] p^n(\mathbf{x}),$$

and

$$T_2 := \sum_{\mathbf{x}} \mathbb{E} \left[W_n \left(\bigcup_{j \neq i} B_j \middle| \mathbf{X}_i \right) \middle| \mathbf{X}_i = \mathbf{x} \right] p^n(\mathbf{x}).$$

Below we show that if we use a typicality decoder with

$$B_i := \{ \mathbf{y} : (\mathbf{x}_i, \mathbf{y}) \in A_n \text{ and } a_n(\mathbf{x}_i) \leq A \},$$

then T_1 and T_2 are each upper bounded by $\lambda/2$. Then (8.1), and therefore (7.9), will be upper bounded by λ , which implies (7.8) is also upper bounded by λ . Hence, as argued in Section 7.2 there will exist an encoder-decoder pair with the required rate and with $P(\varphi(\mathbf{Y}) \neq M) < \lambda$; i.e., R will be achievable under the average probability of error criterion.

To bound T_1 , we need the observation that

$$\begin{aligned} W_n(B_i^c | \mathbf{x}_i) &\leq W_n(\{ \mathbf{y} : (\mathbf{x}_i, \mathbf{y}) \notin A_n \} | \mathbf{x}_i) + W_n(\{ \mathbf{y} : a_n(\mathbf{x}_i) > A \} | \mathbf{x}_i) \\ &= W_n(\{ \mathbf{y} : (\mathbf{x}_i, \mathbf{y}) \notin A_n \} | \mathbf{x}_i) + I_{(A, \infty)}(a_n(\mathbf{x}_i)). \end{aligned}$$

Then

$$\begin{aligned} T_1 &\leq \sum_{\mathbf{x}} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{x}_i, \mathbf{y}) \notin A_n} W_n(\mathbf{y} | \mathbf{x}_i) \middle| \mathbf{X}_i = \mathbf{x} \right] p^n(\mathbf{x}) + P(a_n(\mathbf{X}_i) > A) \\ &\leq \sum_{\mathbf{x}} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{x}, \mathbf{y}) \notin A_n} W_n(\mathbf{y} | \mathbf{x}) \middle| \mathbf{X}_i = \mathbf{x} \right] p^n(\mathbf{x}) + \lambda/4 \\ &= \sum_{\mathbf{x}} \sum_{\mathbf{y}: (\mathbf{x}, \mathbf{y}) \notin A_n} W_n(\mathbf{y} | \mathbf{x}) p^n(\mathbf{x}) + \lambda/4. \end{aligned}$$

Since we are working with a DMC, $W_n(\mathbf{y} | \mathbf{x}) = W^n(\mathbf{y} | \mathbf{x})$. Hence, $W_n(\mathbf{y} | \mathbf{x}) p^n(\mathbf{x}) = W^n(\mathbf{y} | \mathbf{x}) p^n(\mathbf{x}) = P_{XY}^n(\mathbf{x}, \mathbf{y})$. In other words, $T_1 \leq P_{XY}^n(A_n^c) + \lambda/4 < \lambda/2$.

To bound T_2 , write

$$\begin{aligned} T_2 &\leq \sum_{j \neq i} \sum_{\mathbf{x}} \mathbb{E} [W_n(B_j | \mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}] p^n(\mathbf{x}), \quad \text{by the union bound,} \\ &\leq \sum_{j \neq i} \sum_{\mathbf{x}} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{x}_j, \mathbf{y}) \in A_n} W_n(\mathbf{y} | \mathbf{X}_i) \middle| \mathbf{X}_i = \mathbf{x} \right] p^n(\mathbf{x}) \end{aligned}$$

$$\begin{aligned}
&= \sum_{j \neq i} \sum_{\mathbf{x}} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{X}_j, \mathbf{y}) \in A_n} W_n(\mathbf{y}|\mathbf{x}) \middle| \mathbf{X}_i = \mathbf{x} \right] p^n(\mathbf{x}), \quad \text{by the law of subst.,} \\
&= \sum_{j \neq i} \sum_{\mathbf{x}} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{X}_j, \mathbf{y}) \in A_n} W_n(\mathbf{y}|\mathbf{x}) \right] p^n(\mathbf{x}), \quad \text{by independence,} \\
&= \sum_{j \neq i} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{X}_j, \mathbf{y}) \in A_n} \sum_{\mathbf{x}} W^n(\mathbf{y}|\mathbf{x}) p^n(\mathbf{x}) \right] \\
&= \sum_{j \neq i} \mathbb{E} \left[\sum_{\mathbf{y}: (\mathbf{X}_j, \mathbf{y}) \in A_n} P_Y^n(\mathbf{y}) \right] \\
&= \sum_{j \neq i} \sum_{\mathbf{x}'} p^n(\mathbf{x}') \sum_{\mathbf{y}: (\mathbf{x}', \mathbf{y}) \in A_n} P_Y^n(\mathbf{y}) \\
&= (N-1) Q_n(A_n) \\
&\leq (N-1) \exp(-n[I(P_{XY}) - 3\epsilon]), \quad \text{by the joint AEP,} \\
&\leq N \exp(-n[I(P_{XY}) - 3\epsilon]) \\
&\leq \exp(nR) \exp(-n[I(P_{XY}) - 3\epsilon]), \quad \text{since } N := \lfloor \exp(nR) \rfloor \leq \exp(nR), \\
&= \exp(-n[I(p \times W) - R - 3\epsilon]), \quad \text{since } P_{XY}(x, y) := p(x)W(y|x), \\
&< \lambda/2.
\end{aligned}$$

We have now shown that there is a random codebook $\mathbf{X}_1, \dots, \mathbf{X}_N$ with $P(a_n(\mathbf{X}_i) > A) < \lambda/4$ and

$$\mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N W^n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i) \right] < \lambda.$$

By the random coding argument, there is a codebook realization $\mathbf{x}_1, \dots, \mathbf{x}_N$ such that with our typicality decoder,

$$\frac{1}{N} \sum_{i=1}^N W^n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) < \lambda.$$

However, since we did not guarantee that $P(a_n(\mathbf{X}_i) > A) = 0$, it may happen that some $\mathbf{x}_i$ have $a_n(\mathbf{x}_i) > A$. To avoid this question, we use our observation on small averages in Section 5.5.1 to say that there must be at least $N/2$ codewords with

$$W^n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) < 2\lambda.$$

Let G denote the set of all such messages i *excluding the possible case* $i = N$. Then $|G| > N/2 - 1$. For $i \in G$, we claim that $a_n(\mathbf{x}_i) \leq A$; for if $a_n(\mathbf{x}_i) > A$, then $B_i = \emptyset$,

which implies, by (7.6), that $\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} = \mathbf{Y}^n$, which implies the above probability is one, which is not less than 2λ . We also note that

$$\begin{aligned} \frac{\log |G|}{n} &> \frac{\log(N/2 - 1)}{n} \geq \frac{\log(N/2 - N/4)}{n} = \frac{\log N/4}{n} = \frac{\log N}{n} - \frac{\log 4}{n} \\ &> (R - \Delta R/2) - \Delta R/2 \\ &= R - \Delta R. \end{aligned}$$

We thus have a modified encoder-decoder pair (f_G, φ_G) showing that R is achievable under the maximum probability of error and codeword constraint A . $\square$

Problems

1. Show that if M , $\mathbf{X}$, and $\mathbf{Y}$ have the joint pmf in (10.1), then for any function $g(\mathbf{x})$,

$$\mathbb{E}[g(\mathbf{X})] = \frac{1}{N} \sum_{i=1}^N g(\mathbf{x}_i).$$

Remark. If $\mathbf{x} = (x_1, \dots, x_n)$ and $g(\mathbf{x}) = a(x_k)$, then

$$\mathbb{E}[a(X_k)] = \frac{1}{N} \sum_{i=1}^N a(x_{i,k}),$$

where $x_{i,k}$ is the k th component of codeword $\mathbf{x}_i$. If p_k denotes the pmf of X_k , then the above display says that

$$\sum_x a(x) p_k(x) = \frac{1}{N} \sum_{i=1}^N a(x_{i,k}).$$

2. Let $\hat{V}$, $\hat{X}$, and $\hat{Y}$ have the joint pmf of (10.2). Show that if each codeword $\mathbf{x}_i$ satisfies $a_n(\mathbf{x}_i) \leq A$, then $\mathbb{E}[a(\hat{X})] \leq A$. *Hint:* The remark in Problem 10.1 may be helpful.
3. Let $\hat{V}$, $\hat{X}$, and $\hat{Y}$ have the joint pmf of (10.2). Show that

$$I(\hat{X} \wedge \hat{Y}) \geq I(\hat{X} \wedge \hat{Y} | \hat{V}).$$

Hint: Use chain rules for mutual information along with the fact that $\hat{V} \rightarrow \hat{X} \rightarrow \hat{Y}$ is a Markov chain.

CHAPTER 11

Continuous Channels

11.1. Differential Entropy

By analogy with the entropy of probability mass functions, the *differential entropy* of a probability *density* function (pdf) p is

$$H_d(p) := \int p(x) \log \frac{1}{p(x)} dx.$$

If X is a continuous random variable with density p_X , we sometimes write $H_d(X)$ instead of $H_d(p_X)$. In this case, we can also write $H_d(X) = \mathbb{E}[-\log p_X(X)]$.

Example 11.1. Although differential entropy is defined by analogy with entropy, the two are fundamentally different. One way to see this is to approximate the above integral by a Riemann sum. For $\Delta > 0$,

$$\begin{aligned} \int p(x) \log \frac{1}{p(x)} dx &\approx \sum_k \left[p(k\Delta) \log \frac{1}{p(k\Delta)} \right] \Delta \\ &= \sum_k p(k\Delta) \Delta \log \frac{\Delta}{p(k\Delta) \Delta} \\ &= \sum_k p(k\Delta) \Delta \log \frac{1}{p(k\Delta) \Delta} + (\log \Delta) \sum_k p(k\Delta) \Delta. \end{aligned}$$

For small Δ , $p(k\Delta)\Delta$ is approximately a probability that lies in $[0, 1]$. Hence, the first sum is nearly an ordinary entropy, which must be nonnegative. However, the second sum is approximately $(\log \Delta) \int p(x) dx = \log \Delta$, which is a negative number for $\Delta < 1$.

Example 11.2 (Invariance with Respect to Translation). We show that

$$H_d(p(\cdot - x_0)) = H_d(p),$$

which does not depend on the shift x_0 . Write

$$H(p(\cdot - x_0)) = \int p(x - x_0) \log \frac{1}{p(x - x_0)} dx = \int p(y) \log \frac{1}{p(y)} dy = H_d(p),$$

where the second equality follows by making the change of variable $y = x - x_0$.

It is instructive to write^a

$$H_d(p) = \int_{\{x:p(x)>1\}} p(x) \log \frac{1}{p(x)} dx + \int_{\{x:p(x)\leq 1\}} p(x) \log \frac{1}{p(x)} dx,$$

where the left-hand integral is omitted if $p(x) \leq 1$ for all x . Notice that the left-hand integral is negative, while the right-hand one is positive. If at least one of these integrals is finite, then the differential entropy is well defined since we allow $\pm\infty$ as possible values for $H_d(p)$. Otherwise, since the expression $-\infty + \infty$ is undefined, the differential entropy is undefined. In the Problems you will construct a densities whose differential entropy is $+\infty$, $-\infty$, and undefined.

Example 11.3. If X is a continuous random variable satisfying $E[|X|^r] < \infty$ for some $r > 0$, then it can be shown that the above right-hand integral is finite,¹ thus implying $H_d(X) < \infty$. Hence, if $p_X(x) \leq 1$ and $E[|X|^r] < \infty$, then $0 \leq H_d(X) < \infty$.

Example 11.4. We show that if for some $\delta > 0$, $\int p(x)^{1+\delta} dx < \infty$, then the above left-hand integral is finite, thus implying $H_d(p) > -\infty$. Since $\lim_{t \rightarrow \infty} (\log t)/t^\delta = 0$, let t_0 be such that for $t > t_0$, we have $(\log t)/t^\delta \leq 1$. Then write

$$\begin{aligned} \int_{p(x)>1} p(x) \log p(x) dx &= \int_{1 < p(x) \leq t_0} p(x) \log p(x) dx + \int_{p(x)>t_0} p(x) \log p(x) dx \\ &\leq \int_{1 < p(x) \leq t_0} p(x) \log t_0 dx + \int_{p(x)>t_0} p(x) p(x)^\delta dx \\ &\leq \log t_0 \int p(x) dx + \int_{p(x)>t_0} p(x)^{1+\delta} dx \\ &= \log t_0 + \int p(x)^{1+\delta} dx < \infty. \end{aligned}$$

We summarize Examples 11.3 and 11.4 by writing

$$E[|X|^r] < \infty \Rightarrow H_d(X) < \infty \quad \text{and} \quad \int p(x)^{1+\delta} dx < \infty \Rightarrow H_d(X) > -\infty.$$

Example 11.5. Let X have the unbounded density $p(x) = 1/(2\sqrt{x})$ for $0 < x \leq 1$. For $0 < \delta < 1$,

$$\int_0^1 p(x)^{1+\delta} dx = \frac{1}{2^{1+\delta}} \int_0^1 x^{-(1+\delta)/2} dx = \frac{1}{2^\delta(1-\delta)} < \infty.$$

^aRecall that although the values of a pmf always lie between zero and one, a density $p(x)$ may be greater than one for many values of x ; e.g., the uniform density on $[0, 1/2]$ or the Gaussian density with $\sigma < 1/\sqrt{2\pi}$. In fact, a density can be unbounded; e.g., $p(x) = 1/(2\sqrt{x})$ for $0 < x \leq 1$.

For all $r > 0$,

$$\mathbb{E}[|X|^r] = \frac{1}{2} \int_0^1 x^{r-1/2} dx = \frac{1}{2r+1} < \infty.$$

Hence, this density satisfies $-\infty < H_d(p) < \infty$.

Example 11.6. We show that if p is a bounded density, then $H_d(p) > -\infty$. Suppose that $p(x) \leq B < \infty$ for all x . Then

$$\int p(x)^{1+\delta} dx \leq \int p(x) B^\delta dx = B^\delta \int p(x) dx = B^\delta < \infty.$$

Hence, $H_d(p) > -\infty$.

We will see shortly that differential entropy can be negative, even $-\infty$ (Problem 11.6). This property makes it distinctly different from the entropy of pmfs and discrete random variables.

11.2. Divergence for Densities

If p and q are probability densities, then their **Kullback–Leibler informational divergence** or **relative entropy** is

$$D(p\|q) := \int p(x) \log \frac{p(x)}{q(x)} dx,$$

where the integral is understood to be over those x with $p(x) > 0$. Just as in the discrete pmf case, the log inequality can be used to show that for densities, $-D(p\|q) \leq 0$; hence $D(p\|q) \geq 0$.

Example 11.7. Let g denote the Gaussian density with mean m and variance σ^2 . If p also has mean m and variance σ^2 , we show that

$$D(p\|g) = -H_d(p) + \frac{1}{2} \log(2\pi\sigma^2 e),$$

where e is the base of the natural logarithm. Since

$$g(x) = \frac{e^{-(x-m)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma}},$$

we have

$$\log g(x) = -\frac{(x-m)^2}{2\sigma^2} \log e - \frac{1}{2} \log(2\pi\sigma^2),$$

where the factor $\log e$ accounts for the fact that $\log$ may not be the natural logarithm. Now write^b

$$\begin{aligned}
 D(p\|g) &= \int p(x)[\log p(x) - \log g(x)] dx \\
 &= -H_d(p) + \int p(x) \left[\frac{(x-m)^2}{2\sigma^2} \log e + \frac{1}{2} \log(2\pi\sigma^2) \right] dx \\
 &= -H_d(p) + \frac{\sigma^2}{2\sigma^2} \log e + \frac{1}{2} \log(2\pi\sigma^2) \\
 &= -H_d(p) + \frac{1}{2} \log(2\pi\sigma^2 e).
 \end{aligned}$$

From the preceding example and the fact that $D(p\|q) \geq 0$, we see that for any continuous random variable X with mean m and variance σ^2 ,^c

$$H_d(X) \leq \frac{1}{2} \log(2\pi\sigma^2 e), \quad (11.1)$$

with equality if $X \sim N(m, \sigma^2)$. Thus,

$$H_d(g) = \frac{1}{2} \log(2\pi\sigma^2 e).$$

As we expect from Example 11.2, $H_d(g)$ does not depend on the mean of the density. Notice also that for small σ^2 , the above logarithm is negative, while for large σ^2 , the logarithm is positive. Hence, differential entropy can be positive or negative!

The divergence between joint densities p_{XY} and q_{XY} is

$$D(p_{XY}\|q_{XY}) := \iint p_{XY}(x, y) \log \frac{p_{XY}(x, y)}{q_{XY}(x, y)} dx dy.$$

If X and Y are random variables with joint density p_{XY} , then

$$D(p_{XY}\|q_{XY}) = \mathbb{E} \left[\log \frac{p_{XY}(X, Y)}{q_{XY}(X, Y)} \right].$$

By taking $q_{XY}(x, y) = (p_X \times p_Y)(x, y) := p_X(x)p_Y(y)$, we get the mutual information

$$I(X \wedge Y) := D(p_{XY}\|p_X \times p_Y).$$

Again, mutual information is nonnegative, and equal to zero if and only if X and Y are independent.

^b By Example 11.3, we know that $H_d(p) < \infty$. This can be used to justify breaking the integral into a sum of integrals.

^c This inequality captures a special case of Example 11.3; i.e., if X has finite second moment, then $H_d(X) < \infty$.

11.3. Divergence for Mixed Random Variables

Sometimes we have a pair of random variables X and Y where X is discrete and Y is continuous. If $p(x)$ is the marginal pmf of X and $w(y|x)$ is the conditional density of Y given that $X = x$, we call

$$p_{XY}(x, y) := p(x)w(y|x)$$

the joint “density” of X and Y . To justify this terminology, first recall that by definition, w allows us to compute

$$P(Y \in B|X = x) = \int_B w(y|x) dy.$$

Then

$$\begin{aligned} P(X \in A, Y \in B) &= \sum_{x \in A} P(X = x, Y \in B) \\ &= \sum_{x \in A} P(Y \in B|X = x)P(X = x) \\ &= \sum_{x \in A} \int_B w(y|x) dy p(x) \\ &= \sum_{x \in A} \int_B p(x)w(y|x) dy. \end{aligned}$$

Taking $A = X$ shows that

$$P(Y \in B) = \sum_x \int_B p(x)w(y|x) dy = \int_B \left[\sum_x p(x)w(y|x) \right] dy,$$

which implies that the marginal density for Y is

$$p_Y(y) = (pw)(y) := \sum_x p(x)w(y|x).$$

If we also have a mixed “density” $q_{XY}(x, y)$, we put

$$D(p_{XY} \| q_{XY}) := \sum_x \int p_{XY}(x, y) \log \frac{p_{XY}(x, y)}{q_{XY}(x, y)} dy.$$

Taking $q_{XY}(x, y) = (p_X \times p_Y)(x, y) := p_X(x)p_Y(y)$ yields the mutual information

$$I(X \wedge Y) := D(p_{XY} \| p_X \times p_Y).$$

11.4. The Discrete-Time Gaussian Channel

The definitions of achievable rates for a continuous channel under a codeword power constraint are as in Chapter 10, but with $X = Y = \mathbb{R}$, $a(x) = x^2$, and

$$W(B|x) = \int_B w(y|x) dy,$$

where $w(y|x)$ is a transition density; i.e., for each x , $w(\cdot|x)$ is a probability density. The **Gaussian channel** is given by $w(y|x) = g(y-x)$, where g is the $N(0, \sigma^2)$ density

$$g(t) := \frac{e^{-(t/\sigma)^2/2}}{\sqrt{2\pi}\sigma}.$$

This structure corresponds to having a channel whose output is $Y = X + Z$, where X is the channel input symbol and Z is additive noise having the Gaussian density g .

Theorem 11.8 (Coding Theorem for the Gaussian Channel). *A rate R is achievable for the discrete-time, memoryless Gaussian channel with variance σ^2 under the average or maximum probability of error criterion and codeword power constraint A if and only if*

$$0 \leq R \leq \frac{1}{2} \log\left(1 + \frac{A}{\sigma^2}\right).$$

Proof. For the Gaussian channel $w(y|x) = g(y-x)$, first note that

$$H_d(Y|X = x) = H_d(w(\cdot|x)) = H_d(g(\cdot - x)) = H_d(g) = \frac{1}{2} \log(2\pi\sigma^2 e).$$

Hence, if X and Y have conditional density $p_{Y|X}(y|x) = g(y-x)$,

$$H_d(Y|X) = \frac{1}{2} \log(2\pi\sigma^2 e).$$

This holds whether X is a discrete or continuous random variable. Furthermore, whether X is discrete or continuous, Y is continuous. This was shown in the previous section for the case of discrete X . A similar argument works for continuous X . We can now write

$$\begin{aligned} I(X \wedge Y) &= H_d(Y) - H_d(Y|X) = H_d(Y) - \frac{1}{2} \log(2\pi\sigma^2 e) \\ &\leq \frac{1}{2} \log(2\pi \text{var}(Y)e) - \frac{1}{2} \log(2\pi\sigma^2 e), \end{aligned}$$

with equality if Y is Gaussian. Note that to make Y Gaussian, it suffices to take X Gaussian.^d To maximize the mutual information, we need to maximize the variance of Y . We now assume X has finite second moment $\mathbb{E}[X^2] \leq A$ and mean m . By the law of total probability,

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[X] = m,$$

and

$$\mathbb{E}[Y^2] = \mathbb{E}[\mathbb{E}[Y^2|X]] = \mathbb{E}[\sigma^2 + X^2] = \sigma^2 + \mathbb{E}[X^2].$$

Hence, $\text{var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \sigma^2 + \mathbb{E}[X^2] - m^2$ is maximized if X has zero mean. In this case,

$$I(X \wedge Y) \leq \frac{1}{2} \log \left(1 + \frac{\mathbb{E}[X^2]}{\sigma^2} \right) \leq \frac{1}{2} \log \left(1 + \frac{A}{\sigma^2} \right),$$

where the second inequality uses the monotonicity of the logarithm. This bound is achieved if $X \sim N(0, A)$.

The adaptation of the weak converse proof to the Gaussian channel shows that an achievable rate under the average probability of error criterion and codeword power constraint A must satisfy $R \leq \sup I(X \wedge Y)$, where $p_{Y|X}(y|x) = g(y-x)$ and $\mathbb{E}[X^2] \leq A$. This supremum is given by the above right-hand side.

The adaptation of the forward proof to the Gaussian channel can be carried out for $p \sim N(0, A)$,^e thus showing that any rate strictly less than the above right-hand side is achievable. $\square$

11.4.1. Parallel Gaussian Channels

We generalize the foregoing by taking $X = Y = \mathbb{R}^r$ and $a(x) = \|x\|^2$, the square of the Euclidean norm of $x \in \mathbb{R}^r$. Thus, if $x = [x_1, \dots, x_r]^T$,

$$a(x) = \|x\|^2 = \sum_{\ell=1}^r x_\ell^2.$$

The channel output is $Y = X + Z$, where $X \in \mathbb{R}^r$ is the channel input symbol and Z is r -dimensional Gaussian noise whose components are independent, but not identically distributed, with the ℓ th component of Z being $N(0, \sigma_\ell^2)$.

^d In this case, $(pw)(y) = \int p(x)w(y|x)dx = \int p(x)g(y-x)dx$ is the convolution of two Gaussian densities, which corresponds to the density of the sum of two independent Gaussian random variables, which is Gaussian.

^e Using the joint density $p_{XY}(x, y) = p(x)g(y-x)$ with p being Gaussian implies X and Y are jointly Gaussian. Hence, the differential entropies $H_d(p_X)$, $H_d(p_Y)$, and $H_d(p_{XY})$ are all finite in the continuous analog of the joint AEP.

We consider the problem

$$\max_{\mathbb{E}[\|X\|^2] \leq A} I(X \wedge Y).$$

Put $A_\ell := \mathbb{E}[X_\ell^2]$. Then our constraint is that $\sum_\ell A_\ell \leq A$. As before, we can write

$$\begin{aligned} I(X \wedge Y) &= H_d(Y) - H_d(Y|X) = H_d(Y) - \sum_{\ell=1}^r \frac{1}{2} \log(2\pi\sigma_\ell^2 e) \\ &\leq \sum_{\ell=1}^r H_d(Y_\ell) - \frac{1}{2} \log(2\pi\sigma_\ell^2 e) \\ &\leq \sum_{\ell=1}^r \frac{1}{2} \log(2\pi(\sigma_\ell^2 + A_\ell)e) - \frac{1}{2} \log(2\pi\sigma_\ell^2 e) \\ &= \sum_{\ell=1}^r \frac{1}{2} \log\left(1 + \frac{A_\ell}{\sigma_\ell^2}\right). \end{aligned}$$

It remains to maximize this sum as a function of $A_1, \dots, A_r$ over $A_\ell \geq 0$ and subject to the constraint $\sum_\ell A_\ell \leq A$. This is a standard problem using Lagrange multipliers. Although there is no closed-form solution, it is not hard to show that an appropriate Lagrange multiplier exists and to compute it and the optimal solution numerically [8, Ch. 5]. In the case $r = 4$ and $\sigma_1^2 < \dots < \sigma_4^2$, if the Lagrange multiplier is $\lambda > 0$, then the relationship between λ , the noise levels σ_ℓ^2 and the optimal values of A_ℓ is illustrated in Figure 11.1. This picture suggests the colorful terminology of “**water-filling**.”

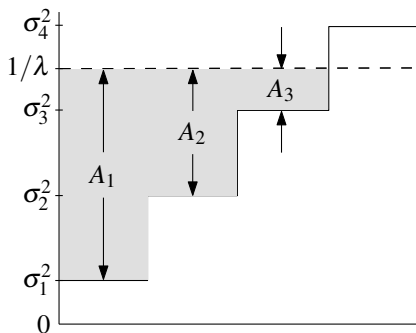


Figure 11.1. Water-filling solution $[A_1, \dots, A_4]^\top$ for achieving capacity of parallel Gaussian channels with noise levels $\sigma_1^2 < \dots < \sigma_4^2$. The Lagrange multiplier is $\lambda > 0$. Here $A_1 + A_2 + A_3 = A$ and $A_4 = 0$.

11.5. Continuous-Time Channels

11.5.1. Projections

Let L^2 denote the set of finite-energy waveforms on a given time interval such as $(-\infty, \infty)$, $[0, \infty)$, or $[0, T]$. By finite energy, we mean

$$\|x\|^2 := \int |x(t)|^2 dt < \infty,$$

where the integral is over the time interval under consideration. In digital communication systems, we usually transmit signaling waveforms from a particular finite-dimensional **signal subspace** $\mathcal{S}$ of L^2 . If $\dim \mathcal{S} = n$, we may write every $s \in \mathcal{S}$ in terms of an **orthonormal basis** $\beta_1, \dots, \beta_n$, where by orthonormal we mean that $\langle \beta_i, \beta_j \rangle = \delta_{ij}$, where the **inner product** is

$$\langle u, v \rangle := \int u(t)v(t) dt.$$

Note that $\langle u, u \rangle = \|u\|^2$.

For $x \in L^2$, its **projection** onto $\mathcal{S}$ is

$$Px := \sum_{k=1}^n \langle x, \beta_k \rangle \beta_k \in \mathcal{S}.$$

It is easy to check that if $s \in \mathcal{S}$, say $s = \sum_{j=1}^n c_j \beta_j$, then $Ps = s$. It follows that for $x \in L^2$, $P(Px) = Px$; i.e., P is **idempotent**. We also have the **orthogonality property**,

$$\langle x - Px, s \rangle = 0, \quad s \in \mathcal{S}.$$

This can be seen by first writing

$$\begin{aligned} \langle x - Px, \beta_j \rangle &= \langle x, \beta_j \rangle - \left\langle \sum_{k=1}^n \langle x, \beta_k \rangle \beta_k, \beta_j \right\rangle \\ &= \langle x, \beta_j \rangle - \sum_{k=1}^n \langle x, \beta_k \rangle \langle \beta_k, \beta_j \rangle \\ &= \langle x, \beta_j \rangle - \langle x, \beta_j \rangle = 0. \end{aligned}$$

If $s = \sum_{j=1}^n c_j \beta_j$, we have by linearity that $\langle x - Px, s \rangle = 0$.

The orthogonality property can be rewritten as $\langle x, s \rangle = \langle Px, s \rangle$. In other words, for $s \in \mathcal{S}$, the inner product of any x with s is the same as the inner product of the projection of x with s .

11.5.2. Approximately-Bandlimited Signaling Waveforms

Consider a digital communication system with signaling waveforms on $[0, T]$. Due to physical channel considerations or to regulatory considerations, suppose we want to use signaling waveforms that are bandlimited with **cutoff frequency** W . Since the only finite-energy waveform that is both time-limited and bandlimited is the zero waveform, we select signaling waveforms from a subspace $\mathcal{S}$ whose elements have “most” of their energy concentrated in the frequency band $[-W, W]$. This is accomplished by using

$$\beta_k(t) := \begin{cases} \psi_{k-1}(2t/T - 1)/\sqrt{\lambda_k}, & 0 \leq t \leq T, \\ 0, & \text{otherwise,} \end{cases}$$

where the $\{\psi_k\}_{k=0}^{\infty}$ are the angular **prolate spheroidal wave functions**, and the corresponding λ_k are strictly decreasing eigenvalues. If we use the signal subspace $\mathcal{S} := \{\beta_1, \dots, \beta_n\}$, where $n \approx 2WT$, then the waveforms in $\mathcal{S}$ will have nearly all of their energy concentrated in the desired frequency band. For details, see [22], [24], or [8, Ch. 8].

11.5.3. White Noise

We say that a **random process** $\Xi(t)$ is **white noise** if

$$\mathbb{E}[\Xi(t_1)\Xi(t_2)] = \frac{\mathcal{N}_0}{2} \delta(t_1 - t_2),$$

where the parameter $\mathcal{N}_0/2$ is the **channel noise power**, and δ is the **Dirac delta** or **unit impulse**.

Lemma 11.9. *If*

$$U = \langle u, \Xi \rangle = \int u(t)\Xi(t) dt \quad \text{and} \quad V = \langle v, \Xi \rangle = \int v(\tau)\Xi(\tau) d\tau,$$

then

$$\mathbb{E}[UV] = \mathbb{E}[\langle u, \Xi \rangle \langle v, \Xi \rangle] = \frac{\mathcal{N}_0}{2} \langle u, v \rangle.$$

Proof. Write

$$\begin{aligned} \mathbb{E}[UV] &= \mathbb{E}\left[\left(\int u(t)\Xi(t) dt\right)\left(\int v(\tau)\Xi(\tau) d\tau\right)\right] \\ &= \int \left(\int u(t)\mathbb{E}[\Xi(t)\Xi(\tau)] dt\right) v(\tau) d\tau \\ &= \frac{\mathcal{N}_0}{2} \int \left(\int u(t)\delta(t - \tau) dt\right) v(\tau) d\tau \\ &= \frac{\mathcal{N}_0}{2} \int u(\tau)v(\tau) d\tau = \frac{\mathcal{N}_0}{2} \langle u, v \rangle. \end{aligned}$$

□

The projection of white noise Ξ onto a signal subspace $\mathcal{S}$ is

$$\widehat{\Xi} := P\Xi = \sum_{k=1}^n \langle \Xi, \beta_k \rangle \beta_k.$$

Proposition 11.10. *If Ξ is a zero-mean, white-noise process, then the random variables $\langle u, \Xi - \widehat{\Xi} \rangle$ and $\langle v, \widehat{\Xi} \rangle$ are uncorrelated.*

Proof. We must show that

$$\mathbb{E}[\langle u, \Xi - \widehat{\Xi} \rangle \langle v, \widehat{\Xi} \rangle] = \mathbb{E}[\langle u, \Xi - \widehat{\Xi} \rangle] \mathbb{E}[\langle v, \widehat{\Xi} \rangle].$$

Since the process has zero mean, the random variables $\langle u, \Xi \rangle$ and $\langle v, \widehat{\Xi} \rangle$ both have zero mean, and so the above right-hand side is zero. To show the left-hand side is zero, write it as

$$\mathbb{E}[\langle u, \Xi - \widehat{\Xi} \rangle \langle v, \widehat{\Xi} \rangle] = \mathbb{E}[\langle u, \Xi \rangle \langle v, \widehat{\Xi} \rangle] - \mathbb{E}[\langle u, \widehat{\Xi} \rangle \langle v, \widehat{\Xi} \rangle].$$

We show that these last two expectations are equal. From the above formula for $\widehat{\Xi}$,

$$\langle v, \widehat{\Xi} \rangle = \sum_{k=1}^n \langle \Xi, \beta_k \rangle \langle v, \beta_k \rangle.$$

Hence,

$$\mathbb{E}[\langle u, \Xi \rangle \langle v, \widehat{\Xi} \rangle] = \sum_{k=1}^n \langle v, \beta_k \rangle \mathbb{E}[\langle u, \Xi \rangle \langle \Xi, \beta_k \rangle] = \frac{\mathcal{N}_0}{2} \sum_{k=1}^n \langle u, \beta_k \rangle \langle v, \beta_k \rangle$$

by the above lemma. We also have

$$\begin{aligned} \mathbb{E}[\langle u, \widehat{\Xi} \rangle \langle v, \widehat{\Xi} \rangle] &= \mathbb{E}\left[\left(\sum_{i=1}^n \langle \Xi, \beta_i \rangle \langle u, \beta_i \rangle\right)\left(\sum_{k=1}^n \langle \Xi, \beta_k \rangle \langle v, \beta_k \rangle\right)\right] \\ &= \sum_{k=1}^n \sum_{i=1}^n \langle u, \beta_i \rangle \langle v, \beta_k \rangle \mathbb{E}[\langle \Xi, \beta_i \rangle \langle \Xi, \beta_k \rangle] \\ &= \sum_{k=1}^n \sum_{i=1}^n \langle u, \beta_i \rangle \langle v, \beta_k \rangle \frac{\mathcal{N}_0}{2} \langle \beta_i, \beta_k \rangle \\ &= \frac{\mathcal{N}_0}{2} \sum_{k=1}^n \langle u, \beta_k \rangle \langle v, \beta_k \rangle. \end{aligned}$$

□

11.5.4. Gaussian Processes

A random process $\Xi(t)$ is said to be **Gaussian** if for all u , $\langle u, \Xi \rangle$ is a scalar Gaussian random variable. Since a random variable with zero variance is constant almost surely, we consider such random variables to be Gaussian with mean equal to their constant value and variance zero. For waveforms $u_1, \dots, u_m$ and scalars $c_1, \dots, c_m$,

$$c_1 \langle u_1, \Xi \rangle + \dots + c_m \langle u_m, \Xi \rangle = \left\langle \sum_{i=1}^m c_i u_i, \Xi \right\rangle.$$

Since the right-hand side is Gaussian, we have that $\langle u_1, \Xi \rangle, \dots, \langle u_m, \Xi \rangle$ are jointly Gaussian random variables. If the u_i are orthogonal, and if Ξ is a zero-mean, white-noise process, then $E[\langle u_i, \Xi \rangle \langle u_j, \Xi \rangle] = \frac{N_0}{2} \langle u_i, u_j \rangle$, and thus $\langle u_1, \Xi \rangle, \dots, \langle u_m, \Xi \rangle$ are uncorrelated. Since they are jointly Gaussian, they are independent [7, pp. 365–366]. By combining this observation and the preceding proposition, we have the following important result.

Theorem 11.11. *If $\Xi(t)$ is a Gaussian white-noise process with zero mean, then $\{\Xi(t) - \widehat{\Xi}(t)\}$ and $\{\widehat{\Xi}(t)\}$ are independent Gaussian random processes.*

Proof. It suffices to show that for every u and v , the random variables $\langle u, \Xi - \widehat{\Xi} \rangle$ and $\langle v, \widehat{\Xi} \rangle$ are jointly Gaussian and uncorrelated. The details are left to Problem 11.13. $\square$

11.5.5. Detection of Signals in Additive White Gaussian Noise

We say that a random process $\Xi(t)$ is **white Gaussian noise** if it is zero mean, white, and Gaussian. We say that a signaling waveform s is transmitted over an **additive white Gaussian noise (AWGN) channel** if the received waveform is

$$Y(t) = s(t) + \Xi(t).$$

If s belongs to a finite-dimensional signal subspace $\mathcal{S}$, then we may project the received waveform onto $\mathcal{S}$. The projection is

$$\widehat{Y} := PY = P(s + \Xi) = Ps + P\Xi = s + \widehat{\Xi},$$

since $s \in \mathcal{S}$ implies $Ps = s$. Notice that

$$Y - \widehat{Y} = (s + \Xi) - (s + \widehat{\Xi}) = \Xi - \widehat{\Xi}.$$

Observe that knowledge of the pair $\hat{Y}$ and $Y - \hat{Y}$ is equivalent to knowledge of Y since the pair can be computed from Y , and Y can be computed from the pair. Furthermore, $Y - \hat{Y}$ does not depend on the signal, and $Y - \hat{Y} = \Xi - \hat{\Xi}$ is independent of $\hat{\Xi}$, which is the additive noise in $\hat{Y} = s + \hat{\Xi}$. Hence, $Y - \hat{Y}$ cannot help us to detect the transmitted signal s .

The foregoing discussion implies that we can base our detection of s on $\hat{Y}$. Fortunately, the waveform $\hat{Y}$ is equivalent to the vector of coefficients^f

$$\mathbf{Y} := [\langle Y, \beta_1 \rangle, \dots, \langle Y, \beta_n \rangle]^T.$$

Hence,

$$\mathbf{Y} = \mathbf{s} + \mathbf{\Xi},$$

where $\mathbf{s} := [\langle s, \beta_1 \rangle, \dots, \langle s, \beta_n \rangle]^T$ is the vector of coefficients of the expansion of $s \in \mathcal{S}$, and $\mathbf{\Xi} := [\langle \Xi, \beta_1 \rangle, \dots, \langle \Xi, \beta_n \rangle]^T$ is the vector of coefficients used to compute $\hat{\Xi}$.

We have shown that problem of detecting a continuous-time waveform from a finite-dimensional waveform signal subspace in additive white Gaussian noise is equivalent to detecting a finite-dimensional column vector in the presence of a column vector of noise random variables that are i.i.d. Gaussian.

11.5.6. The Bandlimited Gaussian Channel

Consider the signal subspace $\mathcal{S}$ of approximately-bandlimited waveforms on $[0, T]$ that were introduced in Section 11.5.2. Recall that if the cutoff frequency is W , then the dimension of $\mathcal{S}$ is $n = 2WT$. For signals $s \in \mathcal{S}$, suppose we have a power constraint of the form

$$\frac{1}{T} \int_0^T |s(t)|^2 dt \leq \mathcal{P}.$$

We can write this more compactly as $\|s\|^2 \leq T\mathcal{P}$. Since $s \in \mathcal{S}$, it is easy to check that

$$\|s\|^2 = \sum_{j=1}^n |\langle s, \beta_j \rangle|^2 = \|\mathbf{s}\|^2.$$

Hence, the power constraint can be rewritten as $\|\mathbf{s}\|^2 \leq T\mathcal{P}$, or, since $n = 2WT$,

$$\frac{1}{n} \|\mathbf{s}\|^2 \leq \frac{\mathcal{P}}{2W}.$$

^fIt is clear that $\hat{Y}$ can be obtained from $\mathbf{Y}$. To obtain $\mathbf{Y}$ from $\hat{Y}$, recall from the end of Section 11.5.1 that $\langle \hat{Y}, \beta_k \rangle = \langle Y, \beta_k \rangle$.

Assuming additive white Gaussian noise Ξ , we observe that the components of the noise vector Ξ satisfy $E[\langle \Xi, \beta_i \rangle \langle \Xi, \beta_j \rangle] = \frac{\mathcal{N}_0}{2} \delta_{ij}$; i.e., they are uncorrelated. Since they are jointly Gaussian and have zero mean, they are i.i.d. $N(0, \sigma^2)$, where $\sigma^2 = \mathcal{N}_0/2$.

The formula $\mathbf{Y} = \mathbf{s} + \Xi$ is thus seen to describe a discrete-time Gaussian channel, whose capacity is

$$\frac{1}{2} \log_2 \left(1 + \frac{\mathcal{P}/(2W)}{\mathcal{N}_0/2} \right) = \frac{1}{2} \log_2 \left(1 + \frac{\mathcal{P}}{W\mathcal{N}_0} \right)$$

measured in bits per channel use. To express this in units of bits per second, observe that n channel uses take T seconds. So we multiply the above formula by n/T and apply the formula $n = 2WT$ to get

$$W \log_2 \left(1 + \frac{\mathcal{P}}{W\mathcal{N}_0} \right) \quad \text{bits per second.}$$

By l'Hôpital's rule [19, p. 109], the **wideband limit** is

$$\lim_{W \rightarrow \infty} W \log_2 \left(1 + \frac{\mathcal{P}}{W\mathcal{N}_0} \right) = \frac{\mathcal{P}}{\mathcal{N}_0} \log_2 e.$$

Thus, although the capacity is strictly increasing as the bandwidth W increases, the capacity does *not* become infinite. The limit also says that for large bandwidth, the capacity increases linearly with the power constraint $\mathcal{P}$.

Notes

Note 11.1. We show that if p is a density with $\int |x|^r p(x) dx < \infty$ for some $r > 0$, then

$$\int_{\{x: p(x) \leq 1\}} p(x) \ln \frac{1}{p(x)} dx < \infty.$$

To do this, partition the range of integration as

$$\begin{aligned} \{x : p(x) \leq 1\} &= \{x : p(x) \leq 1\} \cap [-1, 1] \\ &\quad \cup \{x : p(x) \leq 1\} \cap \{x : |x| > 1 \text{ and } p(x) \leq e^{-|x|^r}\} \\ &\quad \cup \{x : p(x) \leq 1\} \cap \{x : |x| > 1 \text{ and } p(x) > e^{-|x|^r}\}. \end{aligned}$$

Next, we put $f(t) := -t \ln t$ for $t > 0$ and $f(0) := 0$. Then f is strictly increasing on $[0, 1/e]$ and strictly decreasing on $[1/e, 1]$. Also, for $t \in [0, 1]$, $f(t)$ is maximized at

$t = 1/e$ with maximum value $1/e$. With these facts in mind, we can write the integral in question as $I_1 + I_2 + I_3$, where

$$I_1 := \int_{\{x:p(x)\leq 1\}\cap[-1,1]} f(p(x)) dx \leq \int_{\{x:p(x)\leq 1\}\cap[-1,1]} e^{-1} dx \leq \int_{-1}^1 e^{-1} dx = 2/e,$$

$$\begin{aligned} I_2 &:= \int_{\{x:p(x)\leq 1\}\cap\{x:|x|>1 \text{ and } p(x)\leq e^{-|x|^r}\}} f(p(x)) dx \\ &\leq \int_{\{x:p(x)\leq 1\}\cap\{x:|x|>1 \text{ and } p(x)\leq e^{-|x|^r}\}} f(e^{-|x|^r}) dx \\ &= \int_{\{x:p(x)\leq 1\}\cap\{x:|x|>1 \text{ and } p(x)\leq e^{-|x|^r}\}} |x|^r e^{-|x|^r} dx \\ &\leq \int_{-\infty}^{\infty} |x|^r e^{-|x|^r} dx < \infty, \end{aligned}$$

and

$$\begin{aligned} I_3 &:= \int_{\{x:p(x)\leq 1\}\cap\{x:|x|>1 \text{ and } p(x)>e^{-|x|^r}\}} f(p(x)) dx \\ &= \int_{\{x:p(x)\leq 1\}\cap\{x:|x|>1 \text{ and } -\ln p(x)<|x|^r\}} p(x) [-\ln p(x)] dx \\ &= \int_{\{x:p(x)\leq 1\}\cap\{x:|x|>1 \text{ and } -\ln p(x)<|x|^r\}} |x|^r p(x) dx \\ &\leq \int_{-\infty}^{\infty} |x|^r p(x) dx < \infty. \end{aligned}$$

Problems

1. Compute the differential entropy of the uniform density on $[0, T]$. For what values of T is the differential entropy positive? For what values of T is the differential entropy negative?
2. Find the differential entropy of an exponential random variable X with density $p(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. *Hint:* $E[X] = 1/\lambda$.
3. Compute the differential entropy of $p(x) = 1/(2\sqrt{x})$ for $0 < x \leq 1$. *Hint:* $(d/dx)[2\sqrt{x}(\ln x - 2)] = (\ln x)/\sqrt{x}$.
4. Show that the Cauchy density $p(x) = (1/\pi)/(1+x^2)$ has finite differential entropy.
5. Construct a density with infinite differential entropy. *Hint:* Review Problem 3.12.

6. Verify that the following density [2] has differential entropy $-\infty$. For $k \geq 1$, put

$$a_k := k, \quad b_k := k + 2^{-k}e^{-2^k}, \quad \text{and} \quad c_k := e^{2^k}.$$

Define the density $p(x) = c_k$ for $x \in [a_k, b_k]$, and $p(x) = 0$ otherwise.

(a) Show that $\int_0^\infty p(x) dx = 1$.

(b) Show that

$$\int_{a_k}^{b_k} p(x) \ln \frac{1}{p(x)} dx = -1,$$

and thus $H_d(p) = -\infty$.

7. Construct a density for which the differential entropy is undefined.

8. Let q denote the exponential density, $q(x) := \lambda e^{-\lambda x}$ for $x \geq 0$. Let p be any density on $[0, \infty)$ with mean m . Show that $H_d(p) \leq \lambda m \log e - \log \lambda$. Show that equality holds if $p = q$, in which case $H_d(p) = \log(e/\lambda)$.

9. Let X be a continuous random variable with finite second moment. Rearrange (11.1) to show that if the natural logarithm is used to compute the differential entropy, then

$$\text{var}(X) \geq \frac{e^{2H_d(X)}}{2\pi e}.$$

Remark. Since the mean is the minimum mean-squared-error constant estimator of X , it follows that for all constants c ,

$$\mathbb{E}[|X - c|^2] \geq \mathbb{E}[|X - \mathbb{E}[X]|^2] = \text{var}(X) \geq \frac{e^{2H_d(X)}}{2\pi e}.$$

10. Let X and Y be random variables such that X has a conditional density given Y . Assume that this conditional density has finite second moments for all y . Show that if the natural logarithm is used to compute the conditional differential entropy, then

$$\mathbb{E}[|X - \mathbb{E}[X|Y]|^2] \geq \frac{e^{2H_d(X|Y)}}{2\pi e}.$$

Hint: Apply the result of the preceding problem to the conditional density. Then average over Y and apply Jensen's inequality.

Remark. Since the conditional expectation is the minimum mean-squared-error estimator of X based on Y , we have that for all functions φ satisfying $\mathbb{E}[\varphi(X)^2] < \infty$,

$$\mathbb{E}[|X - \varphi(X)|^2] \geq \mathbb{E}[|X - \mathbb{E}[X|Y]|^2] \geq \frac{e^{2H_d(X|Y)}}{2\pi e}.$$

11. Show that the divergence between two joint “densities” of mixed random variables is nonnegative.
12. Let X , Y , and Z be random variables with X being discrete with pmf $p(x)$, Y being continuous with conditional density $w(y|x)$, and Z being discrete with conditional pmf $v(z|y, x)$. In other words, X , Y , and Z have joint “density”

$$p_{XYZ}(x, y, z) := p(x)w(y|x)v(z|y, x).$$

Then the marginal joint “density” of the continuous variable Y and the discrete variable Z is

$$p_{YZ}(y, z) = \sum_x p(x)w(y|x)v(z|y, x).$$

Then

$$I(X \wedge Y) = D(p_{XY} \| p_X \times p_Y) \quad \text{and} \quad I(X \wedge YZ) = D(p_{XYZ} \| p_X \times p_{YZ}),$$

where $(p_X \times p_{YZ})(x, y, z) := p_X(x)p_{YZ}(y, z)$.

(a) Propose a definition of $I(X \wedge Z|Y)$ so that

$$I(X \wedge YZ) = I(X \wedge Y) + I(X \wedge Z|Y).$$

(b) Show that if $v(z|y, x)$ does not depend on x , then $I(X \wedge Z|Y) = 0$.

13. Complete the proof of Theorem 11.11.

CHAPTER 12

Rate-Distortion Theory

12.1. Introduction

The rate-distortion problem begins with a **source alphabet** X and a **reproduction alphabet** Y . For $n = 1, 2, \dots$, we consider mappings $q: \mathsf{X}^n \rightarrow \mathsf{Y}^n$ with the property that q takes $N < |\mathsf{X}^n|$ distinct values in Y^n . The idea is that for $\mathbf{x} \in \mathsf{X}^n$, $q(\mathbf{x}) \in \mathsf{Y}^n$ is an “approximation” or “quantization” of $\mathbf{x}$. To measure the error or distortion, we use a function $d_n: \mathsf{X}^n \times \mathsf{Y}^n \rightarrow [0, \infty)$. If $\mathbf{X} = (X_1, \dots, X_n)$ is an n -tuple of X -valued random variables, we can consider the mean distortion,

$$\mathbb{E}[d_n(\mathbf{X}, q(\mathbf{X}))].$$

Example 12.1 (Probability of Error). Suppose $\mathsf{X} = \mathsf{Y}$ and $d_n(\mathbf{x}, \mathbf{y}) = 1$ if $\mathbf{x} \neq \mathbf{y}$ and $d_n(\mathbf{x}, \mathbf{y}) = 0$ if $\mathbf{x} = \mathbf{y}$. In other words, d_n is the **trivial metric** on X^n . Then

$$\mathbb{E}[d_n(\mathbf{X}, q(\mathbf{X}))] = \mathbb{P}(q(\mathbf{X}) \neq \mathbf{X})$$

is just the probability of error for a block source code.

Example 12.2 (Hamming Distance). The **Hamming distance** between $\mathbf{x}$ and $\mathbf{y}$ is the number of entries in which they differ. Since $\mathbf{1}_{\{x\}^c}(y) = 1$ if $y \neq x$ and is zero otherwise, the formula

$$\sum_{k=1}^n \mathbf{1}_{\{x_k\}^c}(y_k)$$

counts the number of positions in which $\mathbf{x}$ and $\mathbf{y}$ differ. The **normalized Hamming distance** distortion measure is

$$d_n(\mathbf{x}, \mathbf{y}) := \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{\{x_k\}^c}(y_k).$$

Example 12.3 (Squared Error). If $\mathsf{X} = \mathsf{Y} = \mathbb{R}$, then the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ is

$$\|\mathbf{x} - \mathbf{y}\|^2 = \sum_{k=1}^n |x_k - y_k|^2.$$

We call $\|\mathbf{x} - \mathbf{y}\|^2$ the **squared error**. The *normalized squared error* distortion measure is

$$d_n(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \|\mathbf{x} - \mathbf{y}\|^2 = \frac{1}{n} \sum_{k=1}^n |x_k - y_k|^2.$$

Thus,

$$\mathbb{E}[d_n(\mathbf{X}, q(\mathbf{X}))] = \frac{1}{n} \mathbb{E}[\|\mathbf{X} - q(\mathbf{X})\|^2]$$

differs from the mean-squared error formula in Chapter 2 only by the factor $1/n$.

The last two examples illustrate the concept of a **single-letter distortion measure**, that is, a distortion measure of the form

$$d_n(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{k=1}^n d(x_k, y_k),$$

where $d: \mathcal{X} \times \mathcal{Y} \rightarrow [0, \infty)$. The normalized Hamming distance uses $d(x, y) = \mathbf{1}_{\{x\}^c}(y)$, and the normalized squared-error uses $d(x, y) = |x - y|^2$.

12.2. Achievable Rate-Distortion Pairs

Definition 12.4. We say that a pair of nonnegative real numbers (R, D) is an **achievable rate-distortion pair** for a source $X_1, X_2, \dots$, if $\forall \Delta D > 0, \forall \Delta R > 0, \exists n_0$ such that $\forall n \geq n_0, \exists q: \mathcal{X}^n \rightarrow \mathcal{Y}^n$ taking at most N distinct values with

$$\frac{\log N}{n} < R + \Delta R \quad \text{and} \quad \mathbb{E}[d_n(\mathbf{X}, q(\mathbf{X}))] < D + \Delta D,$$

where $\mathbf{X} := (X_1, \dots, X_n)$. The **rate-distortion region** is the set of achievable rate-distortion pairs.

For fixed $R \geq 0$, if (R, D) is achievable, then it is easy to see that for any $D' > D$, the pair (R, D') is also achievable. Similarly, for fixed $D \geq 0$, if (R, D) is achievable, then for any $R' > R$, the pair (R', D) is also achievable. Hence, to determine the rate-distortion region, it suffices to characterize either the **rate-distortion function**

$$R(D) := \inf\{R \geq 0 : (R, D) \text{ is achievable}\},$$

or the **distortion-rate function**

$$D(R) := \inf\{D \geq 0 : (R, D) \text{ is achievable}\}.$$

Typical rate-distortion functions are shown in Figure 12.1. Of course, $R(D)$ and $D(R)$ depend on the distribution of the source $X_1, X_2, \dots$, although we suppress this dependence in the notation.

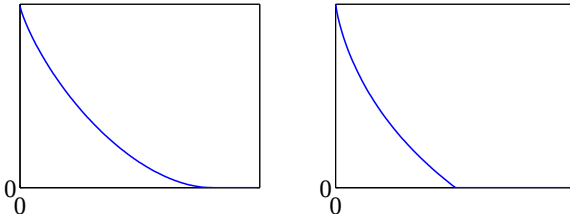


Figure 12.1. Typical rate-distortion functions $R(D)$.

12.2.1. Properties of the Rate-Distortion Function for a DMS

Put

$$\underline{d}(x) := \inf_{y \in \mathcal{Y}} d(x, y) \quad \text{and} \quad \bar{d}(x) := \sup_{y \in \mathcal{Y}} d(x, y).$$

Since d is nonnegative, $\underline{d}(x) \geq 0$. If $\mathcal{Y}$ is a finite set, then $\bar{d}(x) < \infty$. Otherwise, we may have $\bar{d}(x) = \infty$ for some x . If $X_1, X_2, \dots$ is a discrete memoryless $\mathcal{X}$ -valued source with common pmf p , then

$$\mathbb{E}[d_n(\mathbf{X}, q(\mathbf{X}))] \leq \mathbb{E}\left[\frac{1}{n} \sum_{k=1}^n \bar{d}(X_k)\right] = \sum_x p(x) \bar{d}(x) =: D_{\max}.$$

Similarly

$$\mathbb{E}[d_n(\mathbf{X}, q(\mathbf{X}))] \geq \sum_x p(x) \underline{d}(x) =: D_{\min}.$$

We always have $D_{\min} \geq 0$. If $\bar{d}(x)$ is finite for all x , and if $\mathcal{X}$ is a finite set, then $D_{\max} < \infty$. Otherwise, we may have $D_{\max} = \infty$.

If $D_{\max}$ is finite and $D_{\max} \leq D < \infty$, then for all $R \geq 0$, the pair (R, D) is achievable, and so $R(D) = \inf\{R \geq 0\} = 0$. Conversely, if $D < D_{\min}$, then there is no $R \geq 0$ for which (R, D) is achievable,^a and so $R(D) = \inf \emptyset := +\infty$. Thus, the only interesting values of D are those with $D_{\min} \leq D < D_{\max}$, even if $D_{\max} = \infty$.

We conclude this subsection by noting that there is no loss of generality in assuming $\underline{d}(x) = 0$. To see this, consider the modified distortion function $\hat{d}(x, y) := d(x, y) - \underline{d}(x)$. Since

$$\mathbb{E}[\hat{d}_n(\mathbf{X}, q(\mathbf{X}))] = \mathbb{E}[d_n(\mathbf{X}, q(\mathbf{X}))] - D_{\min},$$

(R, D) is achievable under the modified distortion function if and only if $(R, D + D_{\min})$ is achievable under the original distortion function.

^aTo see this, take $\Delta D = D_{\min} - D > 0$ in Definition 12.4.

12.3. Properties of $\tilde{R}(D)$

If X and Y are finite sets and $D \geq D_{\min}$, let

$$\tilde{R}(D) := \inf_{W: \mathbb{E}[d(X,Y)] \leq D} I(X \wedge Y),$$

where in computing the mutual information and the expectation, we use $P_{XY}(x, y) = p(x)W(y|x)$. In other words, to compute $\tilde{R}(D)$, we minimize $I(p \times W)$ as a function of the transition probability W , subject to the inequality constraint on W that

$$\sum_x \sum_y d(x, y) p(x) W(y|x) \leq D.$$

For a discrete memoryless source and a single-letter distortion measure, we will show that for fixed $D \geq 0$, the pair (R, D) is achievable if and only if $R \geq \tilde{R}(D)$. Thus, $R(D) = \tilde{R}(D)$ for a discrete memoryless source.

To indicate the plausibility of the foregoing assertion, we show that $\tilde{R}$ has some of the same properties as $R(D)$. If $P_{XY}(x, y) = p(x)W(y|x)$, then since

$$\mathbb{E}[d(X, Y)] = \sum_x p(x) \sum_y d(x, y) W(y|x),$$

it follows that

$$D_{\min} \leq \mathbb{E}[d(X, Y)] \leq D_{\max}.$$

For future use, we note that when Y is a finite set, $D_{\min}$ can be attained if W is chosen properly. For example, for each x , let y_x satisfy $d(x, y_x) = \underline{d}(x)$. Put

$$W(y|x) := \begin{cases} 1, & y = y_x, \\ 0, & y \neq y_x. \end{cases}$$

Then

$$\mathbb{E}[d(X, Y)] = \sum_x p(x) \sum_y d(x, y) W(y|x) = \sum_x p(x) d(x, y_x) = \sum_x p(x) \underline{d}(x) = D_{\min}.$$

Thus, for $D \geq D_{\min}$, when we compute $\tilde{R}(D)$, the set of W over which we take the infimum is nonempty.

Just as we argued that $R(D)$ is only interesting for $D_{\min} \leq D < D_{\max}$, we can do the same for $\tilde{R}$. If $D < D_{\min}$, then in the definition of $\tilde{R}(D)$, there are no W s over which to take the infimum. In this case, $\tilde{R}(D) = \inf \emptyset := +\infty$. If $D \geq D_{\max}$, then $\mathbb{E}[d(X, Y)] \leq D$ for all transition probabilities W , even ones for which $W(y|x)$ does not depend on x ; hence, for such W , X and Y are independent, which implies

$I(X \wedge Y) = 0$. Thus, $\tilde{R}(D) = 0$ for such D . So the only interesting values of D are those with $D_{\min} \leq D < D_{\max}$.

Note also that with the modified distortion function $\hat{d}$ defined in the previous section,

$$\mathbb{E}[\hat{d}(X, Y)] = \mathbb{E}[d(X, Y)] - D_{\min}.$$

Thus, for $\tilde{R}$, there is no loss of generality in assuming $\underline{d}(x) = 0$.

12.4. Preparation for the Converse

The converse theorem of the next section shows that if (R, D) is an achievable pair for a discrete memoryless source, then $R \geq \tilde{R}(D)$. Hence,

$$R(D) := \inf\{R \geq 0 : (R, D) \text{ is achievable}\} \geq \tilde{R}(D).$$

To carry out this proof, we need some properties of $\tilde{R}$ that we now establish.

Lemma 12.5. *$\tilde{R}(D)$ is nonincreasing and convex on $[D_{\min}, \infty)$.*

Proof. As we increase the value of D , the infimum defining $\tilde{R}(D)$ is over a larger set of W , and so $\tilde{R}(D)$ cannot increase. To prove convexity, we must show that for all $0 \leq \lambda \leq 1$, and any $D_{\min} \leq D_1 < D_2$,

$$\tilde{R}(\lambda D_1 + (1 - \lambda)D_2) \leq \lambda \tilde{R}(D_1) + (1 - \lambda)\tilde{R}(D_2).$$

Suppose $\tilde{R}(D_1) = I(p \times W_1)$ and $\tilde{R}(D_2) = I(p \times W_2)$. Put

$$P_{XY}^\lambda := p \times [\lambda W_1 + (1 - \lambda)W_2],$$

or in more detail,

$$P_{XY}^\lambda(x, y) = p(x)[\lambda W_1(y|x) + (1 - \lambda)W_2(y|x)].$$

Then under P_{XY}^λ ,

$$\begin{aligned} \mathbb{E}[d(X, Y)] &= \sum_x \sum_y P_{XY}^\lambda(x, y) d(x, y) \\ &= \lambda \mathbb{E}_{W_1}[d(X, Y)] + (1 - \lambda) \mathbb{E}_{W_2}[d(X, Y)] \\ &\leq \lambda D_1 + (1 - \lambda)D_2. \end{aligned}$$

Hence,

$$\begin{aligned} \tilde{R}(\lambda D_1 + (1 - \lambda)D_2) &\leq I(P_{XY}^\lambda) \\ &= I(p \times [\lambda W_1 + (1 - \lambda)W_2]) \\ &\leq \lambda I(p \times W_1) + (1 - \lambda)I(p \times W_2) \\ &= \lambda \tilde{R}(D_1) + (1 - \lambda)\tilde{R}(D_2). \end{aligned}$$

□

Corollary 12.6. $\tilde{R}$ is continuous on $(D_{\min}, \infty)$.

Proof. It is well known [28, p. 5] that convexity on an interval implies continuity at all interior points of the interval. $\square$

A convex function on an interval need not be continuous at its endpoints. To see this, consider the function $f(D_{\min}) = 1$ and $f(D) = 0$ for $D > D_{\min}$. Fortunately, the function $\tilde{R}$ is continuous at $D_{\min}$.

Lemma 12.7. $\tilde{R}$ is continuous at $D = D_{\min}$.

Proof. (Based on [4, p. 125].) It suffices to show that for every nonnegative sequence $D_n \rightarrow D_{\min}$, we have $\tilde{R}(D_n) \rightarrow \tilde{R}(D_{\min})$. To show that $\tilde{R}(D_n) \rightarrow \tilde{R}(D_{\min})$, we show that given any subsequence $\tilde{R}(D_{n_k})$, there is a further subsequence such that $\tilde{R}(D_{n_{k_\ell}}) \rightarrow \tilde{R}(D_{\min})$. So suppose $\tilde{R}(D_{n_k})$ is any subsequence. Note that this is a bounded sequence of real numbers since $0 \leq I(p \times W) \leq H(p) \leq \log |X|$. Hence, there is a converging sub-subsequence, $\tilde{R}(D_{n_{k_\ell}})$. Since the infimum defining $\tilde{R}(D)$ is always achieved,^b we have

$$\tilde{R}(D_{n_{k_\ell}}) = I(p \times W_\ell)$$

for some W_ℓ satisfying

$$\sum_x p(x) \sum_y W_\ell(y|x) d(x, y) \leq D_{n_{k_\ell}}. \quad (12.1)$$

Furthermore, since the W_ℓ lie in a sequentially compact set, there is a further subsequence $W_{\ell_i} \rightarrow W_\infty$ for some transition probability W_∞ as $i \rightarrow \infty$. By continuity of $I(p \times W)$ in W ,

$$\lim_{\ell \rightarrow \infty} \tilde{R}(D_{n_{k_\ell}}) = \lim_{i \rightarrow \infty} \tilde{R}(D_{n_{k_{\ell_i}}}) = \lim_{i \rightarrow \infty} I(p \times W_{\ell_i}) = I(p \times W_\infty).$$

Also, from (12.1) with ℓ replaced by ℓ_i , letting $i \rightarrow \infty$ yields

$$\sum_x p(x) \sum_y W_\infty(y|x) d(x, y) \leq D_{\min}.$$

Hence,

$$\tilde{R}(D_{\min}) := \inf_{W: E[d(X, Y)] \leq D_{\min}} I(p \times W) \leq I(p \times W_\infty) = \lim_{i \rightarrow \infty} I(p \times W_{\ell_i}) = \lim_{i \rightarrow \infty} \tilde{R}(D_{n_{k_{\ell_i}}}).$$

Since $\tilde{R}$ is nonincreasing, $\tilde{R}(D_{\min}) \geq \tilde{R}(D_{n_{k_{\ell_i}}})$. Letting $i \rightarrow \infty$ yields the desired result. $\square$

^b The infimum is achieved because $I(p \times W)$ is a continuous function of W , and the set of W over which the infimum is taken is sequentially compact.

12.5. The Converse Theorem

Theorem 12.8. *Assuming finite source and reproduction alphabets, if (R, D) is achievable for a discrete memoryless source, then $R \geq \tilde{R}(D)$.*

Proof. It suffices to restrict attention to $D_{\min} \leq D < D_{\max}$, where $\tilde{R}$ is finite, continuous, and convex there. Let $0 < \Delta D < D_{\max} - D$ and $\Delta R > 0$ be given. Suppose $\mathbf{Y} = q(\mathbf{X})$, where $q: \mathcal{X}^n \rightarrow \mathcal{Y}^n$ takes at most N distinct values,

$$n(R + \Delta R) > \log N \quad \text{and} \quad \mathbb{E}[d_n(\mathbf{X}, q(\mathbf{X}))] < D + \Delta D.$$

Since q takes at most N distinct values, $H(\mathbf{Y}) = H(q(\mathbf{X})) \leq \log N$. Hence,

$$\begin{aligned} n(R + \Delta R) &> H(\mathbf{Y}) \\ &\geq H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}), \quad \text{since } H(\mathbf{Y}|\mathbf{X}) \geq 0, \\ &= I(\mathbf{X} \wedge \mathbf{Y}) \\ &= H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}) \\ &= \sum_{k=1}^n H(X_k) - H(\mathbf{X}|\mathbf{Y}), \quad \text{since the } X_k \text{ are i.i.d.,} \\ &= \sum_{k=1}^n H(X_k) - H(X_k|\mathbf{Y}, X_1, \dots, X_{k-1}), \quad \text{chain rule for cond. entropy,} \\ &\geq \sum_{k=1}^n H(X_k) - H(X_k|Y_k), \quad \text{reducing conditioning increases entropy,} \\ &= \sum_{k=1}^n I(X_k \wedge Y_k) \\ &\geq \sum_{k=1}^n \tilde{R}\left(\mathbb{E}[d(X_k, Y_k)]\right), \quad \text{by definition of } \tilde{R}. \end{aligned}$$

Now write

$$\begin{aligned} R + \Delta R &\geq \frac{1}{n} \sum_{k=1}^n \tilde{R}\left(\mathbb{E}[d(X_k, Y_k)]\right) \\ &\geq \tilde{R}\left(\frac{1}{n} \sum_{k=1}^n \mathbb{E}[d(X_k, Y_k)]\right), \quad \text{since } \tilde{R} \text{ is convex,} \\ &= \tilde{R}\left(\mathbb{E}[d_n(\mathbf{X}, \mathbf{Y})]\right) \\ &\geq \tilde{R}(D + \Delta D), \quad \text{since } \tilde{R} \text{ is nonincreasing.} \end{aligned}$$

Since $\Delta R > 0$ is arbitrary, we must have $R \geq \tilde{R}(D + \Delta D)$. Since $\Delta D > 0$ is arbitrary, and since $\tilde{R}$ is continuous at D , we must have $R \geq \tilde{R}(D)$. $\square$

12.6. Preparation for the Forward Theorem

Lemma 12.9. For $0 \leq x, y \leq 1$ and $n \geq 1$,

$$(1 - xy)^n \leq 1 - x + e^{-yn}.$$

Proof. For fixed $y \in [0, 1]$, put $g(x) := (1 - xy)^n$. Then $g'(x)$ is a nondecreasing function of $x \in [0, 1]$, which implies g itself is a convex function of x . Hence,

$$(1 - xy)^n = g(x) = g((1 - x)0 + x1) \leq (1 - x)g(0) + xg(1) = 1 - x + x(1 - y)^n.$$

Since we show below that $(1 - y)^n \leq e^{-yn}$, we can write

$$(1 - xy)^n \leq 1 - x + xe^{-yn} \leq 1 - x + e^{-yn}.$$

To show that $(1 - y)^n \leq e^{-yn}$, put $f(y) := e^{-y} - 1 + y$. Since $f(0) = 0$ and $f'(y) \geq 0$ for $y \geq 0$, we have $f(y) \geq 0$ for $y \geq 0$.^c Thus, for $0 \leq y \leq 1$, we have $0 \leq 1 - y \leq e^{-y}$, and raising to the n th power yields the result. $\square$

12.6.1. The Joint AEP for Rate Distortion

Given ε and a joint pmf P_{XY} , we define the **distortion-typical sets** for rate distortion (cf. Section 8.1)

$$\begin{aligned} \hat{A}_n := \left\{ (\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n : \right. & \left| \frac{1}{n} \log \frac{1}{P_X^n(\mathbf{x})} - H(P_X) \right| \leq \varepsilon, \\ & \left| \frac{1}{n} \log \frac{1}{P_Y^n(\mathbf{y})} - H(P_Y) \right| \leq \varepsilon, \\ & \left| \frac{1}{n} \log \frac{1}{P_{XY}^n(\mathbf{x}, \mathbf{y})} - H(P_{XY}) \right| \leq \varepsilon, \\ & \text{and } |d_n(\mathbf{x}, \mathbf{y}) - \mathbb{E}[d(X, Y)]| \leq \varepsilon \left. \right\}. \end{aligned}$$

There are two important things to notice about $\hat{A}_n$. First, $\hat{A}_n \subset A_n$, where A_n is the collection of jointly typical pairs defined in Section 8.1. Hence, the probability bounds

^cThis can be seen by the mean-value theorem or by writing $f(y) = \int_0^y 1 - e^{-t} dt \geq 0$ for $y \geq 0$.

on $P_X^n(\mathbf{x})$, $P_Y^n(\mathbf{y})$, and $P_{XY}^n(\mathbf{x}, \mathbf{y})$ there hold here as well for $(\mathbf{x}, \mathbf{y}) \in \hat{A}_n$. In addition, if $(\mathbf{x}, \mathbf{y})$ are distortion typical, then

$$d_n(\mathbf{x}, \mathbf{y}) \leq \mathbb{E}[d(X, Y)] + \varepsilon.$$

If P_{XY} is chosen so that $\mathbb{E}[d(X, Y)] \leq D$, then $d_n(\mathbf{x}, \mathbf{y}) \leq D + \varepsilon$.

Theorem 12.10 (Joint AEP for Rate Distortion). *Let $\hat{A}_n$ be defined as above. For large n , the set $\hat{A}_n$ has high P_{XY}^n probability in the sense that*

$$\lim_{n \rightarrow \infty} P_{XY}^n(\hat{A}_n) = 1.,$$

and for all n and $(\mathbf{x}, \mathbf{y}) \in \hat{A}_n$,

$$P_Y^n(\mathbf{y}) \geq P_{Y|X}^n(\mathbf{y}|\mathbf{x}) \exp(-n[I(P_{XY}) + 3\varepsilon]).$$

Proof. We show that $P_{XY}^n(\hat{A}_n^c) \rightarrow 0$. Let $(X_1, Y_1), (X_2, Y_2), \dots$ be i.i.d. pairs with common probability mass function P_{XY} . Put $\mathbf{X} := (X_1, \dots, X_n)$ and $\mathbf{Y} := (Y_1, \dots, Y_n)$. Then

$$\begin{aligned} \mathbb{P}((\mathbf{X}, \mathbf{Y}) \notin \hat{A}_n) &= \mathbb{P}\left(\left|\frac{1}{n} \sum_{k=1}^n \log \frac{1}{P_X(X_k)} - H(P_X)\right| > \varepsilon, \right. \\ &\quad \text{or } \left|\frac{1}{n} \sum_{k=1}^n \log \frac{1}{P_Y(Y_k)} - H(P_Y)\right| > \varepsilon, \\ &\quad \text{or } \left|\frac{1}{n} \sum_{k=1}^n \log \frac{1}{P_{XY}(X_k, Y_k)} - H(P_{XY})\right| > \varepsilon, \\ &\quad \left. \text{or } \left|\frac{1}{n} \sum_{k=1}^n d(X_k, Y_k) - \mathbb{E}[d(X, Y)]\right| > \varepsilon\right) \\ &\leq \mathbb{P}\left(\left|\frac{1}{n} \sum_{k=1}^n \log \frac{1}{P_X(X_k)} - H(P_X)\right| > \varepsilon\right) \\ &\quad + \mathbb{P}\left(\left|\frac{1}{n} \sum_{k=1}^n \log \frac{1}{P_Y(Y_k)} - H(P_Y)\right| > \varepsilon\right) \\ &\quad + \mathbb{P}\left(\left|\frac{1}{n} \sum_{k=1}^n \log \frac{1}{P_{XY}(X_k, Y_k)} - H(P_{XY})\right| > \varepsilon\right) \\ &\quad + \mathbb{P}\left(\left|\frac{1}{n} \sum_{k=1}^n d(X_k, Y_k) - \mathbb{E}[d(X, Y)]\right| > \varepsilon\right). \end{aligned}$$

In each probability on the right, the terms in the sums are i.i.d., and the expectation of each logarithm is the corresponding entropy. Hence, by the weak law of large numbers, each probability tends to zero as $n \rightarrow \infty$.

For the rest of the proof, use the fact that $\hat{A}_n \subset A_n$, where A_n is the jointly typical set in Section 8.1. Use the probability bounds there to upper bound

$$P_{Y|X}^n(\mathbf{y}|\mathbf{x}) = \frac{P_{XY}^n(\mathbf{x}, \mathbf{y})}{P_X^n(\mathbf{x})} = P_Y^n(\mathbf{y}) \frac{P_{XY}^n(\mathbf{x}, \mathbf{y})}{P_X^n(\mathbf{x})P_Y^n(\mathbf{y})}. \quad \square$$

12.7. The Forward Theorem

Theorem 12.11. *Assuming finite source and reproduction alphabets and a discrete memoryless source, if $R > \tilde{R}(D)$, then (R, D) is achievable.*

Proof. Let p be the pmf of the discrete memoryless source. Let W be such that with $P_{XY}(x, y) = p(x)W(y|x)$,

$$\mathbb{E}[d(X, Y)] \leq D \quad \text{and} \quad \tilde{R}(D) = I(p \times W).$$

We show that if $R > \tilde{R}(D)$, then (R, D) is achievable. Let $\Delta D > 0$ and $\Delta R > 0$ be given. Choose ε so that

$$0 < \varepsilon < \frac{R - I(p \times W)}{3} \quad \text{and} \quad \varepsilon < \frac{\Delta D}{3}.$$

With these choices of P_{XY} and ε , let $\hat{A}_n$ be as above with n so large that

$$\mathbb{P}((\mathbf{X}, \mathbf{Y}) \notin \hat{A}_n) < \varepsilon/d_{\max} \quad \text{and} \quad e^{-\exp(n[R - I(p \times W) - 3\varepsilon])} < \varepsilon/d_{\max},$$

where

$$d_{\max} := \max_{x, y} d(x, y).$$

Let $N := \lceil \exp(nR) \rceil$. Given $\mathbf{y}_1, \dots, \mathbf{y}_N$ each in $\mathcal{Y}^n$, consider the mapping $q: \mathcal{X}^n \rightarrow \mathcal{Y}^n$ defined by $q(\mathbf{x}) = \mathbf{y}_i$ if $(\mathbf{x}, \mathbf{y}_i) \in \hat{A}_n$. If there is more than one such i , use the smallest, and if there is no such i , put $q(\mathbf{x}) = \mathbf{y}_N$. It is convenient to put

$$B := \bigcup_{i=1}^N B_i, \quad \text{where} \quad B_i := \{\mathbf{x} : (\mathbf{x}, \mathbf{y}_i) \in \hat{A}_n\}.$$

Then if $\mathbf{x} \in B$, there is an i with $(\mathbf{x}, \mathbf{y}_i)$ distortion typical, and thus $d_n(\mathbf{x}, q(\mathbf{x})) \leq D + \varepsilon$. Hence,

$$\begin{aligned} \mathbb{E}[d_n(\mathbf{X}, q(\mathbf{X}))] &= \mathbb{E}[d_n(\mathbf{X}, q(\mathbf{X}))\mathbf{1}_B(\mathbf{X})] + \mathbb{E}[d_n(\mathbf{X}, q(\mathbf{X}))\mathbf{1}_{B^c}(\mathbf{X})] \\ &\leq D + \varepsilon + d_{\max}\mathbb{P}(\mathbf{X} \in B^c). \end{aligned} \quad (12.2)$$

To analyze this last probability, write

$$\begin{aligned} P(\mathbf{X} \in B^c) &= \sum_{\mathbf{x}} p^n(\mathbf{x}) \mathbf{1}_{B^c}(\mathbf{x}) = \sum_{\mathbf{x}} p^n(\mathbf{x}) \prod_{i=1}^N \mathbf{1}_{B_i^c}(\mathbf{x}) = \sum_{\mathbf{x}} p^n(\mathbf{x}) \prod_{i=1}^N \mathbf{1}_{\hat{A}_n^c}(\mathbf{x}, \mathbf{y}_i) \\ &=: \alpha(\mathbf{y}_1, \dots, \mathbf{y}_N). \end{aligned}$$

We now employ the random coding argument. Let $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ be i.i.d. with common pmf $(pW)^n(\mathbf{y})$. Then

$$E[\alpha(\mathbf{Y}_1, \dots, \mathbf{Y}_N)] = \sum_{\mathbf{x}} p^n(\mathbf{x}) E[\mathbf{1}_{\hat{A}_n^c}(\mathbf{x}, \mathbf{Y})]^N.$$

Next write

$$E[\mathbf{1}_{\hat{A}_n^c}(\mathbf{x}, \mathbf{Y})] = 1 - E[\mathbf{1}_{\hat{A}_n}(\mathbf{x}, \mathbf{Y})].$$

Then

$$\begin{aligned} E[\mathbf{1}_{\hat{A}_n}(\mathbf{x}, \mathbf{Y})] &= \sum_{\mathbf{y}: (\mathbf{x}, \mathbf{y}) \in \hat{A}_n} P_Y^n(\mathbf{y}) \\ &\geq \sum_{\mathbf{y}: (\mathbf{x}, \mathbf{y}) \in \hat{A}_n} P_{Y|X}^n(\mathbf{y}|\mathbf{x}) \exp(-n[I(P_{XY}) + 3\epsilon]), \quad \text{by Theorem 12.10.} \end{aligned}$$

We now have

$$\begin{aligned} E[\alpha(\mathbf{Y}_1, \dots, \mathbf{Y}_N)] &\leq \sum_{\mathbf{x}} P_X^n(\mathbf{x}) \left[1 - \sum_{\mathbf{y}: (\mathbf{x}, \mathbf{y}) \in \hat{A}_n} P_{Y|X}^n(\mathbf{y}|\mathbf{x}) \exp(-n[I(P_{XY}) + 3\epsilon]) \right]^N \\ &\leq \sum_{\mathbf{x}} P_X^n(\mathbf{x}) \left[1 - \sum_{\mathbf{y}: (\mathbf{x}, \mathbf{y}) \in \hat{A}_n} P_{Y|X}^n(\mathbf{y}|\mathbf{x}) + e^{-N \exp(-n[I(P_{XY}) + 3\epsilon])} \right], \quad \text{by Lemma 12.9,} \\ &= \left[1 - \sum_{(\mathbf{x}, \mathbf{y}) \in \hat{A}_n} P_{XY}^n(\mathbf{x}, \mathbf{y}) \right] + e^{-N \exp(-n[I(X \wedge Y) + 3\epsilon])} \\ &= P((\mathbf{X}, \mathbf{Y}) \in \hat{A}_n^c) + e^{-N \exp(-n[I(X \wedge Y) + 3\epsilon])} \\ &\leq P((\mathbf{X}, \mathbf{Y}) \in \hat{A}_n^c) + e^{-\exp(n[R - I(X \wedge Y) - 3\epsilon])} \\ &< 2\epsilon/d_{\max}. \end{aligned}$$

Hence, there is a realization of $\mathbf{y}_1, \dots, \mathbf{y}_N$ with $\alpha(\mathbf{y}_1, \dots, \mathbf{y}_N) < 2\epsilon/d_{\max}$. Returning to (12.2), we have

$$E[d_n(\mathbf{X}, q(\mathbf{X}))] \leq D + 3\epsilon < D + \Delta D. \quad \square$$

We saw earlier that $D \geq D_{\max} \Rightarrow R(D) = 0$. We also saw that for a DMS with pmf p , $D \geq D_{\max} \Rightarrow \tilde{R}(D) = 0$. In the case of a DMS, we can find *all* D for which $R(D) = 0$. Of course this is equivalent to finding all D for which $\tilde{R}(D) = 0$.

Proposition 12.12. *Assuming finite source and reproduction alphabets, a discrete memoryless source with pmf p satisfies $R(D) = 0 \Leftrightarrow D \geq D_*$, where*

$$D_* := \min_y \sum_x p(x) d(x, y).$$

Proof. We first show that $D \geq D_* \Rightarrow \tilde{R}(D) = 0$. Let y_0 be any value of y achieving the minimum in the definition of D_* ; i.e.,

$$\sum_x p(x) d(x, y_0) = D_*.$$

Put

$$W_0(y) := \begin{cases} 1, & y = y_0, \\ 0, & y \neq y_0. \end{cases}$$

If $P_{XY}(x, y) = p(x)W_0(y)$, then X and Y are independent, and so $I(p \times W_0) = 0$. Also,

$$\mathbb{E}[d(X, Y)] = \sum_x \sum_y p(x) W_0(y) d(x, y) = \sum_x p(x) d(x, y_0) = D_*.$$

These observations show that for $D \geq D_*$,

$$\tilde{R}(D) = \inf_{\mathbb{E}[d(X, Y)] \leq D} I(X \wedge Y) \leq I(p \times W_0) = 0.$$

To prove that $\tilde{R}(D) = 0 \Rightarrow D \geq D_*$, suppose

$$0 = \tilde{R}(D) = \inf_{W: \mathbb{E}[d(X, Y)] \leq D} I(p \times W).$$

Since the infimum must be achieved, there is a W with $\mathbb{E}[d(X, Y)] \leq D$ and $I(p \times W) = 0$. Now $I(X \wedge Y) = 0$ implies X and Y are independent, which means that $W(y|x) = W(y)$ does not depend on x . Hence,

$$\begin{aligned} D \geq \mathbb{E}[d(X, Y)] &= \sum_x \sum_y p(x) W(y) d(x, y) = \sum_y W(y) \left[\sum_x p(x) d(x, y) \right] \\ &\geq \sum_y W(y) D_* = D_*. \end{aligned} \quad \square$$

12.8. Computation of $\tilde{R}(D)$

Example 12.13. Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ equipped with mod-2 addition. Let $p(1) := \beta < 1/2$ so that $p(0) = 1 - \beta > 1/2$. Consider the Hamming distortion $d(x, y) := 0$ if $x = y$ and $d(x, y) := 1$ otherwise. We show that

$$\tilde{R}(D) := \inf_{W: \mathbb{E}[d(X, Y)] \leq D} I(p \times W) = \begin{cases} h(\beta) - h(D), & 0 \leq D \leq \min\{\beta, 1 - \beta\}, \\ 0, & D > \min\{\beta, 1 - \beta\}. \end{cases}$$

Consider

$$\begin{aligned}
 I(p \times W) &= I(X \wedge Y) \\
 &= H(X) - H(X|Y) \\
 &= H(X) - \sum_y P_Y(y) H(P_{X|Y}(\cdot|y)) \\
 &= H(X) - \sum_y P_Y(y) H(P_{X|Y}(\cdot + y|y)), \quad (\text{addition mod 2}), \\
 &= H(X) - \sum_y P_Y(y) H(X - Y|Y = y) \\
 &= H(X) - H(X - Y|Y) \\
 &\geq H(X) - H(X - Y) \\
 &= H(X) - h(\mathbb{P}(X \neq Y)) \\
 &= H(p) - h(\mathbb{P}(X \neq Y)).
 \end{aligned}$$

Now, observe that $\mathbb{E}[d(X, Y)] = \mathbb{E}[\mathbf{1}_{\{X \neq Y\}}] = \mathbb{P}(X \neq Y)$. So, if $\mathbb{E}[d(X, Y)] \leq D$, $\mathbb{P}(X \neq Y) \leq D$. Since $h(\cdot)$ is increasing on $[0, 1/2]$,

$$\begin{aligned}
 I(p \times W) &\geq H(p) - h(\mathbb{P}(X \neq Y)) \\
 &\geq H(p) - h(D), \quad 0 \leq D \leq 1/2.
 \end{aligned}$$

In fact, since $H(p) = h(\beta)$, and $\beta < 1/2$,

$$I(p \times W) \geq h(\beta) - h(D), \quad 0 \leq D \leq \beta.$$

Now, equality will hold if $X - Y$ and Y are independent, and if $\mathbb{E}[d(X, Y)] = \mathbb{P}(X \neq Y) = D$. Let $Z := X - Y$; then $X = Y + Z$, where Z and Y are independent. We want $H(Z) = h(\mathbb{P}(Z = 1)) = h(\mathbb{P}(X \neq Y)) = h(D)$; i.e., we want $\mathbb{P}(Z = 1) = D$.

Let $r = \mathbb{P}(Y = 1)$. If Z and Y are independent with $\mathbb{P}(Z = 1) = D$, and $r = \mathbb{P}(Y = 1)$, can r be chosen so that if $X := Y + Z$, then $\mathbb{P}(X = 1) = \beta$? That is, can we find r such that if $X = Y + Z$, Y , and Z are independent, then

$$\begin{aligned}
 \beta &= \mathbb{P}(X = 1) = \mathbb{P}(X = 1|Y = 0)\mathbb{P}(Y = 0) + \mathbb{P}(X = 1|Y = 1)\mathbb{P}(Y = 1) \\
 &= D(1 - r) + (1 - D)r ?
 \end{aligned}$$

If this holds,

$$r = \frac{\beta - D}{1 - 2D}.$$

Note that $0 \leq D \leq \beta < 1/2$ implies $0 \leq r < 1$. So, given β and D , put $r := (\beta - D)/(1 - 2D)$. Consider independent random variables Y and Z such that $\mathbb{P}(Y = 1) = r$ and $\mathbb{P}(Z = 1) = D$. Put $X := Y + Z$. Then $\mathbb{P}(X = 1) = \beta$, and

$$\mathbb{E}[d(X, Y)] = \mathbb{P}(X \neq Y) = \mathbb{P}(Z = 1) = D \quad \text{and} \quad I(X \wedge Y) = h(\beta) - h(D).$$

Hence, $\tilde{R}(D) = h(\beta) - h(D)$, $0 \leq D \leq \beta < 1/2$.

If $D > \beta$, let $W(0|x) = 1$; then X and Y are independent and $I(X \wedge Y) = 0$; also, for this choice of W , $E[d(X, Y)] = \sum_{x=0}^1 p(x)W(0|x)d(x, 0) = \beta < D$. Hence, $\tilde{R}(D) = 0$ for $D > \beta$.

Similarly, if $\beta \geq 1/2$, we would have argued, $I(p \times W) \geq h(1 - \beta) - h(D)$ for $0 \leq D \leq 1 - \beta$. We could then show that $r := (\beta - D)/(1 - 2D)$ implies $0 < 1 - r \leq 1$, since $(1 - r) = [(1 - \beta) - D]/(1 - 2D)$, where $0 \leq D < 1 - \beta \leq 1/2$. For $D \geq 1 - \beta$, take $W(1|x) = 1$. Then $E[d(X, Y)] = 1 - \beta \leq D$.

Example 12.14. Let $X = Y = \{0, 1, \dots, m-1\}$ with mod- m addition. We again use the Hamming distortion, $d(x, y) := 1$ if $x \neq y$ and $d(x, y) = 0$ if $x = y$. Let $P(X = x) = 1/m$ for all $x \in X$. Find $\tilde{R}(D)$. As in last example, write

$$\begin{aligned} I(p \times W) &= H(X) - H(X|Y) \\ &= H(X) - H(X - Y|Y) \\ &\geq H(X) - H(X - Y), \end{aligned}$$

with equality $\Leftrightarrow X - Y$ and Y are independent. Let $Z := X - Y$. Observe that

$$E[d(X, Y)] = E[\mathbf{1}_{\{X \neq Y\}}] = P(X \neq Y) = P(X - Y \neq 0) = P(Z \neq 0).$$

Next,

$$\begin{aligned} H(X - Y) &= H(Z) = P(Z = 0) \log \frac{1}{P(Z = 0)} + \sum_{z=1}^{m-1} P(Z = z) \log \frac{1}{P(Z = z)} \\ &\leq P(Z = 0) \log \frac{1}{P(Z = 0)} + P(Z \neq 0) \log \frac{m-1}{P(Z \neq 0)} \\ &= h(P(Z \neq 0)) + P(Z \neq 0) \log(m-1), \end{aligned}$$

by the log-sum inequality, with equality $\Leftrightarrow$

$$P(Z = z) = \frac{P(Z \neq 0)}{m-1}, \quad \text{for } z \neq 0. \quad (12.3)$$

Since $h(\theta) + \theta \log(m-1)$ is strictly increasing for $0 \leq \theta \leq (m-1)/m$, we have

$$E[d(X, Y)] = P(Z \neq 0) \leq D \leq \frac{m-1}{m} \Rightarrow H(Z) \leq h(D) + D \log(m-1),$$

with equality if and only if (12.3) holds with $P(Z \neq 0) = D$. Now suppose also that Y and Z are independent with

$$P(Y = y) = r(y), \quad P(Z = 0) = 1 - D, \quad \text{and} \quad P(Z = z) = \frac{D}{m-1}, \quad z \neq 0.$$

Can we find a probability mass function r such that if $X := Y + Z$, then $P(X = x) = 1/m$ for all $x \in X$? By Problem 12.7, it suffices to take r to be the uniform distribution. With this choice of r and $P(Z = 0) = 1 - D$, we have $E[d(X, Y)] = D$ and

$$\begin{aligned} I(X \wedge Y) &= H(X) - H(X|Y) \\ &= \log |X| - [h(D) + D \log(|X| - 1)], \end{aligned}$$

which gives,

$$\tilde{R}(D) = \begin{cases} \log |X| - [h(D) + D \log(|X| - 1)], & 0 \leq D < \frac{|X|-1}{|X|}, \\ 0, & D \geq \frac{|X|-1}{|X|}. \end{cases}$$

The case $D \geq (|X| - 1)/|X|$ can be handled by taking $W(y|x) := 1/|X|$. Then

$$\begin{aligned} E[d(X, Y)] &= \sum_x p(x) \sum_y W(y|x) d(x, y) \\ &= \sum_x p(x) \sum_{y \neq x} W(y|x) \\ &= \sum_x p(x) \frac{|X| - 1}{|X|} \\ &\leq D. \end{aligned}$$

Furthermore, $W(y|x) = 1/|X|$ implies X and Y are independent, which implies that $I(X \wedge Y) = 0$. Since the mutual information cannot be negative, we have shown that the infimum must be zero, that is, $\tilde{R}(D) = 0$ for $D \geq (|X| - 1)/|X|$. A sketch of $\tilde{R}(D)$ for the case $|X| = 5$ is shown at the left in Figure 12.1.

12.9. The Shannon Lower Bound

Consider a distortion measure $d(x, y)$ such that for all $y, y' \in Y$, $\{d(\cdot, y)\}$ and $\{d(\cdot, y')\}$ are permutations of each other. Put

$$\eta(D) := \sup_{q: \sum_x q(x) d_c(x) \leq D} H(q)$$

where $d_c(x) := d(x, y_0)$ for some fixed $y_0 \in Y$. Since $H(q) = H(q')$ when q' is a permutation of q , we see that $\eta(D)$ does not depend on y_0 .

Theorem 12.15 (Shannon Lower Bound). *Let d have the permutation property of the preceding paragraph. Assuming finite source and reproduction alphabets and a discrete memoryless source with pmf p ,*

$$R(D) \geq H(p) - \eta(D),$$

where η is nondecreasing and concave.

Proof. It is clear that η is a nondecreasing function of D . We next show that η is a concave function of D . Suppose $\eta(D_1) = H(q_1)$ and $\eta(D_2) = H(q_2)$, where $\sum_x q_1(x)d_c(x) \leq D_1$ and $\sum_x q_2(x)d_c(x) \leq D_2$. Put $q = \lambda q_1 + (1 - \lambda)q_2$ where $0 \leq \lambda \leq 1$. Then,

$$\begin{aligned} \sum_x q(x)d_c(x) &= \lambda \sum_x q_1(x)d_c(x) + (1 - \lambda) \sum_x q_2(x)d_c(x) \\ &\leq \lambda D_1 + (1 - \lambda)D_2. \end{aligned}$$

Furthermore,

$$\begin{aligned} \eta(\lambda D_1 + (1 - \lambda)D_2) &\geq H(q) \\ &\geq \lambda H(q_1) + (1 - \lambda)H(q_2), \quad \text{since } H \text{ is concave,} \\ &= \lambda \eta(D_1) + (1 - \lambda)\eta(D_2). \end{aligned}$$

We now establish the lower bound. Suppose that $E[d(X, Y)] \leq D$. Then

$$\begin{aligned} D &\geq \sum_x \sum_y P_{XY}(x, y)d(x, y) \\ &= \sum_y P_Y(y) \left[\underbrace{\sum_x P_{X|Y}(x|y)d(x, y)}_{=: \hat{d}(y)} \right]. \end{aligned}$$

With $\hat{d}(y)$ as above,

$$\begin{aligned} I(X \wedge Y) &= H(X) - H(X|Y) \\ &= H(X) - \sum_y P_Y(y)H(X|Y = y) \\ &= H(X) - \sum_y P_Y(y)H(P_{X|Y}(\cdot|y)) \\ &\geq H(X) - \sum_y P_Y(y)\eta(\hat{d}(y)), \quad \text{by def. of } \eta, \\ &\geq H(X) - \eta(\sum_y P_Y(y)\hat{d}(y)), \quad \text{since } \eta \text{ is concave,} \\ &= H(X) - \eta(E[d(X, Y)]), \\ &\geq H(X) - \eta(D), \quad \text{since } \eta \text{ is nondecreasing.} \end{aligned} \quad \square$$

Remark. It can be shown that the Shannon lower bound holds with equality if d has the additional property that for any $x, x' \in \mathcal{X}$, $\{d(x, \cdot)\}$ and $\{d(x', \cdot)\}$ are permutations of each other.

Problems

1. Derive the chain rule for conditional entropy,

$$H(UV|Y) = H(U|Y) + H(V|YU).$$

2. Show that if $N := \lceil \exp(nR) \rceil$, then

$$\lim_{n \rightarrow \infty} \frac{\log N}{n} = R.$$

3. Prove the second part of Theorem 12.10.

4. Show that the rate-distortion region is closed.

5. Fix $0 \leq D \leq \beta < 1/2$, and put $r := (\beta - D)/(1 - 2D)$. Let Y and Z be independent, $\{0, 1\}$ -valued random variables with $P(Y = 1) = r$ and $P(Z = 1) = D$. If $X := Y + Z \bmod 2$, find $P(Y = y|X = x)$ for $x, y \in \{0, 1\}$.

6. Let X and Y be countable sets. Suppose that for some $y_0 \in Y$, we have $W(y_0|x) = 1$ for all $x \in X$. Show that X and Y are independent.

7. Let $X := \{0, 1, \dots, m-1\}$ with addition mod- m . Let Y and Z be independent X -valued random variables. If the pmf of Y is uniform, show that the pmf of $X := Y + Z$ is uniform.

8. Show that the function $g(\theta) := h(\theta) + \theta \log(m-1)$ is increasing for $0 \leq \theta \leq (m-1)/m$.

9. For the source and distortion function in Example 12.14, compute $D_{\max}$ as in Section 12.2.1 and compute D_* using Proposition 12.12. Does $D_{\max} = D_*$?

10. Consider a binary source taking values 0 and 1 with equal probability $1/2$. Find $\tilde{R}(D)$ for the distortion function

$$d(x, y) := \begin{cases} 0, & y = x, \\ 1, & x = 1, y = 0, \\ \infty, & x = 0, y = 1. \end{cases}$$

A sketch of $\tilde{R}(D)$ is shown at the right in Figure 12.1.

11. If X is a continuous random variable, put

$$\tilde{R}(D) := \inf_{w: E[d(X, Y)] \leq D} I(p \times w),$$

where $d(x, y) := |x - y|^2$ and w is a conditional “density.” If p is a Gaussian density with zero mean and variance σ^2 , find $\tilde{R}(D)$.

CHAPTER 13

Source Coding Revisited – Discrete Stationary Sources

13.1. Stationarity and the Entropy Rate

Let X be a finite set. Let $X_1, X_2, \dots$ be an X -valued, discrete-time random process such that for all positive integers n and k ,

$$P(X_{k+1} = x_1, \dots, X_{k+n} = x_n) = P(X_1 = x_1, \dots, X_n = x_n),$$

where $x_1, \dots, x_n$ are in X . In this case, the process is said to be **(strictly) stationary**. For such processes, we show later that

- (a) $H(X_n | X_1, \dots, X_{n-1})$ is nonincreasing with n .
 - (b) $\frac{1}{n}H(X_1, \dots, X_n) \geq H(X_n | X_1, \dots, X_{n-1})$.
 - (c) $\frac{1}{n}H(X_1, \dots, X_n)$ is nonincreasing with n .
 - (d) $\lim_{n \rightarrow \infty} \frac{1}{n}H(X_1, \dots, X_n) = \lim_{n \rightarrow \infty} H(X_n | X_1, \dots, X_{n-1})$.
- The fact that entropy is bounded below by zero combined with properties (a) and (c) implies that these limits exist.

The **entropy rate** of a stationary process $X_1, X_2, \dots$ is

$$\mathcal{H} := \lim_{n \rightarrow \infty} \frac{1}{n}H(X_1, \dots, X_n).$$

We then have the following generalization of the Variable-Length Source Coding Theorem 3.21. (Note: If $X_1, X_2, \dots$ are i.i.d. with common pmf p , then $\mathcal{H} = H(p)$.)

Theorem 13.1 (Variable-Length Source Coding Theorem for Stationary Source). *Let X be a finite set, and let $X_1, X_2, \dots$ be a stationary X -valued source. Then for every blocklength $n \geq 1$, there is a prefix code $C_n: \mathsf{X}^n \rightarrow \mathsf{D}^*$ such that*

$$\frac{\frac{1}{n}H(X_1 \cdots X_n)}{\log D} \leq \frac{E[\ell(C_n(X_1, \dots, X_n))]}{n} < \frac{\frac{1}{n}H(X_1 \cdots X_n)}{\log D} + \frac{1}{n}.$$

Furthermore, given $\varepsilon > 0$, for all sufficiently large n ,

$$\frac{\mathcal{H}}{\log D} \leq \frac{E[\ell(C_n(X_1, \dots, X_n))]}{n} < \frac{\mathcal{H}}{\log D} + \varepsilon.$$

13.2. Ergodic Sources

A stationary process $X_1, X_2, \dots$ that is **ergodic** has the additional property that for every function $f(x)$,^a

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(X_k) = \mathbb{E}[f(X_1)], \quad \text{a.s.},$$

whenever $\mathbb{E}[|f(X_1)|] < \infty$. For example, an i.i.d. sequence has this property by the **strong law of large numbers**.

Example 13.2 (A Stationary Process That Is Not Ergodic). Let Z be a binary random variable taking the values 0 and 1 with equal probability. Put $X_k := Z$ for all k . Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n X_k = Z,$$

which is either zero or one, and is therefore never equal to $\mathbb{E}[X_1] = \mathbb{E}[Z] = 1/2$!

An ergodic process also has the property that if $f(x_1, \dots, x_m)$ is a function of m variables, then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n f(X_k, X_{k+1}, \dots, X_{k+(m-1)}) = \mathbb{E}[f(X_1, \dots, X_m)], \quad \text{a.s.},$$

whenever $\mathbb{E}[|f(X_1, \dots, X_m)|] < \infty$.

13.3. The Shannon–McMillan–Breiman Theorem

For $\Delta R > 0$, put

$$A_n := \left\{ (x_1, \dots, x_n) \in X^n : \left| \frac{1}{n} \log \frac{1}{p_n(x_1, \dots, x_n)} - \mathcal{H} \right| \leq \Delta R \right\},$$

where $p_n(x_1, \dots, x_n) := \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$. Then $\mathbf{x} = (x_1, \dots, x_n) \in A_n$ implies

$$\exp(-n[\mathcal{H} + \Delta R]) \leq p_n(\mathbf{x}) \leq \exp(-n[\mathcal{H} - \Delta R]).$$

^aThe abbreviation “a.s.” stands for **almost surely**.

With $\mathbf{X} = (X_1, \dots, X_n)$, it would be nice to have $\lim_{n \rightarrow \infty} P(\mathbf{X} \notin A_n) = 0$. If the X_k were i.i.d., we could use the weak law of large numbers. For stationary ergodic processes, we use the the following result.

Theorem 13.3 (Shannon–McMillan–Breiman). *Let $\mathbf{X}$ be a finite set, and let $X_1, X_2, \dots$ be a stationary, ergodic, $\mathbf{X}$ -valued process. Then*

$$\lim_{n \rightarrow \infty} P((X_1, \dots, X_n) \notin A_n) = 0.$$

Proof. See below. □

With the Shannon–McMillan–Breiman Theorem, it is straightforward to adapt the proof of Shannon’s Source Coding Theorem 4.3 to prove the following result.

Theorem 13.4 (Source Coding Theorem for a Stationary Source). *Let $\mathbf{X}$ be a finite set, and let $X_1, X_2, \dots$ be a stationary, ergodic, $\mathbf{X}$ -valued source. Then every $R \geq \mathcal{H}$ is achievable, and every $R < \mathcal{H}$ is not achievable.*

Proof. Apply the Shannon–McMillan–Breiman Theorem. □

The remainder of this section is devoted to proving the Shannon–McMillan–Breiman theorem.

Lemma 13.5. *Let $\mathbf{X}$ be a finite set, and let $X_1, X_2, \dots$ be a stationary, ergodic, $\mathbf{X}$ -valued process. For $m \geq 1$,*

$$\lim_{\substack{n \rightarrow \infty \\ n > m}} \frac{1}{n} \log Q_m(X_1, \dots, X_n) = -H(X_{m+1} | X_m, \dots, X_1), \quad \text{a.s.,}$$

where for $n > m$,

$$Q_m(x_1, \dots, x_n) := P(X_1 = x_1, \dots, X_m = x_m) \cdot \prod_{k=m+1}^n P(X_k = x_k | X_{k-1} = x_{k-1}, \dots, X_{k-m} = x_{k-m}).$$

Proof. Write,

$$\begin{aligned} \frac{1}{n} \log Q_m(X_1, \dots, X_n) &= \frac{1}{n} \log p_m(X_1, \dots, X_m) \\ &\quad + \frac{1}{n} \sum_{k=m+1}^n \log P_{X_k | X_{k-1} \dots X_{k-m}}(X_k | X_{k-1}, \dots, X_{k-m}). \end{aligned}$$

Clearly, $\frac{1}{n} \log p_m(X_1, \dots, X_m) \rightarrow 0$ a.s. as $n \rightarrow \infty$. Since

$$\begin{aligned} \mathbb{E}[\log P_{X_k|X_{k-1}, \dots, X_{k-m}}(X_k|X_{k-1}, \dots, X_{k-m})] &= -H(X_k|X_{k-1}, \dots, X_{k-m}) \\ &= -H(X_{m+1}|X_m, \dots, X_1), \end{aligned}$$

and since the source is ergodic, the term

$$\frac{1}{n} \sum_{k=m+1}^n \log P_{X_k|X_{k-1}, \dots, X_{k-m}}(X_k|X_{k-1}, \dots, X_{k-m}) \rightarrow -H(X_{m+1}|X_m, \dots, X_1)$$

almost surely. □

Lemma 13.6. *Let X be a finite set, and let $X_1, X_2, \dots$ be a stationary X -valued process. For any $\Delta R > 0$ and $\varepsilon > 0$, for sufficiently large m and n , $n > m$,*

$$\mathbb{P}\left(\left|\frac{1}{n} \log Q_m(X_1, \dots, X_n) - \frac{1}{n} \log p_n(X_1, \dots, X_n)\right| > \Delta R\right) < \varepsilon.$$

Proof. Write, by Markov's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n} \log \frac{Q_m(X_1, \dots, X_n)}{p_n(X_1, \dots, X_n)}\right| > \Delta R\right) \leq \frac{1}{n\Delta R} \mathbb{E}\left[\left|\log \frac{Q_m(X_1, \dots, X_n)}{p_n(X_1, \dots, X_n)}\right|\right].$$

We now use the fact that for $\theta > 0$, $|\log \theta| = 2[\log \theta]_+ - \log \theta \leq 2\theta(\log e)/e - \log \theta$, where $[t]_+ := t$ for $t \geq 0$ and $[t]_+ := 0$ for $t < 0$, and where the inequality $[\log \theta]_+ \leq \theta(\log e)/e$ is proved in Problem 13.2. So,

$$\mathbb{E}\left[\left|\log \frac{Q_m(X_1, \dots, X_n)}{p_n(X_1, \dots, X_n)}\right|\right] \leq \frac{2 \log e}{e} \mathbb{E}\left[\frac{Q_m(X_1, \dots, X_n)}{p_n(X_1, \dots, X_n)}\right] - \mathbb{E}\left[\log \frac{Q_m(X_1, \dots, X_n)}{p_n(X_1, \dots, X_n)}\right].$$

The first expectation on the right-hand side is simply

$$\sum_{x_1, \dots, x_n} Q_m(x_1, \dots, x_n) = 1,$$

and the second expectation is (including the minus sign)

$$-H(X_1 \cdots X_n) + H(X_1 \cdots X_m) + (n-m)H(X_{m+1}|X_m \cdots X_1).$$

So,

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{n} \log \frac{Q_m(X_1, \dots, X_n)}{p_n(X_1, \dots, X_n)}\right| > \Delta R\right) &\leq \frac{1}{\Delta R} \left\{ \frac{2 \log e}{ne} - \frac{1}{n} H(X_1 \cdots X_n) + \frac{1}{n} H(X_1 \cdots X_m) \right. \\ &\quad \left. + \left(1 - \frac{m}{n}\right) H(X_{m+1}|X_m \cdots X_1) \right\}, \end{aligned}$$

which is less than ε for m and n both large enough. □

Proof of the Shannon–McMillan–Breiman Theorem. We must show that

$$\frac{1}{n} \log \frac{1}{p_n(X_1, \dots, X_n)} \xrightarrow{\text{in prob.}} \mathcal{H}.$$

For $n > m$, write

$$\frac{1}{n} \log \frac{1}{p_n(X_1, \dots, X_n)} - \mathcal{H} \quad (13.1)$$

$$= \frac{1}{n} \log \frac{1}{p_n(X_1, \dots, X_n)} - \frac{1}{n} \log \frac{1}{Q_m(X_1, \dots, X_n)} \quad (13.2)$$

$$+ \frac{1}{n} \log \frac{1}{Q_m(X_1, \dots, X_n)} - H(X_{m+1}|X_m \cdots X_1) \quad (13.3)$$

$$+ H(X_{m+1}|X_m \cdots X_1) - \mathcal{H}. \quad (13.4)$$

Let $\varepsilon > 0$ be given. Fix m so large that the absolute value of the term in (13.4) is less than or equal to $\Delta R/3$. Next let $n > m$ be so large that $P(|(13.3)| > \Delta R/3) < \varepsilon/2$; this is possible by Lemma 13.5. Finally, by Lemma 13.6, let $n > m$ be so large that $P(|(13.2)| > \Delta R/3) < \varepsilon/2$. Since

$$\{|(13.2)| \leq \Delta R/3\} \cap \{|(13.3)| \leq \Delta R/3\} \cap \{|(13.4)| \leq \Delta R/3\} \subset \{|(13.1)| \leq \Delta R\},$$

it follows that $P(|(13.1)| > \Delta R)$ is upper bounded by

$$\underbrace{P(|(13.2)| > \Delta R/3)}_{< \varepsilon/2} + \underbrace{P(|(13.3)| > \Delta R/3)}_{< \varepsilon/2} + \underbrace{P(|(13.4)| > \Delta R/3)}_{=0} < \varepsilon. \quad \square$$

Remark. (i) The almost sure convergence of Lemma 13.5 implies convergence in probability. (ii) If a_m and a are real numbers (deterministic), and if $|a_m - a| \leq \Delta R/3$, then $P(|a_m - a| > \Delta R/3) = P(\emptyset) = 0$.

Proof of Facts (a)–(d).

(a): Since dropping conditioning increases entropy,

$$\begin{aligned} H(X_{n+1}|X_n \cdots X_1) &\leq H(X_{n+1}|X_n \cdots X_2) \\ &= H(X_n|X_{n-1} \cdots X_1), \quad \text{by stationarity.} \end{aligned}$$

(b): Write

$$\begin{aligned} \frac{1}{n} H(X_1 \cdots X_n) &= \frac{1}{n} [H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_{n-1} \cdots X_1)] \\ &= \frac{1}{n} [H(X_n) + H(X_n|X_{n-1}) + \cdots + H(X_n|X_{n-1} \cdots X_1)] \\ &\geq \frac{1}{n} \cdot nH(X_n|X_{n-1} \cdots X_1). \end{aligned}$$

(c): Write

$$\begin{aligned}\frac{1}{n}H(X_1 \cdots X_n) &= \frac{1}{n} [H(X_1 \cdots X_{n-1}) + H(X_n | X_{n-1} \cdots X_1)] \\ &\leq \frac{n-1}{n} \frac{1}{n-1} H(X_1 \cdots X_{n-1}) + \frac{1}{n^2} H(X_1 \cdots X_n), \quad \text{by (b)}.\end{aligned}$$

Or,

$$\left(1 - \frac{1}{n}\right) \frac{1}{n} H(X_1 \cdots X_n) \leq \left(1 - \frac{1}{n}\right) \frac{1}{n-1} H(X_1 \cdots X_{n-1}).$$

(d): Since entropy and conditional entropy are lower bounded by zero, the monotonicity in (a) and (c) guarantee the existence of the limits. For $m \geq 1$,

$$\begin{aligned}\frac{1}{n+m}H(X_1 \cdots X_{n+m}) &= \frac{1}{n+m}H(X_1 \cdots X_{n-1}) \\ &\quad + \frac{1}{n+m} [H(X_n | X_{n-1} \cdots X_1) + H(X_{n+1} | X_n \cdots X_1) \\ &\quad + \cdots + H(X_{n+m} | X_{n+m-1} \cdots X_1)] \\ &\leq \frac{1}{n+m}H(X_1 \cdots X_{n-1}) + \frac{m+1}{n+m}H(X_n | X_{n-1} \cdots X_1).\end{aligned}$$

Letting $m \rightarrow \infty$ yields

$$\mathcal{H} \leq H(X_n | X_{n-1} \cdots X_1) \leq \frac{1}{n} H(X_1 \cdots X_n), \quad \text{by (b)}.$$

Letting $n \rightarrow \infty$ yields

$$\mathcal{H} \leq \lim_{n \rightarrow \infty} H(X_n | X_{n-1} \cdots X_1) \leq \mathcal{H}. \quad \square$$

13.4. Universal Data Compression

13.4.1. The Lempel–Ziv Algorithm

Definition 13.7. Let $\mathcal{X} := \{0, 1\}$ and consider the sequence $\mathbf{x} \in \mathcal{X}^n$. The procedure of parsing $\mathbf{x}$ into the shortest substrings that have not appeared so far is called the **Lempel–Ziv parsing**.

Example 13.8. Suppose $\mathbf{x} = 1011010100010$. We parse it using Lempel–Ziv parsing as

shortest phrases list = 1, 0, 11, 01, 010, 00, 10.

Note that each phrase in the list is unique.

The **Lempel–Ziv Algorithm** consists of two steps.

Parsing: Parse the sequence $\mathbf{x}$ using the Lempel–Ziv parsing to obtain the “shortest phrases list.”

Coding: Represent each phrase found in $\mathbf{x}$, say $x_i \cdots x_{j-1}x_j$, by the ordered pair (p, x_j) , where p is the position of $x_i \cdots x_{j-1}$ in the “shortest phrases list.” Thus, to decode (p, x_j) we concatenate the p th entry in the “shortest phrases list” with the bit x_j . Positions in the “shortest phrases list” are numbered starting at $p = 1$. One-bit phrases of the form x_j are coded as $(0, x_j)$ and are decoded as x_j without looking at the “shortest phrases list.”

Example 13.9. Let $\mathbf{x}$ and its “shortest phrases list” be as in the previous example. Applying coding procedure to $\mathbf{x}$, the corresponding code pairs are

$$(0, 1), (0, 0), (1, 1), (2, 1), (4, 0), (2, 0), (1, 0).$$

If $\mathbf{x} \in X^n$, we denote the number of code pairs by $c_n(\mathbf{x})$. For the sequence $\mathbf{x}$ of Example 13.8, $n = 13$ and $c_{13}(\mathbf{x}) = 7$.

Example 13.10. Let $\mathbf{x} = 101101010001010$. This is same as the previous example, with an additional two bits appended at the end. The result of Lempel–Ziv parsing is

$$\text{shortest phrases list} = 1, 0, 11, 01, 010, 00, 10.$$

Note that the last phrase 10 in $\mathbf{x}$ is already present in the “shortest phrases list,” so it is not repeated there. The code pairs in this case are

$$(0, 1), (0, 0), (1, 1), (2, 1), (4, 0), (2, 0), (1, 0), (1, 0).$$

The code corresponding to the last phrase 10 is the same as the one for the preceding 10 phrase. Since this time $\mathbf{x} \in X^{15}$ and there are 8 code pairs, $c_{15}(\mathbf{x}) = 8$. Notice that the length of the “shortest phrases list” may be shorter than the number of code pairs $c_n(\mathbf{x})$.

13.4.2. Length of Lempel–Ziv Codes

Since the length of the “shortest phrases list” is at most $c_n(\mathbf{x})$, pointer values range at most over $0, 1, \dots, c_n(\mathbf{x})$. Hence, to represent each pointer requires at most $\lceil \log_2(c_n(\mathbf{x}) + 1) \rceil$ bits. Each code pair requires this number of bits plus 1. Since there are a total of $c_n(\mathbf{x})$ code pairs, to represent all of them requires at most

$$\ell_{\text{LZ}}(\mathbf{x}) := \underbrace{c_n(\mathbf{x})}_{\text{number of code pairs}} \underbrace{\left\{ \lceil \log_2(c_n(\mathbf{x}) + 1) \rceil + 1 \right\}}_{\text{bits/code pair}} \text{ bits}.$$

Theorem 13.11. *If $\{X_k\}$ is a stationary, ergodic, binary source, then*

$$\overline{\lim}_{n \rightarrow \infty} \frac{\ell_{\text{LZ}}(X_1 \cdots X_n)}{n} \leq \mathcal{H}, \quad \text{a.s.}$$

Problems

1. Let X be a finite set, and let $X_1, X_2, \dots$ be a time-homogeneous Markov chain; i.e., $P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_1 = x_1) = p(x_n | x_{n-1})$ for some transition probability $p(\cdot | \cdot)$ that does not depend on n . Evaluate

$$\lim_{n \rightarrow \infty} H(X_n | X_{n-1}, \dots, X_1).$$

2. For any real number t ,

$$t_+ := \begin{cases} t, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

- (a) Show that $2t_+ - t = |t|$.
- (b) Show that for $t > 0$, $(\log t)_+ \leq t(\log e)/e$.

CHAPTER 14

The Discrete Memoryless Multiple-Access Channel

Let X , Y , and Z be finite sets. Source 1 transmits message $i \in \{1, \dots, N\}$ and source 2 transmits message $j \in \{1, \dots, M\}$. Message i goes through encoder f to produce codeword $\mathbf{x}_i \in X^n$, and message j goes through encoder g to produce codeword $\mathbf{y}_j \in Y^n$. The **multiple-access channel** (MAC) W_n accepts the codewords $\mathbf{x}_i, \mathbf{y}_j$ and produces output $\mathbf{z} \in Z^n$, which goes through decoder ϕ to produce an estimate of transmitted source messages $(\hat{i}, \hat{j})$.

For a discrete memoryless DMMAC,

$$W_n(\mathbf{z}|\mathbf{x}, \mathbf{y}) = \prod_{k=1}^n W(z_k|x_k, y_k) =: W^n(\mathbf{z}|\mathbf{x}, \mathbf{y}).$$

For example, let X , Y , and V be binary random variables with (X, Y) independent of V . Put $Z := X + Y + V$ with mod-2 addition. Then $W(z|x, y) = \rho_V(z - x - y)$.

The **encoders** are mappings

$$f: \{1, \dots, N\} \rightarrow X^n \quad \text{and} \quad g: \{1, \dots, M\} \rightarrow Y^n.$$

The **decoder** is the mapping

$$\phi: Z^n \rightarrow \{1, \dots, N\} \times \{1, \dots, M\}.$$

The **rate pair** of the code (f, g, ϕ) is

$$\left(\frac{\log N}{n}, \frac{\log M}{n} \right) \in \mathbb{R}_{+}^2.$$

We use the following probabilistic model. Let α be a $\{1, \dots, N\}$ -valued uniform random variable, β a $\{1, \dots, M\}$ -valued uniform random variable, and $\mathbf{Z}$ be a Z^n -valued random variable with

$$P(\alpha = i, \beta = j, \mathbf{Z} = \mathbf{z}) = \frac{1}{NM} W_n(\mathbf{z}|\mathbf{x}_i, \mathbf{y}_j),$$

where $\mathbf{x}_i = f(i)$ and $\mathbf{y}_j = g(j)$. The probability of error is easily computed using the laws of total probability and substitution. Write

$$e_n(f, g, \phi) := P(\phi(\mathbf{z}) \neq (\alpha, \beta))$$

$$\begin{aligned}
&= E[P(\varphi(\mathbf{z}) \neq (\alpha, \beta) | \alpha, \beta)] \\
&= \sum_{i=1}^N \sum_{j=1}^M \frac{1}{N} \frac{1}{M} P(\varphi(\mathbf{z}) \neq (i, j) | \alpha = i, \beta = j) \\
&= \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M \underbrace{W_n(\{\mathbf{z} : \varphi(\mathbf{z}) \neq (i, j)\} | \mathbf{x}_i, \mathbf{y}_j)}_{e_n^{ij}(f, g, \varphi)}
\end{aligned}$$

Definition 14.1. A pair of nonnegative real numbers, (R_1, R_2) is an **achievable rate pair** under the **average probability of error criterion**, if $\forall \lambda > 0, \forall \Delta R > 0, \exists n_0$ such that $\forall n \geq n_0, \exists$ codewords $\mathbf{x}_1, \dots, \mathbf{x}_N$ each in $\mathcal{X}^n$ and $\mathbf{y}_1, \dots, \mathbf{y}_M$ each in $\mathcal{Y}^n$ and there exists a decoder $\varphi : \mathcal{Z}^n \rightarrow \{1, \dots, N\} \times \{1, \dots, M\}$ with

$$\frac{\log N}{n} > R_1 - \Delta R, \quad \frac{\log M}{n} > R_2 - \Delta R, \quad \text{and} \quad e_n(f, g, \varphi) < \lambda.$$

The set of achievable rate pairs, $\mathcal{C} := \{(R_1, R_2) : (R_1, R_2) \text{ is achievable}\}$ is called the **capacity region**.

Remark. For single-user channels, we have by Theorem 5.10 that the capacity region under the maximum probability of error criterion is equal to the capacity region under the average probability of error criterion. For the MAC, this is *not* the case [4, pp. 271–272].

14.1. The Weak Converse

Let (R_1, R_2) be achievable. Then $\forall \lambda > 0, \forall \Delta R > 0$, and for sufficiently large n , $\exists$ codewords $\mathbf{x}_1, \dots, \mathbf{x}_N$ each in $\mathcal{X}^n$ and $\mathbf{y}_1, \dots, \mathbf{y}_M$ each in $\mathcal{Y}^n$ and there exists a decoder $\varphi : \mathcal{Z}^n \rightarrow \{1, \dots, N\} \times \{1, \dots, M\}$ with

$$\frac{\log N}{n} > R_1 - \Delta R, \quad \frac{\log M}{n} > R_2 - \Delta R, \quad \text{and} \quad e_n(f, g, \varphi) < \lambda.$$

Consider random variables $\alpha, \beta, \mathbf{X}, \mathbf{Y}$, and $\mathbf{Z}$ with joint pmf

$$P(\alpha = i, \beta = j, \mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}, \mathbf{Z} = \mathbf{z}) = \frac{1}{NM} \mathbf{1}_{\{\mathbf{x}_i\}}(\mathbf{x}) \mathbf{1}_{\{\mathbf{y}_j\}}(\mathbf{y}) W^n(\mathbf{z} | \mathbf{x}, \mathbf{y}). \quad (14.1)$$

Then

$$P(\varphi(\mathbf{Z}) \neq (\alpha, \beta)) < \lambda.$$

Observe that $\varphi(\mathbf{Z}) = (\varphi_1(\mathbf{Z}), \varphi_2(\mathbf{Z}))$ implies

$$P(\varphi_1(\mathbf{Z}) \neq \alpha) \leq P(\varphi(\mathbf{Z}) \neq (\alpha, \beta)) < \lambda$$

and

$$P(\varphi_2(\mathbf{Z}) \neq \beta) \leq P(\varphi(\mathbf{Z}) \neq (\alpha, \beta)) < \lambda$$

as well. By the Corollary to Fano's inequality,

$$\log NM = H(\alpha, \beta) \leq \lambda \log NM + \log 2 + I(\alpha\beta \wedge \varphi(\mathbf{Z})).$$

Rearranging yields

$$(1 - \lambda) \log NM \leq I(\alpha\beta \wedge \varphi(\mathbf{Z})) + \log 2,$$

which implies

$$(1 - \lambda)[n(R_1 - \Delta R) + n(R_2 - \Delta R)] \leq I(\alpha\beta \wedge \varphi(\mathbf{Z})) + \log 2.$$

Hence,

$$R_1 + R_2 \leq \frac{I(\alpha\beta \wedge \varphi(\mathbf{Z})) + \log 2}{n(1 - \lambda)} + 2\Delta R.$$

Observe that

$$\begin{aligned} I(\alpha\beta \wedge \varphi(\mathbf{Z})) &\leq I(\mathbf{XY} \wedge \mathbf{Z}), && \text{by the Data Processing Lemma,} \\ &= H(\mathbf{Z}) - H(\mathbf{Z}|\mathbf{X}, \mathbf{Y}) \\ &= H(\mathbf{Z}) - \sum_{k=1}^n H(Z_k|X_k Y_k), && \text{since the channel is memoryless,} \\ &\leq \sum_{k=1}^n H(Z_k) - H(Z_k|X_k Y_k), && \text{by (6.1),} \\ &= \sum_{k=1}^n I(X_k Y_k \wedge Z_k) \\ &= n \cdot \frac{1}{n} \sum_{k=1}^n I(X_k Y_k \wedge Z_k), \end{aligned}$$

where

$$\begin{aligned} P_{X_k, Y_k, Z_k}(x, y, z) &= p_k(x) q_k(y) W(z|x, y) \\ &= p(x|k) q(y|k) W(z|x, y), \end{aligned}$$

and

$$p_k(x) := P(X_k = x) \quad \text{and} \quad q_k(y) := P(Y_k = y).$$

Note that the independence of X_k and Y_k follows from (14.1). If we put

$$P_{VXYZ}(k, x, y, z) := \frac{1}{k} p(x|k) q(y|k) W(z|x, y),$$

then

$$n \sum_{k=1}^n I(X_k Y_k \wedge Z_k) \frac{1}{n} = n \sum_{k=1}^n I(XY \wedge Z|V = k) P(V = k) = nI(XY \wedge Z|V).$$

Hence,

$$\begin{aligned} R_1 + R_2 &\leq \frac{nI(XY \wedge Z|V) + \log 2}{n(1 - \lambda)} + 2\Delta R \\ &= \frac{I(XY \wedge Z|V)}{1 - \lambda} + \frac{\log 2}{n(1 - \lambda)} + 2\Delta R. \end{aligned}$$

In a similar way, we obtain,

$$R_1 \leq \frac{I(\alpha \wedge \varphi_1(\mathbf{Z})) + \log 2}{n(1 - \lambda)} + \Delta R$$

and

$$R_2 \leq \frac{I(\beta \wedge \varphi_2(\mathbf{Z})) + \log 2}{n(1 - \lambda)} + \Delta R.$$

Now write, by the Data Processing Lemma,

$$I(\alpha \wedge \varphi_1(\mathbf{Z})) \leq I(\mathbf{X} \wedge \mathbf{Z}),$$

since from (14.1),

$$P(\alpha = i, \mathbf{X} = x, \mathbf{Z} = z) = \frac{1}{N} \mathbf{1}_{\mathbf{X}_i}(\mathbf{x}) \underbrace{\frac{1}{M} \sum_{j=1}^M W^n(\mathbf{z}|\mathbf{x}, \mathbf{y}_j)}_{=: \tilde{W}_n(\mathbf{z}|\mathbf{x})}.$$

So, $\alpha \rightarrow \mathbf{X} \rightarrow \mathbf{Z}$ is a Markov chain. Thus, Data Processing applies. Next,

$$\begin{aligned} I(\mathbf{X} \wedge \mathbf{Z}) &\leq I(\mathbf{X} \wedge \mathbf{Y}\mathbf{Z}) \\ &= \underbrace{I(\mathbf{X} \wedge \mathbf{Y})}_{=0 \text{ since } X \text{ and } Y \text{ are independent by (14.1)}} + I(\mathbf{X} \wedge \mathbf{Z}|\mathbf{Y}). \end{aligned}$$

Then

$$\begin{aligned}
 I(\mathbf{X} \wedge \mathbf{Z} | \mathbf{Y}) &= H(\mathbf{Z} | \mathbf{Y}) - H(\mathbf{Z} | \mathbf{XY}) \\
 &= H(\mathbf{Z} | \mathbf{Y}) - \sum_{k=1}^n H(Z_k | X_k Y_k) \\
 &\leq \sum_{k=1}^n H(Z_k | \mathbf{Y}) - \sum_{k=1}^n H(Z_k | X_k Y_k) \\
 &\leq \sum_{k=1}^n H(Z_k | Y_k) - \sum_{k=1}^n H(Z_k | X_k Y_k) \\
 &= \sum_{k=1}^n I(X_k \wedge Z_k | Y_k) \\
 &= n \sum_{k=1}^n I(X_k \wedge Z_k | Y, V = k) \cdot \frac{1}{n}.
 \end{aligned}$$

If

$$P_{VXYZ}(v, x, y, z) := \frac{1}{n} p_k(x) q_k(y) w(z | x, y),$$

then $I(\mathbf{X} \wedge \mathbf{Z} | \mathbf{Y}) \leq nI(X \wedge Z | YV)$, and

$$R_1 \leq \frac{I(X \wedge Z | YV)}{1 - \lambda} + \frac{\log 2}{n(1 - \lambda)} + \Delta R.$$

Similarly,

$$R_2 \leq \frac{I(Y \wedge Z | XV)}{1 - \lambda} + \frac{\log 2}{n(1 - \lambda)} + \Delta R.$$

From above, we repeat

$$R_1 + R_2 \leq \frac{I(XY \wedge Z | V)}{1 - \lambda} + \frac{\log 2}{n(1 - \lambda)} + 2\Delta R.$$

Definition 14.2. If V is a finite set, and if $\gamma, p(\cdot | v)$ and $q(\cdot | v)$, where $v \in V$, are probability mass functions on V, X , and Y , respectively, let

$$\begin{aligned}
 \mathcal{R}(\gamma, p, q) := \{ (R_1, R_2) : & 0 \leq R_1 \leq I(X \wedge Z | YV), \\
 & 0 \leq R_2 \leq I(Y \wedge Z | XV), \\
 & R_1 + R_2 \leq I(XY \wedge Z | V) \},
 \end{aligned}$$

where it is understood that

$$P_{VXYZ}(v, x, y, z) = \gamma(v)p(x|v)q(y|v)W(z|x, y).$$

Also, let

$$\mathcal{R}^* := \bigcup_{\text{all finite sets } \mathbf{V}} \left(\bigcup_{\gamma, p, q} \mathcal{R}(\gamma, p, q) \right)$$

We have shown that if (R_1, R_2) is achievable, then $(R_1, R_2) \in \mathcal{R}^*$; i.e., $C \subset \mathcal{R}^*$. The forward theorem, which we do not prove, shows that $\mathcal{R}^* \subset C$. We will then have $C = \mathcal{R}^*$. As we will see, $\mathcal{R}^*$ is convex and closed.

CHAPTER 15

Analysis of ML Decoding

15.1. First Steps

We showed in Section 5.4 that the probability of a decoding error using equally-likely messages is

$$\frac{1}{N} \sum_{i=1}^N W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i).$$

If we use random codewords $\mathbf{X}_1, \dots, \mathbf{X}_N$, then

$$\mathbb{E} \left[\frac{1}{N} \sum_{i=1}^N W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i) \right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E} [W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)].$$

Our approach is to obtain a bound on $\mathbb{E} [W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)]$ that does *not* depend on i . To this end, we first compute the conditional expectation

$$\mathbb{E} [W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}].$$

We will then have by the law of total probability for expectation that

$$\mathbb{E} [W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)] = \sum_{\mathbf{x}} \mathbb{E} [W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}] p_n(\mathbf{x}), \quad (15.1)$$

where p_n is the probability mass function of $\mathbf{X}_i$, which we will take to be the same for all i .

15.2. ML Decoder Analysis

Since the ML decoder satisfies (7.7),

$$W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{x}_i) \leq W_n \left(\bigcup_{j \neq i} B_j \middle| \mathbf{x}_i \right) = \sum_{\mathbf{y}} \mathbf{1}_{[\bigcup_{j \neq i} B_j]}(\mathbf{y}) W_n(\mathbf{y} | \mathbf{x}_i).$$

If we now replace $\mathbf{x}_1, \dots, \mathbf{x}_N$ by random codewords $\mathbf{X}_1, \dots, \mathbf{X}_N$ and take conditional expectations, we find that

$$\mathbb{E} [W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}] \leq \sum_{\mathbf{y}} \mathbb{E} [\mathbf{1}_{[\bigcup_{j \neq i} B_j]}(\mathbf{y}) W_n(\mathbf{y} | \mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}]$$

$$\begin{aligned}
&= \sum_{\mathbf{y}} \mathbb{E} \left[\mathbf{1}_{[\cup_{j \neq i} B_j]}(\mathbf{y}) W_n(\mathbf{y}|\mathbf{x}_i) \middle| \mathbf{X}_i = \mathbf{x} \right] \\
&= \sum_{\mathbf{y}} \mathbb{E} \left[\mathbf{1}_{[\cup_{j \neq i} B_j]}(\mathbf{y}) \middle| \mathbf{X}_i = \mathbf{x} \right] W_n(\mathbf{y}|\mathbf{x}_i), \quad (15.2)
\end{aligned}$$

where now ϕ and the B_j are random. In particular,

$$B_j = \left\{ \mathbf{y} : W_n(\mathbf{y}|\mathbf{X}_j) \geq W_n(\mathbf{y}|\mathbf{X}_l) \text{ for all } l \right\}.$$

Note also that

$$B_j \subset \left\{ \mathbf{y} : W_n(\mathbf{y}|\mathbf{X}_j) \geq W_n(\mathbf{y}|\mathbf{X}_i) \right\} =: \widehat{B}_j. \quad (15.3)$$

We now follow [6, p. 136] and write, for any $0 < \rho \leq 1$,

$$\begin{aligned}
\mathbb{E} \left[\mathbf{1}_{[\cup_{j \neq i} B_j]}(\mathbf{y}) \middle| \mathbf{X}_i = \mathbf{x} \right] &= \mathbb{E} \left[\mathbf{1}_{[\cup_{j \neq i} B_j]}(\mathbf{y})^\rho \middle| \mathbf{X}_i = \mathbf{x} \right], \quad \text{since } 0 = 0^\rho \text{ and } 1 = 1^\rho, \\
&\leq \left(\mathbb{E} \left[\mathbf{1}_{[\cup_{j \neq i} B_j]}(\mathbf{y}) \middle| \mathbf{X}_i = \mathbf{x} \right] \right)^\rho, \quad \text{by Jensen's inequality,} \\
&\leq \left(\sum_{j \neq i} \mathbb{E} [\mathbf{1}_{B_j}(\mathbf{y}) | \mathbf{X}_i = \mathbf{x}] \right)^\rho, \quad \text{since } (\cdot)^\rho \text{ is increasing,} \\
&\leq \left(\sum_{j \neq i} \mathbb{E} [\mathbf{1}_{\widehat{B}_j}(\mathbf{y}) | \mathbf{X}_i = \mathbf{x}] \right)^\rho, \quad \text{by (15.3).}
\end{aligned}$$

We can write this last conditional expectation as a conditional probability. Next, let $s > 0$. Assuming that $\mathbf{X}_i$ and $\mathbf{X}_j$ are independent, we can write

$$\begin{aligned}
\mathbb{P} \left(W_n(\mathbf{y}|\mathbf{X}_j) \geq W_n(\mathbf{y}|\mathbf{X}_i) \middle| \mathbf{X}_i = \mathbf{x} \right) &= \mathbb{P} \left(W_n(\mathbf{y}|\mathbf{X}_j) \geq W_n(\mathbf{y}|\mathbf{x}) \middle| \mathbf{X}_i = \mathbf{x} \right) \\
&= \mathbb{P} \left(W_n(\mathbf{y}|\mathbf{X}_j) \geq W_n(\mathbf{y}|\mathbf{x}) \right) \\
&= \sum_{\mathbf{x}'} p_n(\mathbf{x}') \mathbf{1}_{[W_n(\mathbf{y}|\mathbf{x}), \infty)}(W_n(\mathbf{y}|\mathbf{x}')) \\
&\leq \sum_{\mathbf{x}'} p_n(\mathbf{x}') \left[\frac{W_n(\mathbf{y}|\mathbf{x}')}{W_n(\mathbf{y}|\mathbf{x})} \right]^s.
\end{aligned}$$

We can now write

$$\begin{aligned}
\mathbb{E} \left[\mathbf{1}_{[\cup_{j \neq i} B_j]}(\mathbf{y}) \middle| \mathbf{X}_i = \mathbf{x} \right] &\leq \left[\sum_{j \neq i} \sum_{\mathbf{x}'} p_n(\mathbf{x}') \left[\frac{W_n(\mathbf{y}|\mathbf{x}')}{W_n(\mathbf{y}|\mathbf{x})} \right]^s \right]^\rho \\
&= (N-1)^\rho W_n(\mathbf{y}|\mathbf{x})^{-\rho s} \left[\sum_{\mathbf{x}'} p_n(\mathbf{x}') W_n(\mathbf{y}|\mathbf{x}')^s \right]^\rho.
\end{aligned}$$

Putting this together with (15.2) and (7.9), we find that

$$\begin{aligned} & \mathbb{E}[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)] \\ & \leq (N-1)^\rho \sum_{\mathbf{y}} \left[\sum_{\mathbf{x}} p_n(\mathbf{x}) W_n(\mathbf{y} | \mathbf{x})^{1-\rho s} \right] \left[\sum_{\mathbf{x}'} p_n(\mathbf{x}') W_n(\mathbf{y} | \mathbf{x}')^s \right]^\rho. \end{aligned}$$

Specializing to $s = 1/(1+\rho)$, we obtain

$$\mathbb{E}[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)] \leq (N-1)^\rho \sum_{\mathbf{y}} \left[\sum_{\mathbf{x}} p_n(\mathbf{x}) W_n(\mathbf{y} | \mathbf{x})^{1/(1+\rho)} \right]^{1+\rho}. \quad (15.4)$$

Note that since the left-hand side is at most one, (15.4) is a valid bound even for $\rho = 0$. In the case of a discrete memoryless channel, $W_n = W^n$. If we also take $p_n = p^n$, then the above sum over $\mathbf{x}$ becomes

$$\prod_{k=1}^n \left(\sum_x p(x) W(y_k | x)^{1/(1+\rho)} \right).$$

Denoting this product by $\lambda(y_1) \cdots \lambda(y_n)$, (15.4) becomes

$$\mathbb{E}[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)] \leq (N-1)^\rho \sum_{\mathbf{y}} \lambda(y_1)^{1+\rho} \cdots \lambda(y_n)^{1+\rho}.$$

Hence, for a DMC with $p_n = p^n$, we have

$$\mathbb{E}[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)] \leq (N-1)^\rho \left\{ \sum_{\mathbf{y}} \left[\sum_x p(x) W(y | x)^{1/(1+\rho)} \right]^{1+\rho} \right\}^n.$$

If $N = \lfloor \exp(nR) \rfloor \leq \exp(nR)$, we obtain the further upper bound

$$\mathbb{E}[W_n(\{\mathbf{y} : \varphi(\mathbf{y}) \neq i\} | \mathbf{X}_i)] \leq \exp\{-n[E_o(\rho, p) - \rho R]\},$$

where

$$E_o(\rho, p) := -\log \sum_{\mathbf{y}} \left[\sum_x p(x) W(y | x)^{1/(1+\rho)} \right]^{1+\rho}.$$

If there is some $0 \leq \rho \leq 1$ for which $E_o(\rho, p) - \rho R > 0$, then the error probability decays exponentially fast. This is indeed the case if $0 \leq R < I(p \times W)$. The outline of the argument is as follows; proofs of claims are established in the Notes.¹ Let $f(\rho) = E_o(\rho, p) - \rho R$. Then f is continuously differentiable with $f(0) = 0$ and $f'(0) = I(p \times W) - R > 0$. Since $f(0) = 0$, we can write $f(\rho) = \int_0^\rho f'(t) dt$. Since

$f'(0) > 0$ and f' is continuous, there is a $\rho^* > 0$ with $f'(\rho) > f'(0)/2$ for $0 \leq \rho \leq \rho^*$. Thus,

$$f(\rho^*) = \int_0^{\rho^*} f'(\rho) d\rho \geq \rho^* f'(0)/2 > 0.$$

The foregoing analysis shows that the error probability decays exponentially fast. For fixed p , the best bound is $\exp[-nE_r(R, p)]$, where

$$E_r(R, p) := \sup_{0 \leq \rho \leq 1} [E_o(\rho, p) - \rho R]. \quad (15.5)$$

To obtain the best bound overall, we optimize over p as well. Hence, the best bound is $\exp[-nE_r(R)]$, where

$$E_r(R) := \sup_p E_r(R, p)$$

is called the **random coding exponent**.

Notes

Note 15.1. Proofs of Claims in Section 15.2. We begin by studying $E_o(\rho, p)$. If we put

$$(pW^\theta)(y) := \sum_x p(x)W(y|x)^\theta, \quad \theta > 0, \quad (15.6)$$

and

$$\gamma_p(\rho) := \sum_y (pW^{1/(1+\rho)})(y)^{1+\rho}, \quad \rho \geq 0, \quad (15.7)$$

then

$$E_o(\rho, p) = -\log \gamma_p(\rho).$$

Remark 15.1. The expression $W(y|x)^\theta$ is understood as $e^{\theta \ln W(y|x)}$ if $W(y|x) > 0$ and zero otherwise. Hence, $W(y|x)^\theta$ is an infinitely differentiable function of θ for each x and y .

When $\theta = 1$, we write $(pW)(y)$ instead of $(pW^1)(y)$. For a given y , if $(pW^\theta)(y) = 0$ for some $\theta > 0$, then each term in (15.6) must be zero. This happens if and only if $W(y|x) = 0$ for all x with $p(x) > 0$. In this case, $(pW^\theta)(y) = 0$ for all $\theta > 0$, including $\theta = 1$.

Remark 15.2. By the preceding remark, each term in (15.7) is either the zero function or a positive, infinitely differentiable function of ρ . We now show that at least one term in (15.7) is positive. Suppose otherwise that $\gamma_p(\rho) = 0$ for some $\rho \geq 0$. Then every term in (15.7) is zero, which implies that for every y , $(pW^{1/(1+\rho)})(y)^{1+\rho} =$

0, which implies $(pW^{1/(1+\rho)})(y) = 0$. By the preceding remark, this last equality must also hold for $1/(1+\rho) = 1$; thus, $(pW)(y) = 0$. We would then conclude that

$$0 = \sum_y (pW)(y) = \sum_y \sum_x p(x)W(y|x) = 1,$$

which is a contradiction. Hence, $\gamma_p(\rho) > 0$.

Since $1/(1+\rho) = 1$ corresponds to $\rho = 0$, the foregoing calculation shows that $\gamma_p(0) = 1$. Since $\gamma_p(\rho) > 0$, $E_o(\rho, p)$ is finite for all $\rho \geq 0$. Furthermore, since $\gamma_p(0) = 1$, $E_o(0, p) = 0$.

Remark 15.3. We now analyze the derivative

$$\frac{\partial}{\partial \rho} E_o(\rho, p) = \frac{-\gamma'_p(\rho) \log e}{\gamma_p(\rho)}.$$

We showed in Remark 15.2 that $\gamma_p(0) = 1$. Hence, the above derivative at $\rho = 0$ is just $-\gamma'_p(0) \log e$. To compute $\gamma'_p(\rho)$, use the formula

$$(pW^{1/(1+\rho)})(y)^{1+\rho} = e^{(1+\rho) \ln(pW^{1/(1+\rho)})(y)}$$

to show that

$$\begin{aligned} \gamma'_p(\rho) &= \sum_y (pW^{1/(1+\rho)})(y)^{1+\rho} \cdot \left[\ln(pW^{1/(1+\rho)})(y) \right. \\ &\quad \left. + \frac{(1+\rho)}{(pW^{1/(1+\rho)})(y)} \cdot \frac{\partial}{\partial \rho} (pW^{1/(1+\rho)})(y) \right], \end{aligned}$$

where

$$\frac{\partial}{\partial \rho} (pW^{1/(1+\rho)})(y) = \sum_x p(x)W(y|x)^{1/(1+\rho)} \cdot \frac{-\ln W(y|x)}{(1+\rho)^2}.$$

In particular,

$$\gamma'_p(0) = \sum_y (pW)(y) \ln(pW)(y) - \sum_x \sum_y p(x)W(y|x) \ln W(y|x)$$

and so $\gamma'_p(0) \log e = -H(Y) + H(Y|X)$. Hence,

$$\left. \frac{\partial}{\partial \rho} E_o(\rho, p) \right|_{\rho=0} = I(p \times W).$$

Additional Results about $E_o(\rho, p)$

Remark 15.4. Lyapunov's inequality implies $(pW^{1/(1+\rho)})(y)^{1+\rho}$ is a nonincreasing function of ρ . Hence, $\gamma_p(\rho)$ is also nonincreasing, with $\gamma_p(\rho) \leq \gamma_p(0) = 1$. Since $\gamma_p(\rho)$ is nonincreasing, $E_o(\rho, p)$ is nondecreasing, which implies

$$\frac{\partial}{\partial \rho} E_o(\rho, p) \geq 0.$$

In fact, $\gamma_p(\rho)$ is strictly decreasing (and thus $E_o(\rho, p)$ is strictly increasing). This will follow if we can show that for at least one y , $(pW^{1/(1+\rho)})(y)^{1+\rho}$ is strictly decreasing. Suppose otherwise that for each y , $(pW^{1/(1+\rho)})(y)^{1+\rho}$ is not strictly decreasing. Then for each y , Lyapunov's inequality implies that $W(y|x) = c_y$ for all x with $p(x) > 0$ and some constant c_y . It is easy to check that this implies $I(p \times W) = 0$, which contradicts the assumption that $I(p \times W) > 0$.

Remark 15.5. We show that $E_o(\rho, p)$ is a concave function of ρ . Fix nonnegative $\rho_1 \neq \rho_2$, and for $0 < \lambda < 1$, put $\rho = \lambda\rho_1 + (1 - \lambda)\rho_2$. Then

$$1 + \rho = \lambda(1 + \rho_1) + (1 - \lambda)(1 + \rho_2).$$

Put

$$\mu = \frac{(1 + \rho_1)\lambda}{1 + \rho} \quad \text{so that} \quad 1 - \mu = \frac{(1 + \rho_2)(1 - \lambda)}{1 + \rho}. \quad (15.8)$$

We can apply Hölder's inequality to write

$$\begin{aligned} (pW^{1/(1+\rho)})(y) &= \sum_x p(x)W(y|x)^{1/(1+\rho)} \\ &= \sum_x p(x) [W(y|x)^{\mu/(1+\rho_1)} \cdot W(y|x)^{(1-\mu)/(1+\rho_2)}] \\ &\leq \left[\sum_x p(x)W(y|x)^{1/(1+\rho_1)} \right]^\mu \left[\sum_x p(x)W(y|x)^{1/(1+\rho_2)} \right]^{1-\mu} \\ &= (pW^{1/(1+\rho_1)})(y)^\mu \cdot (pW^{1/(1+\rho_2)})(y)^{1-\mu}. \end{aligned} \quad (15.9)$$

Raising both sides to the $1 + \rho$ power and using (15.8), we have

$$(pW^{1/(1+\rho)})(y)^{1+\rho} \leq (pW^{1/(1+\rho_1)})(y)^{(1+\rho_1)\lambda} \cdot (pW^{1/(1+\rho_2)})(y)^{(1+\rho_2)(1-\lambda)}.$$

Summing over y yields

$$\gamma_p(\rho) \leq \sum_y (pW^{1/(1+\rho_1)})(y)^{(1+\rho_1)\lambda} \cdot (pW^{1/(1+\rho_2)})(y)^{(1+\rho_2)(1-\lambda)}.$$

Applying Hölder's inequality again, we find that

$$\gamma_p(\rho) \leq \left[\sum_y (pW^{1/(1+\rho_1)})(y)^{(1+\rho_1)} \right]^\lambda \left[\sum_y (pW^{1/(1+\rho_2)})(y)^{(1+\rho_2)} \right]^{1-\lambda}. \quad (15.10)$$

Taking logarithms and multiplying by negative one yields

$$E_o(\rho, p) \geq \lambda E_o(\rho_1, p) + (1 - \lambda) E_o(\rho_2, p).$$

Remark 15.6. Since $E_o(\rho, p)$ is concave, its derivative is nonincreasing. Again consider the function $f(\rho) = E_o(\rho, p) - \rho R$. Its derivative is also nonincreasing. Write $f(\rho) = \int_0^\rho f'(t) dt$. If f' is nonnegative on $[0, 1]$, then $\rho = 1$ maximizes f . Note that $f'(1) \geq 0$ if and only if

$$R \leq \left. \frac{\partial}{\partial \rho} E_o(\rho, p) \right|_{\rho=1} =: R_{\text{cr}}(p),$$

which is called the **critical rate** for p . For $R_{\text{cr}}(p) < R < I(p \times W)$, we can use the intermediate-value property of derivatives [19, p. 108, Th. 5.12] to conclude that there is a ρ_0 with $f'(\rho_0) = 0$; i.e., ρ_0 maximizes f .

Remark 15.7. Strict concavity of $E_o(\rho, p)$ fails if and only if for some $\rho_1 \neq \rho_2$ and some $0 < \lambda < 1$ we have

$$\boxed{\text{equality in (15.9) for every } y} \quad \text{and} \quad \boxed{\text{equality in (15.10).}}$$

Now, for fixed y , (15.9) holds with equality if and only if there are nonnegative constants a_y and b_y , not both zero, such that

$$a_y W(y|x)^{1/(1+\rho_1)} = b_y W(y|x)^{1/(1+\rho_2)}, \quad \text{for } x \text{ with } p(x) > 0.$$

Hence, if $p(x)W(y|x) > 0$, we can solve for $W(y|x) = c(y)$ for some function $c(y) > 0$. Next, equality holds in (15.10) if and only if there are nonnegative constants α and β , not both zero, such that

$$\alpha (pW^{1/(1+\rho_1)})(y)^{(1+\rho_1)} = \beta (pW^{1/(1+\rho_2)})(y)^{(1+\rho_2)}, \quad \text{for all } y,$$

or more explicitly,

$$\alpha \left[\sum_x p(x) W(y|x)^{1/(1+\rho_1)} \right]^{1+\rho_1} = \beta \left[\sum_x p(x) W(y|x)^{1/(1+\rho_2)} \right]^{1+\rho_2}.$$

Using $p(x)W(y|x) = p(x)c(y)$, the foregoing simplifies to

$$\alpha c(y) \left[\sum_{x:W(y|x)>0} p(x) \right]^{1+\rho_1} = \beta c(y) \left[\sum_{x:W(y|x)>0} p(x) \right]^{1+\rho_2}.$$

Since $c(y) > 0$, it can be canceled, and thus the quantity in brackets is a constant that does not depend on y (or x). We can exploit this observation to write

$$\begin{aligned} I(p \times W) &= \sum_x \sum_y p(x)W(y|x) \log \frac{W(y|x)}{(pW)(y)} \\ &= \sum_{x,y:p(x)W(y|x)>0} p(x)W(y|x) \log \frac{W(y|x)}{(pW)(y)} \\ &= \sum_{x,y:p(x)W(y|x)>0} p(x)W(y|x) \log \frac{c(y)}{c(y) \sum_{x':W(y|x')>0} p(x')} \\ &= \sum_{x,y:p(x)W(y|x)>0} p(x)W(y|x) \log \frac{1}{\sum_{x':W(y|x')>0} p(x')} \\ &= \log \frac{1}{\sum_{x':W(y|x')>0} p(x')}, \end{aligned}$$

where the last step follows because the log factor does not depend on x or y .

APPENDIX A

Continuity: Pointwise vs. Uniform

A.1. Metrics

Let X be a nonempty set, and let $\rho: X \times X \rightarrow [0, \infty)$ have the following properties for any points x, y , and z in X :

- (i) $\rho(x, y) \geq 0$, and $\rho(x, y) = 0$ if and only if $x = y$.
- (ii) $\rho(x, y) = \rho(y, x)$.
- (iii) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$. This is known as the **triangle inequality**.

The function ρ is called a **metric**, and the pair (X, ρ) is called a **metric space**. When ρ is understood, we just say that X is a metric space.

Example A.1. On the real numbers $\mathbb{R}$, if we put $\rho(x, y) := |x - y|$, where $|\cdot|$ denotes the usual absolute value; i.e.,

$$|x| := \begin{cases} x, & x \geq 0, \\ -x, & x < 0, \end{cases}$$

then ρ is a metric. More generally, for $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, if we put $\|\mathbf{x}\| := \sqrt{\sum_{k=1}^n x_k^2}$, then $\rho(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|$ is a metric.

Example A.2 (The Trivial Metric). On any nonempty set X , the **trivial metric** is

$$\rho(x, y) := \begin{cases} 1, & x \neq y, \\ 0, & x = y. \end{cases}$$

A.2. Pointwise Continuity

Consider a function $f: X \rightarrow Y$, where X is equipped with metric ρ and Y is equipped with metric m . We say that f is **continuous at the point** $x_0 \in X$ if

$$\begin{aligned} \forall \varepsilon > 0, \exists \delta > 0, \forall x \in X, \\ \rho(x, x_0) < \delta \Rightarrow m(f(x), f(x_0)) < \varepsilon. \end{aligned} \tag{A.1}$$

Letting B_ρ and B_m denote open balls in X and Y , respectively,^a we can rewrite the above implication as

$$x \in B_\rho(x_0, \delta) \Rightarrow f(x) \in B_m(f(x_0), \varepsilon).$$

In other words, $f(x)$ is close to $f(x_0)$ whenever x is close enough to x_0 . “Close enough” is determined by δ , which depends on the point x_0 .

If a function is continuous at every point in a subset of the whole space, then we say that the function is **continuous on the subset**.

Example A.3. If we take ρ to be the trivial metric on X , then every function is continuous. To see this, let $\varepsilon > 0$ be given, and fix any $0 < \delta < 1$. Then

$$\rho(x, x_0) < \delta \Rightarrow \rho(x, x_0) = 0 \Rightarrow x = x_0 \Rightarrow m(f(x), f(x_0)) = m(f(x), f(x)) = 0 < \varepsilon.$$

Notice that this result does not depend on what metric is used on Y .

A.3. Uniform Continuity

Sometimes we have a stronger kind of continuity called **uniform continuity**. We say that f is **uniformly continuous** if $f(x)$ and $f(x_0)$ are close whenever x and x_0 are close enough, and “close enough” does *not* depend on either of the two points. More precisely, we say f is **uniformly continuous** if

$$\begin{aligned} \forall \varepsilon > 0, \exists \delta > 0, \forall x, x_0 \in X, \\ \rho(x, x_0) < \delta \Rightarrow m(f(x), f(x_0)) < \varepsilon. \end{aligned}$$

Notice that both x and x_0 appear *after* the existence of δ is mentioned. This is in contrast with (A.1), where x_0 is fixed *before* the existence of δ is mentioned.

Sometimes a pointwise-continuous function is uniformly continuous:

Theorem A.4. *A continuous function on a closed and bounded subset of $\mathbb{R}^n$ is uniformly continuous.*

Proof. Omitted. □

^a Here $B_\rho(x_0, \delta) := \{x \in X : \rho(x, x_0) < \delta\}$ and $B_m(y_0, \varepsilon) := \{y \in Y : m(y, y_0) < \varepsilon\}$.

Example A.5. Consider the function $f: [0, 1] \rightarrow \mathbb{R}$ defined by $f(t) := t \ln(1/t)$ for $t \in (0, 1]$ and $f(0) := 0$. It is easy to show that f is continuous for $t \in (0, 1]$. Using l'Hôpital's rule, f is continuous at $t = 0$ as well. Since f is continuous on $[0, 1]$, by the theorem, f is uniformly continuous on $[0, 1]$.

Bibliography

- [1] M. Baer. (2008 July 22). “A simple countable infinite-entropy distribution,” [Online]. Available: <https://www.mbbaer.com/Hinf.pdf>
- [2] B. R. Barmish, Private Communication, Mar. 2010.
- [3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley, 2006. [Online]. Available: UW–Library.
- [4] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981. 2nd Ed. New York: Cambridge Univ. Press, 2011. [Online]. Available: UW–Library.
- [5] L. Finesso and P. Spreij, “Nonnegative matrix factorization and I-divergence alternating minimization,” *Linear Algebra and its Appl.*, vol. 416, pp. 270–287, 2006.
- [6] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
- [7] J. A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*. Cambridge, UK: Cambridge Univ. Press, 2006.
- [8] J. A. Gubner. “ECE 735 Course Notes.” [Online]. Available: <http://gubner.ece.wisc.edu/ece735/ece735notesDir/>
- [9] D. A. Huffman, “A method for the construction of minimum-redundancy codes,” *Proc. IRE*, vol. 40, no. 9, pp. 1098–1101, Sep. 1952. [Online]. Available: IEEE Xplor through UW–Library.
- [10] J. C. Kieffer, “Exponential rate of convergence for Lloyd’s method I,” *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 205–210, Mar. 1982. [Online]. Available: IEEE Xplor through UW–Library.
- [11] J. C. Kieffer, “Uniqueness of locally optimal quantizer for log-concave density and convex error weighting function,” *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 42–47, Jan. 1983. [Online]. Available: IEEE Xplor through UW–Library.
- [12] H. Komiya, “Elementary proof for Sion’s minimax theorem,” *Kodai Math. J.*, vol. 11, no. 1, pp. 5–7, 1988. [Online]. Available: Project Euclid.
- [13] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” *Adv. Neural Inform. Processing Syst. (NIPS)*, 2000.
- [14] C.-J. Lin, “On the convergence of multiplicative update algorithms for nonnegative matrix factorization,” *IEEE Trans. Neural Netw.*, vol. 18, no. 6, pp. 1589–1596, Nov. 2007.
- [15] K. Liu, V. Raghavan, and A. M. Sayeed, “Capacity scaling and spectral efficiency in wide-band correlated MIMO channels,” *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2504–2526, Oct. 2003.
- [16] S. P. Lloyd, “Least squares quantization in PCM,” unpublished Bell Laboratories Memorandum, July 31, 1957; also *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 129–137, Mar. 1982. [Online]. Available: IEEE Xplor through UW–Library.
- [17] J. Max, “Quantizing for minimum distortion,” *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960. [Online]. Available: IEEE Xplor through UW–Library.
- [18] R. J. McEliece, *The Theory of Information and Coding*. Reading, MA: Addison-Wesley, 1977.
- [19] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976.
- [20] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, July, Oct. 1948. [Online]. Available: BSTJ Archive.
- [21] C. E. Shannon, “Communication in the presence of noise,” *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949. Reprinted in *Proc. IEEE*, vol. 86, no. 2, pp. 447–457, Feb. 1998. [Online]. Available: IEEE Xplor through UW–Library.
- [22] D. Slepian and H. O. Pollack, “Prolate spheroidal wave functions, Fourier analysis and uncertainty — I,” *Bell Syst. Tech. J.*, vol. 40, pp. 43–64, 1961. [Online]. Available: BSTJ Archive.
- [23] D. Slepian, “On bandwidth,” *Proc. IEEE*, vol. 64, no. 3, pp. 292–300, Mar. 1976. [Online]. Available: IEEE Xplor through UW–Library.

- [24] D. Slepian, "Some comments on Fourier analysis, uncertainty and modeling," *SIAM Rev.*, vol. 25, no. 3, pp. 379–393, July 1983.
- [25] A. V. Trushkin, "Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 187–198, Mar. 1982. [Online]. Available: IEEE *Xplor* through UW–Library.
- [26] A. V. Trushkin, "Monotony of Lloyd's method II for log-concave density and convex error weighting function," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 2, pp. 380–383, Mar. 1984. [Online]. Available: IEEE *Xplor* through UW–Library.
- [27] A. V. Trushkin, "On the design of an optimal quantizer," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1180–1194, Mar. 1993. [Online]. Available: IEEE *Xplor* through UW–Library.
- [28] J. van Tiel, *Convex Analysis: An Introductory Text*. Chichester: Wiley, 1984.
- [29] S. Verdú, "On channel capacity per unit cost," *IEEE Trans. Inform. Theory*, vol. 36, no. 5, pp. 1019–1030, Sept. 1990. [Online]. Available: IEEE *Xplor* through UW–Library.
- [30] S. Verdú, "Spectral efficiency in the wideband regime," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1319–1343, June 2002. [Online]. Available: IEEE *Xplor* through UW–Library.

Index

- 20 questions, 34
- achievable rate, 51
 - for a channel code, 51, 52
 - under a codeword constraint, 85
 - for a MAC, 136
 - for a source code, 37
 - λ -achievable, 64
- achievable rate-distortion pair, 111
- additive white Gaussian noise (AWGN), 48
- additive white Gaussian noise (AWGN) channel, 104
- AEP, *see* asymptotic equipartition property
- almost surely, 128
- alphabet, 11
- a.s., *see* almost surely
- asymptotic equipartition property (AEP), 40, 73
 - joint, 74
 - for rate distortion, 118
- average probability of error criterion, 51
- AWGN, *see* additive white Gaussian noise
- bandlimited waveform, 6
- bandwidth, 6
- BEC, *see* binary erasure channel
- binary entropy function, 32
- binary erasure channel (BEC), 82
- binary symmetric channel (BSC), 46, 48
 - and repetition code, 50
 - capacity, 81
- bit, 20
- Bolzano–Weierstrass theorem, 44
- BSC, *see* binary symmetric channel
- capacity, 3, 53
 - of BEC, 83
 - of BSC, 81
 - region, 53
 - for a MAC, 136
- capacity region
 - under a codeword constraint, 85
- capacity-cost function, 86
- cardinality, 12
- cascade of DMCs, 84
- chain rule
 - for entropy, 21
 - for mutual information, 66
- channel, 1, 45
 - coding system, 50
 - decoder, 2, 3
 - encoder, 2, 3
 - input symbols, 45
 - multiple access (MAC), 135
 - output symbols, 45
- channel noise power, 102
- codebook, 12, 49
- codeword, 11, 48
 - alphabet, 11
- codeword constraint
 - inactive, 85
- coding theorem
 - DMC forward part, 76
 - with constraints, 88
 - DMC weak converse, 63
 - with constraints, 87
 - for channels, 3
 - for fixed-length source codes, 3, 37, 129
 - for variable-length source codes, 27, 127
 - Gaussian Channel
 - discrete-time, 98
- communication system, 1
 - digital, 1
- concatenation, 13
- concave function, 4
- conditional entropy, 21
- continuity
 - at a point, 149
 - of entropy, 44
 - on a set, 150
 - uniform, 150
- converse theorem
 - DMC
 - strong, 65
 - weak, 63
- convex
 - function, 4
 - set, 41
- critical rate, 147
- crossover probability, 46
- cutoff frequency, 102
- data processing
 - inequality, 61, 67
 - lemma, 61
- decoder, 1

- for a channel code, 49
 - for a MAC, 135
 - modified, 54
- decoding sets
 - for a channel code, 49
- demodulator, 1
- destination, 1
- differential entropy, 93
 - pdf with infinite differential entropy, 107
 - undefined, 94
- digital communication system, 1
- Dirac delta, 102
- discrete memoryless channel (DMC), 46, 48
 - cascade, 84
 - forward theorem, 76
 - under a codeword constraint, 88
 - parallel, 84
 - strong converse theorem, 65
 - weak converse theorem, 63
 - under a codeword constraint, 87
- discrete memoryless source (DMS), 22
- distance, 43
- distortion-rate function, 111
- distortion-typical sets, 117
- divergence, 32, 59
 - and mutual information, 59
 - for probability densities, 95
 - for probability mass functions, 26
- DMC, *see* discrete memoryless channel
- DMS, *see* discrete memoryless source
- encoder, 1
 - for a channel code, 48
 - for a MAC, 135
 - modified, 54
- entropy, 3, 20
 - as a measure of uncertainty, 20
 - chain rule, 21
 - concavity, 43
 - conditional, 21
 - continuity, 44
 - differential, *see* differential entropy
 - joint, 21
 - of a DMS, 22
 - of the geometric distribution, 33
 - pmf with infinite entropy, 33
 - rate, 127
- entropy rate, 127
- erasure channel, *see* binary erasure channel
- ergodic process, 128
- expected length, 12
 - extension of a code, 13
- Fano's inequality, 62
- feedback, 46
- finite geometric series, 19
- forward theorem, 53
 - DMC, 65, 76
 - under a codeword constraint, 88
- Fourier transform, 6
- Gaussian channel
 - continuous time
 - bandlimited, 105
 - discrete-time, 47, 98
 - parallel, 99
- Gaussian density, 47
- Gaussian random process, 104
- geometric distribution
 - entropy, 33
 - mean, 33
- geometric series
 - finite, 19
- Hamming distance, 110
 - normalized, 110
- idempotent, 101
- indicator function, 7
- informational divergence, *see* divergence
- inner product, 101
- instantaneous code, 14
- Jensen's inequality, 4, 108
- joint AEP, 74
 - for rate distortion, 118
- joint entropy, 21
- joint typicality, 73
 - for rate distortion, 117
- Kraft inequality
 - prefix-free codes, 16
- Kullback–Leibler informational divergence, *see* divergence
- ℓ^1 , 43
- L^2 , 101
- λ -achievability, 64
- Lempel–Ziv
 - algorithm, 133
 - parsing, 132
- l'Hôpital's rule, 5, 106

- log inequality, 22, 31
- log-sum inequality, 31, 32
- MAC, *see* multiple-access channel
- majority rule decoding, 50
- Markov chain, 59
- mathematical induction, 5
- Matlab commands
 - . *, 32
 - find, 34
 - linspace, 32
 - log2, 34
 - nargin, 34
 - plot, 32
 - prod, 34
 - repmat, 56
 - reshape, 34
 - size, 34
 - varargin, 34
- Matlab M-files
 - entropy2, 34
 - lincmb, 9
 - relent, 35
 - smpthm, 9
 - sscript, 9
- matrix factorization, 32
- maximum probability of error criterion, 52
- maximum-likelihood (ML) decoding, 69
- mean-value theorem, 22
- memoryless channel, 46
- messages, 48
- metric, 110, 149
 - trivial, 149
- metric space, 149
- ML, *see* maximum-likelihood
- modified decoder, 54
- modified encoder, 54
- modulator, 1
- multiple-access channel (MAC), 135
- mutual information
 - chain rule, 66
 - for discrete random variables, 58
 - for jointly continuous random variables, 96
 - in terms of divergence, 59
 - mixed random variables, 97
- nat, 20, 49
- nonnegative matrix factorization, 32
- normal density, *see* Gaussian density
- normal equations, 8
- normalized Hamming distance, 110, 111
- normalized squared error, 111
- one's complement, 12
- one-to-one, 13
- orthogonality principle, 8
- orthogonality property, 101
- orthonormal basis, 101
- parallel DMCs, 84
- parallel Gaussian channel, 99
- partition, 7
- prefix-free code, 14
- probability mass function, 12
- projection, 101
- prolate spheroidal wave functions, 102
- quantization, 2, 110
- quantizer, 7
- random coding argument, 72
- random coding exponent, 144
- random process, 102
 - Gaussian, 104
 - stationary, 127
 - white noise, 102
- random variable, 12
- range, 12
- rate
 - of a channel code, 3, 49
 - of a source encoder, 2
- rate pair, 135
- rate-distortion, 3
- rate-distortion function, 111
- rate-distortion region, 111
- receiver, 1
- relative entropy, *see* divergence
- repetition code, 50
- reproduction alphabet, 110
- sampling frequency, 6
- sampling theorem, 2, 6
- self-punctuating code, 13
- Shannon code, 24
 - under mismatch, 26
- Shannon's channel coding theorem, 3
- Shannon's source coding theorem, 3, 37, 129
- Shannon–McMillan–Breiman theorem, 129
- signal subspace, 101
- sinc function, 6
- single-letter distortion measure, 111
- source

- alphabet, 110
- code, 12
- coding theorem
 - fixed-length codes, 3, 37, 129
 - variable-length codes, 27, 127
- decoder, 2, 3
- encoder, 2
- location, 1
- symbols, 12
- Source-Channel Separation theorem, 2, 4
- squared error, 111
 - normalized, 111
- stationary channel, 46
- stationary process, 127
- strong converse, 53
 - for the DMC, 58
- strong law of large numbers, 128
- symmetric channel, 80
 - weak, 81
- transition probability, 43, 45
- transmitter, 1
- triangle inequality, 43
 - for metric spaces, 149
- trivial metric, 110, 149
- twenty questions, *see* 20 questions
- typical sequence, 40
- typicality
 - decoding, 76
 - joint, 73
- uniform continuity, 150
- unique decodability, 13
- unit impulse, 102
- unit simplex, 42
- variable-length source coding theorem, 27
 - for a stationary source, 127
- variational distance, 43
- vector quantizer, 7
- Voronoi region, 7
- water-filling, 100
- weak converse, 53
 - for the DMC, 58, 63
 - under a codeword constraint, 87
- weak law of large numbers (WLLN), 38, 75, 89, 118
- weakly symmetric channel, 81
- white Gaussian noise, 104
- white noise, 102
- wideband limit, 106
- WLLN, *see* weak law of large numbers