Friedrich-Alexander Universität Erlangen-Nürnberg

Institute for Digital Communications

Tutorials on

Information Theory and Coding

A supplementary manuscript for the lecture

Information Theory and Coding

consistent with the textbook

Information Theory, Inference and Learning Algorithms by *David MacKay*

Current Edition: Summer 2022

Dr.-Ing. Ali Bereyhi

This manuscript shall be used only for *non-profitable* purposes

Preface

This manuscript provides exercises with comprehensive solutions to fundamental topics in information theory. It is hence suitable for a basic course on *Information Theory and Coding*. The goal is to help students having a better understanding of the information-theoretic concepts through exercises. The content of this manuscript have been collected and prepared through multiple years of teaching tutorials on the *Information Theory and Coding* course of Prof. Dr.-Ing. Ralf Müller at the Friedrich-Alexander University of Erlangen-Nuremberg. It is hence consistent with textbook of the course *Information Theory, Inference and Learning Algorithms* written by *David MacKay*. Some exercises are directly borrowed from the textbook, some are inspired by them, and some are designed completely independently.

I would like to kindly thank my colleague Sebastian Lotter for his useful comments on the technical contents of the manuscript, as well as Hassan Nazim Bicer and Arda Buglagil who helped typesetting this manuscript.

I hope that this writing helps the readers to have a better and deeper understanding of basic concepts in information theory. As this is an initial version of the manuscript, it contains several typos and writing mistakes. I would hence be grateful to receive any comments or feedback on the manuscript. You can always reach me by sending me an email at ali.bereyhi@fau.de or contacting me in person.

Ali Bereyhi
Winter 2021, Erlangen, Germany[1]

[1] Last revision on Summer 2022.

Contents

How to Use

As indicated in the preface, this manuscript is consistent with MacKay's book. Therefore, wherever we indicate *textbook*, we are referring to

> MacKay, David JC., *Information Theory, Inference and Learning Algorithms*,
> Cambridge University Press, 2003.

In each chapter, we first quickly go through the main concepts. The following section then gives you *only* the exercises and the one after gives you the exercises *with solutions*. Some of these exercises are taken directly from the textbook, some are partially related and some are designed independently. In case that the exercise is related to an exercise in the textbook, an explanation is given in this regard. Such an explanation is marked by ♣.

To study the exercises, I would suggest to do the following:

1. Try to solve each exercise after the corresponding tutorial session without looking at the solutions.

2. If you have done your best with the exercises; then, go through the solutions. I would *strongly* suggest to read all solutions, *even those you have solved correctly.* This would help you understanding all details.

At the end of each chapter, you are given by homework. The homework contains new exercise out of the textbook and some important exercises form the textbook. I would suggest to examine yourself at the end of each chapter by going through the homework.

The last two chapters contain sample exams given in the previous semesters at FAU. In Chapter 10, recent exams are given with solutions. I would suggest that you go through respective problems in this chapter, after you finish with studying each topic. The final chapter then gives you some older exams without solutions. These exams could be used for self-test to check whether you are ready for the exam or not.

The notation of the manuscript is consistent with that of the textbook. Namely, uppercase letters denote random variables and lowercase letters are deterministic outcomes. $\mathcal{E}\left[\cdot\right]$ denotes mathematical expectation. $P_X\left(x\right)$, $f_X\left(x\right)$, and $F_X\left(x\right)$ denote the probability mass function, the probability density function and the cumulative distribution function of X, respectively. Other notations are defined within the text.

Chapter 1

Preliminaries

This chapter tries to address two main aspects

1. It tries to give some motivation to information theory by clarifying the importance of information theory and its key applications.
2. It goes through the basic definitions in information theory.

The contents of this chapter follows the first two chapters of the textbook.

1.1 Brief Review of Main Concepts

There are few key concepts in this chapters

- Stirling's approximation states that

$$\binom{N}{c} \approx 2^{NH_2\left(\frac{c}{N}\right)}$$

when N is large.

> $\rightsquigarrow$ REMINDER:
>
> Remember that $H_2\left(x\right)$ is given by Eq. (1.14) in page 2 of the textbook as
>
> $$H_2\left(x\right) = x \log \frac{1}{x} + \left(1-x\right) \log \frac{1}{1-x}$$
>
> for $0 \leq x \leq 1$.

- The entropy of a *discrete* random variable X which is defined in *bits* as

$$H\left(X\right) = \sum_{x \in \mathcal{A}_X} \Pr\left\{X = x\right\} \log_2 \frac{1}{\Pr\left\{X = x\right\}}.$$

Note that

- $\Pr\{X = x\}$ is the probability of X being the outcome x.
- $\mathcal{A}_X$ is the set of all possible outcomes for X known as the alphabet of X.

- For independent X and Y, we have

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

With these in mind, let us start with the exercises.

1.2 Exercises

1.2.1 Channel encoding and decoding

♣ Exercise 1 covers Exercise 1.3. in page 8.

1. Let $\mathcal{R}_N$ denote the repetition code with N repetitions where N is an *odd* integer number. This code is employed for encoding over a binary symmetric channel (BSC) with flipping probability f. The decoder uses majority vote to decode the received sequence.

   (a) For an arbitrary N, determine the error probability $p_{\mathrm{b}}(N)$ and the transmission rate $R(N)$.

   (b) Use Stirling's approximation and the *highest order approximation* to approximate the error probability.

   Highest order approximation: For $f < 0.5$ and some constant c, we have

   $$\sum_{n=N/2+c}^{N} 2^{NH_2(n/N)} f^n (1-f)^{N-n} \approx 2^N \left[f(1-f)\right]^{N/2}$$

   when N is large enough.

   (c) Let $f = 0.1$. Plot the error probability-rate diagram using the approximated term for the error probability.

1.2.2 Basic definitions and tools in information theory

♣ Exercise 1 covers partially Exercise 2.27. in page 37.
Exercise 2 covers Exercise 2.21., Exercise 2.25. and Exercise 2.26. in page 37.

1. Consider binary sequence B_1, B_2 in which B_1 and B_2 are independent Bernoulli random variables with

$$\Pr\{B_n = 1\} = 1 - \Pr\{B_n = 0\} = q.$$

Let random variable $X = \mathrm{decimal}(B_1 B_2)$ be the decimal representation, e.g.

$$2 = \mathrm{decimal}(\mathtt{10}).$$

   (a) Determine the entropy of X.

   (b) Conclude the result directly from the basic properties of the entropy function.

2. Let X be a discrete random variable with alphabet $\mathcal{A}_X = \{x_1, \dots, x_M\}$ and distribution $P_X(x)$, i.e.

$$\Pr\{X = x_m\} = P_X(x_m) \qquad \text{for } m \in \{1, \dots, M\}.$$

Let $Q_X(x)$ be another distribution on $\mathcal{A}_X$.

   (a) Determine the expectation of random variables

$$Y = \frac{1}{P_X(X)} \qquad \text{and} \qquad Z = \frac{Q_X(X)}{P_X(X)}.$$

   (b) Show that the function $f(\cdot) : (0, +\infty) \mapsto (-\infty, +\infty)$

$$f(x) = -\log x$$

   is a convex function.

> ⇝ REMINDER:
>
> Remember that $f(\cdot)$ is a convex function if and only if $f''(x) > 0$ for all domain points.

   (c) Either show that $H(X) \leq \log M$ or prove Gibbs' inequality.

1.3 Solutions to Exercises

1.3.1 Channel encoding and decoding

♣ Exercise 1 covers Exercise 1.3. in page 8.

1. Let $\mathcal{R}_N$ denote the repetition code with N repetitions where N is an *odd* integer number. This code is employed for encoding over a BSC with flipping probability f. The decoder uses majority vote to decode the received sequence.

   (a) For an arbitrary N, determine the error probability $p_{\mathrm{b}}(N)$ and the transmission rate $R(N)$.

(b) Use Stirling's approximation and the *highest order approximation* to approximate the error probability.

Highest order approximation: For $f < 0.5$ and some constant c, we have

$$\sum_{n=N/2+c}^{N} 2^{NH_2(n/N)} f^n (1-f)^{N-n} \approx 2^N \left[f(1-f)\right]^{N/2}$$

when N is large enough.

(c) Let $f = 0.1$. Plot the error probability-rate diagram using the approximated term for the error probability.

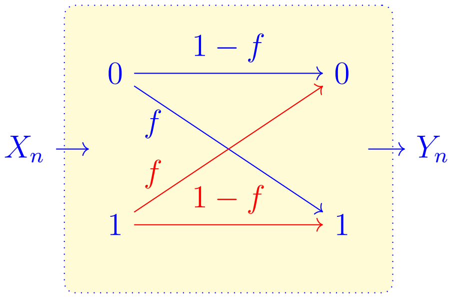
♠ **Solution:**

(a) Let us denote the symbol, which we intend to transmit over the channel, by M. M is a binary bit, meaning that it is either 0 or 1.

Encoding the message: The repetition code $\mathcal{R}_N$ constructs the *codeword* X^N by repeating message M for N times. This means

$$X^N = X_1, \dots, X_N = \underbrace{M, \dots, M}_{N \text{ times}}.$$

X^N is then transmitted over the BSC within N time slots.
Receiving the codeword: In time slot n, where $n \in \{1, \dots, N\}$, the receiver receives Y_n which is either 0 or 1; see the diagram below.



BSC

Since $X_n = M$ for any $n \in \{1, \dots, N\}$, we can conclude that

$$\begin{cases} Y_n = M & \text{with probability } 1-f \\ Y_n \neq M & \text{with probability } f \end{cases}.$$

This means that

$$\Pr\{Y_n \neq M | X_n = M\} = 1 - \Pr\{Y_n = M | X_n = M\} = f$$

for any realization of M, i.e. $M = 0, 1$. Noting that

$$\Pr\{M = 0\} = \Pr\{M = 1\} = \frac{1}{2},$$

we can conclude by chain rule ans the fact that $X_n = M$ for all n that

$$\begin{aligned}\Pr\{Y_n \neq M\} &= \Pr\{M = 0\} \Pr\{Y_n \neq M | M = 0\} + \\ &\qquad \Pr\{M = 1\} \Pr\{Y_n \neq M | M = 1\} \\ &= \Pr\{M = 0\} \Pr\{Y_n \neq M | X_n = 0\} + \\ &\qquad \Pr\{M = 1\} \Pr\{Y_n \neq M | X_n = 1\} \\ &= \frac{1}{2} \times f + \frac{1}{2} \times f = f,\end{aligned}$$

and hence

$$\Pr\{Y_n = M\} = 1 - f.$$

Decoding the message: For decoding, the *majority vote* is used.

MAJORITY VOTE: The decoder determines N_0 and N_1, where

$$\begin{cases} N_1: & \text{\# of bits in } Y^N \text{ which are } 1 \\ N_0: & \text{\# of bits in } Y^N \text{ which are } 0 \end{cases},$$

and decodes the transmitted symbol as

$$\hat{M} = \begin{cases} 0 & \text{if } N_0 > N_1 \\ 1 & \text{if } N_1 > N_0 \end{cases}.$$

Noting that $N_1 + N_0 = N$ and N is an *odd* integer, the decoding rule can be written as

$$\hat{M} = \begin{cases} 0 & \text{if } N_0 \geq (N+1)/2 \\ 1 & \text{if } N_1 \geq (N+1)/2 \end{cases}.$$

Analysis of the error probability: Given the transmitted symbol M, the error occurs, either if $M = 0$ and $N_1 \geq (N+1)/2$, or if $M = 1$ and $N_0 \geq (N+1)/2$. In other words, the error occurs, if the number of bit flips in the BSC are more than, or equal by, $(N+1)/2$. To determine the error probability, let us define the random variable $\mathcal{E}$ as

$$\mathcal{E} = \text{\# of bit flips in } Y^N.$$

The error probability is then given by

$$p_{\mathrm{b}}(N) = \sum_{c=(N+1)/2}^{N} \Pr\{\mathcal{E} = c\}.$$

We now note that Y^N is a binary sequence whose elements are M with probability $1-f$ and are the flipped version of M with probability f. Hence, $\mathcal{E}$ is a binomial random variable[1] whose probability of being equal to c for any $c \in \{1, \dots, N\}$ is

$$\Pr\{\mathcal{E} = c\} = \binom{N}{c} f^c (1-f)^{N-c}.$$

Consequently, the error probability reads

$$p_{\mathrm{b}}(N) = \sum_{c=(N+1)/2}^{N} \binom{N}{c} f^c (1-f)^{N-c}. \tag{EQ:A1}$$

Data transmission rate: In order to determine data rate $R(N)$, we note that by $\mathcal{R}_N$, we transmit only one bit, i.e. M, within N transmission time intervals. Hence, the rate reads

$$R(N) = \frac{\#\text{ of transmitted bits}}{\#\text{ of time intervals}} = \frac{1}{N} \qquad \text{bits/transmission}$$

(b) Stirling's approximation states that

$$\binom{N}{c} \approx 2^{NH_2\left(\frac{c}{N}\right)}$$

when N is large. Substituting into (EQ:A1), we get

$$p_{\mathrm{b}}(N) \sum_{c=(N+1)/2}^{N} \approx 2^{NH_c\left(\frac{c}{N}\right)} f^c (1-f)^{N-c}.$$

Using the highest order term approximation, we have

$$\begin{aligned} p_{\mathrm{b}}(N) &\approx 2^N f^{\frac{N}{2}} (1-f)^{\frac{N}{2}} \\ &\approx [4f(1-f)]^{\frac{N}{2}} = \tilde{p}_{\mathrm{b}}(N). \end{aligned} \tag{EQ:B1}$$

Note that (EQ:B1) is not necessarily good approximation for small N. For example, setting $N=1$ and $f=0.1$, $p_{\mathrm{b}}(1) = f = 0.1$ while the approximation in (EQ:B1) gives $\tilde{p}_{\mathrm{b}}(1) \approx 0.6$. Nevertheless, for large enough N, this approximation is accurate.

The accuracy of the approximation is shown in Fig. 1.3.1, where the difference between the exact error probability given by (EQ:A1) and the approximation in (EQ:B1) has been sketched in terms of N. As it shows, for $N=99$, the approximation is different from the exact error term only 1.06×10^{-22} which is sufficiently small.

(c) Using the results in Parts (a) and (b), the error probability-rate diagram is given by plotting the pair $(R(N), p_b(N))$ for various values of N. Fig. 1.3.2 shows this

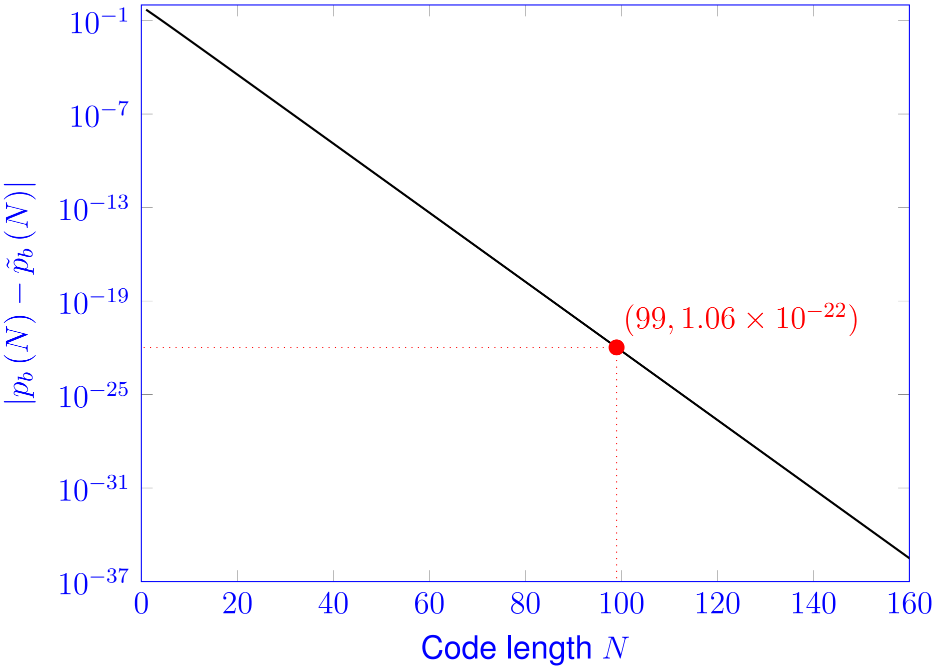


Figure 1.3.1: The difference between the exact value of the error probability, i.e. $p_{\mathrm{b}}(N)$ and its approximation $\tilde{p}_{\mathrm{b}}(N)$ versus the code length N.

diagram for $f = 0.1$, considering both exact and approximated terms for error probability.

In Chapter 9, you will learn that the capacity of this channel is

$$C = 1 - H_2(f) = 0.531.$$

This means that by using a *very good* channel code, the error probability for any $R < 0.531$ would be theoretically zero[2]. Considering the diagram in Fig.1.3.2, this indicates that the curve on the left side of the capacity line could be effectively pushed down to zero.

Shannon explains the reason of not having zero error probability on the left side via repetition code as following: This is not happening by the repetition code, since this code is not *good enough*.

⇝ REMARK:

The main reasons that you are studying information theory is to learn

- what is this capacity limit for a given channel?
- how we could push the diagram enough down on the left side of the capacity line?

[1] The distribution of this random variable is discussed in the first chapter of the textbook

[2] In practice, the error probability might be not exactly zero, but significantly small.

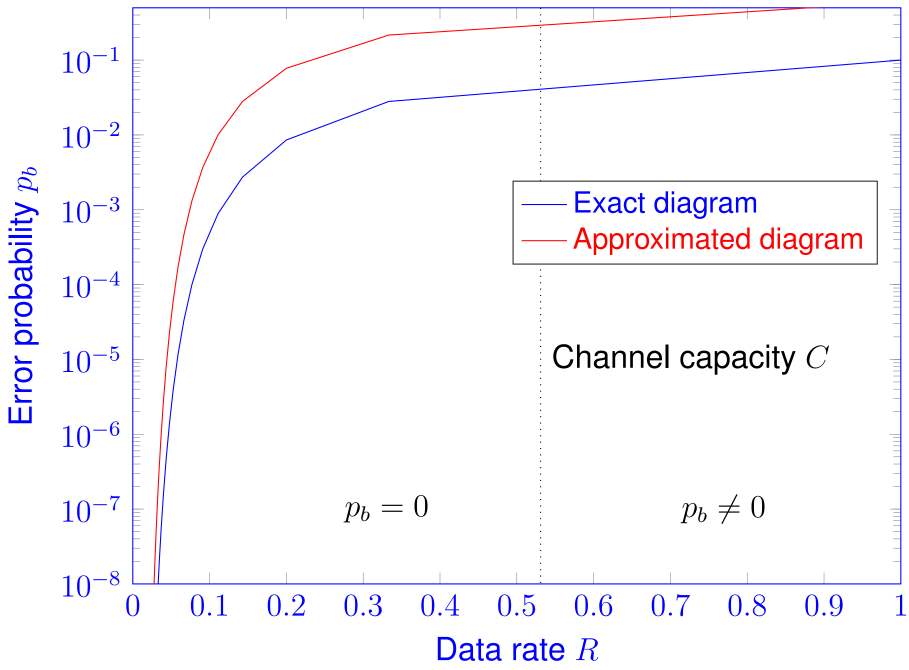


Figure 1.3.2: The error probability-rate diagram for repetition code.

1.3.2 Basic definitions and tools in information theory

♣ Exercise 1 covers partially Exercise 2.27. in page 37.
Exercise 2 covers Exercise 2.21., Exercise 2.25. and Exercise 2.26. in page 37.

1. Consider binary sequence B_1, B_2 in which B_1 and B_2 are independent Bernoulli random variables with

$$\Pr\{B_n = 1\} = 1 - \Pr\{B_n = 0\} = q.$$

Let random variable $X = \text{decimal}(B_1 B_2)$ be the decimal representation, e.g. $2 = \text{decimal}(\texttt{10})$.

(a) Determine the entropy of X.

(b) Conclude the result directly from the basic properties of the entropy function.

♠ **Solution:**

(a) Noting that random sequence B_1, B_2 are binary, the alphabet of $B_1 B_2$ contains four different outcomes $\{\texttt{00}, \texttt{01}, \texttt{10}, \texttt{11}\}$. Hence, the alphabet of random variable X is

$$\mathcal{A}_X = \{0, 1, 2, 3\}.$$

Noting that B_1 and B_2 are independent, the distribution of X is determined as

$$\begin{aligned}
\Pr\{X=0\} &= \Pr\{B_1=0,B_2=0\} = \Pr\{B_1=0\}\Pr\{B_2=0\} = (1-q)^2,\\
\Pr\{X=1\} &= \Pr\{B_1=0,B_2=1\} = \Pr\{B_1=0\}\Pr\{B_2=1\} = (1-q)\,q,\\
\Pr\{X=2\} &= \Pr\{B_1=1,B_2=0\} = \Pr\{B_1=1\}\Pr\{B_2=0\} = q\,(1-q),\\
\Pr\{X=3\} &= \Pr\{B_1=1,B_2=1\} = \Pr\{B_1=1\}\Pr\{B_2=1\} = q^2.
\end{aligned}$$

Consequently, the entropy of X is

$$\begin{aligned}
H(X) &= \sum_{x\in\mathcal{A}_X} \Pr\{X=c\}\log_2\frac{1}{\Pr\{X=x\}}\\
&= 2\,(1-q)^2\log_2\frac{1}{1-q} + 2\,q\,(1-q)\log_2\frac{1}{q\,(1-q)} + 2\,q^2\log_2\frac{1}{q}\\
&= 2\,(1-q)\log_2\frac{1}{1-q} + 2\,q\log_2\frac{1}{q}\\
&= 2\,H_2(q)\,.
\end{aligned}$$

(b) The result of Part (a) could be directly derived using the properties of the entropy function. From Eq. (2.39) in page 33 of the textbook, we know that

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

when X and Y are independent. Hence, we can write

$$\begin{aligned}
H(X) = H(B_1,B_2) &= H(B_1) + H(B_2)\\
&= H_2(q) + H_2(q) = 2\,H_2(q)
\end{aligned}$$

which concludes the same result.

2. Let X be a discrete random variable with alphabet $\mathcal{A}_X = \{x_1,\dots,x_M\}$ and distribution $P_X(x)$, i.e.

$$\Pr\{X=x_m\} = P_X(x_m) \qquad \text{for } m\in\{1,\dots,M\}\,.$$

Let $Q_X(x)$ be another distribution on $\mathcal{A}_X$.

(a) Determine the expectation of random variables

$$Y = \frac{1}{P_X(X)} \qquad \text{and} \qquad Z = \frac{Q_X(X)}{P_X(X)}.$$

(b) Show that the function $f(\cdot):(0,+\infty)\mapsto(-\infty,+\infty)$

$$f(x) = -\log x$$

is a convex function.

> $\rightsquigarrow$ REMINDER:
>
> Remember that $f(\cdot)$ is a convex function if and only if $f''(x) > 0$ for all domain points.

(c) Either show that $H(X)\leq\log M$ or prove Gibbs' inequality.

♠ **Solution:**

(a) Let us first determine the alphabet and distribution of Y and Z. When x_m is the outcome of random variable X, the outcomes of Y and Z are y_m and z_m, respectively, which can be written as

$$\begin{aligned} y_m &= \frac{1}{P_X(x_m)} \\ z_m &= \frac{Q_X(x_m)}{P_X(x_m)}. \end{aligned}$$

Hence we can conclude that

$$\begin{aligned} \mathcal{A}_Y = \{y_1, \ldots, y_M\} &= \left\{ \frac{1}{P_X(x_1)}, \ldots, \frac{1}{P_X(x_M)} \right\} \\ \mathcal{A}_Z = \{z_1, \ldots, z_M\} &= \left\{ \frac{Q_X(x_1)}{P_X(x_1)}, \ldots, \frac{Q_X(x_M)}{P_X(x_M)} \right\} \end{aligned}$$

The probability of each outcome happening can be further found as

$$\begin{aligned} P_Y(y_m) &= \Pr\{Y = y_m\} = \Pr\{X = x_m\} = P_X(x_m) \\ P_Z(z_m) &= \Pr\{Z = z_m\} = \Pr\{X = x_m\} = P_X(x_m). \end{aligned}$$

As a result, the expectations of Y and Z read

$$\begin{aligned} \mathcal{E}[Y] &= \sum_{m=1}^{M} y_m P_Y(y_m) = \sum_{m=1}^{M} \frac{1}{P_X(x_m)} P_X(x_m) = M \\ \mathcal{E}[Z] &= \sum_{m=1}^{M} z_m P_Z(z_m) = \sum_{m=1}^{M} \frac{Q_X(x_m)}{P_X(x_m)} P_X(x_m) = \sum_{m=1}^{M} Q_X(x_m) \overset{(\mathrm{A})}{=} 1. \end{aligned}$$

where equality (A) comes from the fact that $Q_X(x)$ is a probability distribution on $\mathcal{A}_X$, and hence its sum over all $x_m \in \mathcal{A}_X$ equals to one.

(b) Using the hint given in the reminder, we can write

$$f''(x) = \frac{\ln 2}{x^2} > 0 \qquad \forall x > 0.$$

Hence, we can conclude that $f(\cdot)$ is a convex function.

(c) PROOF OF $H(X) \leq M$:

Consider random variable Y in Part (a). Since $f(\cdot)$ is a convex function, we can use Jensen's inequality and write

$$f(\mathcal{E}[Y]) \leq \mathcal{E}[f(Y)]$$

where "=" holds if Y is a deterministic variable, i.e. all the entries in $\mathcal{A}_Y$ are the same. Noting that $\mathcal{E}[Y] = M$, and substituting the definition of Y into Jensen's inequality, we have

$$-\log M \leq -\mathcal{E}[\log Y].$$

We now calculate $\mathcal{E}\left[\log Y\right]$ by writing its definition:

$$\begin{aligned}\mathcal{E}\left[\log Y\right] &= \sum_{m=1}^{M} \log\left(y_m\right) P_Y\left(y_m\right)\\ &= \sum_{m=1}^{M} \log\left(y_m\right) P_X\left(x_m\right)\\ &= \sum_{m=1}^{M} \log\left(\frac{1}{P_X\left(x_m\right)}\right) P_X\left(x_m\right) = H\left(X\right)\end{aligned}$$

Consequently, we conclude that

$$-\log M \leq -H\left(X\right) \Rightarrow \log M \geq H\left(X\right)$$

where "=" holds if

$$\frac{1}{P_X\left(x_m\right)} = \frac{1}{P_X\left(x_t\right)}$$

for all $m \neq t$. This means that the equality holds, when X is *uniformly distributed*.

PROOF OF Gibbs' inequality:

To prove Gibbs' inequality, we consider random variable Z in Part (a) and use Jensen's inequality which says

$$f\left(\mathcal{E}\left[Z\right]\right) \leq \mathcal{E}\left[f\left(Z\right)\right]$$

where "=" holds if all the entries in $\mathcal{A}_Z$ are the same. Noting that $\mathcal{E}\left[Z\right] = 1$, we have

$$0 \leq -\mathcal{E}\left[\log Z\right] \overset{\text{(B)}}{=} D_{\text{KL}}\left(P_X \| Q_X\right)$$

where (B) is concluded by the similar approach as the one we took to calculate[3] $\mathcal{E}\left[\log Y\right]$. Here, "=" holds if

$$\frac{Q_X\left(x_m\right)}{P_X\left(x_m\right)} = \frac{Q_X\left(x_t\right)}{P_X\left(x_t\right)}$$

for all $m \neq t$. This means that the equality holds, when $P_X = Q_X$. This proves Gibb's inequality.

[3] You are asked to show this in your homework.

1.4 Homework

The following exercises are suggested for further practice.

1.4.1 Primary exercises

1. Let X be a discrete random variable with alphabet $\mathcal{A}_X = \{x_1, \dots, x_M\}$ and distribution $P_X(x)$, i.e.

$$\Pr\{X = x_m\} = P_X(x_m) \qquad m \in \{1, \dots, M\}.$$

Assume that $Q_X(x)$ is another distribution defined on the alphabet $\mathcal{A}_X$. This means that for any $x_m \in \mathcal{A}_x$, we have $Q_X(x_m) \geq 0$ and

$$\sum_{m=1}^{M} Q_X(x_m) = 1.$$

(a) Find the alphabet and probability distribution of the following random variable

$$Z = \log \frac{P_X(X)}{Q_X(X)}.$$

(b) Show that

$$\mathcal{E}[Z] = D_{\text{KL}}(P_X \| Q_X)$$

where $D_{\text{KL}}(P_X \| Q_X)$ denotes the Kullback-Leibler divergence between $P_X(x)$ and $Q_X(x)$.

♠ **Hint:** Part (a) has been solved in Tutorial 1. For Part (b) follow the same steps as those taken in Tutorial 1 to show that $\mathcal{E}[\log P_X(X)] = -H(X)$.

1.4.2 Further exercises from the textbook

- Chapter 1: Exercise 1.12.
- Chapter 2: Exercise 2.4., Exercise 2.5., Exercise 2.7., Exercise 2.8., Exercise 2.14., Exercise 2.17., Exercise 2.18., Exercise 2.19., Exercise 2.20., Exercise 2.21., Exercise 2.34., Exercise 2.36., Exercise 2.37.

1.5 Fun Facts

Thomas M. Cover (August 7, 1938 – March 26, 2012) is one of the most well-known information theorists who co-authored the book *Elements of Information Theory* and lectured the information theory course for several decades at Stanford University. Prof. Cover once said

I don't normally prove theorems, but when I do I use Jensen's inequality!

Chapter 2

Bayesian Inference

This chapter investigates the concept of Bayesian estimation. The contents of this chapter are consistent with Chapter 2 of the textbook. We further have some exercises on Chebyshev's inequality which is later on used in the source coding theorem.

2.1 Brief Review of Main Concepts

There are two types of inference problems: *Forward* and *backward*. In a forward problem, you are asked to calculate the probability of a set out outcomes happening. This is the most basic inference problem. However, in practice, we mostly deal with backward or *inverse* inference in which we should *estimate* what was the outcome of a random variable, given an observation. A standard approach to do it is to use the *Bayes rule* which is called Bayesian inference.

In the Bayesian framework, an inverse problem consists of multiple elements: You are given by the outcome of a random variable D known as *data* and want to estimate another random variable C which depends on the data. You are either given by the distribution of C, or you postulate one. This distribution is called the *prior* distribution.

To estimate C from the observation, we calculate the so-called *posterior* probability via the Bayes rule which indicates

$$\Pr\{C = c | D = d\} = \frac{\Pr\{C = c, D = d\}}{\Pr\{D = d\}} \qquad = \frac{\Pr\{D = d | C = c\} \Pr\{C = c\}}{\Pr\{D = d\}}$$

In the Bayes rule, we have

- $\Pr\{C = c\}$ which is the *prior* distribution.
- $\Pr\{D = d | C = c\}$ which is the *likelihood* of $\{C = c\}$.

⇝ REMINDER:

Note that this is *not* a probability, meaning that in general

$$\sum_{c \in \mathcal{A}_C} \Pr\{D = d | C = c\} \neq 1.$$

The likelihood is derived by solving a forward inference problem.

- $\Pr\{D=d\}$ is the *marginal* probability and is determined by marginalization as follows:

  $$\Pr\{D=d\} = \sum_{c\in\mathcal{A}_C} \Pr\{D=d, C=c\} = \sum_{c\in\mathcal{A}_C} \Pr\{D=d|C=c\}\Pr\{C=c\}$$

  in terms of the likelihood and the prior distribution

The optimal Bayesian estimate is then the outcome of C which has the highest posterior probability. These concepts get further clear throughout the exercises.

2.2 Exercises

2.2.1 Basics of Bayesian inference

♣ Exercise 1 covers Exercise 3.12. in page 58.
Exercise 2 covers Exercise 3.1. in page 47.

1. A bag contains a pen which is either *red or blue*. I put a red pen inside the bag and shake it. Then, I take a pen out of the bag which is red. You are supposed to guess the color of the pen which was *initially* in the bag.

   (a) What kind of probability problem is this? Forward or backward?

   (b) What is the probability that the pen, which was initially in the bag, is also red?

   (c) What was the result, if I have asked you to guess the color in the first place, i.e., before putting the second pen and shaking the bag?

2. Two images, namely image A and image B, are saved on a computer. Each image has 20 pixels, and each pixel has a certain level of light intensity. The intensity level is an integer number between 1 and 10.

   The frequency of each intensity level for images A and B are given in the following table.

Intensity level	1	2	3	4	5	6	7	8	9	10
# of pixels in image A	6	4	3	2	1	1	1	1	1	0
# of pixels in image B	3	3	2	2	2	2	2	2	1	1

   A portable memory is capable of saving the intensity level of 7 pixels. The computer chooses one of the images at random, and copies 7 random pixels on the portable memory. This memory is then given to you and contains the following intensity levels:

   $$d = \begin{array}{|c|c|c|c|c|c|c|}\hline 5 & 3 & 9 & 3 & 8 & 4 & 7 \\ \hline\end{array}$$

   You are asked to *recognize the image* based on data d.

(a) Formulate the problem as a Bayesian inference problem and specify the *likelihood* and the *posterior*.

(b) Determine the probability that the computer has chosen image A.

(c) Which image is recognized via the Bayesian detector?

2.2.2 Chebyshev's inequality

♣ Exercise 1 covers Exercise 4.19. in page 85.

1. *Chernoff bound* is widely used in communications, in order to bound the error probability. It is an inequality which states that

For a random variable X and a constant a, we have

$$\Pr\{X \geq a\} \leq e^{-sa}\mathcal{E}\left[e^{sX}\right]$$

for any $s > 0$.

Show this inequality by means of Chebyshev's inequality.

2.3 Solutions to Exercises

2.3.1 Basics of Bayesian inference

♣ Exercise 1 covers Exercise 3.12. in page 58.
Exercise 2 covers Exercise 3.1. in page 47.

1. A bag contains a pen which is either *red or blue*. I put a red pen inside the bag and shake it. Then, I take a pen out of the bag which is red. You are supposed to guess the color of the pen which was *initially* in the bag.

(a) What kind of probability problem is this? Forward or backward?

(b) What is the probability that the pen, which was initially in the bag, is also red?

(c) What was the result, if I have asked you to guess the color in the first place, i.e., before putting the second pen and shaking the bag?

♠ **Solution:**

(a) In this problem, we are given by a set of observations and are supposed to infer the probability of prior unobserved events conditioned to these observations. It is hence a *backward*, or as called in the textbook *inverse*, probability problem.

(b) To calculate the probability, we need to formulate this problem in the formal Bayesian framework. To this end, let us denote the *data* being observed after taking one pen out of the bag, by random variable D. The alphabet of this random variable is

$$\mathcal{A}_D = \{\text{red}, \text{blue}\} \tag{2.3.1}$$

Moreover, let the color of the initial pen be C which is again a random variable with alphabet

$$\mathcal{A}_C = \{\text{red}, \text{blue}\}. \tag{2.3.2}$$

Our observation is that

$$D = \text{red}$$

and we are supposed to determine the following probability

$$\Pr\{C = \text{red}|D = \text{red}\}.$$

The above probability is called, in the context of Bayesian inference, the *posterior* probability, since it is determined conditioned to our observation $D = \text{red}$. Using the Bayes rule, we have

$$\begin{aligned}\Pr\{C = \text{red}|D = \text{red}\} &= \frac{\Pr\{C = \text{red}, D = \text{red}\}}{\Pr\{D = \text{red}\}} \\ &= \frac{\Pr\{D = \text{red}|C = \text{red}\}\Pr\{C = \text{red}\}}{\Pr\{D = \text{red}\}} \end{aligned} \tag{EQ:A2}$$

The terms in (EQ:A2) can be determined explicitly from the information we have in hand, namely

- $\Pr\{C = \text{red}\}$ is the *prior* probability of $\{C = \text{red}\}$. It is called prior, as it is determined prior to our observation. More precisely, it is the probability of the first pen being red, when you had no observation at all. Before observing the color of the second pen, we expect the first pen to be red or blue with the same probability. Hence, we could conclude that

$$\Pr\{C = \text{red}\} = \Pr\{C = \text{blue}\} = \frac{1}{2}.$$

- $\Pr\{D = \text{red}|C = \text{red}\}$ is called the *likelihood* of $\{C = \text{red}\}$. Note that this is *not* a probability, meaning that in general

$$\Pr\{D = \text{red}|C = \text{red}\} + \Pr\{D = \text{red}|C = \text{blue}\} \neq 1.$$

The likelihood is in fact determines how likely is to have such an observation, i.e. $D = \text{red}$ for the given assumption on C, i.e. assuming $C = \text{red}$. To derive the likelihood term, we need to solve the following *forward* probability problem: Assume the color of the first pen is red, what is the probability that we observe a red pen after adding a red pen and shaking the bag? Noting that in this case, we have two red pens, we can conclude that

$$\Pr\{D = \text{red}|C = \text{red}\} = 1.$$

Similarly, we can conclude that

$$\Pr\{D = \text{red}|C = \text{blue}\} = \frac{1}{2}.$$

Note once again that likelihood is not a probability distribution on C, as we have

$$\Pr\{D = \text{red}|C = \text{red}\} + \Pr\{D = \text{red}|C = \text{blue}\} = \frac{3}{2}.$$

- $\Pr\{D = \text{red}\}$ is the *marginal* probability of $D = \text{red}$. This probability is called marginal, since it is determined by marginalization as follows: For a given choice of $d \in \{\text{red}, \text{blue}\}$, the sum rule in page 24 says

$$\Pr\{D = d\} = \sum_{c\in\{\text{red},\text{blue}\}} \Pr\{D = d, C = c\}. \tag{2.3.3}$$

Using the chain rule, we have

$$\Pr\{D = d, C = c\} = \Pr\{D = d|C = c\} \Pr\{C = c\}$$

Hence, we can write

$$\begin{aligned}\Pr\{D = d\} &= \Pr\{D = d, C = \text{red}\} + \Pr\{D = d, C = \text{blue}\} \\ &= \Pr\{D = d|C = \text{red}\} \Pr\{C = \text{red}\} + \\ &\quad\ \Pr\{D = d|C = \text{blue}\} \Pr\{C = \text{blue}\}. \end{aligned} \tag{EQ:B2}$$

Using the likelihood terms and the prior probabilities of C, we can determine $\Pr\{D = \text{red}\}$ from (EQ:B2). Since we use the sum rule, and marginalize the joint probability distribution $\Pr\{D = d, C = c\}$ over all possible choices of c, this probability is called a marginal probability. Substituting into (EQ:B2), we could conclude that

$$\begin{aligned}\Pr\{D = \text{red}\} = \Pr\{D = \text{red}|C = \text{red}\} \Pr\{C = \text{red}\} + \\ \Pr\{D = \text{red}|C = \text{blue}\} \Pr\{C = \text{blue}\} \\ = 1 \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{3}{4}.\end{aligned}$$

Clearly, we also have

$$\Pr\{D = \text{blue}\} = 1 - \Pr\{D = \text{red}\} = \frac{1}{4}.$$

Using the prior distribution, likelihood and the marginal distribution of D, we can write

$$\Pr\{C = \text{red}|D = \text{red}\} = \frac{\Pr\{D = \text{red}|C = \text{red}\}\Pr\{C = \text{red}\}}{\Pr\{D = \text{red}\}}$$
$$= \frac{1 \times \frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}.$$

As a result, we further have

$$\Pr\{C = \text{blue}|D = \text{red}\} = 1 - \Pr\{C = \text{red}|D = \text{red}\} = \frac{1}{3}.$$

Consequently, if you were asked to guess a color for the initial pen, Bayesian inference would suggest you to guess red, since it is more probable to be.

(c) If you were asked to guess the color without putting a red pen into the bag and shaking it, you had no information, and hence, the probability of being red was given by the prior distribution of C, i.e.

$$\Pr\{C = \text{red}\} = \Pr\{C = \text{blue}\} = \frac{1}{2}.$$

$\rightsquigarrow$ REMARK:

Here, you are acquiring some information from your observation. If I had not done this procedure (I mean putting a red pen in the bag, shaking the bag, taking one pen out and showing you that pen), there were no difference to choose either red or blue. This sounds like a magic, but it's not. It is just Bayesian inference!

2. Two images, namely image A and image B, are saved on a computer. Each image has 20 pixels, and each pixel has a certain level of light intensity. The intensity level is an integer number between 1 and 10.

The frequency of each intensity level for images A and B are given in the following table.

Intensity level	1	2	3	4	5	6	7	8	9	10
# of pixels in image A	6	4	3	2	1	1	1	1	1	0
# of pixels in image B	3	3	2	2	2	2	2	2	1	1

A portable memory is capable of saving the intensity level of 7 pixels. The computer chooses one of the images at random, and copies 7 random pixels on the portable memory. This memory is then given to you and contains the following intensity levels:

$$d = \begin{array}{|c|c|c|c|c|c|c|}\hline 5 & 3 & 9 & 3 & 8 & 4 & 7 \\ \hline\end{array}$$

You are asked to *recognize the image* based on data d.

(a) Formulate the problem as a Bayesian inference problem and denote the *likelihood* and the *posterior*.

(b) Determine the probability that the computer had chosen image A.

(c) Which image is recognized?

♠ **Solution:**

(a) In this problem, our data D is a vector of 7 digits. We hence show it as

$$D = [D_1, \dots, D_7].$$

Each of the entries is an integer between 1 and 10, i.e. for $j \in \{1, \dots, 7\}$,

$$\mathcal{A}_{D_j} = \{1, \dots, 10\}.$$

Thus, the alphabet of D is

$$\mathcal{A}_D = \{1, \dots, 10\}^7.$$

Let random variable I be the image which had been chosen by the computer. This variable is either A or B; hence,

$$\mathcal{A}_I = \{\mathrm{A}, \mathrm{B}\}.$$

We have observed

$$D = d = [5, 3, 9, 3, 8, 4, 7]$$

and intend to recognize the image. Hence, we need to find the *posterior* distribution

$$\Pr\{I = \mathrm{A} | D = d\}$$

which reads

$$\Pr\{I = \mathrm{A} | D = d\} = \frac{\Pr\{I = \mathrm{A}, D = d\}}{\Pr\{D = d\}}$$
$$= \frac{\Pr\{D = d | I = \mathrm{A}\} \Pr\{I = \mathrm{A}\}}{\Pr\{D = d\}} \qquad \text{(EQ:C2)}$$

In (EQ:C2), $\Pr\{I = \mathrm{A} | D = d\}$ is the posterior, and $\Pr\{D = d | I = \mathrm{A}\}$ is the likelihood of $\{I = \mathrm{A}\}$.

(b) To determine posterior $\Pr\{I = \mathrm{A} | D = d\}$, we have to find the prior probabilities, the likelihoods and the marginal probability of $D = d$.

- For prior, we note that the computer has chosen the image at random without any preference. Hence, prior to observation $D = d$, images A and B are same likely to be chosen. Thus,

$$\Pr\{I = \mathrm{A}\} = \Pr\{I = \mathrm{B}\} = \frac{1}{2}.$$

- The likelihood term $\Pr\{D=d|I=\mathrm{A}\}$ is determined by solving this *forward* probability problem: Assume image A is chosen, what is the probability of $\{D=d\}$? To answer this question, we note that the entries of D are chosen independently. Therefore, we can write

  $$\begin{aligned}\Pr\{D=d|I=\mathrm{A}\} &= \Pr\{[D_1,\dots,D_7]=[5,3,9,3,8,4,7]|I=\mathrm{A}\}\\ &= \Pr\{D_1=5|I=\mathrm{A}\}\times\dots\times\Pr\{D_7=7|I=\mathrm{A}\}.\end{aligned}$$

  The probability terms $\Pr\{D_j=d_j|I=\mathrm{A}\}$ can be determined from the table given in the problem. For instance, to determine $\Pr\{D_1=5|I=\mathrm{A}\}$, we note that among the 20 pixels of image A, a pixel with intensity level 5 occurs with frequency 1; hence,

  $$\Pr\{D_1=5|I=\mathrm{A}\} = \frac{\text{\# of pixels with intensity level 5}}{\text{Total \# of pixels}} = \frac{1}{20}.$$

  Determining the whole terms, one could finally conclude that

  $$\Pr\{D=d|I=\mathrm{A}\} = \frac{1\times3\times1\times3\times1\times2\times1}{20^7} = \frac{18}{20^7} = \frac{9}{64}\times10^{-7}.$$

  For the likelihood term $\Pr\{D=d|I=\mathrm{B}\}$, we can take the same steps and conclude that

  $$\Pr\{D=d|I=\mathrm{B}\} = \frac{2\times2\times1\times2\times2\times2\times2}{20^7} = \frac{2^6}{20^7} = \frac{1}{2}\times10^{-7}.$$

- The marginal probability $\Pr\{D=d\}$ is furthermore given by the sum rule

  $$\begin{aligned}\Pr\{D=d\} &= \Pr\{D=d|I=\mathrm{A}\}\Pr\{I=\mathrm{A}\}+\\ &\quad \Pr\{D=d|I=\mathrm{B}\}\Pr\{I=\mathrm{B}\}\\ &= \frac{9}{64}\times10^{-7}\times\frac{1}{2}+\frac{1}{2}\times10^{-7}\times\frac{1}{2} = \frac{41}{128}\times10^{-7}.\end{aligned}$$

Thus, the posterior probability of $\{I=\mathrm{A}\}$ is calculated as

$$\begin{aligned}\Pr\{I=\mathrm{A}|D=d\} &= \frac{\Pr\{D=d|I=\mathrm{A}\}\Pr\{I=\mathrm{A}\}}{\Pr\{D=d\}}\\ &= \frac{\dfrac{9}{64}\times10^{-7}\times\frac{1}{2}}{\dfrac{41}{128}\times10^{-7}} = \frac{9}{41}.\end{aligned}$$

the posterior probability of $\{I=\mathrm{B}\}$ is moreover given by

$$\Pr\{I=\mathrm{B}|D=d\} = 1-\Pr\{I=\mathrm{A}|D=d\} = \frac{32}{41}.$$

(c) By Bayesian estimation, we recognize the image to be B, since

$$\Pr\{I=\mathrm{B}|D=d\} > \Pr\{I=\mathrm{A}|D=d\}.$$

2.3.2 Chebyshev's inequality

♣ Exercise 1 covers Exercise 4.19. in page 85.

1. *Chernoff bound* is widely used in communications, in order to bound the error probability. It is an inequality which states that

   For a random variable X and a constant a, we have

   $$\Pr\left\{X \geq a\right\} \leq e^{-sa}\mathcal{E}\left[e^{sX}\right]$$

   for any $s > 0$.

   Show this inequality by means of Chebyshev's inequality.

♠ **Solution:** Chebyshev's inequality indicates that for any random variable $T \geq 0$ and constant $\alpha > 0$,

$$\Pr\left\{T \geq \alpha\right\} \leq \frac{\mathcal{E}\left\{T\right\}}{\alpha}.$$

See page 81 of the textbook. We now define random variable T as a function of X as

$$T = e^{sX}$$

for some constant s. Moreover, we set

$$\alpha = e^{sa}$$

for some real a. Noting that $e^u > 0$ for any choice of u, we have

$$\Pr\left\{e^{sX} \geq e^{sa}\right\} \leq \frac{\mathcal{E}\left[e^{sX}\right]}{e^{sa}} = e^{-sa}\mathcal{E}\left[e^{sX}\right]. \tag{EQ:D2}$$

Now, consider the inequality $e^{sX} \geq e^{sa}$, we can take logarithm in the natural base from the both sides of the inequality, without any change in the inequality. Hence, we have

$$e^{sX} \geq e^{sa} \Leftrightarrow sX \geq sa.$$

For $s > 0$, we can further divide the both sides of the inequality by s without any change in the inequality. Thus, we have

$$e^{sX} \geq e^{sa} \Leftrightarrow sX \geq sa \Leftrightarrow X \geq a$$

when $s > 0$. Substituting in (EQ:D2), Chernoff bound is concluded.

2.4 Homework

In Tutorial 2, we studied Chapter 3 of [**?**]. The following exercises are suggested for further practice.

2.4.1 Primary exercises

1. Consider Exercise 2 in Section 2.1 of Tutorial 2.

   (a) Solve this probability problem once again assuming the following data is observed

   $$d = \begin{array}{|c|c|c|c|c|c|c|}\hline 5 & 3 & 9 & 4 & 5 & 10 & 1 \\ \hline \end{array}\,.$$

   (b) Illustrate why the final recognition given by Bayesian inference was initially obvious.

   ♠ **Hint:** Pay attention to the frequency of intensity level 10 in both images.

2.4.2 Further exercises from the textbook

- Chapter 3: Exercise 3.5., Exercise 3.11., Exercise 3.13., Exercise 3.15.

2.5 Fun Facts

Assume that Exercise 1 in Section 2.2.1 is a game, in which you need to guess the color of the remained pen. You win €100 if you guess the right color and lose €100 if you guess it wrong. If you play this game 30,000 times, and you always choose the red color, you will win almost **€1,000,000** at the end! You can show this by the *law of large numbers*.

Chapter 3

AEP and Source Coding Theorem

In this chapter, we study the asymptotic equipartition property (AEP) of identically and independently distributed (i.i.d.) sequences. We then practice the first theorem of Shannon. The discussions in this tutorial covers Chapter 4 of textbook. It is worth mentioning that this chapter only considers the *Source Coding Theorem*. The source coding algorithms will be practiced in the next chapter.

3.1 Brief Review of Main Concepts

3.1.1 Asymptotic Equipartition Property

AEP is a key tool to understand the theorems of Shannon. In a nutshell, AEP states that for a given i.i.d. sequence X^N with very large length N, the set of all possible outcomes are partitioned roughly into two subsets: A set of *typical* sequences which happen with probability of almost 1 and the set of *non-typical* sequences which happen with probability of almost 0. In other words, no matter what, there are some sequences which never happen (non-typical), and some sequences which one of them always happen (typical).

Formally AEP is presented as follows: Considering the random variable X, the β-typical set is defined as

$$T_{N_\beta} = \left\{ x^N : |\frac{1}{N} \log \frac{1}{P\left(x^N\right)} - H| < \beta \right\}. \tag{3.1.1}$$

in Eq. (4.29), page 80 of the textbook. The three main properties of T_{N_β} are as follows

1. The number of sequences in T_{N_β} satisfies

   $$\left(1 - \frac{\theta}{\beta^2 N}\right) 2^{N(H-\beta)} \leq |T_{N_\beta}| \leq 2^{N(H+\beta)}$$

   for some constant θ.

2. The probability of *a single β-typical sequence x^N* satisfies

   $$2^{-N(H+\beta)} \leq P\left(x^N\right) \leq 2^{-N(H-\beta)}$$

3. The probability of T_{N_β} happening satisfies

$$1 - \frac{\theta}{\beta^2 N} \leq \Pr\left\{X^N \in T_{N_\beta}\right\} \leq 1$$

for some constant θ.

For very small choices of β and very large N, these properties simply indicate that

1. The number of sequences in T_{N_β} reads

$$|T_{N_\beta}| \approx 2^{NH_2(q)}.$$

2. The probability of *a single β-typical sequence* x^N is

$$P\left(x^N\right) \approx 2^{-NH_2(q)}.$$

3. The probability of T_{N_β} happening reads

$$\Pr\left\{X^N \in T_{N_\beta}\right\} \approx 1.$$

These simplified form clarifies the intuitive idea behind AEP. It gets more clear, as we go through the exercises.

3.1.2 Source Coding Theorem

We assume a source generate an i.i.d. sequence X^N and intend to represent this sequence with a binary sequence of length $B\left(N\right)$. In this case, the compression rate is defined as

$$R\left(N\right) = \frac{B\left(N\right)}{N}.$$

The first theorem of Shannon indicates two things:

1. For very large N, X^N is encoded without any loss, if

$$R\left(N\right) > H\left(X\right).$$

This means that in this case, there *exist a lossless* source code which encodes X^N to a binary sequence of length $B\left(N\right)$. It is also guaranteed that we get back from the binary sequence to X^N *uniquely*.

2. For very large N, if

$$R\left(N\right) < H\left(X\right),$$

it is *impossible* to compress X^N into $B\left(N\right)$ bits without any loss.

3.2 Exercises

3.2.1 Asymptotic equipartition principle

1. Consider i.i.d. sequence $X^N = X_1, \dots, X_N$ whose entries are Bernoulli variables with

$$\Pr\{X = 1\} = 1 - \Pr\{X = 0\} = q.$$

   (a) Which sequences are typical for X^N?

   (b) Write the β-typical set T_{N_β} for this sequence.

   (c) Compare T_{N_β} to your intuitive definition in Part (a), when β is almost zero.

   (d) Write the three main properties of T_{N_β} that you learned in Chapter 4.

   (e) Using your definition in Part (a), derive

   - the number of typical sequences,
   - the probability of one typical sequence occurring,
   - the probability of X^N being in the typical set,

   and approximate it for large N using Stirling's approximation. Compare the result with Part (d).

3.2.2 Shannon's first theorem

♣ Exercise 1 covers Exercise 4.5. in page 74 and partially Exercise 4.15. and Exercise 4.16. in page 85.

1. Consider a ternary random variable X whose alphabet is $\mathcal{A}_X = \{a, b, c\}$ and is uniformly distributed, i.e.

$$\Pr\{X = a\} = \Pr\{X = b\} = \Pr\{X = c\} = \frac{1}{3}.$$

   (a) What is the minimum compression ratio suggested by Shannon?

   (b) X^N is an i.i.d. sequence. Let $B(N)$ denote the number of bits required to encode X^N via a binary code. Determine $B(N)$ for a *simple* binary code.

   (c) Define

$$R(N) = \frac{B(N)}{N},$$

   and determine the asymptotic limit

$$R = \lim_{N \uparrow \infty} R(N).$$

   Compare the result with Shannon's limit.

3.3 Solutions to Exercises

3.3.1 Asymptotic equipartition principle

1. Consider i.i.d. sequence $X^N = X_1, \dots, X_N$ whose entries are Bernoulli variables with

$$\Pr\{X=1\} = 1 - \Pr\{X=0\} = q.$$

   (a) Which sequences are typical for X^N?

   (b) Write the β-typical set T_{N_β} for this sequence.

   (c) Compare T_{N_β} to your intuitive definition in Part (a), when β is almost zero.

   (d) Write the three main properties of T_{N_β} that you learned in Chapter 4.

   (e) Using your definition in Part (a), derive

   - the number of typical sequences,
   - the probability of one typical sequence occurring,
   - the probability of X^N being in the typical set,

   and approximate it for large N using Stirling's approximation. Compare the result with Part (d).

♠ **Solution:**

To answer this question let us define for a given binary sequence x^N

$$N_0\left(x^N\right) = \text{Number of entries in } x^N \text{ which are zero}$$
$$N_1\left(x^N\right) = \text{Number of entries in } x^N \text{ which are one}$$

(a) Considering the distribution of X, any binary sequence x^N for which

$$\frac{N_0\left(x^N\right)}{N} \approx 1-q \qquad \text{and} \qquad \frac{N_1\left(x^N\right)}{N} \approx q$$

is *intuitively* typical.

(b) The definition of β-typical set is given in Eq. (4.29) in page 80 as

$$T_{N_\beta} = \left\{ x^N : \left| \frac{1}{N} \log \frac{1}{P\left(x^N\right)} - H \right| < \beta \right\}. \qquad \text{(EQ:A3)}$$

Here, H is the entropy of each entry of X^N which reads

$$H = H(X) = H_2(q) = q \log \frac{1}{q} + (1-q) \log \frac{1}{1-q}.$$

Moreover, $P\left(x^N\right)$ for a sequence x^N is given by

$$P\left(x^N\right) = \prod_{n=1}^{N} P\left(x_n\right) = \left[\Pr\{X=0\}\right]^{N_0\left(x^N\right)} \left[\Pr\{X=1\}\right]^{N_1\left(x^N\right)}$$
$$= (1-q)^{N_0\left(x^N\right)} q^{N_1\left(x^N\right)}$$

due to the fact that X^N is i.i.d.. Substituting into (EQ:A3), we have

$$\begin{aligned} T_{N_\beta} &= \left\{ x^N : \left| \frac{1}{N} \log \frac{1}{P\left(x^N\right)} - H \right| < \beta \right\} \\ &= \left\{ x^N : \left| \frac{1}{N} \log \frac{1}{(1-q)^{N_0(x^N)} q^{N_1(x^N)}} - q \log \frac{1}{q} - (1-q) \log \frac{1}{1-q} \right| < \beta \right\} \\ &= \left\{ x^N : \left| \frac{N_1\left(x^N\right)}{N} \log \frac{1}{q} + \frac{N_0\left(x^N\right)}{N} \log \frac{1}{1-q} - q \log \frac{1}{q} - (1-q) \log \frac{1}{1-q} \right| < \beta \right\} \\ &= \left\{ x^N : \left| \left(\frac{N_1\left(x^N\right)}{N} - q \right) \log \frac{1}{q} + \left(\frac{N_0\left(x^N\right)}{N} - (1-q) \right) \log \frac{1}{1-q} \right| < \beta \right\}. \end{aligned}$$

Hence the typical set consists of all binary sequences x^N for which

$$\left| f_1\left(x^N\right) \log \frac{1}{q} + f_0\left(x^N\right) \log \frac{1}{1-q} \right| < \beta \qquad \text{(EQ:B3)}$$

where we define

$$\begin{aligned} f_1\left(x^N\right) &:= \frac{N_1\left(x^N\right)}{N} - q \\ f_0\left(x^N\right) &:= \frac{N_0\left(x^N\right)}{N} - (1-q). \end{aligned}$$

(c) When β is almost zero, (EQ:B3) is satisfied if

$$f_1\left(x^N\right) \approx 0 \qquad \text{and} \qquad f_0\left(x^N\right) \approx 0.$$

This means that X^N is typical in this case, if

$$\frac{N_0\left(x^N\right)}{N} \approx 1-q \qquad \text{and} \qquad \frac{N_1\left(x^N\right)}{N} \approx q.$$

This recovers the intuitive definition in Part (a).

(d) The three main properties of T_{N_β} indicated in the lecture are

i. The number of sequences in T_{N_β} satisfies

$$\left(1 - \frac{\theta}{\beta^2 N}\right) 2^{N(H-\beta)} \leq \left|T_{N_\beta}\right| \leq 2^{N(H+\beta)}$$

for some constant θ.

ii. The probability of *a single β-typical sequence* x^N satisfies

$$2^{-N(H+\beta)} \leq P\left(x^N\right) \leq 2^{-N(H-\beta)}$$

iii. The probability of T_{N_β} happening satisfies

$$1 - \frac{\theta}{\beta^2 N} \leq \Pr\left\{X^N \in T_{N_\beta}\right\} \leq 1$$

for some constant θ.

Considering the distribution of X, for very small choices of β and very large N, these properties simply indicate that

i. The number of sequences in T_{N_β} reads

$$|T_{N_\beta}| \approx 2^{NH_2(q)}. \tag{PR:1}$$

ii. The probability of *a single* β*-typical sequence* x^N is

$$P\left(x^N\right) \approx 2^{-NH_2(q)}. \tag{PR:2}$$

iii. The probability of T_{N_β} happening reads

$$\Pr\left\{X^N \in T_{N_\beta}\right\} \approx 1. \tag{PR:3}$$

(e) In Part (a), we indicated that x^N is *intuitively* typical if

$$N_0\left(x^N\right) \approx (1-q)\,N \qquad \text{and} \qquad N_1\left(x^N\right) \approx qN.$$

By this definition, we can write

i. The number of typical sequences approximately equals to the number of binary sequences of length N whose number of one symbols is qN. Hence,

$$\#\text{ of typical seq. } \approx \binom{N}{qN}.$$

Using Stirling's approximation, we have

$$\#\text{ of typical seq. } \approx \binom{N}{qN} \approx 2^{NH_2(q)}.$$

ii. For the probability of a typical sequence, we can write

$$P\left(x^N\right) = q^{N_1\left(x^N\right)}\left(1-q\right)^{N_0\left(x^N\right)}.$$

Noting that for a typical sequence $N_0\left(x^N\right) \approx (1-q)\,N$ and $N_1\left(x^N\right) \approx qN$, we have

$$P\left(x^N\right) \approx q^{qN}\left(1-q\right)^{(1-q)N}.$$

Using the fact that $q = 2^{\log q}$, we can write

$$\begin{aligned} P\left(x^N\right) &\approx q^{qN}\left(1-q\right)^{(1-q)N} \\ &= 2^{N(q\log q + (1-q)\log(1-q))} \\ &= 2^{-NH_2(q)}. \end{aligned}$$

iii. Assuming that X^N is generated randomly, the probability of X^N being typical is then given by

$$\begin{aligned} \Pr\left\{X^N \in T_{N_\beta}\right\} &\approx \#\text{ of typical seq. } \times \Pr\left(X^N = \text{a typical seq.}\right) \\ &\approx 2^{NH_2(q)} \times 2^{-NH_2(q)} = 1. \end{aligned}$$

One can observe that these results recover the properties in (PR:1)-(PR:3).

3.3.2 Shannon's first theorem

♣ Exercise 1 covers Exercise 4.5. in page 74 and partially Exercise 4.15. and Exercise 4.16. in page 85.

1. Consider a ternary random variable X whose alphabet is $\mathcal{A}_X = \{a, b, c\}$ and is uniformly distributed, i.e.

$$\Pr\{X = a\} = \Pr\{X = b\} = \Pr\{X = c\} = \frac{1}{3}.$$

(a) What is the minimum compression ratio suggested by Shannon?

(b) X^N is an i.i.d. sequence. Let $B(N)$ denote the number of bits required to encode X^N via a binary code. Determine $B(N)$ for a *simple* binary code.

(c) Define

$$R(N) = \frac{B(N)}{N},$$

and determine the asymptotic limit

$$R = \lim_{N\uparrow\infty} R(N).$$

Compare the result with Shannon's limit.

♠ **Solution:**

(a) Consider an i.i.d. sequence X^N. Assume that you want to represent this sequence with a binary sequence of length $B(N)$. In this case, the compression rate is defined as

$$R(N) = \frac{B(N)}{N}.$$

The first theorem of Shannon indicates two things:

i. For very large N, you can encode your source without any loss, if

$$R(N) > H(X).$$

This means that in this case, there *exist a lossless uniquely decodeable* source code which encodes X^N to a binary sequence of length $B(N)$.

ii. For very large N, if

$$R(N) < H(X),$$

it is *impossible* to compress X^N into $B(N)$ bits without any loss.

Based on Shannon's theorem, we can conclude that the minimum possible compression rate is

$$H(X) = \log 3.$$

(b) For this ternary source, there exists in total 3^N different realizations of X^N. Hence, to encode the source without any loss, we need

$$B(N) = \lceil \log 3^N \rceil$$

bits to make sure that there exists at least one distinct binary sequence for each realization of X^N.

(c) The compression rate in this case reads

$$R(N) = \frac{\lceil \log 3^N \rceil}{N}.$$

Noting that $x \leq \lceil x \rceil \leq x + 1$, we can write

$$\frac{\log 3^N}{N} \leq R(N) \leq \frac{\log 3^N + 1}{N}.$$

Noting that $\log x^N = N \log x$, we can conclude that

$$\log 3 \leq R(N) \leq \log 3 + \frac{1}{N}.$$

Taking the limit of $N \uparrow \infty$, we finally have

$$\lim_{N\uparrow\infty} \log 3 \leq \lim_{N\uparrow\infty} R(N) \leq \lim_{N\uparrow\infty} \log 3 + \frac{1}{N}$$

or equivalently

$$\lim_{N\uparrow\infty} R(N) = \log 3.$$

This is exactly the same limit as Shannon said.

⇝ REMARK:

Shannon used a more generic version of the same idea to prove his first theorem. He called this idea *block coding* meaning that he does not encode the sequence symbol-by-symbol, but he waits long enough and encode the whole stream together.

3.4 Homework

The following exercises are suggested for further practice.

3.4.1 Primary exercises

1. Assume X is a uniform random variable taken from alphabet $\mathcal{A}_X = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$. Let X^N be an i.i.d. sequence of X.

   (a) Let $N = 9$, give an example of a typical sequence.

   (b) Let N be a multiple of 3. Define $\mathcal{N}_{\text{Typ}}$ as the number of sequences in which symbols a, b and c occur with same frequencies. Determine $\mathcal{N}_{\text{Typ}}$.

   ♠ **Solution:** With basic combinatorics, you can show that

$$\mathcal{N}_{\text{Typ}} = \frac{N!}{[(N/3)!]^3} = \binom{N}{N/3}\binom{2N/3}{N/3}.$$

2. Using Stirling's approximation formula, approximate $\mathcal{N}_{\text{Typ}}$ determined in the previous question, for large values of N. Compare the result with

$$\mathcal{N}_{\text{Typ}} \approx 2^{NH(X)}$$

   suggested by the AEP.

3.4.2 Further exercises from the textbook

- Chapter 4: Exercise 4.11., Exercise 4.12., Exercise 4.15. and Exercise 4.16.

Chapter 4

Source Coding Algorithms

This chapter investigates algorithmic approaches for source coding namely the symbol and stream codes. The contents are consistent with Chapters 5 and 6 of the textbook.

4.1 Brief Review of Main Concepts

There are in general two approaches to compress a sequence X^N:

- *Symbol coding* in which we encode every single symbol in X^N separately. This means we do the following

  $$X_1X_2 \dots X_N \mapsto C\left(X_1\right) C\left(X_2\right) \dots C\left(X_N\right)$$

  where $C\left(\cdot\right)$ is the coding algorithm. In this chapter, we consider the Huffman algorithm for symbol coding.

- *Stream coding* in which we encode the whole stream at once, i.e.,

  $$X_1X_2 \dots X_N \mapsto C\left(X_1X_2 \dots X_N\right)$$

  where $C\left(\cdot\right)$ is the coding algorithm. In this chapter, we consider two stream coding algorithms, namely Lempel-Ziv and arithmetic coding.

The details on each algorithm is given throughout the solutions.

Along with source coding, there are several transforms which are used in practice for better representation of a source sequence. These transforms takes X^N and sort its entries in a specific manner. The new sorted sequence has this feature that it is more effectively compressed by a stream code. The most well-known transform of such type is the Burrows-Wheeler transform which we address it in this chapter. The transform are re-transform is comprehensively illustrated in the solutions.

> ⇝ REMARK:
>
> It is important to note that the Burrows-Wheeler transform is only a transform which re-sort a given sequence. It does **not** perform any compression or source coding. The source coding task is usually performed by a stream code after the transform on the transformed sequence.

4.2 Exercises

4.2.1 Symbol codes

♣ Exercise 1 covers Exercise 5.8. in page 93 and partially Exercise 5.19. and Exercise 5.20. in page 102.
Exercise 2 covers Exercise 5.29. in page 103 and partially Exercise 5.21. in page 102.
Exercise 3 covers Exercise 5.22. in page 102 and Exercise 5.25. in page 103.

1. Consider a source X with alphabet $\mathcal{A}_X = \{a, b, c\}$ whose distribution is

$$\Pr\{X = a\} = \Pr\{X = b\} = \frac{1}{4}, \qquad \text{and} \qquad \Pr\{X = c\} = \frac{1}{2}.$$

The following two symbol codes are proposed to encode the outcomes of X:

X	a	b	c
$C_1(X)$	00	01	1
$C_2(X)$	0	01	1

(a) Which code is *not* uniquely decodable?

(b) Is code $C_1(X)$ optimal? What is its expected length $L(C_1, X)$?

(c) Consider i.i.d. sequence X^N with $N = 100$. Assume that X^N is β-typical with $\beta \approx 0$. Use the asymptotic equipartition property (AEP) to approximately determine the length of $C_1\left(X^N\right)$. Compare this length with the expected length of the code.

2. Consider random variable X in the previous exercise. Make a Huffman code for X^2.

3. Consider random variable X whose alphabet is $\mathcal{A}_X = \{a_1, \dots, a_I\}$. Let the probability distribution of X be

$$p_i = \Pr\{X = a_i\} = 2^{-l_i}$$

for some integers $\{l_1, \dots, l_I\}$.

(a) Using Kraft's inequality, show that there exist a uniquely decodeable code $C(X)$ for which

$$\text{Length}\,(C(a_i)) = l_i.$$

(b) Determine the expected length of this code $L(C, X)$, and compare it with Shannon's limit.

4.2.2 Arithmetic codes

♣ Exercise 1 covers partially Exercise 6.3. in page 118 and Exercise 6.7. in page 123.

1. Consider the following correlated source:
   The source generates a sequence of a and b. For the first symbol, a and b occur with same probabilities. When we have sequence s^N of length N which has N_{a} symbols a and N_{b} symbols b, the probability of the next symbol being a is given by

$$\Pr\left(S_{N+1} = \mathsf{a} | S^N = \mathbf{s}^N\right) = \begin{cases} \dfrac{1}{2^{N_{\mathsf{a}}-N_{\mathsf{b}}+1}} & N_{\mathsf{a}} \geq N_{\mathsf{b}} \\ 1 - \dfrac{1}{2^{N_{\mathsf{b}}-N_{\mathsf{a}}+1}} & N_{\mathsf{b}} > N_{\mathsf{a}} \end{cases}.$$

   Use arithmetic coding to find a binary source code for sequence

   abab

4.2.3 Lempel-Ziv codes

♣ Exercise 1 covers Exercise 6.5. in page 120.
Exercise 2 covers Exercise 6.6. in page 120.

1. Consider the following binary sequence

   00000000000001000000000000

   (a) Encode this sequence using the *standard (basic)* Lempel-Ziv algorithm.

   (b) Optimize your coding by omitting the redundant bits.

2. The following sequence represents a string encoded by the *standard (basic)* Lempel-Ziv algorithm:

   00101011101100100100011010101000011

   Decode this string to find the source sequence.

4.2.4 Burrows-Wheeler transform

1. Consider the string

   papa_pan

   (a) Find the Burrows-Wheeler transform of the string.

   (b) Explain how you can utilize the transformed version to compress the string.

   (c) Invert the transformed string to find back the original string.

4.3 Solutions to Exercises

4.3.1 Symbol codes

♣ Exercise 1 covers Exercise 5.8. in page 93 and partially Exercise 5.19. and Exercise 5.20. in page 102.
Exercise 2 covers Exercise 5.29. in page 103 and partially Exercise 5.21. in page 102.
Exercise 3 covers Exercise 5.22. in page 102 and Exercise 5.25. in page 103.

1. Consider a source X with alphabet $\mathcal{A}_X = \{a, b, c\}$ whose distribution is

$$\Pr\{X = a\} = \Pr\{X = b\} = \frac{1}{4}, \qquad \text{and} \qquad \Pr\{X = c\} = \frac{1}{2}.$$

The following two symbol codes are proposed to encode the outcomes of X:

X	a	b	c
$C_1(X)$	00	01	1
$C_2(X)$	0	01	1

(a) Which code is *not* uniquely decodeable?

(b) Is code $C_1(X)$ optimal? What is its expected length $L(C_1, X)$?

(c) Consider i.i.d. sequence X^N with $N = 100$. Assume that X^N is β-typical with $\beta \approx 0$. Use the asymptotic equipartition property (AEP) to approximately determine the length of $C_1(X^N)$. Compare this length with the expected length of the code.

♠ **Solution:**

(a) $C_2(X)$ is not uniquely decodeable. To show this, we can take two approaches:

i. The direct approach is to give an example of a codeword which is not decoded uniquely. To this end, we can write the codeword 01 which is decoded as both b and ac.

ii. The second approach is to use the Kraft's inequality given in page 95. Kraft's inequality indicates that for source X with $|\mathcal{A}_X| = I$, there exists a uniquely decodeable symbol code with code lengths $\{l_1, \dots, l_I\}$, if

$$\sum_{i=1}^{I} 2^{-l_i} \leq 1.$$

for $C_2(X)$, we have $I = 3$ and $l_1 = 1$, $l_2 = 2$ and $l_3 = 1$. Hence,

$$\sum_{i=1}^{I} 2^{-l_i} = \frac{1}{2} + \frac{1}{4} + \frac{1}{2} = \frac{3}{2} > 1.$$

This means that for a source X with three outcomes, there exists no uniquely decodeable symbol code with lengths $l_1 = 1$, $l_2 = 2$ and $l_3 = 1$. Consequently, any example of such symbol codes, including $C_2(X)$, is not uniquely decodeable.

$\rightsquigarrow$ REMARK:

Note that Kraft's inequality only gives statement on the existence of a uniquely decodeable symbol code. This means that if a symbol code *does not satisfy Kraft's* inequality, it is concluded that the code *is not uniquely decodeable*. However, if a particular symbol code satisfy it, we **cannot** conclude that it is uniquely decodeable.

(b) Based on the definition in page 97 of the textbook, a uniquely decodeable source code is optimal if

$$l_i = \log \frac{1}{p_i}$$

for $i = 1, \dots, I$. In our example, $I = 3$, we further have

$$l_1 = 2 = \log \frac{1}{p_1} \quad \text{and} \quad l_2 = 2 = \log \frac{1}{p_2} \quad \text{and} \quad l_3 = 1 = \log \frac{1}{p_3}.$$

Hence, the code is *optimal*.

$\rightsquigarrow$ REMARK:

Note that for optimal codes, Kraft's inequality becomes an equality. We see this in exercise 3 of this section.

The expected length is moreover given by Eq. (5.13) in page 97 which reads

$$L(C_1, X) = \sum_{i=1}^{3} p_i l_i = \sum_{i=1}^{3} p_i \log \frac{1}{p_i} = H(X) = 1.5.$$

(c) Consider sequence X^N with $N = 100$. Let us denote the number of symbols a, b and c occurring in X^N by N_a, N_b and N_c, respectively. From Exercise 1 in Section 2.3 in Tutorial 2, we know that when X^N is β-typical sequence with $\beta \approx 0$, we can write

$$N_a \approx N/4 = 25 \quad \text{and} \quad N_b \approx N/4 = 25 \quad \text{and} \quad N_5 \approx N/2 = 50.$$

In this case, the length of $C_1\left(X^N\right)$ is given by

$$\text{Length}\left(C_1\left(X^N\right)\right) \approx N_a \times 2 + N_b \times 2 + N_c \times 1 = 150.$$

This means that a sequence of length $N = 100$ is encoded to a binary sequence whose length is approximately $B(N) = 150$. As a result, the compression rate reads

$$R = \frac{B(N)}{N} = \frac{150}{100} = 1.5 = H(X).$$

This observation indicates that the compression rate achieved by $C_1(X)$ is equal to the source entropy. From the first theorem of Shannon, we know that this is the minimum possible compression rate. Hence, we can understand why $C_1(X)$ is called *optimal*.

2. Consider random variable X in the previous exercise. Make a Huffman code for X^2.

♠ **Solution:** X^2 has $3^2 = 9$ possible outcomes which read

X^2	aa	ab	ac	ba	bb	bc	ca	cb	cc
$P\left(X^2\right)$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{4}$

Hence, the Huffman tree for this source is

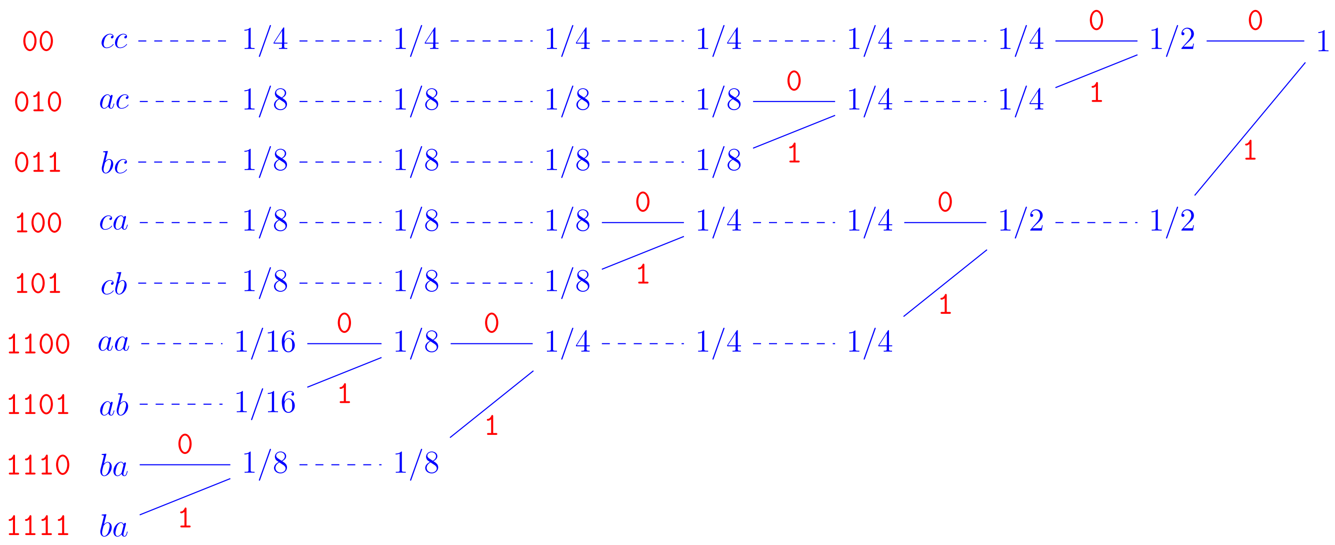

00 cc 1/4 1/4 1/4 1/4 1/4 1/4 0 1/2 0 1
010 ac 1/8 1/8 1/8 1/8 0 1/4 1/4 1
011 bc 1/8 1/8 1/8 1/8 1 1
100 ca 1/8 1/8 1/8 0 1/4 1/4 0 1/2 1/2
101 cb 1/8 1/8 1/8 1
1100 aa 1/16 0 1/8 0 1/4 1/4 1/4 1
1101 ab 1/16 1
1110 ba 0 1/8 1/8 1
1111 ba 1

Figure 4.3.1: Huffman coding algorithm.

The codewords have been shown with the red color next to the outcomes.

⇝ REMARK:

While constructing the Huffman tree, keep the following items in mind:

(a) To have a prefix code, the codewords are read from right to left on the tree.

(b) In each step, 0 and 1 are written in the same order.

3. Consider random variable X whose alphabet is $\mathcal{A}_X = \{a_1, \ldots, a_I\}$. Let the probability distribution of X be

$$p_i = \Pr\{X = a_i\} = 2^{-l_i}$$

for some integers $\{l_1, \ldots, l_I\}$.

(a) Using Kraft's inequality, show that there exist a uniquely decodeable code $C(X)$ for which

$$\text{Length}\,(C(a_i)) = l_i.$$

(b) Determine the expected length of this code $L(C, X)$, and compare it with Shannon's limit.

♠ **Solution:**

(a) Kraft's inequality indicates that, if

$$\sum_{i=1}^{I} 2^{-l_i} \le 1,$$

then, there exists a uniquely decodeable symbol code whose code-lengths are $l_1, \ldots, l_I$. Using the fact that $p_i = 2^{-l_i}$, we can write $l_i = -\log p_i$. Substituting into Kraft's inequality, we have

$$\sum_{i=1}^{I} 2^{\log p_i} = \sum_{i=1}^{I} p_i = 1.$$

⇝ REMARK:

Based on the definition, this code is *optimal*. For such a code, Kraft's inequality always becomes an *equality*.

(b) The expected length for such a code reads

$$\sum_{i=1}^{I} p_i l_i = \sum_{i=1}^{I} p_i \left(-\log p_i\right) = \sum_{i=1}^{I} p_i \log \frac{1}{p_i} = H(X).$$

This is equal to the limiting compression rate suggested by Shannon's theorem. Such an observation is obvious, since such a symbol code is *optimal*; see page 97 for the definition of an optimal symbol code.

4.3.2 Arithmetic codes

♣ Exercise 1 covers partially Exercise 6.3. in page 118 and Exercise 6.7. in page 123.

1. Consider the following correlated source:
The source generates a sequence of a and b. For the first symbol, a and b occur with same probabilities. When we have sequence s^N of length N which has $N_\mathtt{a}$ symbols a and $N_\mathtt{b}$ symbols b, the probability of the next symbol being a is given by

$$\Pr\left(S_{N+1} = \mathsf{a} | S^N = \mathsf{s}^N\right) = \begin{cases} \dfrac{1}{2^{N_\mathtt{a} - N_\mathtt{b} + 1}} & N_\mathtt{a} \ge N_\mathtt{b} \\ 1 - \dfrac{1}{2^{N_\mathtt{b} - N_\mathtt{a} + 1}} & N_\mathtt{b} > N_\mathtt{a} \end{cases}.$$

Use arithmetic coding to find a binary source code for sequence

abab

♠ **Solution:** To determine the arithmetic code, we first calculate all conditional probabilities. Starting with the first symbol we have

$$\begin{aligned}\Pr\{S_1=\mathtt{a}\} &= \frac{1}{2}\\ \Pr\{S_2=\mathtt{b}|S_1=\mathtt{a}\} &= 1-P(\mathtt{a}|\mathtt{a})=\frac{3}{4}\\ \Pr\{S_3=\mathtt{a}|S_1S_2=\mathtt{ab}\} &= \frac{1}{2}\\ \Pr\{S_4=\mathtt{b}|S_1S_2S_3=\mathtt{aba}\} &= 1-P(\mathtt{a}|\mathtt{aba})=\frac{3}{4}.\end{aligned}$$

We now find the interval which corresponds to abab

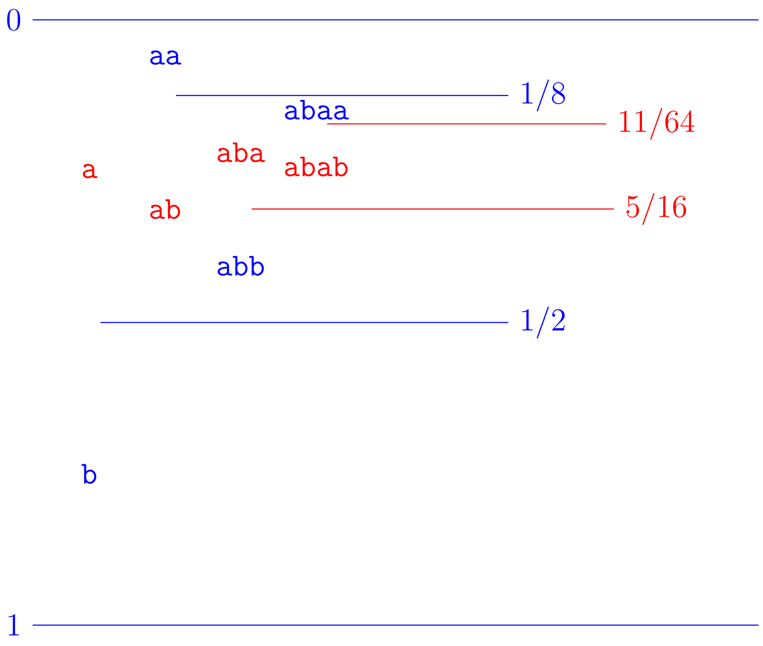


Figure 4.3.2: First part of arithmetic coding

As it is shown in the figure, the corresponding interval of abab is

$$I_{\mathtt{abab}}=\left[\frac{11}{64},\frac{20}{64}\right].$$

Now, we find a binary sequence whose interval lies into this interval.

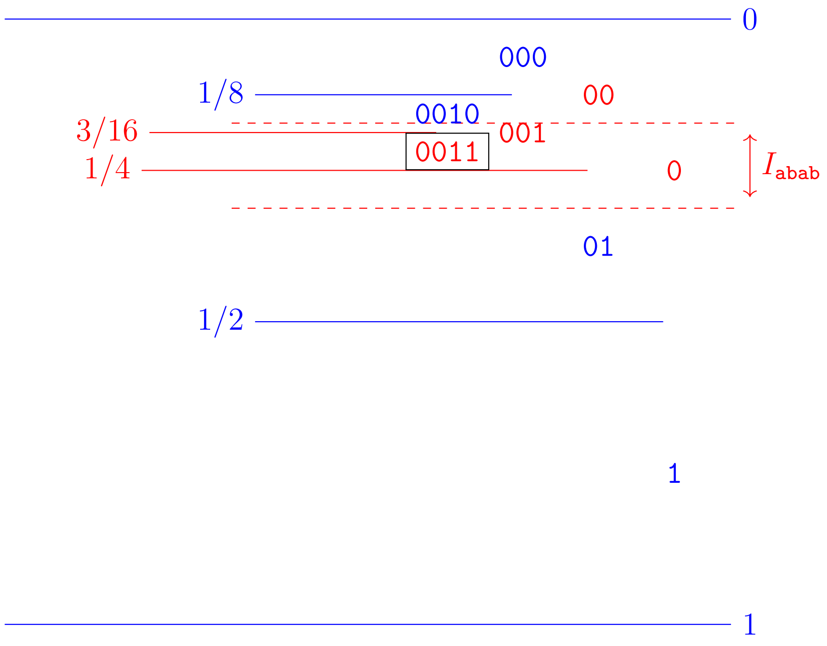


Figure 4.3.3: Second part of arithmetic coding

As it is shown in the second figure, the interval for `0011` is

$$I_{\texttt{0011}} = \left[\frac{12}{64}, \frac{16}{64}\right]$$

which lies into $I_{\texttt{abab}}$, i.e. $I_{\texttt{0011}} \subseteq I_{\texttt{abab}}$. Hence, `0011` is an arithmetic code for `abab`.

⇝ REMARK:

As in Huffman coding, there are multiple arithmetic codes for a given source stream. These different codes are found by considering different prior assumptions. For instance in this exercise, we could have alternatively mapped the upper intervals to `b` and the lower intervals to `a`. which would have resulted in a different interval for `abab`. The corresponding arithmetic code would have been then different. This fact however does not impact the decoding procedure. In fact, for a given set of prior assumption, any encoded stream is uniquely decoded.

4.3.3 Lempel-Ziv codes

♣ Exercise 1 covers Exercise 6.5. in page 120.
Exercise 2 covers Exercise 6.6. in page 120.

1. Consider the following binary sequence

$$\texttt{000000000000100000000000}$$

(a) Encode this sequence using the *standard (basic)* Lempel-Ziv algorithm.

(b) Optimize your coding by omitting the redundant bits.

♠ **Solution:**

(a) The standard Lempel-Ziv coding for this source is as follows:

- We start Lempel-Ziv coding by parsing the stream into sub-strings, such that each sub-string has not been repeated before. We further assume the empty sub-string, denoted by λ, happening prior to all symbols. Hence the sub-strings are

$$\lambda,\ \texttt{0},\ \texttt{00},\ \texttt{000},\ \texttt{0000},\ \texttt{001},\ \texttt{00000},\ \texttt{000000}$$

- We now construct the codebook by indexing each of the sub-strings with a pointer: Let S denote i-th sub-string starting from λ. Then, the pointer of S is

$$p(S) = i - 1.$$

This means that

S	λ	0	00	000	0000	001	00000	000000
$p(S)$	0	1	2	3	4	5	6	7

- Now we encode each sub-string as following: Consider sub-string S, this sub-string is written as

$$S = S_0, \texttt{b}$$

where S_0 is a sub-string occurred before S, i.e. $p(S_0) \leq p(S)$, and $\texttt{b}$ is a new bit. We encode, S as

$$C(S) = \mathrm{Binary}\left[p(S_0)\right], \texttt{b}$$

where $\mathrm{Binary}\left[p(S_0)\right]$ is the binary representation of $p(S_0)$ using $\lceil \log p(S) \rceil$ bits. This means

S	λ	0	00	000	0000	001	00000	000000
$S_0, \texttt{b}$	λ, λ	λ, 0	0, 0	00, 0	000, 0	00, 1	0000, 0	00000, 0
$p(S_0), \texttt{b}$	$0, \lambda$	0, 0	1, 0	2, 0	3, 0	2, 1	4, 0	6, 0
$p(S)$	0	1	2	3	4	5	6	7
$\lceil \log p(S) \rceil$	λ	0	1	2	2	3	3	3
$\mathrm{Binary}\left[p(S_0)\right]$	λ	0	1	10	11	010	100	110
$C(S)$	λ	0	1, 0	10, 0	11, 0	010, 1	100, 0	110, 0

⇝ REMARK:

Keep in mind that $p(S_0)$ is represented by $\lceil \log p(S) \rceil$ bits.

Hence, the encoded stream is

$$\texttt{010100110010110001100}$$

(b) To reduce the length of the encoded sequence, we note that

i. Both of the children of sub-string 2, i.e., $S = \texttt{00}$, are in the code-book (sub-strings 3 and 5). We hence can shrink the code-book by removing sub-string $S = \texttt{00}$ and assigning it to the previous pointer. This means, we can optimize the code-book as

S	λ	0 and 00	000	0000	001	00000	000000
$p(S)$	0	1	2	3	4	5	6

With this new code-book, the source stream is encoded as

$$\texttt{001010001101101010}$$

ii. In this new code-book, the new bit in sub-string 4 carry no information. This bit is denoted in red. To see this point, note that the sub-strings, $S = \texttt{00}$ and $S = \texttt{000}$ have already occurred. This means that if another sub-string with prefix $S = \texttt{00}$ occurs, it should be $\texttt{001}$. We can hence further remove this redundant bit and encode the source stream into

$$\texttt{00101000101101010}$$

⇝ REMARK:

Note that this optimized version of the Lempel-Ziv considers a set of assumptions for code-book reduction. These assumptions are required to be known in advance for decoding. This means that if you intend to decode the encoded sequence, you require to be informed about them.

2. The following sequence represents a string encoded by the *standard (basic)* Lempel-Ziv algorithm:

$$\texttt{00101011101100100100011010101000011}$$

Decode this string to find the source sequence.

♠ **Solution:** For decoding, we should reverse the procedure of standard Lempel-Ziv coding.

- We start decoding by finding the encoded sub-strings. To this end, we remember that the sub-string S after encoding is of the length $\lceil \log p(S) \rceil + 1$. Hence, the encoded sub-strings are

$$\lambda,\ \texttt{0},\ \texttt{01},\ \texttt{010},\ \texttt{111},\ \texttt{0110},\ \texttt{0100},\ \texttt{1000},\ \texttt{1101},\ \texttt{01010},\ \texttt{00011}$$

- We now construct the codebook stepwise as following:

(a) Start with $S_1 = \lambda$ which reads $p(\lambda) = 0$.

(b) The i-th encoded sub-frame $C(S_i)$ has $p(S_i) = i - 1$ and consists of two parts: the last bit which is the new bit and the previous $\lceil \log p(S) \rceil$ bits which are the binary representation of the pointer of a previous sub-string.

For the given example we can write

(a) $S_1 = \lambda$ has $p(S_1) = 0$.

(b) Next encoded sub-string is $C(S_2) =$ 0 and has $p(S_2) = 1$.
It consists of two parts: empty part with length $\lceil \log p(S_2) \rceil = 0$ bits which represents the pointer of $S_1 = \lambda$, and the last bit 0 which is a new bit. Thus, $S_2 = \lambda,$ 0.

(c) Next encoded sub-string is $C(S_3) =$ 01 and has $p(S_3) = 2$.
It consists of two parts: part 0 with length $\lceil \log p(S_3) \rceil = 1$ bits which represents the pointer of $S_1 = \lambda$, and the last bit 1 which is a new bit. Thus, $S_3 = \lambda,$ 1.

(d) Next encoded sub-string is $C(S_4) =$ 010 and has $p(S_4) = 3$.
It consists of two parts: part 01 with length $\lceil \log p(S_4) \rceil = 2$ bits which represents the pointer of $S_2 =$ 0, and the last bit 0 which is a new bit. Thus, $S_4 =$ 00.

(e) Next encoded sub-string is $C(S_5) =$ 111 and has $p(S_5) = 4$.
It consists of two parts: part 11 with length $\lceil \log p(S_5) \rceil = 2$ bits which represents the pointer of $S_4 =$ 00, and the last bit 1 which is a new bit. Thus, $S_5 =$ 001.

(f) Next encoded sub-string is $C(S_6) =$ 0110 and has $p(S_6) = 5$.
It consists of two parts: part 011 with length $\lceil \log p(S_6) \rceil = 3$ bits which represents the pointer of $S_4 =$ 00, and the last bit 0 which is a new bit. Thus, $S_6 =$ 000.

(g) Next encoded sub-string is $C(S_7) =$ 0100 and has $p(S_7) = 6$.
It consists of two parts: part 010 with length $\lceil \log p(S_7) \rceil = 3$ bits which represents the pointer of $S_3 =$ 1, and the last bit 0 which is a new bit. Thus, $S_7 =$ 10.

(h) Next encoded sub-string is $C(S_8) =$ 1000 and has $p(S_8) = 7$.
It consists of two parts: part 100 with length $\lceil \log p(S_8) \rceil = 3$ bits which represents the pointer of $S_5 =$ 001, and the last bit 0 which is a new bit. Thus, $S_8 =$ 0010.

(i) Next encoded sub-string is $C(S_9) =$ 1101 and has $p(S_9) = 8$.
It consists of two parts: part 110 with length $\lceil \log p(S_9) \rceil = 3$ bits which represents the pointer of $S_7 =$ 10, and the last bit 1 which is a new bit. Thus, $S_9 =$ 101.

(j) Next encoded sub-string is $C(S_{10}) =$ 01010 and has $p(S_{10}) = 9$.
It consists of two parts: part 0101 with length $\lceil \log p(S_{10}) \rceil = 4$ bits which represents the pointer of $S_6 =$ 0000, and the last bit 0 which is a new bit. Thus, $S_{10} =$ 0000.

(k) Next encoded sub-string is $C(S_{11}) =$ 00011 and has $p(S_{11}) = 10$.
It consists of two parts: part 00011 with length $\lceil \log p(S_{11}) \rceil = 4$ bits which

represents the pointer of $S_6 = 0001$, and the last bit 1 which is a new bit. Thus, $S_{11} = 01$.

Hence, the codebook reads

S	λ	0	1	00	001	000	10	0010	101	0000	01
$p(S)$	0	1	2	3	4	5	6	7	8	9	10

- The decoded stream then reads

0100001000100010101000001

4.3.4 Burrows-Wheeler transform

1. Consider the string

papa_pan

(a) Find the Burrows-Wheeler transform of the string.

(b) Explain how you can utilize the transformed version to compress the string.

(c) Invert the transformed string to find back the original string.

♠ **Solution:**

(a) To transform papa_pan, we follow the following steps:

- We first add a pointer at the point where it ends. We show this pointer with $. Hence,

papa_pan$

- We now write all possible cyclic shifts of the stream

papa_pan$
$papa_pan
n$papa_pa
an$papa_p
pan$papa_
_pan$papa
a_pan$pap
pa_pan$pa
apa_pan$p

- Then we sort all the shifts considering $\$$ to be least valued, then _ and then letters with respect to their order. Thus, we have

$$\begin{array}{l}
\$\,p\,a\,p\,a\,\_\,p\,a\,n \\
\_\,p\,a\,n\,\$\,p\,a\,p\,a \\
a\,\_\,p\,a\,n\,\$\,p\,a\,p \\
a\,n\,\$\,p\,a\,p\,a\,\_\,p \\
a\,p\,a\,\_\,p\,a\,n\,\$\,p \\
n\,\$\,p\,a\,p\,a\,\_\,p\,a \\
p\,a\,\_\,p\,a\,n\,\$\,p\,a \\
p\,a\,n\,\$\,p\,a\,p\,a\,\_ \\
p\,a\,p\,a\,\_\,p\,a\,n\,\$
\end{array} \tag{BWT}$$

- The Burrows-Wheeler transform is the last column of this sorted matrix. This means that

$$\mathrm{BWT}\left(\mathrm{papa\_pan}\right) = \mathrm{napppaa\_\$}$$

(b) As it is observed, in the transformed sequence, similar letters are next to each other. This property will reduce the length of the encoded stream when we use some stream codes, such as Lempel-Ziv codes, for encoding.

> ⇝ REMARK:
>
> Burrows-Wheeler is only a transform and does not do any source coding.

(c) To invert the transformed stream, we take the following steps:

- We start from the transformed. By sorting this stream we have

$$\begin{array}{c}
\$ \\
\_ \\
a \\
a \\
a \\
n \\
p \\
p \\
p
\end{array}$$

which is the first column of (BWT).

- Now, we put this new sorted column on the right of the transformed stream which results in

$$\begin{array}{cc}
\texttt{n} & \$ \\
\texttt{a} & \_ \\
\texttt{p} & \texttt{a} \\
\texttt{p} & \texttt{a} \\
\texttt{p} & \texttt{a} \\
\texttt{a} & \texttt{n} \\
\texttt{a} & \texttt{p} \\
\_ & \texttt{p} \\
\$ & \texttt{p}
\end{array}$$

By sorting the rows of this new matrix, we have

$$\begin{array}{cc}
\$ & \texttt{p} \\
\_ & \texttt{p} \\
\texttt{a} & \_ \\
\texttt{a} & \texttt{n} \\
\texttt{a} & \texttt{p} \\
\texttt{n} & \$ \\
\texttt{p} & \texttt{a} \\
\texttt{p} & \texttt{a} \\
\texttt{p} & \texttt{a}
\end{array}$$

By comparing the new matrix to (BWT), it is observed that this matrix recovers the first two columns of (BWT).

- Repeat the previous step, until the resulting matrix is a square matrix. This means that
  i. Put the two columns matrix in the previous step next to the transformed stream at right, and then sort the rows of this new three-columns matrix.
  ii. Again, put the sorted three-columns matrix on the right of the transformed stream and sort the rows of this four-column matrix.
  iii. Keep on doing so, until the sorted matrix has same number of columns and rows.
- Take the first row of the matrix and remove the pointer, i.e. $\$$ which results in

$$\mathrm{BWT}^{-1}\left(\texttt{napppaa\_\$}\right) = \texttt{papa\_pan}$$

4.4 Homework

The following exercises are suggested for further practice.

4.4.1 Primary exercises

1. Give an example of a symbol code for source X with $\mathcal{A}_X = \{x_1, \dots, x_I\}$, which satisfies Kraft inequality, but *is not uniquely decodeable*.

2. Consider source X with $\mathcal{A}_X = \{x_1, \dots, x_I\}$. Let $p_i = \Pr\{X = x_i\}$. Start from the following definition:

   Uniquely decodeable symbol code $C(X)$ is strictly optimal, if $l_i = -\log p_i$, where l_i is the length of $C(x_i)$.

   Show that

   If uniquely decodeable symbol code $C(X)$ is strictly optimal; then, the Kraft inequality becomes an equality.

3. Let sequence

$$x^N = x_1, \dots, x_N$$

be a realization of an i.i.d. sequence whose entries are distributed with $P(x)$. Assume that we use arithmetic codes to encode x^N. Let the length of the codeword be L.

(a) Show that

$$L - \log\left(\frac{1}{P(x^N)}\right) \leq 2$$

(b) Is the codeword necessarily unique?

♠ **Solution:**

(a) Using arithmetic coding, the interval corresponding to x^N, i.e. I_{x^N}, is an interval of length $P(x^N)$. In the worst case, this interval could be just inside the interval of a binary sequence; see the following figure.

Let us denote this new interval with $I_{\mathtt{b}^M}$, where $\mathtt{b}^M$ is the corresponding binary sequence which is of length M. Noting that length of $I_{\mathtt{b}^M}$ is more than $P(x^N)$, we can conclude that the number of binary sequences are less than

$$\# \text{ of binary sequences } \leq \frac{1}{P(x^N)}.$$

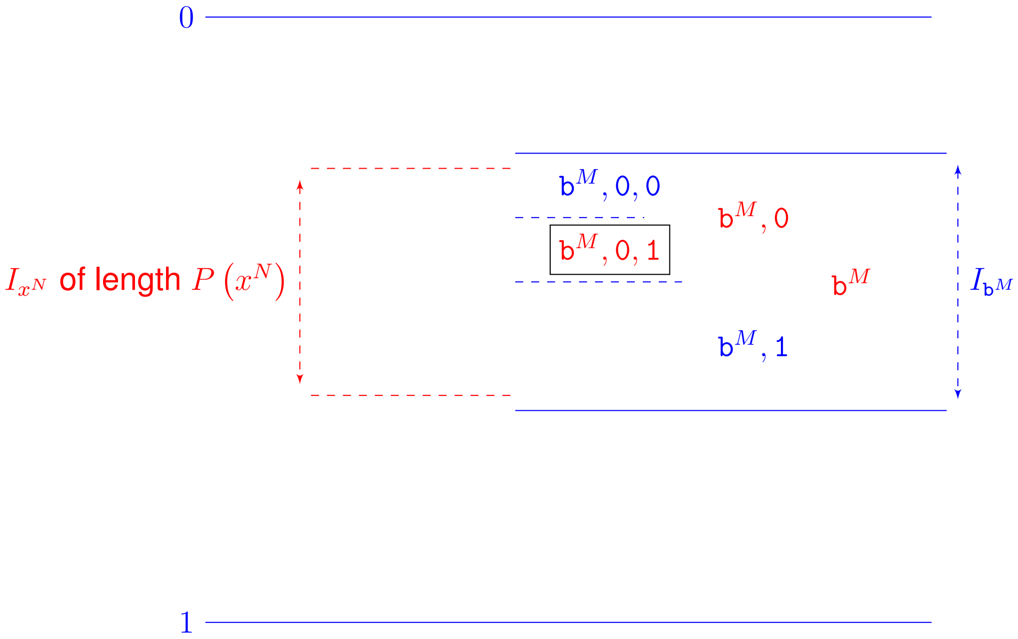


Figure 4.4.1: Worst-case scenario for arithmetic coding.

Furthermore, we know that the number of intervals for binary sequences of length M is 2^M. Hence, we could conclude that

$$2^M \leq \frac{1}{P\left(x^N\right)},$$

or equivalently

$$M \leq \log \frac{1}{P\left(x^N\right)}.$$

As indicated in the figure, to find the binary code in this worst case, we need to pursue binary interval construction for two further steps. Hence, the resulted codeword will be of length

$$L = M + 2 \leq \log \frac{1}{P\left(x^N\right)} + 2.$$

As a result, we have

$$L - \log \frac{1}{P\left(x^N\right)} \leq 2.$$

(b) Not necessarily.
For instance, in the worst-case scenario considered here, one could have given $\mathtt{b}^M,\mathtt{1},\mathtt{0}$ as the codeword, by choosing $\mathtt{b}^M,\mathtt{1}$ in the first step after breaking the interval of $\mathtt{b}^M$ into two equally-sized sub-intervals.

4. Consider the previous exercise. Show that

$$L - \log \frac{1}{P\left(x^N\right)} \geq 0.$$

♠ **Solution:** Follow the same approach as the one given above. You should however consider the best-case scenario instead of the worst case.

5. Find the Burrows-Wheeler transform of

   `pfefferpfanne`

   Recover the string by inverting the transform.

♠ **Solution:** The Burrows-Wheeler transform is

`efnffppfenar$e`

4.4.2 Further exercises from the textbook

- Chapter 5: Exercise 5.14., Exercise 5.19., Exercise 5.20., Exercise 5.21., Exercise 5.28., Exercise 5.31. and Exercise 5.32.
- Chapter 6: Exercise 6.8., Exercise 6.10., Exercise 6.15., Exercise 6.16. and Exercise 6.22.

4.4.3 Practicing by MATLAB

If you wish, you could do some practice with MATLAB:

- Take a string of letters as the input.
- Determine the frequency of each letter, and consider it as its probability.
- Use these probabilities and encode the string via Huffman coding.

4.5 Fun Facts

- David A. Huffman was a Ph.D. student at MIT. In 1951, he took the course “information theory” which was being lectured by Prof. Robert M. Fano. For final evaluation, he was given two choices, either giving a final exam or doing a term paper on *finding the most efficient binary code*. He took the second choice. However, he found out that the topic is complicated, and actually, Fano and Shannon had already worked on it. He was just to give up and start studying for the exam that the idea of Huffman coding came to his mind!

- The software `bzip` uses Burrows-Wheeler transform to put frequently recurring characters next to each other and then uses arithmetic coding to encode the string. In version `bzip2`, the encoder is changed to Huffman, due to patent restrictions.
- Check `http://guanine.evolbio.mpg.de/cgi-bin/bwt/bwt.cgi.pl` to find Burrows-Wheeler transforms and their inverses online.

Chapter 5

Information of Multiple Processes

This chapter starts with more advanced information theoretic definitions and tools such as conditional entropy and mutual information. The contents of this chapter are consistent with Chapter 8 of the textbook.

5.1 Brief Review of Main Concepts

The basic entropy definition is fundamental; however, it is insufficient when it comes to analysis of jointly dependent random variables. We hence define two new information-theoretic metrics:

- Conditional entropy of Y given X which describe the amount of ambiguity remained in Y if we know X and is defined as

$$H\left(Y|X\right) = \sum_{\substack{x\in\mathcal{A}_X\\ y\in\mathcal{A}_Y}} P\left(x,y\right)\log\frac{1}{P\left(y|x\right)}.$$

- Mutual information of X and Y which describe the amount of information obtained about Y if we know X and is defined as

$$I\left(X;Y\right) = H\left(Y\right) - H\left(Y|X\right)$$

There are various tricks and techniques to calculate these parameters and interpret them. We will discuss them throughout the solutions in this chapter.

5.2 Exercises

5.2.1 Conditional entropy and mutual information

♣ Exercise 1 covers Exercise 8.1. in page 140.
Exercise 2 covers Exercise 8.3. in page 140.
Exercise 3 covers Exercise 8.4. in page 140.

1. Let random variables X and Y be $X = (U, V)$ and $Y = (W, V)$, where U, V and W are *independent* random variables with alphabets $\mathcal{A}_U$, $\mathcal{A}_V$ and $\mathcal{A}_W$, respectively.

   (a) Are X and Y independent?

   (b) Calculate $H(X)$, $H(Y)$ and $H(X, Y)$. Compare $H(X, Y)$ to $H(X) + H(Y)$.

   (c) Calculate $H(Y|X)$, and compare $H(X, Y)$ to $H(X) + H(Y|X)$.

   (d) Calculate $I(X; Y)$.

2. *Chain rule* is one of the basic tools in information theory which enables us to expand a *multi-letter* expression as the sum of multiple *single-letter* terms. To illustrate this rule, we focus on a simple case with two variables.

   Let X and Y be two *dependent* random variables with alphabets $\mathcal{A}_X$ and $\mathcal{A}_Y$, respectively.

   (a) Show that

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

   (b) Using this result, show that $H(Y) \geq H(Y|X)$.

3. Consider *dependent* variables X and Y.

   (a) Starting from $I(X; Y) = H(X) - H(X|Y)$, find two equivalent definitions for the mutual information.

   (b) Show that $I(X; Y) = I(Y; X)$.

   (c) Show that $I(X; Y) \geq 0$.

4. *Chain rule* can be further derived for the mutual information. Let X, Y and Z be three *dependent* variables.

   (a) Write the definitions of $I(X, Z; Y)$ and $I(X; Y|Z)$.

   (b) Show that

   $$I(X, Z; Y) = I(Z; Y) + I(X; Y|Z).$$

5.2.2 Information theoretic identities and inequalities

♣ Exercise 1 covers partially Exercise 8.7. in page 141.
Exercise 2 covers Exercise 8.9. in page 141.

1. Show that $I(X; Y) = 0$ if and only if X and Y are *independent*.

2. The *data-processing* theorem is a key inequality in information theory. We study it in this exercise.

Let X, Y and Z be *dependent* random variables with alphabets $\mathcal{A}_X$, $\mathcal{A}_Y$ and $\mathcal{A}_Z$, respectively. Assume that

$$X \rightarrow Y \rightarrow Z$$

form a Markov chain.

(a) Factorize the joint probability distribution of (X, Y, Z) by considering the property of Markov chains.

(b) Show that $H(Z|Y, X) = H(Z|Y)$.

(c) Determine $I(X; Z|Y)$.

(d) Show that

$$I(X; Z) \leq I(X; Y).$$

5.3 Solutions to Exercises

5.3.1 Conditional entropy and mutual information

♣ Exercise 1 covers Exercise 8.1. in page 140.
Exercise 2 covers Exercise 8.3. in page 140.
Exercise 3 covers Exercise 8.4. in page 140.

1. Let random variables X and Y be $X = (U, V)$ and $Y = (W, V)$, where U, V and W are *independent* random variables with alphabets $\mathcal{A}_U$, $\mathcal{A}_V$ and $\mathcal{A}_W$, respectively.

   (a) Are X and Y independent?

   (b) Calculate $H(X)$, $H(Y)$ and $H(X, Y)$. Compare $H(X, Y)$ to $H(X) + H(Y)$.

   (c) Calculate $H(Y|X)$, and compare $H(X, Y)$ to $H(X) + H(Y|X)$.

   (d) Calculate $I(X; Y)$.

♠ **Solution:**

(a) Clearly, X and Y are *dependent*, since the random variable V is common between them. This statement however can be justified by following the definition of independent random variables. We know that the random variables X and Y are independent, if

$$P(x, y) = P(x)\, P(y)\,.$$

Here, $P(x, y)$ denotes the *joint distribution* of X and Y, and $P(x)$ and $P(y)$ are the *marginal distributions* of X and Y, respectively. For the marginal distributions we can write

$$P(x) \overset{\text{(a)}}{=} P(u, v) = P(u)\, P(v)\,,$$

$$P(y) \overset{\text{(b)}}{=} P(w, v) = P(w)\, P(v)\,,$$

where (a) and (b) follow the fact that U, V and W are independent. In this case, we have

$$P(x)\, P(y) = P(u)\, P(w)\, P(v)^2 .$$

The joint distribution of X and Y is moreover given by

$$P(x, y) = P(u, w, v) = P(u)\, P(w)\, P(v) .$$

Comparing this to the product of the marginal distributions, we see that in this case

$$P(x, y) \neq P(x)\, P(y) .$$

Hence X and Y are dependent.

(b) To calculate the entropies, one could use directly the definition of the entropy. However, using the known identities, one could determine them more straightforwardly. Staring with $H(X)$, we have

$$H(X) = H(U, V) \overset{\text{(c)}}{=} H(U) + H(V)$$

where (c) follows the fact that U and V are independent; check equation (2.39) in page 33 of the textbook. Similarly, we can write $H(Y)$ and $H(X, Y)$

$$\begin{aligned} H(Y) &= H(W, V) = H(W) + H(V), \\ H(X, Y) &= H(U, W, V) = H(U) + H(W) + H(V). \end{aligned}$$

Compare $H(X, Y)$ to $H(X) + H(Y)$, we can see that

$$H(X) + H(Y) = H(U) + H(W) + 2H(V) > H(X, Y) = H(U) + H(W) + H(V). \quad \text{(EQ:A5)}$$

The above observation is intuitive. In fact, $H(X)$ evaluate the amount of information obtained by observing X. Similarly, $H(Y)$ denotes the number of bits, we would get to know, if we see Y. However, $H(X, Y)$ describes the number of information bits being known when we see X and Y *together*. Noting that V is common between X and Y and repeats in both the $H(X)$ and $H(Y)$, we conclude the number of information bits after seeing both X and Y together is less than $H(X) + H(Y)$.

The inequality in (EQ:A5) is in fact generic. The generalized form of this inequality is

⇝ REMARK:

For random variables $X_1, \dots, X_N$, we have

$$\sum_{n=1}^{N} H(X_n) \geq H(X_1, \dots, X_N) .$$

The equality happens if and only if $X_1, \dots, X_N$ are jointly independent.

(c) To calculate $H(Y|X)$, we follow the definition in equation (8.4) in page 138. Doing so, we have

$$H(Y|X) = \sum_{\substack{x \in \mathcal{A}_X \\ y \in \mathcal{A}_Y}} P(x,y) \log \frac{1}{P(y|x)}. \tag{EQ:B5}$$

Noting that $X = (U, V)$ and $Y = (W, V)$, we have

$$P(y|x) = P(w, v|u, v).$$

By using the chain rule for the probability distributions, given in equation (2.6) page 24, we have

$$P(y|x) = P(w|u, v) P(v|w, u, v).$$

Since W is independent of U and V, we can write

$$P(w|u, v) = P(w).$$

Moreover, by the definition, it is clear that[1]

$$P(v|w, u, v) = 1.$$

Consequently, we can write

$$P(y|x) = P(w).$$

Furthermore, we know that $P(x, y) = P(w, u, v)$. Substituting into (EQ:B5), we have

$$H(Y|X) = \sum_{\substack{w \in \mathcal{A}_W \\ u \in \mathcal{A}_U \\ v \in \mathcal{A}_V}} P(w,u,v) \log \frac{1}{P(w)} = \sum_{w \in \mathcal{A}_W} \underbrace{\sum_{\substack{u \in \mathcal{A}_U \\ v \in \mathcal{A}_V}} P(w,u,v)}_{P(w)} \log \frac{1}{P(w)}$$

$$= \sum_{w \in \mathcal{A}_W} P(w) \log \frac{1}{P(w)} = H(W).$$

This result is again intuitive. In fact, $H(Y|X)$ represents the following quantity:

> ⇝ REMARK:
>
> Assume you have seen X, but not Y. The conditional entropy $H(Y|X)$ denotes the number of **extra** information bits which you will know after you see also Y.

[1] Note that here v is conditioned to itself, so the probability of $V = v$ given that $V = v$ is always one.

Following this interpretation, you know that by seeing X you have already known U and V. As $Y = (W, V)$, after seeing Y you only get $H(W)$ extra information bits.

To compare $H(X, Y)$ with $H(X) + H(Y|X)$, we write

$$H(X) + H(Y|X) = H(U) + H(V) + H(W) = H(X, Y).$$

This is again a generic identity in information theory known as the *chain rule* for entropy. The general form of this chain rule says

> ⇝ REMARK:
>
> For random variables $X_1, \ldots, X_N$, we have
>
> $$H(X_1, \ldots, X_N) = H(X_1) + \sum_{n=2}^{N} H(X_n | X_1, \ldots, X_{n-1}).$$

(d) To calculate $I(X; Y)$, we invoke the definition in equation (8.8), page (139). For the mutual information, we have

$$\begin{aligned} I(X; Y) = I(Y; X) &= H(Y) - H(Y|X) \\ &= H(W) + H(V) - H(W) = H(V). \end{aligned}$$

This result is again intuitive, since $I(X; Y)$ is interpreted as follows:

> ⇝ REMARK:
>
> Assume you have **not** seen X, **nor** Y. The mutual information $I(X; Y)$ denotes the number of information bits you would know about Y, if you see X.

In our example, if we see X, we would know about U and V. Since $Y = (W, V)$, by seeing X we would also know $H(V)$ bits of information about Y.

2. *Chain rule* is one of the basic tools in information theory which enables us to expand a *multi-letter* expression as the sum of multiple *single-letter* terms. To illustrate this rule, we focus on a simple case with two variables.

   Let X and Y be two *dependent* random variables with alphabets $\mathcal{A}_X$ and $\mathcal{A}_Y$, respectively.

   (a) Show that

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

   (b) Using this result, show that $H(Y) \geq H(Y|X)$.

♠ **Solution:**

(a) To show the chain rule, we start by the definitions of $H(X,Y)$ which is

$$H(X,Y) = \sum_{\substack{x\in\mathcal{A}_X\\ y\in\mathcal{A}_Y}} P(x,y)\log\frac{1}{P(x,y)}.$$

By the chain rule for the distributions, we have $P(x,y) = P(x)P(y|x)$. Hence, $H(X,Y)$ reduces to

$$\begin{aligned}
H(X,Y) &= \sum_{\substack{x\in\mathcal{A}_X\\ y\in\mathcal{A}_Y}} P(x,y)\log\frac{1}{P(x)P(y|x)}\\
&= \sum_{\substack{x\in\mathcal{A}_X\\ y\in\mathcal{A}_Y}} P(x,y)\left(\log\frac{1}{P(x)} + \log\frac{1}{P(y|x)}\right)\\
&= \sum_{\substack{x\in\mathcal{A}_X\\ y\in\mathcal{A}_Y}} P(x,y)\log\frac{1}{P(x)} + \underbrace{\sum_{\substack{x\in\mathcal{A}_X\\ y\in\mathcal{A}_Y}} P(x,y)\log\frac{1}{P(y|x)}}_{H(Y|X)}\\
&= \sum_{x\in\mathcal{A}_X}\underbrace{\sum_{y\in\mathcal{A}_Y} P(x,y)}_{P(x)}\log\frac{1}{P(x)} + H(Y|X)\\
&= \underbrace{\sum_{x\in\mathcal{A}_X} P(x)\log\frac{1}{P(x)}}_{H(X)} + H(Y|X) = H(X) + H(Y|X).
\end{aligned}$$

Alternatively, starting with the other form of the chain rule for distribution, i.e. $P(x,y) = P(y)P(x|y)$, we can conclude the alternative form of the chain rule for entropy which says

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

(b) To show that $H(Y) \geq H(Y|X)$, we start with the identity

$$H(X) + H(Y) \geq H(X,Y)$$

which we learned in the previous tutorial. Using the chain rule, we have

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

This inequality is a famous inequality from the literature, which says

⇝ REMARK:

Conditioning always reduces the entropy, i.e. $H(Y) \geq H(Y|X)$.

3. Consider *dependent* variables X and Y.

   (a) Starting from $I(X;Y) = H(X) - H(X|Y)$, find two equivalent definitions for the mutual information.

   (b) Show that $I(X;Y) = I(Y;X)$.

   (c) Show that $I(X;Y) \geq 0$.

♠ **Solution:**

(a) Alternative definitions can be derived for the mutual information by means of the chain rule. To this end, we start with

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

By adding and subtracting $H(Y)$, we have

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

Hence, the first equivalent definition is

$$\boxed{I(X;Y) = H(X) + H(Y) - H(X,Y)}$$

We now again use the chain rule, but in a different order. We write

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

Thus, the second equivalent definition is

$$\boxed{I(X;Y) = H(Y) - H(Y|X)}$$

(b) The symmetry of the mutual information is simply followed by the alternative definitions. To this end, we start with the basic form $I(X;Y) = H(X) - H(X|Y)$. In this case,

$$I(Y;X) = H(Y) - H(Y|X) \overset{\text{(a)}}{=} I(X;Y).$$

where (a) comes from the second equivalent definition.

(c) Noting that $H(Y) \geq H(Y|X)$, we conclude that $I(X;Y) \geq 0$.

4. *Chain rule* can be further derived for the mutual information. Let X, Y and Z be three *dependent* variables.

   (a) Write the definitions of $I(X,Z;Y)$ and $I(X;Y|Z)$.

(b) Show that

$$I(X,Z;Y) = I(Z;Y) + I(X;Y|Z).$$

♠ **Solution:**

(a) The definition of $I(X,Z;Y)$ is simply given by considering (X,Z) together, as a single random variable. Assume $\tilde{X} = (X,Z)$, then

$$\begin{aligned} I(X,Z;Y) = I(\tilde{X};Y) &= H(\tilde{X}) - H(\tilde{X}|Y) \\ &= H(X,Z) - H(X,Z|Y). \end{aligned}$$

Alternative definitions moreover are given by

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

The definition of the conditional mutual information is given by equation (8.10) in page 139 which reads

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

Similar to $I(X;Y)$, for the conditional mutual information, we can further write the equivalent definitions

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

(b) To show the chain rule for the mutual information, we start with the definition of $I(X,Z;Y)$,

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

Using the chain rule for entropy, we have

$$\begin{aligned} I(X,Z;Y) &= H(Z) + H(X|Z) - (H(Z|Y) + H(X|Y,Z)) \\ &= \underbrace{H(Z) - H(Z|Y)}_{I(Z;Y)} + \underbrace{H(X|Z) - H(X|Y,Z)}_{I(X;Y|Z)} \\ &= I(Z;Y) + I(X;Y|Z). \end{aligned}$$

Using other equivalent definitions, an alternative form of the rule can be concluded, which says

$$I(X,Z;Y) = I(X;Y) + I(Z;Y|X).$$

The general form of the chain rule for mutual information reads

⇝ REMARK:

For random variables $X_1, \dots, X_N$ and Y, we have

$$I(X_1, \dots, X_N; Y) = I(X_1;Y) + \sum_{n=2}^{N} I(X_n; Y|X_1, \dots, X_{n-1}).$$

5.3.2 Information theoretic identities and inequalities

♣ Exercise 1 covers partially Exercise 8.7. in page 141.
Exercise 2 covers Exercise 8.9. in page 141.

1. Show that $I(X;Y) = 0$ if and only if X and Y are *independent*.

♠ **Solution:** To show this, we need to show two statements

- If X and Y are *independent*; then, $I(X;Y) = 0$.
- If $I(X;Y) = 0$; then, X and Y are *independent*.

For the first statement, we note that when X and Y are *independent*,

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

Using the definition of the mutual information, we have

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = 0.$$

The second statement is also simple to prove. We know that $I(X;Y) = 0$, this means that

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

From equation (2.39) in page 33, we know that this happens if and only if X and Y are independent. Hence, we can conclude that X and Y are independent.

$\rightsquigarrow$ REMARK:

Following the equivalent definitions of the mutual information, we can also say

- $H(X) = H(X|Y)$ if and only if X and Y are *independent*.
- $H(Y) = H(Y|X)$ if and only if X and Y are *independent*.

2. The *data-processing* theorem is a key inequality in information theory. We study it in this exercise.

Let X, Y and Z be *dependent* random variables with alphabets $\mathcal{A}_X$, $\mathcal{A}_Y$ and $\mathcal{A}_Z$, respectively. Assume that

$$X \to Y \to Z$$

form a Markov chain.

(a) Factorize the joint probability distribution of (X,Y,Z) by considering the property of Markov chains.

(b) Show that $H(Z|Y,X) = H(Z|Y)$.

(c) Determine $I(X;Z|Y)$.

(d) Show that

$$I(X;Z) \leq I(X;Y).$$

♠ **Solution:**

(a) By the definition, $X \to Y \to Z$ form a Markov chain if

$$P(z|y,x) = P(z|y).$$

⇝ REMARK:

The Markov chain **does NOT** conclude that X and Z are independent.

Consequently, the joint probability distribution of (X,Y,Z) can be written as

$$P(x,y,z) = P(x)P(y|x)P(z|y,x) = P(x)P(y|x)P(z|y).$$

(b) Starting with the definition of $H(Z|Y,X)$, we have

$$\begin{aligned} H(Z|Y,X) = \sum_{\substack{x\in\mathcal{A}_X\\ y\in\mathcal{A}_Y\\ z\in\mathcal{A}_Z}} P(x,y,z)\log\frac{1}{P(z|y,x)} &= \sum_{\substack{x\in\mathcal{A}_X\\ y\in\mathcal{A}_Y\\ z\in\mathcal{A}_Z}} P(x,y,z)\log\frac{1}{P(z|y)} \\ &= \sum_{\substack{y\in\mathcal{A}_Y\\ z\in\mathcal{A}_Z}} \underbrace{\sum_{x\in\mathcal{A}_X} P(x,y,z)}_{P(y,z)}\log\frac{1}{P(z|y)} \\ &= \sum_{\substack{y\in\mathcal{A}_Y\\ z\in\mathcal{A}_Z}} P(y,z)\log\frac{1}{P(z|y)} = H(Z|Y). \end{aligned}$$

(c) From the definition of the conditional mutual information, we have

$$I(X;Z|Y) = H(Z|Y) - H(Z|Y,X) = 0.$$

This is in fact a generic rule which says

⇝ REMARK:

When $X \to Y \to Z$, we have always

$$I(X;Z|Y) = 0.$$

Tip to remember:

The mutual information between the guys at the tails is zero conditioned to the guy at the middle.

(d) To show this inequality, we use chain rule and write

$$I(X;Z,Y) = I(X;Y) + \underbrace{I(X;Z|Y)}_{0} = I(X;Y). \qquad \text{(EQ:A6)}$$

An alternative form of the chain rule says

$$I(X;Z,Y) = I(X;Z) + I(X;Y|Z). \qquad \text{(EQ:B6)}$$

Comparing (EQ:A6) and (EQ:B6), we have

$$I(X;Y) = I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0} \geq I(X;Z).$$

This is the so-called *data processing inequality*. The name comes from the following fact:

⇝ REMARK:

Assume X is a data which has been impaired by some noise. The noisy observation is Y. If you process this observation and get Z, you have

$$X \to Y \to Z.$$

This means that regardless of the processor, we would always have

$$I(X;Z) \leq I(X;Y).$$

In other words, by processing observation Y, you would never add any extra information about X. You can only loose information.

5.4 Homework

The following exercises are suggested for further practice.

5.4.1 Primary exercises

1. Let $Y = f(X)$ where $f(\cdot)$ is a deterministic function.

   (a) Show that $H(Y|X) = 0$.

   (b) Determine the mutual information $I(X;Y)$.

2. Let X and Y be two Bernoulli random variables with

$$\Pr\{X=1\}=1-\Pr\{X=0\}=\alpha$$
$$\Pr\{Y=1\}=1-\Pr\{Y=0\}=\beta$$

and let $Z=X\oplus Y$. Determine the following items:

(a) Conditional distribution $P(z|x)$.

(b) Conditional distribution $P(z|y)$.

(c) Conditional distribution $P(z|x,y)$.

3. In the previous exercise, find the conditions under which Z independent of X.

♠ **Solution:** Z and X are independent, if $\beta=0.5$.

4. Show the general form of the chain rule for mutual information which states that

$$I\left(X^N;Y\right)=\sum_{n=1}^{N} I\left(X_i;Y|X^{i-1}\right)$$

where $X^{i-1}=X_1,\dots X_{i-1}$ and $X^0=\emptyset$, i.e., nothing.

Skip if you want: The next exercise is only for interested readers and is not a typical question in the exam.

5. *Fano's inequality* is one of fundamental inequalities in information theory. It states that:

Let $X\to Y\to\hat{X}$ form a Markov chain and assume that $X,\hat{X}\in\mathcal{A}_X$. Define the error probability

$$P_E=\Pr\left\{X\neq\hat{X}\right\}$$

Then, the error probability is bounded from below via the following inequality

$$H\left(X|Y\right)\leq H_2\left(P_E\right)+P_E\log|\mathcal{A}_X|$$

where $|\mathcal{A}_X|$ denotes the number of elements in set $\mathcal{A}_X$.

In this exercise we intend to prove this inequality. To do so, follow the following steps:

(a) Define the Bernoulli random variable E as

$$E=\begin{cases}1 & X\neq\hat{X}\\ 0 & X=\hat{X}\end{cases}.$$

In this case, we have $\Pr\{E=1\}=1-\Pr\{E=0\}=P_E$.

(b) Use the chain rule and write the both possible expansions of $H\left(X, E|\hat{X}\right)$ to show the following equality

$$H\left(X|\hat{X}\right) = H\left(E|\hat{X}\right) + H\left(X|E,\hat{X}\right). \tag{EQ:A}$$

(c) Show that

$$H\left(E|X\right) \leq H_2\left(P_E\right) \tag{EQ:B}$$

$$H\left(X|E,\hat{X}\right) \leq P_E \log|A_X| \tag{EQ:C}$$

(d) Use (EQ:B) and (EQ:C) along with the equality in (EQ:A) to bound $H(X|\hat{X})$ from above.

(e) Use the data processing inequality to bound $H(X|\hat{X})$ from below.

(f) Conclude Fano's inequality.

5.4.2 Further exercises from the textbook

- Chapter 8: Exercise 8.6., Exercise 8.7., Exercise 8.8., Exercise 8.10.

5.5 Fun Facts

Although Shannon is well known for developing information theory, he had a large scope of contributions, some of which have significantly impacted our current era:

- When he was 21, he wrote his master thesis *A Symbolic Analysis of Relay and Switching Circuits*. In this thesis, he proved that any Boolean algebra problem is solved by a network of switches. This thesis introduced the idea of digital computers!

- Shannon was the first one who developed a device based on what we nowadays know as *machine learning*. Shannon's mouse, known as *Theseus*, could solve a maze. It was a primary experiment in artificial intelligence.

- In 1949, Shannon wrote the first paper which discussed the problem of playing chess with a computer. He also determined an accurate approximation for the complexity of playing chess with computer using brute force approach. This number is 10^{120} which is known as Shannon number.

You can learn more about Shannon by reading the bibliographical book *A Mind at Play* written by Jimmy Soni and Rob Goodman in 2017.

Chapter 6

Channel Coding Theorem

In this chapter, we investigate the concept of channel capacity. To this end, we study the mathematical model of a communication channel and get to know the concept of channel capacity. We then practice some exercises on calculation of the channel capacity. This chapter is consistent with Chapters 9 and 10 of the textbook.

6.1 Brief Review of Main Concepts

Consider a noisy communication channel. When you send a bit over it, you receive a noisy version of it. So, you cannot say for sure what was the transmitted bit. To deal with this problem, you use a channel code: The *information bits* are first mapped to a sequence with (generally) larger number of bits which we call *codeword*. We then send the codeword over the channel. If we design our codeword well; then, under some conditions, we could always estimate the original *information bits* from the received noisy codeword correctly. The key question to answer is the following: How many bits at most can be transmitted over this channel in *a single transmission*, such that we estimate the information bits (almost surely) correctly.

Shannon answers this question in his second theorem, known as the channel coding theorem. In a nutshell, Shannon's answer is as follows: Assume we transmit a codeword X^N which is i.i.d. sequence with distribution $P\left(x\right)$. We then receive the noisy version Y^N. In this case, in each transmission we transmit $I\left(X;Y\right)$ bits of information. The maximum number of bits per transmission is hence given by

$$C = \max_{P(x)} I\left(X;Y\right).$$

Shannon calls this parameter the channel capacity and shows that this is in fact the maximum error-free transmission rate we could have in a channel.

6.1.1 The classic approach for determining the channel capacity

The following steps describe the standard approach for determination of the channel capacity:

(a) First write the definition of the mutual information which is

$$I\left(X;Y\right)=H\left(Y\right)-H\left(Y|X\right).$$

You need to determine two terms $H\left(Y\right)$ and $H\left(Y|X\right)$.

(b) Assume a general distribution for the input X.

(c) Determine the conditional entropy by first calculating $H\left(Y|X=x\right)$ for all outcomes $X=x$ and then writing

$$H\left(Y|X\right)=\sum_{x\in\mathcal{A}_X}\Pr\left\{X=x\right\}H\left(Y|X=x\right).$$

(d) Determine marginal distribution of Y and then calculate $H\left(Y\right)$.

Throughout the exercises, we will learn how exactly it works.

6.2 Exercises

6.2.1 Mathematical model for communication channels

♣ Exercise 1 covers Exercise 9.2. and Exercise 9.4. in page 149, and partially Exercise 9.18. in page 155.

1. Consider

   (C_1) a BSC with flipping probability f,

   (C_2) a binary erasure channel (BEC) with erasing probability e, and

   (C_3) an additive white Gaussian noise (AWGN) channel with noise variance σ^2.

   For these channels, do the following items:

   (a) Show the input-output relation of these channels via graphs.

   (b) Write the conditional distributions of these channels.

6.2.2 Capacity of a communication channel

♣ Exercise 1 covers Exercise 9.13. in page 151 and Exercise 10.12. in page 172.
Exercise 2 covers Exercise 9.12. in page 151.
Exercise 4 covers Exercise 9.4. in page 149.
Exercise 5 covers Exercise 9.17. in page 155.

1. Consider a BEC channel with erasing probability e. Let the input to this channel be Bernoulli variable X with

$$\Pr\left\{X=1\right\}=1-\Pr\left\{X=0\right\}=q.$$

(a) Determine the distribution of the output symbol Y.

(b) Calculate $I(X;Y)$.

(c) If you want to maximize the mutual information, over which variable do you optimize? Specify the domain of this variable.

(d) Determine the capacity of this channel.

2. Find the capacity of a BSC with flipping probability f. Determine the input distribution which achieves this capacity.

3. Consider a channel with 3 possible inputs and 4 possible outputs whose transition probability matrix is

$$\begin{bmatrix} 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 \end{bmatrix}.$$

Denote the possible inputs with x_1, x_2, x_3 and the outputs with y_1, y_2, y_3, y_4.

(a) Sketch the input-output relation for this channels in a graph.

(b) Find the capacity of this channel and the input distribution which achieves it.

4. Determine the capacity of a Z channel with error probability f.

6.2.3 Operational meaning of the channel capacity

> ⇝ REMARK:
>
> Note that the following exercise is *not* a typical exercise you would get in the exam, and is mainly designed for your better understanding of the channel coding theorem.

1. Consider a BSC with flipping probability $f = 0.05$.

(a) Restate the capacity of this channel, and determine it for $f = 0.05$.

We intend to transmit message D reliably over this channel. This message is a random variable which takes values $d_1, \ldots, d_M$, uniformly. In order to transmit over this channel, we put a *channel encoder* $f_N(\cdot)$ before the channel, and a *channel decoder* $g_N(\cdot)$ after the channel. The channel encoder maps message D into a binary codeword $X^N = f_N(D)$ which is of length N. X^N is transmitted over the channel symbol-by-symbol. We denote the received bits at the receive side by Y^N. The decoder then recovers $\hat{D} = g_N\left(Y^N\right)$ as an estimate of D using the channel decoder.

(b) What is the data transmission rate R that we achieve by such a setting? State the unit of R.

(c) Define error probability P_{E}, and write it in terms of Y^N, X^N and M.

Now consider the following three instances of the setting:

- INSTANCE A. D has $M = 2$ possible values. The code-length is $N = 3$, and the channel encoder reads

$$f_N(d_1) = \mathtt{000}, \qquad f_N(d_2) = \mathtt{111}.$$

To decode received signal Y^N, the decoder uses major vote meaning that it decodes $\hat{D} = d_1$ if the number of zeros in Y^N, $\hat{D} = d_2$ otherwise. Hence, $g_N(\cdot)$ reads

$$g_N\left(Y^N\right) = \begin{cases} d_1 & N_0 > N_1 \\ d_2 & N_1 > N_0 \end{cases}$$

where N_0 and N_1 are the numbers of zeros and ones in Y^N, respectively.

For this instance,

(d) What is the transmission rate?

(e) Determine distribution of Y^N, when $D = d_1$ is sent over the channel.

(f) Determine the error probability, when $D = d_1$ is sent over the channel.

- INSTANCE B. D has $M = 4$ possible values. The code-length is $N = 3$, and the channel encoder reads

$$f_N(d_1) = \mathtt{001}, \qquad f_N(d_2) = \mathtt{100}, \qquad f_N(d_3) = \mathtt{010}, \qquad f_N(d_4) = \mathtt{111}.$$

The decoder moreover reads

$$g_N\left(Y^N\right) = \begin{cases} d_1 & Y^N = \mathtt{000} \text{ or } \mathtt{001} \\ d_2 & Y^N = \mathtt{100} \text{ or } \mathtt{101} \\ d_3 & Y^N = \mathtt{010} \text{ or } \mathtt{011} \\ d_4 & Y^N = \mathtt{110} \text{ or } \mathtt{111} \end{cases}$$

For this instance,

(g) What is the transmission rate?

(h) Determine the error probability, when $D = d_1$ is sent over the channel, and compare it to instance A.

- INSTANCE C. D has $M = 8$ possible values. The code-length is $N = 3$, and the channel encoder reads

$$f_N(d_i) = \mathtt{binary}(i-1)$$

where $\mathtt{binary}(i)$ determines the binary representation of i. The decoder moreover reads

$$g_N\left(Y^N\right) = d_i \qquad \text{if} \qquad Y^N = f_N(d_i).$$

For this instance,

(i) What is the transmission rate?

(j) Determine the error probability, when $D = d_1$ is sent over the channel. Compare it to instances A and B.

(k) Plot the rate-error diagram for $N = 3$. Sketch the line $P_{\mathrm{E}} = 0.1$.

Now, let the code-length be $N = 10^4$. Using the Shannon's second theorem, answer the following question.

(l) What is approximately the maximum value for M for which $P_{\mathrm{E}} \leq \epsilon$, for some ϵ close to zero.

6.3 Solutions to Exercises

6.3.1 Mathematical model for communication channels

♣ Exercise 1 covers Exercise 9.2. and Exercise 9.4. in page 149, and partially Exercise 9.18. in page 155.

1. Consider

(C_1) a BSC with flipping probability f,

(C_2) a BEC with erasing probability e, and

(C_3) an AWGN channel with noise variance σ^2.

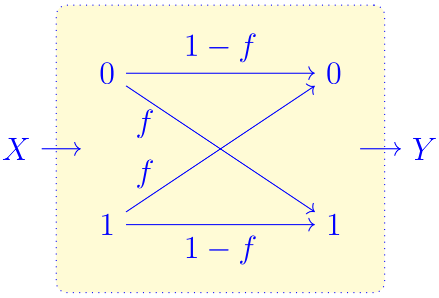
For these channels, do the following items:

(a) Show the input-output relation of these channels via graphs.

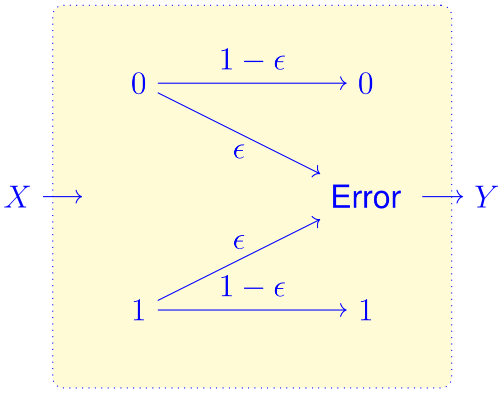
(b) Write the conditional distributions of these channels.

♠ **Solution:**

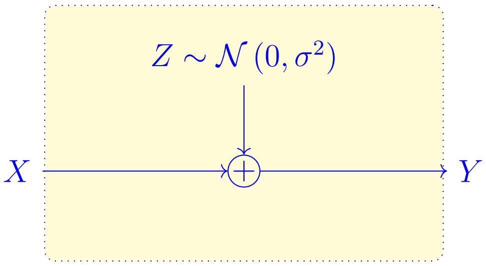
(a) The channel diagram for the BSC is as follows:



For the BEC, the diagram reads:



For the Gaussian channel with binary input, we further have



(b) The conditional distribution of the channel, also referred to as the *transition rule*, is often showed by $P(y|x)$. Depending on the type of output, this conditional distribution is interpreted as follows:

- When the channel has a *discrete output*, like the BSC and BEC, $P(y|x)$ determines the probability of output Y be $Y = y$, when we set the input to $X = x$, i.e.

$$P(y|x) = \Pr\{Y = y | X = x\}.$$

  In this case, $P(y|x)$ is interpreted as a set of *probability mass functions* on Y which changes from one realization of X to another.

> ⇝ REMARK:
>
> For discrete Y, $P(y|x)$ is a distinct probability mass function for each value of x.

- When the channel has a *continuous output*, like the Gaussian channel, $P(y|x)$ determines the *probability density function* of output Y at point $Y = y$, when we set the input to $X = x$, i.e.

$$P(y|x) = f_Y(y|X = x).$$

  In this case, $P(y|x)$ is as a set of *probability density functions* on Y which changes from one realization of X to another.

$\rightsquigarrow$ REMARK:

For continuous Y, $P\left(y|x\right)$ is a distinct probability density function for each value of x.

Considering the BSC, the input has two possible outcomes $X = 0$ and $X = 1$. Hence, $P\left(y|x\right)$ is the set of following two *probability mass functions*

$$P\left(y|0\right) = \Pr\left\{Y = y|X = 0\right\} = \begin{cases} 1-f & y=0 \\ f & y=1 \end{cases}, \qquad \text{and}$$

$$P\left(y|1\right) = \Pr\left\{Y = y|X = 1\right\} = \begin{cases} f & y=0 \\ 1-f & y=1 \end{cases}.$$

Similarly for the BEC, $P\left(y|x\right)$ is the set of following two *probability mass functions*

$$P\left(y|0\right) = \Pr\left\{Y = y|X = 0\right\} = \begin{cases} 1-\epsilon & y=0 \\ \epsilon & y=\text{Error} \\ 0 & y=1 \end{cases}, \qquad \text{and}$$

$$P\left(y|1\right) = \Pr\left\{Y = y|X = 1\right\} = \begin{cases} 0 & y=0 \\ \epsilon & y=\text{Error} \\ 1-\epsilon & y=1 \end{cases}.$$

For the third example of the Gaussian channel, input has two possible outcomes $X = +1$ and $X = -1$. When $X = +1$, $Y = 1 + Z$ which is a Gaussian variable with mean $+1$ and variance σ^2. Similarly, when $X = -1$, $Y = -1 + Z$ which is a Gaussian variable with mean -1 and variance σ^2. Thus, $P\left(y|x\right)$ is the set of following two *probability density functions*

$$P\left(y|+1\right) = f_Y\left(y|X = +1\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\left(y-1\right)^2}{\sigma^2}\right\}, \qquad \text{and}$$

$$P\left(y|-1\right) = f_Y\left(y|X = -1\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\left(y+1\right)^2}{\sigma^2}\right\}.$$

Equivalently, we can write in this case

$$P\left(y|x\right) = f_Y\left(y|X = x\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\left(y-x\right)^2}{\sigma^2}\right\}.$$

for $x = \pm 1$.

6.3.2 Capacity of a communication channel

♣ Exercise 1 covers Exercise 9.13. in page 151 and Exercise 10.12. in page 172.
Exercise 2 covers Exercise 9.12. in page 151.
Exercise 4 covers Exercise 9.4. in page 149.
Exercise 5 covers Exercise 9.17. in page 155.

1. Consider a BEC channel with erasing probability e. Let the input to this channel be Bernoulli variable X with

$$\Pr\left\{X=1\right\} = 1 - \Pr\left\{X=0\right\} = q.$$

(a) Determine the distribution of the output symbol Y.

(b) Calculate $I\left(X;Y\right)$.

(c) If you want to maximize the mutual information, over which variable do you optimize? Specify the domain of this variable.

(d) Determine the capacity of this channel.

♠ **Solution:**

(a) The conditional distribution of this channel is given in the previous exercise. To find the marginal distribution of Y, we use the sum rule in page 24. We hence have

$$\begin{aligned}\Pr\left\{Y=y\right\} &= \sum_{x=0}^{1} \Pr\left\{Y=y, X=x\right\}\\ &= \sum_{x=0}^{1} \underbrace{\Pr\left\{Y=y|X=x\right\}}_{P(y|x)} \Pr\left\{X=x\right\}\\ &= P\left(y|0\right)\left(1-q\right) + P\left(y|1\right)q.\end{aligned}$$

As a result, we could write

$$\begin{aligned}\Pr\left\{Y=0\right\} &= \left(1-\epsilon\right)\left(1-q\right)\\ \Pr\left\{Y=\text{Error}\right\} &= \epsilon\left(1-q\right) + \epsilon q = \epsilon\\ \Pr\left\{Y=1\right\} &= \left(1-\epsilon\right)q.\end{aligned}$$

(b) To calculate $I\left(X;Y\right)$, we write the definition of the mutual information which is

$$I\left(X;Y\right) = H\left(Y\right) - H\left(Y|X\right).$$

We start with determining $H\left(Y|X\right)$. In this respect, we use the definitions in equations (8.3) and (8.4), page 138, and write

$$H\left(Y|X\right) = \sum_{x\in\mathcal{A}_X} \Pr\left\{X=x\right\} H\left(Y|X=x\right).$$

$H\left(Y|X=x\right)$ is defined in (8.3) and is the entropy determined for the conditional distribution $\Pr\left\{Y|X=x\right\}$.

⇝ REMARK:

Remember that for each outcome $X = x$, the conditional distribution $P(y|x)$ is a distinct distribution on Y. The entropy of the distribution corresponding to $X = x$ is $H(Y|X = x)$. The conditional entropy $H(Y|X)$ is then the average of these entropies on X.

As we discussed in the previous exercise, in the BEC, we have the two distributions:

- For the outcome $X = 0$, it reads

$$P(y|0) = \Pr\{Y = y|X = 0\} = \begin{cases} 1-\epsilon & y = 0 \\ \epsilon & y = \text{Error} \\ 0 & y = 1 \end{cases}.$$

  This is a Bernoulli distribution. Hence, the entropy term $H(Y|X = 0)$ is

$$H(Y|X = 0) = H_2(\epsilon).$$

- For the outcome $X = 1$, the distribution of Y is

$$P(y|1) = \Pr\{Y = y|X = 1\} = \begin{cases} 0 & y = 0 \\ \epsilon & y = Error \\ 1-\epsilon & y = 1 \end{cases}.$$

  This is again a Bernoulli distribution, and thus

$$H(Y|X = 1) = H_2(\epsilon).$$

Consequently, $H(Y|X)$ is given by

$$\begin{aligned} H(Y|X) &= \Pr\{X = 0\} H(Y|X = 0) + \Pr\{X = 1\} H(Y|X = 1) \\ &= (1-q) H_2(\epsilon) + q H_2(\epsilon) = H_2(\epsilon). \end{aligned}$$

So, we have concluded that

$$\boxed{H(Y|X) = H_2(\epsilon)}$$

Now, we determine the entropy term $H(Y)$. Considering the distribution derived in Part (a), we have

$$\begin{aligned}
H\left(Y\right) = &\Pr\left\{Y=0\right\}\log\frac{1}{\Pr\left\{Y=0\right\}} + \Pr\left\{Y=\text{Error}\right\}\log\frac{1}{\Pr\left\{Y=\text{Error}\right\}} \\
&+ \Pr\left\{Y=1\right\}\log\frac{1}{\Pr\left\{Y=1\right\}} \\
&= \left(1-\epsilon\right)\left(1-q\right)\log\frac{1}{\left(1-\epsilon\right)\left(1-q\right)} + \epsilon\log\frac{1}{\epsilon} + \left(1-\epsilon\right)q\log\frac{1}{\left(1-\epsilon\right)q} \\
&= \left(1-\epsilon\right)\left[\left(1-q\right)\log\frac{1}{1-q} + \left(1-q\right)\log\frac{1}{1-\epsilon} + q\log\frac{1}{q} + q\log\frac{1}{1-\epsilon}\right] + \epsilon\log\frac{1}{\epsilon} \\
&= \left(1-\epsilon\right)\left[H_2\left(q\right) + \left(1-q\right)\log\frac{1}{1-\epsilon} + q\log\frac{1}{1-\epsilon}\right] + \epsilon\log\frac{1}{\epsilon} \\
&= \left(1-\epsilon\right)H_2\left(q\right) + \underbrace{\left(1-\epsilon\right)\log\frac{1}{1-\epsilon} + \epsilon\log\frac{1}{\epsilon}}_{H_2\left(\epsilon\right)}.
\end{aligned}$$

Hence, we have

$$\boxed{H\left(Y\right) = \left(1-\epsilon\right)H_2\left(q\right) + H_2\left(\epsilon\right)}$$

As the result, the mutual information reads

$$\boxed{I\left(X;Y\right) = \left(1-\epsilon\right)H_2\left(q\right)}$$

(c) Noting that ϵ is fixed and given by the channel, we can conclude that the mutual information is a function of variable q whose domain is $[0, 0.5]$. This variable in fact describes the input distribution. We hence can write

$$I\left(X;Y\right) = f\left(q\right) = \left(1-\epsilon\right)H_2\left(q\right).$$

(d) The capacity of the channel is defined as the maximum mutual information, maximized over all input distributions. This means that

$$\begin{aligned}
C &= \max_{\text{all } P(x)} I\left(X;Y\right) \\
&= \max_{q\in[0,0.5]} f\left(q\right) \\
&= \max_{q\in[0,0.5]} \left(1-\epsilon\right)H_2\left(q\right)
\end{aligned}$$

To maximize $f\left(q\right)$, we note that $f\left(q\right)$ is maximized when $H_2\left(q\right)$ meets its maximum. We know that $\max_q H_2\left(\epsilon \circledast q\right) = 1$. This concludes that

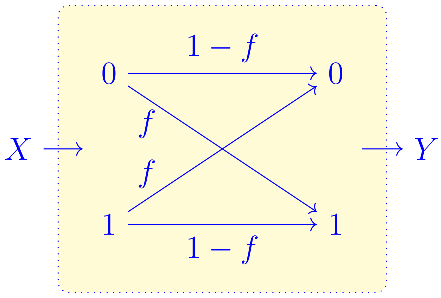
$$\boxed{C = 1-\epsilon}$$

Moreover, we know that $H_2\left(0.5\right) = 1$. This means that the capacity C is achieved when

$$\boxed{q = 0.5}$$

2. Find the capacity of a BSC with flipping probability f. Determine the input distribution which achieves this capacity.

♠ **Solution:** The diagram for the BSC is is given below. To determine the capacity of the channel, we follow the *classic approach*. This approach is illustrated step-by-step at the end of this tutorial.



(a) As the first step, we write the definition of the mutual information which is

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

We note that we need to determine two terms $H(Y)$ and $H(Y|X)$. These terms should be determined for an arbitrary input distribution, substituted in the definition, and finally maximized over all possible choices of input distributions.

(b) In the next step, we consider a general distribution for the input X. For instance, in this example, we have two outcomes for X, $X=0$ and $X=1$. Thus, a general input distribution is

$$\Pr\{X=1\} = 1 - \Pr\{X=0\} = q$$

for some $q \in [0, 0.5]$.

(c) In the third step, we determine the conditional entropy. To this end, we first calculate $H(Y|X=x)$ for all outcomes $X=x$. $H(Y|X)$ is then given by

$$H(Y|X) = \sum_{x\in\mathcal{A}_X} \Pr\{X=x\} H(Y|X=x).$$

In this example $\mathcal{A}_X = \{0,1\}$. Hence, we should determine $H(Y|X=0)$ and $H(Y|X=1)$. For $H(Y|X=0)$, we consider the distribution of Y when $X=0$. In this case, we have

$$P(y|0) = \Pr\{Y=y|X=0\} = \begin{cases} 1-f & y=0 \\ f & y=1 \end{cases}.$$

Thus, $H(Y|X=0) = H_2(f)$. For $H(Y|X=1)$, we further note that

$$P(y|1) = \Pr\{Y=y|X=1\} = \begin{cases} f & y=0 \\ 1-f & y=1 \end{cases}$$

which means that $H\left(Y|X=1\right)=H_2\left(f\right)$. Therefore,

$$\begin{aligned}H\left(Y|X\right)&=\Pr\left\{X=0\right\}H\left(Y|X=0\right)+\Pr\left\{X=1\right\}H\left(Y|X=1\right)\\&=\left(1-q\right)H_2\left(f\right)+qH_2\left(f\right)=H_2\left(f\right).\end{aligned}$$

(d) The next step is to determine marginal distribution of Y and then calculate $H\left(Y\right)$. In this respect, we use the sum rule which says

$$\Pr\left\{Y=y\right\}=\sum_{x\in\mathcal{A}_X}P\left(y|x\right)\Pr\left\{X=x\right\}.$$

In our example, we have

$$\Pr\left\{Y=1\right\}=P\left(1|0\right)\left(1-q\right)+P\left(1|1\right)q=f\left(1-q\right)+\left(1-f\right)q.$$

It is common in the information theory literature to define the following notation:

$$f\circledast q\coloneqq f\left(1-q\right)+\left(1-f\right)q.$$

Thus, we can compactly write

$$\Pr\left\{Y=1\right\}=f\circledast q.$$

Since Y is a Bernoulli variable, we can write

$$\Pr\left\{Y=0\right\}=1-\Pr\left\{Y=1\right\}=1-f\circledast q.$$

Hence,

$$H\left(Y\right)=H_2\left(f\circledast q\right).$$

(e) The last step is to substitute the entropy terms into the definition of the mutual information, and then maximize it over all possible input distributions. In our example, we have

$$I\left(X;Y\right)=H_2\left(f\circledast q\right)-H_2\left(f\right).$$

Here, the input distribution is specified by the value of q; thus, we should maximize the mutual information with respect to q. We hence have

$$C=\max_{q\in[0,0.5]}H_2\left(f\circledast q\right)-H_2\left(f\right).$$

Following the same argument as in the previous exercise, we conclude that the mutual information is maximized when $H_2\left(f\circledast q\right)$ is maximized. As $\max_q H_2\left(f\circledast q\right)=1$, we can write

$$C=1-H_2\left(f\right)$$

and the value of the q for which the mutual information is maximized is

$$\begin{aligned}f\circledast q=0.5&\Rightarrow f\left(1-q\right)+\left(1-f\right)q=0.5\\&\Rightarrow q\left(1-2f\right)=0.5-f\Rightarrow\boxed{q=0.5}.\end{aligned}$$

3. Consider a channel with 3 possible inputs and 4 possible outputs whose transition probability matrix is

$$\begin{bmatrix} 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 \end{bmatrix}.$$

Denote the possible inputs with x_1, x_2, x_3 and the outputs with y_1, y_2, y_3, y_4.

(a) Sketch the input-output relation for this channels in a graph.

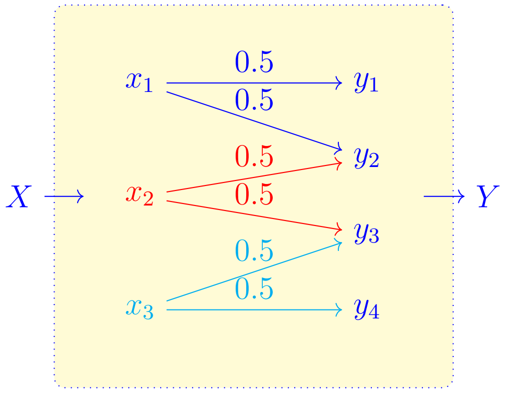
(b) Find the capacity of this channel and the input distribution which achieves it.

♠ **Solution:**

(a) Let us denote the input outcomes by $\{x_1, x_2, x_3\}$ and the output outcomes by $\{y_1, y_2, y_3, y_4\}$. This symbols are shown in the following matrix.

$$\begin{array}{c c} & \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{array}{c} y_1 \\ y_2 \\ y_3 \\ y_4 \end{array} & \begin{bmatrix} 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 \end{bmatrix} \end{array}$$

The above matrix describes the conditional distribution of Y given the input X. This means that given each input outcome, Y is distributed due to the corresponding column in the matrix. The graph for this channel is shown in the figure. The links corresponding to each column is denoted by the same color.



(b) To derive the channel capacity, we follow the classic approach:

(a) We write the definition of the mutual information as

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

We need to determine two terms $H(Y)$ and $H(Y|X)$.

(b) We consider a general distribution for the input X. In this example, we have three outcomes for X, $X = x_1$, $X = x_2$, and $X = x_3$. Thus, a general input distribution is

$$\Pr\{X = x_1\} = q_1 \text{ and } \Pr\{X = x_2\} = q_2 \text{ and } \Pr\{X = x_3\} = 1 - q_1 - q_2.$$

for some $q_1 \geq 0$ and $q_2 \geq 0$, such that $q_1 + q_2 \in [0, 1]$.

(c) Now, we determine the conditional entropy. To this end, we first calculate $H(Y|X = x_i)$ for $i = 1, 2, 3$. For $H(Y|X = x_1)$, we consider the distribution of Y when $X = x_1$ which is

$$P(y|x_1) = \Pr\{Y = y|X = x_1\} = \begin{cases} 0.5 & y = y_1 \\ 0.5 & y = y_2 \\ 0 & y = y_3 \\ 0 & y = y_4 \end{cases}.$$

Thus, $H(Y|X = x_1)$ reads

$$H(Y|X = x_1) = 0.5 \log \frac{1}{0.5} + 0.5 \log \frac{1}{0.5} = H_2(0.5) = 1$$

For $H(Y|X = x_2)$, we consider the distribution of Y when $X = x_2$ which is

$$P(y|x_2) = \Pr\{Y = y|X = x_2\} = \begin{cases} 0 & y = y_1 \\ 0.5 & y = y_2 \\ 0.5 & y = y_3 \\ 0 & y = y_4 \end{cases}.$$

Similarly, we can conclude that $H(Y|X = x_2) = H_2(0.5) = 1$.
Finally for $H(Y|X = x_3)$, we consider the distribution of Y when $X = x_3$ which is

$$P(y|x_3) = \Pr\{Y = y|X = x_3\} = \begin{cases} 0 & y = y_1 \\ 0 & y = y_2 \\ 0.5 & y = y_3 \\ 0.5 & y = y_4 \end{cases}.$$

Similarly, we have $H(Y|X = x_3) = H_2(0.5) = 1$.

As the result, we have

$$\begin{aligned} H(Y|X) = {} & \Pr\{X = x_1\} H(Y|X = x_1) + \Pr\{X = x_2\} H(Y|X = x_2) \\ & + \Pr\{X = x_3\} H(Y|X = x_3) \\ = {} & q_1 H_2(0.5) + q_2 H_2(0.5) + (1 - q_1 - q_2) H_2(0.5) = H_2(0.5) = 1 \end{aligned}$$

(d) The next step is to calculate $H(Y)$. We know that in this case $H(Y)$ is found as a function of q_1 and q_2. This means that, $H(Y) = f_{\text{Ent}}(q_1, q_2)$. To find this function, we need to determine the marginal distribution of Y. Using the sum rule, we have

$$\Pr\{Y = y_1\} = \sum_{i=1}^{3} P(y_1|x_i) = 0.5q_1$$

$$\Pr\{Y = y_2\} = \sum_{i=1}^{3} P(y_2|x_i) = 0.5q_1 + 0.5q_2$$

$$\Pr\{Y = y_3\} = \sum_{i=1}^{3} P(y_3|x_i) = 0.5q_2 + 0.5(1 - q_1 - q_2) = 0.5(1 - q_1)$$

$$\Pr\{Y = y_4\} = \sum_{i=1}^{3} P(y_4|x_i) = 0.5(1 - q_1 - q_2)$$

Thus, we have

$$\begin{aligned} H(Y) &= f_{\text{Ent}}(q_1, q_2) \\ &= 0.5\bigg(q_1 \log \frac{2}{q_1} + (q_1 + q_2) \log \frac{2}{q_1 + q_2} \\ &\qquad + (1 - q_1) \log \frac{2}{1 - q_1} + (1 - q_1 - q_2) \log \frac{2}{1 - q_1 - q_2} \bigg) \end{aligned}$$

(e) The last step is to substitute the entropy terms into the definition of the mutual information, and then maximize it over all possible input distributions. In this example, we have

$$I(X;Y) = f_{\text{Ent}}(q_1, q_2) - 1.$$

Here, the input distribution is specified by q_1 and q_2. Thus, we should maximize the mutual information with respect to q_1 and q_2. Following the fact that $H(Y|X)$ is constant, this means that we should maximize $f_{\text{Ent}}(q_1, q_2)$. Although this task looks complicated, one can do it in a simple way by using the properties of the entropy:

We know that $f_{\text{Ent}}(q_1, q_2) = H(Y)$. We further know that the entropy $H(Y)$ always is bounded as

$$0 \leq H(Y) \leq \log|\mathcal{A}_Y|$$

where $|\mathcal{A}_Y|$ is the number of the outcomes at the output. The maximum value occurs when Y is uniformly distributed. For this example, we have $|\mathcal{A}_Y| = 4$. Therefore,

$$0 \leq H(Y) \leq 2 \qquad \Rightarrow \qquad 0 \leq f_{\text{Ent}}(q_1, q_2) \leq 2.$$

This means that

$$\max_{q_1,q_2} f_{\text{Ent}}(q_1,q_2) = 2$$

and thus

$$C = \max_{q_1,q_2} f_{\text{Ent}}(q_1,q_2) - 1 = 1.$$

To achieve this maximum value, Y should be uniform. This means that

$$\Pr\{Y = y_j\} = 0.25$$

for $j = 1,2,3,4$. Using the marginal distribution of Y, we have

$$\Pr\{Y = y_1\} = \sum_{i=1}^{3} P(y_1|x_i) = 0.5q_1 = 0.25 \Rightarrow \boxed{q_1 = 0.5}$$
$$\Pr\{Y = y_2\} = \sum_{i=1}^{3} P(y_2|x_i) = 0.5q_1 + 0.5q_2 = 0.25 \Rightarrow \boxed{q_2 = 0}.$$

Substituting in the two other probabilities, we have

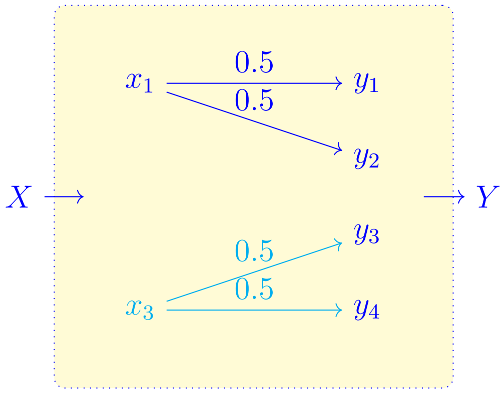
$$\Pr\{Y = y_3\} = \sum_{i=1}^{3} P(y_3|x_i) = 0.5(1-q_1) = 0.25$$
$$\Pr\{Y = y_4\} = \sum_{i=1}^{3} P(y_4|x_i) = 0.5(1-q_1-q_2) = 0.25$$

Thus, by setting

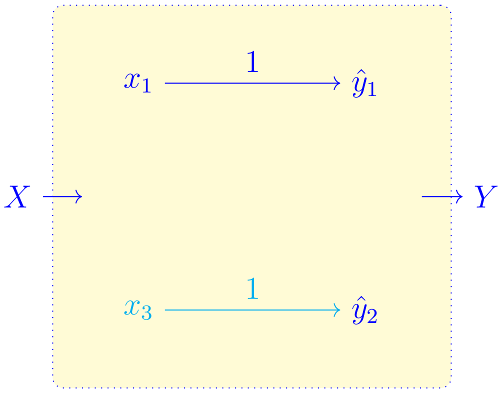
$\Pr\{X = x_1\} = 0.5$ and $\Pr\{X = x_2\} = 0$ and $\Pr\{X = x_3\} = 0.5$

the capacity is achieved.

The result is intuitive. In fact by removing the input outcome x_2, the channel reduces to



By defining $\hat{y}_1 := (y_1, y_2)$ and $\hat{y}_2 := (y_3, y_4)$, the channel becomes



which is a noise-free channel transmitting one bit of information in each transmission.

⇝ REMARK:

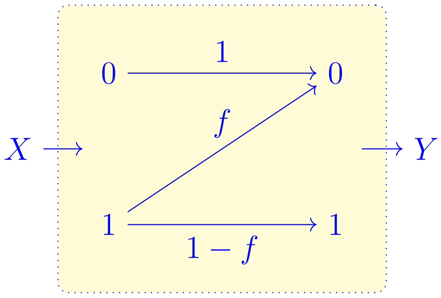
As it is observed in Step (e), It is not required to calculate $f_{\text{Ent}}(q_1, q_2)$ explicitly. This happens in a lot of channel capacity problems.

★Tip:

It is suggested that you assume $H(Y)$ in Step (d) to be some function of the input distribution, but do not determine it explicitly. Then consider Step (e) and check if you could find the maximum mutual information and its corresponding distribution without explicit calculation of $H(Y)$ (Similar to this example). If this was not possible; then, you could go back and determine the explicit expression for $H(Y)$.

4. Determine the capacity of a Z channel with error probability f.

♠ **Solution:** The graph of a Z channel with error probability f is shown below.



In this channel symbol 0 is always passed without error, but 1 could be flipped with probability f. It is called Z channel, as it looks like the letter Z.

To determine the capacity of the channel, we follow the classic approach illustrated in Tutorial 5.

(a) Write the definition of the mutual information which is

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

We need to determine two terms $H(Y)$ and $H(Y|X)$.

(b) Assume a general distribution for the input X. Here, X is binary. Hence, the general distribution is

$$\Pr\{X=1\} = 1 - \Pr\{X=0\} = q$$

for some $q \in [0, 0.5]$.

(c) Determine the conditional entropy. To this end, we first calculate $H(Y|X=0)$ and $H(Y|X=1)$. For $H(Y|X=0)$, we have

$$P(y|0) = \Pr\{Y=y|X=0\} = \begin{cases} 1 & y=0 \\ 0 & y=1 \end{cases}.$$

Thus, $H(Y|X=0)=0$. For $H(Y|X=1)$, we note that

$$P(y|1) = \Pr\{Y=y|X=1\} = \begin{cases} f & y=0 \\ 1-f & y=1 \end{cases}.$$

Consequently, we have $H(Y|X=1) = H_2(f)$. The conditional entropy then reads

$$\begin{aligned} H(Y|X) &= \Pr\{X=0\}\, H(Y|X=0) + \Pr\{X=1\}\, H(Y|X=1) \\ &= qH_2(f). \end{aligned}$$

(d) Determine marginal distribution of Y. For this channel, we have

$$\begin{aligned} \Pr\{Y=1\} &= \Pr\{X=0\}\Pr\{Y=1|X=0\} + \Pr\{X=1\}\Pr\{Y=1|X=1\} \\ &= \Pr\{X=0\}\, P(1|0) + \Pr\{X=1\}\, P(1|1) \\ &= (1-q)\times 0 + q(1-f) = q(1-f). \end{aligned}$$

Consequently,

$$\Pr\{Y=0\} = 1 - \Pr\{Y=0\} = 1 - q(1-f),$$

and

$$H(Y) = H_2(q(1-f)).$$

(e) Substitute the entropy terms into the definition of the mutual information, and then maximize it over all possible input distributions. For this channel, we have

$$I(X;Y) = f(q) = H_2(q(1-f)) - qH_2(f).$$

Thus, the capacity is

$$C = \max_{q\in[0,0.5]} f(q).$$

For function $f(q)$, we have

$$f'(q) = (1-f)\log\frac{1-q(1-f)}{q(1-f)} - H_2(f),$$
$$f''(q) = -\frac{1-f}{\ln 2}\left(\frac{1}{q} + \frac{1-f}{1-q(1-f)}\right)$$

where $\ln 2 = 0.6931$ is the logarithm of 2 in the natural base. As $f''(q) < 0$ for all $q \in [0, 0.5]$, we can conclude that $f(q)$ is concave in this interval, and hence

$$C = \max_{q\in[0,0.5]} f(q) = f(q^\star)$$

where $f'(q^\star) = 0$. By setting the derivative to zero, we find $q^\star$ as

$$q^\star = \frac{1}{(1-f)\,S(f)}$$

where

$$S(f) = 1 + 2^{\frac{H_2(f)}{1-f}}$$

Consequently, the capacity is

$$C = H_2\left(\frac{1}{S(f)}\right) - \frac{H_2(f)}{(1-f)\,S(f)}.$$

⇝ REMARK:

As it is observed in this exercise, the derivation of the capacity is not always intuitive. It is hence suggested to use the classic approach in problems in which you cannot come up with an intuitive solution.

⇝ REMARK:

The distribution which achieves the capacity is NOT always uniform. As indicated in pages 170 and 171 of the book, the uniform distribution is the optimal distribution, ONLY in *symmetric channels*.

6.3.3 Operational meaning of the channel capacity

⇝ REMARK:

Note that the following exercise is *not* a typical exercise you would get in the exam, and is mainly designed for your better understanding of the channel coding theorem.

1. Consider a BSC with flipping probability $f = 0.05$.

   (a) Restate the capacity of this channel, and determine it for $f = 0.05$.

   ♠ **Solution:** As shown in Tutorial 5, the capacity of the BSC reads

$$C = 1 - H_2(f).$$

   For $f = 0.05$, the value of the capacity is $C = 0.7136$.

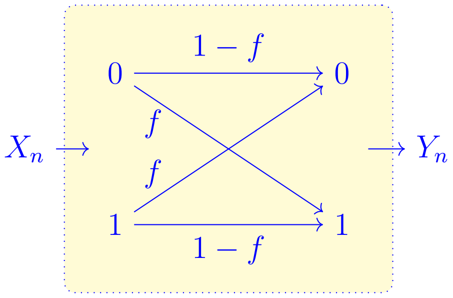
   We intend to transmit message D reliably over this channel. This message is a random variable which takes values $d_1, \dots, d_M$, uniformly. In order to transmit over this channel, we put a *channel encoder* $f_N(\cdot)$ before the channel, and a *channel decoder* $g_N(\cdot)$ after the channel. The channel encoder maps message D into a binary codeword $X^N = f_N(D)$ which is of length N. X^N is transmitted over the channel symbol-by-symbol. We denote the received bits at the receive side by Y^N. The decoder then recovers $\hat{D} = g_N(Y^N)$ as an estimate of D using the channel decoder.

   (b) What is the data transmission rate R that we achieve by such a setting? State the unit of R.

   ♠ **Solution:** When we use this setting, we transmit the message D after sending N symbols $X_1, \dots, X_N$. As D has M possible outcomes and is uniform, we transmit

$$H(D) = \log M$$

   bits of information after sending N symbols. Each symbol X_n is sent over by one *transmission* over the BSC; see the figure below.



   Thus, the *data transmission rate* is

$$R = \frac{\text{\# of transmitted information bits}}{\text{\# of bit transmissions}} = \frac{\log M}{N}.$$

   For example, if the message has $M = 2$ possible outcomes $d_1 =$ Apple and $d_2 =$ Orange, and we use codewords of length $N = 3$. We should wait $N = 3$ time intervals to transmit this message. This message has itself $\log 2 = 1$ bit of information. This means that we transmit 1 bit of information within $N = 3$ transmission, or equivalently, we transmit in average

$$R = \frac{1}{3}$$

bits of information in each transmission. Hence, the transmission rate is $R = 1/3$. The unit of this parameter is then *bits / transmission*. In some texts, it is also stated as *bits/channel-use*, in which by *channel-use* it is meant a single symbol transmission over the channel.

(c) Define error probability $P_{\rm E}$, and write it in terms of Y^N, X^N and M.

♠ **Solution:** The error happens, if the transmitted symbol D is different from $\hat{D}$ estimated by the decoder at the receive side. Therefore, the error probability is defined as

$$\begin{aligned}
P_{\rm E} &= \Pr\left\{\hat{D} \neq D\right\} \\
&= \sum_{i=1}^{M} \Pr\left\{\hat{D} \neq d_i, D = d_i\right\} && \text{(Sum Rule)} \\
&= \sum_{i=1}^{M} \Pr\left\{\hat{D} \neq d_i | D = d_i\right\} \Pr\left\{D = d_i\right\} && \text{(Bayes Rule)} \\
&= \frac{1}{M} \sum_{i=1}^{M} P_{\rm E}\left(d_i\right) && \left(\Pr\left\{D = d_i\right\} = \frac{1}{M}\right)
\end{aligned}$$

where

$$P_{\rm E}\left(d_i\right) := \Pr\left\{\hat{D} \neq d_i | D = d_i\right\}$$

is the error probability when we send the message $D = d_i$ over the channel. We can further expand $P_{\rm E}\left(d_i\right)$ as follows:

$$\begin{aligned}
P_{\rm E}\left(d_i\right) &= \Pr\left\{\hat{D} \neq d_i | D = d_i\right\} \\
&= \Pr\left\{g_N\left(Y^N\right) \neq d_i | D = d_i\right\} && \left(\hat{D} = g_N\left(Y^N\right)\right) \\
&= \Pr\left\{g_N\left(Y^N\right) \neq d_i | X^N = f_N\left(d_i\right)\right\} && \left(D = d_i \Rightarrow X^N = f_N\left(d_i\right)\right)
\end{aligned}$$

Therefore,

$$\boxed{P_{\rm E} = \frac{1}{M} \sum_{i=1}^{M} \Pr\left\{g_N\left(Y^N\right) \neq d_i | X^N = f_N\left(d_i\right)\right\}}$$

Now consider the following three instances of the setting:

- INSTANCE A. D has $M = 2$ possible values. The code-length is $N = 3$, and the channel encoder reads

$$f_N\left(d_1\right) = 000, \qquad f_N\left(d_2\right) = 111.$$

To decode received signal Y^N, the decoder uses major vote meaning that it decodes $\hat{D} = d_1$ if the number of zeros in Y^N, $\hat{D} = d_2$ otherwise. Hence, $g_N(\cdot)$ reads

$$g_N\left(Y^N\right) = \begin{cases} d_1 & N_0 > N_1 \\ d_2 & N_1 > N_0 \end{cases}$$

where N_0 and N_1 are the numbers of zeros and ones in Y^N, respectively.

For this instance,

(d) What is the transmission rate?

♠ **Solution:** In this case, we have

$$R_{\text{A}} = \frac{\log M}{N} = \frac{1}{3} \text{ bits/transmission}$$

(e) Determine distribution of Y^N, when $D = d_1$ is sent over the channel.

♠ **Solution:** To send message $D = d_1$, the encoder transmits $X^N = f_N(d_1) = 000$ over the channel. After three transmissions, the sequence

$$Y^N = Y_1 Y_2 Y_3$$

is received. For symbol Y_1, we have

$$\Pr\{Y_1 = y | D = d_1\} = \Pr\{Y_1 = y | X = 0\} = \begin{cases} 1 - f & y = 0 \\ f & y = 1 \end{cases}.$$

Similar thing occurs for Y_2 and Y_3. Thus,

$$\Pr\left\{Y^N = y^N | D = d_1\right\} = (1-f)^{N_0} f^{N_1}$$

where N_0 and N_1 are the numbers of symbols 0 and symbols 1 in y^N.

(f) Determine the error probability, when $D = d_1$ is sent over the channel.

♠ **Solution:** When we transmit $D = d_1$, the error probability is given by

$$P_{\text{E,A}}(d_1) = \Pr\left\{\hat{D} \neq d_1 | D = d_1\right\}.$$

As shown in Part (c) this error probability is written as

$$P_{\text{E,A}}(d_1) = \Pr\left\{g_N\left(Y^N\right) \neq d_1 | X^N = 000\right\}.$$

The event $g_N\left(Y^N\right) \neq d_1$ happens when $N_1 > N_0$ in Y^N. This means that we have error, if we receive

$$Y^N = 011 \text{ or } 101 \text{ or } 110 \text{ or } 111.$$

As a result, using the result of Part (e), we have

$$\begin{aligned} P_{\mathrm{E,A}}(d_1) =& \Pr\left\{Y^N = 011|D=d_1\right\} + \Pr\left\{Y^N = 101|D=d_1\right\} \\ &+ \Pr\left\{Y^N = 110|D=d_1\right\} + \Pr\left\{Y^N = 111|D=d_1\right\} \\ =& 3(1-f)f^2 + f^3 = f^2(3-2f) \end{aligned}$$

Therefore, in instance A

$$\boxed{P_{\mathrm{E,A}}(d_1) = f^2(3-2f)}$$

- INSTANCE B. D has $M = 4$ possible values. The code-length is $N = 3$, and the channel encoder reads

  $$f_N(d_1) = 001, \qquad f_N(d_2) = 100, \qquad f_N(d_3) = 010, \qquad f_N(d_4) = 111.$$

  The decoder moreover reads

  $$g_N(Y^N) = \begin{cases} d_1 & Y^N = 000 \text{ or } 001 \\ d_2 & Y^N = 100 \text{ or } 101 \\ d_3 & Y^N = 010 \text{ or } 011 \\ d_4 & Y^N = 110 \text{ or } 111 \end{cases}$$

For this instance,

(g) What is the transmission rate?

♠ **Solution:** In this case, we have

$$R_{\mathrm{B}} = \frac{\log M}{N} = \frac{2}{3} \text{ bits/transmission}$$

(h) Determine the error probability, when $D = d_1$ is sent over the channel, and compare it to instance A.

♠ **Solution:** When we transmit $D = d_1$, the error probability is given by

$$\begin{aligned} P_{\mathrm{E,B}}(d_1) &= \Pr\left\{\hat{D} \neq d_1|D=d_1\right\} \\ &= \Pr\left\{g_N(Y^N) \neq d_1|X^N = 001\right\} \\ &= 1 - \Pr\left\{g_N(Y^N) = d_1|X^N = 001\right\}. \end{aligned}$$

The event $g_N(Y^N) = d_1$ happens when we receive

$$Y^N = 000 \text{ or } 001.$$

Hence, the error probability reads

$$\begin{aligned} P_{\mathrm{E,B}}\left(d_1\right) =& 1-\left(\Pr\left\{Y^N=\mathtt{000}|X^N=\mathtt{001}\right\}+\Pr\left\{Y^N=\mathtt{001}|X^N=\mathtt{001}\right\}\right) \\ =& 1-\left(f\left(1-f\right)^2+\left(1-f\right)^3\right)=1-\left(1-f\right)^2 \end{aligned}$$

Therefore, in instance B

$$\boxed{P_{\mathrm{E,B}}\left(d_1\right)=f\left(2-f\right).}$$

For any $f\in[0,1]$, we have $P_{\mathrm{E,B}}\left(d_1\right)\geq P_{\mathrm{E,A}}\left(d_1\right)$.

- INSTANCE C. D has $M=8$ possible values. The code-length is $N=3$, and the channel encoder reads

  $$f_N\left(d_i\right)=\mathtt{binary}\left(i-1\right)$$

  where $\mathtt{binary}\left(i\right)$ determines the binary representation of i. The decoder moreover reads

  $$g_N\left(Y^N\right)=d_i \qquad \text{if} \qquad Y^N=f_N\left(d_i\right).$$

For this instance,

(i) What is the transmission rate?

♠ **Solution:** In this case, we have

$$R_{\mathrm{C}}=\frac{\log M}{N}=\frac{3}{3}=1 \text{ bits/transmission}$$

(j) Determine the error probability, when $D=d_1$ is sent over the channel. Compare it to instances A and B.

♠ **Solution:** When we transmit $D=d_1$, the error probability is given by

$$\begin{aligned} P_{\mathrm{E,C}}\left(d_1\right) &= \Pr\left\{\hat{D}\neq d_1|D=d_1\right\} \\ &= \Pr\left\{g_N\left(Y^N\right)\neq d_1|X^N=\mathtt{binary}\left(0\right)=\mathtt{000}\right\} \\ &= 1-\Pr\left\{g_N\left(Y^N\right)=d_1|X^N=\mathtt{000}\right\}. \end{aligned}$$

The event $g_N\left(Y^N\right)=d_1$ in this instance happens only when we receive

$$Y^N=f_N\left(d_1\right)=\mathtt{000}.$$

Hence, the error probability reads

$$\begin{aligned} P_{\mathrm{E,C}}\left(d_1\right) =& 1-\Pr\left\{Y^N=\mathtt{000}|X^N=\mathtt{000}\right\} \\ =& 1-\left(1-f\right)^3 \end{aligned}$$

Therefore, in instance C

$$\boxed{P_{\mathrm{E,C}}\left(d_1\right)=f\left(3-3f+f^2\right).}$$

For any $f\in[0,1]$, we have $P_{\mathrm{E,C}}\left(d_1\right)\geq P_{\mathrm{E,B}}\left(d_1\right)\geq P_{\mathrm{E,A}}\left(d_1\right)$. The error probability of all three instances is sketched against f in the following figure.

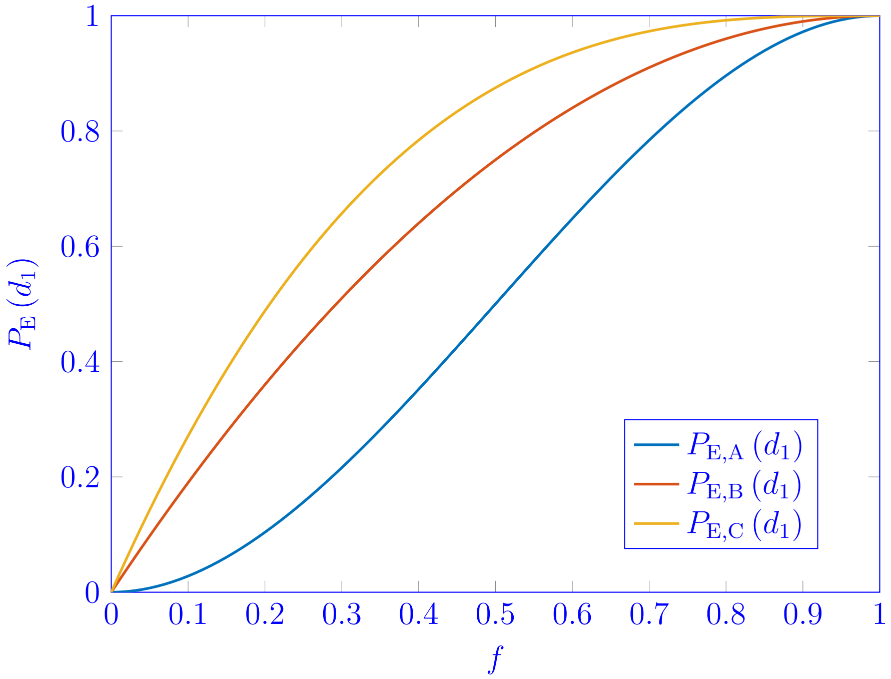


Figure 6.3.1: Error probability versus f.

(k) Plot the rate-error diagram for $N = 3$. Sketch the line $P_E = 0.1$.

♠ **Solution:** The diagram has been plotted below for $f = 0.05$. As the diagram shows, the coding scheme in instances A-C gives error probability $P_E(d_1) \leq 0.1$, if $R \leq R^\star = 0.6851$. Shannon has proved in this second theorem that if we are able to choose N as large as we wish, and search over all possible coding schemes, then the best possible curve would look like the dashed red curve. In this curve, for any arbitrary small $\epsilon > 0$, the line $P_E(d_1) = \epsilon$ meets the curve at the point $R = C = 0.7136$; hence for any $R < C$, we can achieve a significantly small error probability. Nevertheless, if we transmit with rates higher than C, we will have a considerably large error rate. Based on this observation, Shannon called C the capacity of this channel.

Now, let the code-length be $N = 10^4$. Using the Shannon's second theorem, answer the following question.

(l) What is approximately the maximum value for M for which $P_E \leq \epsilon$, for some ϵ close to zero.

♠ **Solution:** Based on the optimal diagram in the previous part, we know that Shannon's curve meets the line $P_E = \epsilon$ for almost any ϵ around $R = C$. This means that the maximum value for R with which $P_E \leq \epsilon$ is $R_{\max} \approx C$. Noting that

$$R = \frac{\log M}{N},$$

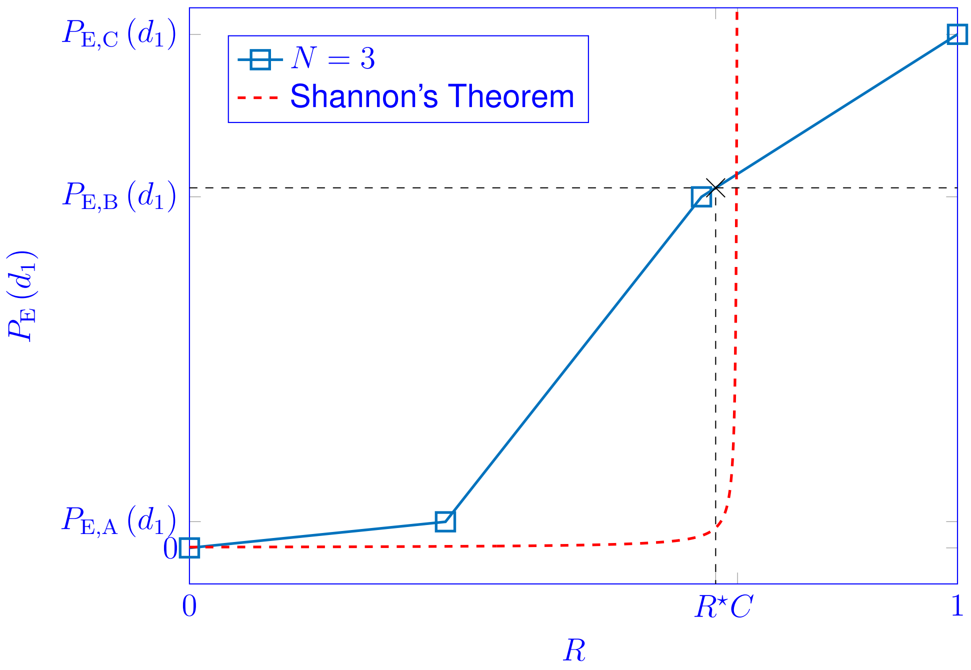


Figure 6.3.2: Error probability versus rate.

we can write

$$\frac{\log M_{\max}}{N} \approx C \qquad \Rightarrow \qquad \boxed{M_{\max} \approx 2^{NC}}.$$

6.4 Homework

The following exercises are suggested for further practice.

6.4.1 Primary exercises

1. Determine the capacity of a channel with five inputs and ten outputs whose transition matrix is

$$\begin{bmatrix} 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0.25 & 0.25 \end{bmatrix}$$

♠ **Solution:** Let us denote the input outcomes of this channel with x_i for $i = 1, \ldots, 5$ and the output outcomes with y_j for $j = 1, \ldots, 10$. In this case the conditional probabilities are given by the following labeling:

$$\mathbf{P}_0 = \begin{array}{c} \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{bmatrix} 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0.25 & 0.25 \end{bmatrix} \end{array}$$

There are two possible ways to find the capacity of this channel:

- **Solution 1:** Following the definition of symmetric channels in page 170 of the book, we see that this is a symmetric channel. In fact, in this matrix all the rows have two entries 0.25 and three entries 0. Similarly, all the columns have four entries 0.25 and six entries 0. This means that all each row is a permutation of other rows, and each column is a permutation of other columns.

As indicated in Exercise 10.10. in page 171, in a symmetric channel the capacity is always achieved via the *uniform* input distribution. Therefore, to find the capacity, we set the input to be uniform, i.e.

$$\Pr\{X = x_i\} = 0.2 \qquad \text{for } i = 1, \ldots, 5.$$

The capacity is then given by calculating the mutual information between X and Y when X is distributed uniformly. To this end, we first find the conditional entropy. For a given input outcome $X = x_i$, $P(y|x_i)$ is given by one of the columns of the transition matrix. Noting that all the columns have four 0.25 and six 0, we have

$$\begin{aligned} H(Y|X = x_i) &= \sum_{j=1}^{10} P(y_j|x_i) \log \frac{1}{P(y_j|x_i)} \\ &= 4 \times \left(0.25 \log \frac{1}{0.25}\right) = 2. \end{aligned}$$

Consequently,

$$H(Y|X) = \sum_{i=1}^{5} \Pr\{X = x_i\} H(Y|X = x_i) = 2.$$

To find $H(Y)$, we determine the marginal distribution of Y, i.e. $P(y)$. This is simply done by multiplying the transition matrix of the channel with the vector of uniform input distribution. This means that

$$\begin{bmatrix} P(y_1) \\ P(y_2) \\ P(y_3) \\ P(y_4) \\ P(y_5) \\ P(y_6) \\ P(y_7) \\ P(y_8) \\ P(y_9) \\ P(y_{10}) \end{bmatrix} = \mathbf{P}_0 \begin{bmatrix} P(x_1) \\ P(x_2) \\ P(x_3) \\ P(x_4) \\ P(x_5) \end{bmatrix} = \begin{bmatrix} 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0.25 & 0.25 \end{bmatrix} \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}$$

$\rightsquigarrow$ REMARK:

Consider a channel with I input and J output symbols. Let $\mathbf{P}_{\text{Channel}}$ be the transition matrix of the channel, and define the probability vectors $\mathbf{p}_X$ and $\mathbf{p}_Y$ as

$$\mathbf{p}_X := \begin{bmatrix} P(x_1) \\ \vdots \\ P(x_I) \end{bmatrix} \qquad \text{and} \qquad \mathbf{p}_Y := \begin{bmatrix} P(y_1) \\ \vdots \\ P(y_J) \end{bmatrix}.$$

Then, we have

$$\boxed{\mathbf{p}_Y = \mathbf{P}_{\text{Channel}}\, \mathbf{p}_X}$$

Since Y is uniformly distributed, its entropy reads

$$H(Y) = \log 10.$$

As a result, the capacity is

$$C = H(Y) - H(Y|X) = \log 10 - 2 = \log \frac{10}{4}.$$

- **Solution 3:** The second solution is to take the classic approach. In this case, we have
  (a) Write the definition of the mutual information which is

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

  We need to determine two terms $H(Y)$ and $H(Y|X)$.

  (b) Assume a general distribution for the input X. For this channel, this is given by the vector

  $$\mathbf{p}_X = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ 1 - q_1 - q_2 - q_3 - q_4 \end{bmatrix}$$

  for some $q_1, \dots, q_4 \geq 0$ which satisfy $\sum_{i=1}^{4} q_i \leq 1$.

  (c) Determine the conditional entropy. This has been done in previous solution where we found

  $$\boxed{H(Y|X) = 2}$$

(d) Determine marginal distribution of Y. This is simply done by

$$\mathbf{p}_Y = \begin{bmatrix} P(y_1) \\ P(y_2) \\ P(y_3) \\ P(y_4) \\ P(y_5) \\ P(y_6) \\ P(y_7) \\ P(y_8) \\ P(y_9) \\ P(y_{10}) \end{bmatrix} = \mathbf{P}_0 \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ 1 - q_1 - q_2 - q_3 - q_4 \end{bmatrix}.$$

The entropy of Y also reds

$$H(Y) = \sum_{j=1}^{10} P(y_j) \log \frac{1}{P(y_j)}.$$

We follow the tip given in Tutorial 5 and avoid explicit calculation of $H(Y)$. At this point, we only indicate that

$$H(Y) = f_{\text{Ent}}(\mathbf{p}_X).$$

We later derive the explicit expression, **if it was needed**.

(e) Substitute the entropy terms into the definition of the mutual information, and then maximize it over all possible input distributions. For this channel, we have

$$I(X;Y) = f_{\text{Ent}}(\mathbf{p}_X) - 2.$$

Hence, the capacity reads

$$C = \max_{\mathbf{p}_X} f_{\text{Ent}}(\mathbf{p}_X) - 2.$$

Here, we need to maximize $f_{\text{Ent}}(\mathbf{p}_X)$ over all possible choices of $\mathbf{p}_X$. Noting that

$$f_{\text{Ent}}(\mathbf{p}_X) = H(Y) \leq \log|\mathcal{A}_Y| = \log 10,$$

we conclude that $\max_{\mathbf{p}_X} f_{\text{Ent}}(\mathbf{p}_X) = \log 10$, and thus

$$C = \log 10 - 2 = \log \frac{10}{4}.$$

To find the input distribution which achieves the capacity, we note that $f_{\text{Ent}}(\mathbf{p}_X) = H(Y) = \log|\mathcal{A}_Y|$, when Y is uniformly distributed. Hence, we need to find the choice of $\mathbf{p}_X$ which results in a uniform output distribution. To this end, we consider the following system of equations:

$$\begin{bmatrix} 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0 & 0 & 0 & 0.25 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0.25 & 0.25 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ 1 - q_1 - q_2 - q_3 - q_4 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}.$$

It is a simple linear equation whose solution is $q_1 = \ldots = q_4 = 0.2$. Hence, the capacity achieving input distribution is

$$\mathbf{p}_X = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}$$

As you see, the explicit derivation of $f_{\text{Ent}}(\mathbf{p}_X)$ **was not required**.

6.4.2 Further exercises from the textbook

- Chapter 9: Exercise 9.10., Exercise 9.16., Exercise 9.17., Exercise 9.18.
- Chapter 10: Exercise 10.4., Exercise 10.5., Exercise 10.6., Exercise 10.7., Exercise 10.8., Exercise 10.10., Exercise 10.11., Exercise 10.12..

6.5 Fun Facts

Apart from the channel capacity, there is also the so-called “Shannon capacity of a *graph*” which is again defined by Shannon in 1956. In contrast to the source and channel coding problems, the Shannon capacity of a graph is left open. The best upper bound for this parameter is given by László Lovász in 1979, i.e. 23 years after it was introduced by Shannon. You can find Lovász’s result on https://ieeexplore.ieee.org/document/1055985.

Chapter 7

Gaussian Channels

While solving exercises on channel coding, you might have noticed problems in which you need to calculate the entropy of a continuous random variable. In this chapter, we are going to address this issue by introducing the concept of *differential entropy*. We then keep on with the channel coding theorem. This chapter is consistent with the remaining parts of Chapter 10 and Chapter 11 of textbook.

7.1 Brief Review of Main Concepts

Before we start, it is good to have a quick overview on some basic definitions regarding continuous random variables.

7.1.1 Continuous random variables

The distribution of a real-valued continuous random variable X is characterized by two functions:

- Cumulative distribution function (CDF) which is defined as

$$F_X\left(x\right) = \Pr\left\{X \leq x\right\}.$$

  This function has these properties:

  1. The function is increasing meaning that $x_1 \leq x_2 \Rightarrow F_X\left(x_1\right) \leq F_X\left(x_2\right)$
  2. The following limits exist

$$\lim_{x \to +\infty} F_X\left(x\right) = \Pr\left\{X \leq +\infty\right\} = 1$$
$$\lim_{x \to -\infty} F_X\left(x\right) = \Pr\left\{X \leq 0\infty\right\} = 0$$

- Probability density function (PDF) which is defined as

$$f_X\left(x\right) = \frac{\mathrm{d}F_X\left(x\right)}{\mathrm{d}x}.$$

The key property of the PDF is

$$\boxed{\int_a^b f_X(x)\,\mathrm{d}x = \Pr\{a < X \le b\}}$$

⇝ REMARK:

PDF at a given point x is the *probability density* of X at point x which can be more than 1. This value is **not** a probability!

7.1.2 Differential entropy

Standard definition of entropy yields always infinity for a continuous random variable; you will see it in the first exercise of this chapter. We hence define as alternative notion of entropy in this case which is called the *differential* entropy:

$$h(X) = \int_{\mathcal{A}_X} f_X(x) \log \frac{1}{f_X(x)} \mathrm{d}x$$

One should note that the new notion of differential entropy $h(X)$ is different from the entropy. Intuitively, the differential entropy calculates the difference between the entropy of a continuous random variable and the entropy of the reference random variable which is uniformly distributed on $[0,1]$. Although they both have infinite entropy, their difference is finite. You will see it in details in the exercises.

The notion of differential entropy extends to conditional differential entropy which is defined as

$$h(Y|X) = \int_{\mathcal{A}_Y} \int_{\mathcal{A}_X} f_{X,Y}(x,y) \log f_{Y|X}(y|x)\,\mathrm{d}x\mathrm{d}y$$

The mutual information between two continuous random variables X and Y is then defined as

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

Using these definitions, we can repeat whatever we have done before, by replacing the entropy with the differential entropy wherever we have a continuous random variable. This will be clarified through the exercises.

7.2 Exercises

7.2.1 Differential entropy

1. Consider *continuous* random variable X uniformly distributed in the interval $[0,1)$.

(a) Let the *discrete* random variable Q_X be a uniform quantization of X with M levels, such that

$$Q_X = m \qquad \text{if} \qquad X \in \left[\frac{m}{M}, \frac{m+1}{M}\right)$$

for $m \in \{0, \dots M-1\}$. Determine $H(Q_X)$.

(b) What is the limit of $H(Q_X)$, when M tends to infinity?

(c) Determine the *differential entropy* of X, i.e. $h(X)$.

2. Consider two *independent* real Gaussian random variables X and Z, where

$$X \sim \mathcal{N}\left(0, \sigma_X^2\right) \qquad \text{and} \qquad Z \sim \mathcal{N}\left(0, \sigma_Z^2\right).$$

Let the random variable Y be

$$Y = X + Z.$$

(a) What is the distribution of Y?

(b) Determine the differential entropies $h(X)$, $h(Z)$ and $h(Y)$.

(c) Determine the conditional differential entropy $h(Y|X)$.

(d) Determine the mutual information $I(X;Y)$.

7.2.2 Capacity of Gaussian channel

♣ Exercise 1 covers Exercise 11.1. and Exercise 11.2. in page 181.

1. Consider the following AWGN channel

$$Y = X + Z$$

where X and Y are the real-valued input and input symbol, respectively, and Z is

$$Z \sim \mathcal{N}\left(0, \sigma_Z^2\right).$$

(a) For this channel, define the capacity without assuming any limitation on the transmitter.

(b) Show that *when we have no power constraint*, the channel has infinite capacity.

(c) Formulate the capacity, when we are restricted to have an average transmit power less than P.

(d) For which distribution, the capacity of the channel with restricted transmit power is achieved?

> **Hint:** For a fixed variance, the differential entropy is maximized by a Gaussian random variable.

(e) Determine the capacity of this channel with transmit power constraint.

7.2.3 Further practice on channel coding

1. Consider a channel with input symbol $X \in \{0, A\}$ where $A \geq 0$ is a real number. The input of this channel is related to the output symbol Y as

$$Y = X + Z$$

where Z is uniformly distributed over $[0, 1]$. This means that the probability density function of Z is

$$f\left(z\right) = \begin{cases} 1 & 0 \leq z \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

(a) Determine the channel capacity.

(b) Find the input distribution with which the capacity of this channel is achieved.

(c) Plot the channel capacity in terms of A.

7.3 Solutions to Exercises

7.3.1 Differential entropy

1. Consider *continuous* random variable X uniformly distributed in the interval $[0, 1)$.

(a) Let the *discrete* random variable Q_X be a uniform quantization of X with M levels, such that

$$Q_X = m \qquad \text{if} \qquad X \in \left[\frac{m}{M}, \frac{m+1}{M}\right)$$

for $m \in \{0, \dots M-1\}$. Determine $H\left(Q_X\right)$.

(b) What is the limit of $H\left(Q_X\right)$, when M tends to infinity?

(c) Determine the *differential entropy* of X, i.e. $h\left(X\right)$.

♠ **Solution:**

(a) In order to determine the entropy of Q_X, we need to find its distribution first. To this end, we note that

$$\begin{aligned} \Pr\left\{Q_X = m\right\} &= \Pr\left\{\frac{m}{M} < X \leq \frac{m+1}{M}\right\} \\ &= \int_{m/M}^{(m+1)/M} f_X\left(x\right) \mathrm{d}x \end{aligned}$$

As X is uniformly distributed, we have

$$f_X\left(x\right) = \begin{cases} 1 & x \in (0, 1] \\ 0 & \text{otherwise} \end{cases}.$$

Therefore,

$$\Pr\{Q_X = m\} = \int_{m/M}^{(m+1)/M} 1 \,\mathrm{d}x = \frac{1}{M}.$$

This means that Q_X is a uniform random variable, and therefore

$$H(Q_X) = \log M$$

(b) When M tends to infinity, we have

$$\lim_{M\to\infty} H(Q_X) = \lim_{M\to\infty} \log M = \infty$$

This means that for very precise quantization, the entropy is infinite. This is the case for all continuous random variables. We hence cannot define *entropy* for these variables. This is the reason that we define the *differential entropy* in this case.

(c) From page 180 of the textbook, we can write the differential entropy of a continuous random variable as

$$h(X) = \int_{\mathcal{A}_X} f_X(x) \log \frac{1}{f_X(x)} \mathrm{d}x.$$

For the uniform random variable, we have

$$h(X) = \int_0^1 1 \ \log 1 \ \mathrm{d}x = 0.$$

⇝ REMARK:

Differential entropy $h(X)$ is different from entropy $H(X)$.

2. Consider two *independent* real Gaussian random variables X and Z, where

$$X \sim \mathcal{N}(0, \sigma_X^2) \qquad \text{and} \qquad Z \sim \mathcal{N}(0, \sigma_Z^2).$$

Let the random variable Y be

$$Y = X + Z.$$

(a) What is the distribution of Y?

(b) Determine the differential entropies $h(X)$, $h(Z)$ and $h(Y)$.

(c) Determine the conditional differential entropy $h(Y|X)$.

(d) Determine the mutual information $I(X;Y)$.

♠ **Solution:**

(a) As X and Z are both Gaussian, we can conclude that Y is Gaussian too.

⇝ Reminder:

Sum of multiple Gaussian random variables is a Gaussian random variable.

In this case,

$$\mathcal{E}\left\{Y\right\} = \mathcal{E}\left\{X+Z\right\} = \mathcal{E}\left\{X\right\} + \mathcal{E}\left\{Z\right\} = 0,$$

and

$$\begin{aligned}\sigma_Y^2 = \mathcal{E}\left\{(Y-\mathcal{E}\left\{Y\right\})^2\right\} = \mathcal{E}\left\{Y^2\right\} \\ &= \mathcal{E}\left\{(X+Z)^2\right\} = \mathcal{E}\left\{X^2\right\} + 2\mathcal{E}\left\{XZ\right\} + \mathcal{E}\left\{Z^2\right\} \\ &= \sigma_X^2 + 2\mathcal{E}\left\{XZ\right\} + \sigma_Z^2\end{aligned}$$

We now remember that

⇝ Reminder:

For two independent random variables X and Z, we have

$$\boxed{\mathcal{E}\left\{XZ\right\} = \mathcal{E}\left\{X\right\}\mathcal{E}\left\{Z\right\}}$$

Therefore, $\mathcal{E}\left\{XZ\right\} = \mathcal{E}\left\{X\right\}\mathcal{E}\left\{Z\right\} = 0$ and

$$\sigma_Y^2 = \sigma_X^2 + \sigma_Z^2.$$

This means that $Y \sim \mathcal{N}\left(0, \sigma_X^2 + \sigma_Z^2\right)$.

⇝ Remark:

Let $X_1, \ldots, X_K$ be K jointly independent Gaussian random variable, such that

$$X_k \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right).$$

Then, the sum random variable reads

$$Y = \sum_{k=1}^{K} X_k \sim \mathcal{N}\left(\sum_{k=1}^{K}\mu_k, \sum_{k=1}^{K}\sigma_k^2\right).$$

(b) As they are all Gaussian random variables, it is sufficient to find the differential entropy for one. Writing the definition of differential entropy, we have

$$h\left(X\right) = \int_{\mathcal{A}_X} f_X\left(x\right) \log \frac{1}{f_X\left(x\right)} \mathrm{d}x.$$

The PDF of the random variable X is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left\{-\frac{x^2}{2\sigma_X^2}\right\}.$$

Therefore, we have

$$\begin{aligned} h(X) &= \int_{-\infty}^{\infty} f_X(x) \log\left(\sqrt{2\pi\sigma_X^2} \exp\left\{\frac{x^2}{2\sigma_X^2}\right\}\right) \mathrm{d}x \\ &= \int_{-\infty}^{\infty} f_X(x) \log\left(\sqrt{2\pi\sigma_X^2}\right) \mathrm{d}x + \int_{-\infty}^{\infty} f_X(x) \log\left(\exp\left\{\frac{x^2}{2\sigma_X^2}\right\}\right) \mathrm{d}x \\ &= \frac{1}{2} \log 2\pi\sigma_X^2 \underbrace{\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x}_{1} + \frac{\log e}{2\sigma_X^2} \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x. \end{aligned}$$

As $\mathcal{E}\{X\} = 0$,

$$\int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x = \mathcal{E}\left\{X^2\right\} = \mathcal{E}\left\{(X - \mathcal{E}\{X\})^2\right\} = \sigma_X^2.$$

Thus,

$$h(X) = \frac{1}{2} \log 2\pi\sigma_X^2 + \frac{\log e}{2\sigma_X^2} \underbrace{\int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x}_{\sigma_X^2} = \frac{1}{2} \log 2\pi\sigma_X^2 + \frac{1}{2} \log e.$$

Using basic properties of logarithm function we can conclude that

$$\boxed{h(X) = \frac{1}{2} \log 2\pi e\sigma_X^2}$$

Similarly, we can write

$$\begin{aligned} h(Z) &= \frac{1}{2} \log 2\pi e\sigma_Z^2 \\ h(Y) &= \frac{1}{2} \log 2\pi e\sigma_Y^2 \end{aligned}$$

$\rightsquigarrow$ REMARK:

It is shown that for any random variable X which satisfies

$$\mathcal{E}\left\{(X - \mathcal{E}\{X\})^2\right\} \leq \sigma^2,$$

we have

$$h(X) \leq \frac{1}{2} \log 2\pi e\sigma^2.$$

This means that for a given fix variance, Gaussian random variable has the maximum differential entropy.

(c) Analogous to the case for discrete random variables, we can define the conditional differential entropy as

$$\boxed{h\left(Y|X\right)=\int_{\mathcal{A}_Y}\int_{\mathcal{A}_X} f_{X,Y}\left(x,y\right)\log f_{Y|X}\left(y|x\right)\mathrm{d}x\mathrm{d}y}$$

Here, $f_{X,Y}\left(x,y\right)$ is the joint PDF of X and Y, and $f_{Y|X}\left(y|x\right)$ is the conditional PDF defined as

$$f_{Y|X}\left(y|x\right)=\frac{f_{X,Y}\left(x,y\right)}{f_X\left(x\right)}.$$

Similar to the conditional entropy, for the conditional differential entropy we have

$$\boxed{h\left(Y|X\right)=\int_{\mathcal{A}_X} f_X\left(x\right)h\left(Y|X=x\right)\mathrm{d}x}$$

where

$$h\left(Y|X=x\right)=\int_{\mathcal{A}_Y} f_{Y|X}\left(y|x\right)\log f_{Y|X}\left(y|x\right)\mathrm{d}y.$$

In order to determine the conditional differential entropy, we first need to determine the conditional PDF. To this end, we first determine the conditional CDF as follows

$$\begin{aligned}F_{Y|X}\left(y|x\right)&=\Pr\left\{Y\leq y|X=x\right\}\\&=\Pr\left\{X+Z\leq y|X=x\right\}=\Pr\left\{x+Z\leq y\right\}=\Pr\left\{Z\leq y-x\right\}\end{aligned}$$

Noting that Z is a Gaussian variable, we can write

$$F_{Y|X}\left(y|x\right)=\int_{-\infty}^{y-x} f_Z\left(z\right)\mathrm{d}z=\frac{1}{\sqrt{2\pi\sigma_Z^2}}\int_{-\infty}^{y-x}\exp\left\{-\frac{z^2}{2\sigma_Z^2}\right\}\mathrm{d}z$$

Therefore,

$$\begin{aligned}f_{Y|X}\left(y|x\right)&=\frac{\partial}{\partial y}F_{Y|X}\left(y|x\right)\\&=\frac{\partial}{\partial y}\int_{-\infty}^{y-x} f_Z\left(z\right)\mathrm{d}z\\&=f_Z\left(y-x\right)=\frac{1}{\sqrt{2\pi\sigma_Z^2}}\exp\left\{-\frac{\left(y-x\right)^2}{2\sigma_Z^2}\right\}\sim\mathcal{N}\left(x,\sigma_Z^2\right).\end{aligned}$$

This means that given $X=x$, Y is distributed Gaussian with mean x and variance σ_Z^2. As the result, $h\left(Y|X=x\right)$ is the differential entropy of a Gaussian random variable with mean x and variance σ_Z^2. For this distribution, the differential entropy is derived similar to Part (b) which is

$$h\left(Y|X=x\right)=\frac{1}{2}\log 2\pi e\sigma_Z^2.$$

⇝ REMARK:

For the differential entropy, we have in general

$$\boxed{h(X) = h(X + \mu)}$$

For any constant μ.

The conditional entropy is hence given by

$$h(Y|X) = \int_{\mathcal{A}_X} f_X(x)\, h(Y|X = x)\, \mathrm{d}x = \int_{-\infty}^{+\infty} f_X(x) \log 2\pi e \sigma_Z^2 \mathrm{d}x = \frac{1}{2} \log 2\pi e \sigma_Z^2.$$

(d) The mutual information is defined as

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

Using the results of previous parts, we have

$$\begin{aligned} I(X;Y) &= \frac{1}{2} \log 2\pi e \sigma_Y^2 - \frac{1}{2} \log 2\pi e \sigma_Z^2 \\ &= \frac{1}{2} \log \frac{\sigma_Y^2}{\sigma_Z^2} = \frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right). \end{aligned}$$

7.3.2 Capacity of Gaussian channel

♣ Exercise 1 covers Exercise 11.1. and Exercise 11.2. in page 181.

1. Consider the following AWGN channel

$$Y = X + Z$$

where X and Y are the real-valued input and input symbol, respectively, and Z is

$$Z \sim \mathcal{N}\left(0, \sigma_Z^2\right).$$

(a) For this channel, define the capacity without assuming any limitation on the transmitter.

(b) Show that *when we have no power constraint*, the channel has infinite capacity.

(c) Formulate the capacity, when we are restricted to have an average transmit power less than P.

(d) For which distribution, the capacity of the channel with restricted transmit power is achieved?

Hint: For a fixed variance, the differential entropy is maximized by a Gaussian random variable.

(e) Determine the capacity of this channel with transmit power constraint.

♠ **Solution:**

(a) As for channels with discrete input and output symbols, the channel capacity is defined as

$$C = \max_{f_X(x)} I(X;Y).$$

Here, the maximization is done over all possible PDFs.

(b) Assume that X_0 is a Gaussian random variable with zero mean and variance σ_X^2. Let us denote, the output of the channel for this input, as

$$Y_0 = X_0 + Z$$

Since the channel capacity is defined as the maximum of the mutual information, we could say that

$$I(X_0;Y_0) \leq \max_{f_X(x)} I(X;Y) = C.$$

As shown in the precious exercise, $I(X_0;Y_0)$ reads

$$I(X_0;Y_0) = \frac{1}{2}\log\left(1+\frac{\sigma_X^2}{\sigma_Z^2}\right).$$

Hence, we can conclude that

$$\frac{1}{2}\log\left(1+\frac{\sigma_X^2}{\sigma_Z^2}\right) \leq C.$$

Since we have no constraint on the power, the transmit power, i.e. $\mathcal{E}\{X^2\} = \sigma_X^2$ can be chosen unboundedly large. Thus, we can say

$$C \geq \lim_{\sigma_X^2 \to \infty} \frac{1}{2}\log\left(1+\frac{\sigma_X^2}{\sigma_Z^2}\right) = \infty.$$

This means that the capacity of the channel in this case is infinite.

> ⇝ REMARK:
>
> This result is intuitively true. In fact, when we have no constraint on the transmit power, we can transmit infinitely high power signals. In this case, the impact of noise is almost removed and hence we can transmit as much information as we wish without any error.

(c) When we have power constraint P on the channel input, we should restrict the optimization to the PDFs whose $\mathcal{E}\{X^2\}$ are less than or equal to P. This means that we should search over all PDFs $f_X(x)$ which satisfy

$$\mathcal{E}\{X^2\} = \int_{\mathcal{A}_X} x^2 f_X(x)\,\mathrm{d}x \leq P$$

Hence, the channel capacity in this case reads

$$\boxed{C = \max_{f_X(x)} I(X;Y) \qquad \text{subject to } \int_{\mathcal{A}_X} x^2 f_X(x)\,\mathrm{d}x \leq P.}$$

In other words, if we define the set $\mathcal{S}_X(P)$ as the set of PDFs $f_X(x)$ whose $\mathcal{E}\{X^2\} \leq P$, i.e.

$$\mathcal{S}_X(P) = \left\{ \text{all } f_X(x) : \int_{\mathcal{A}_X} x^2 f_X(x)\,\mathrm{d}x \leq P \right\},$$

then, the capacity is given by

$$C = \max_{f_X(x) \in \mathcal{S}_X(P)} I(X;Y).$$

(d) The direct approach for the derivation of the channel capacity and its corresponding PDF is given in page 189 of the book. We however give an solution by following the classic approach:

(a) Write the mutual information as

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

(b) Consider a general distribution on X. In this example, it would be $X \sim f_X(x)$ which satisfies

$$\int_{\mathcal{A}_X} x^2 f_X(x)\,\mathrm{d}x \leq P.$$

(c) Determine the conditional differential entropy which in this example reads

$$h(Y|X) = \int_{\mathcal{A}_X} h(Y|X=x) f_X(x)\,\mathrm{d}x.$$

As shown in the previous exercise, for the outcome $X = x$, the output Y is conditionally a Gaussian random variable with mean x and variance σ_Z^2. Thus,

$$h(Y|X=x) = \frac{1}{2}\log 2\pi e \sigma_Z^2,$$

and

$$h(Y|X) = \int_{\mathcal{A}_X} h(Y|X=x) f_X(x)\,\mathrm{d}x = \frac{1}{2}\log 2\pi e \sigma_Z^2.$$

(d) Now, we should determine the output differential entropy. As indicated in the approach, we use the tip in Tutorial 5, and just consider this differential entropy as a function of the input distribution. We later might derive the explicit expression, if it would be needed. Hence, we could write

$$h(Y) = f_{\text{Ent}}(f_X(x)).$$

Noting that X is independent of Z and its variance is less than P, we could further write

$$\mathcal{E}\left\{Y\right\} = \mathcal{E}\left\{X+Z\right\} = \mathcal{E}\left\{X\right\} + \underbrace{\mathcal{E}\left\{Z\right\}}_{0} = \mathcal{E}\left\{X\right\},$$

and

$$\begin{aligned}\mathcal{E}\left\{Y^2\right\} = \mathcal{E}\left\{(X+Z)^2\right\} &= \mathcal{E}\left\{X^2\right\} + 2\mathcal{E}\left\{XZ\right\} + \mathcal{E}\left\{Z^2\right\} \\ &= \mathcal{E}\left\{X^2\right\} + 2\mathcal{E}\left\{X\right\}\underbrace{\mathcal{E}\left\{Z\right\}}_{0} + \mathcal{E}\left\{Z^2\right\} \leq P + \sigma_Z^2.\end{aligned}$$

Therefore, we conclude that $f_{\rm Ent}\left(f_X\left(x\right)\right)$ is the differential entropy of a random variable with mean $\mathcal{E}\left\{X\right\}$ and $\mathcal{E}\left\{Y^2\right\}$ less than or equal to $P+\sigma_Z^2$.

(e) We now substitute the derived expressions into the definition of the mutual information, and maximize it over all possible input distributions. To this end, we write

$$\begin{aligned}I\left(X;Y\right) &= h\left(Y\right) - h\left(Y|X\right) \\ &= f_{\rm Ent}\left(f_X\left(x\right)\right) - \frac{1}{2}\log 2\pi e \sigma_Z^2.\end{aligned}$$

Therefore, the capacity is given by maximizing $f_{\rm Ent}\left(f_X\left(x\right)\right)$ over all possible input PDFs. To this end, we refer to the remark given in Exercise 2 of the previous section:

> ⇝ REMINDER:
>
> We know that the differential entropy for a random variable with restricted variance is maximized, when the random variable is Gaussian with the given upper bound on the variance.

We hence can conclude that $f_{\rm Ent}\left(f_X\left(x\right)\right)$ is maximized, when Y is a zero-mean Gaussian variable with variance $P+\sigma_Z^2$. This means that

$$\max_{f_X(x)\in\mathcal{S}_X(P)} f_{\rm Ent}\left(f_X\left(x\right)\right) = \frac{1}{2}\log 2\pi e\left(P+\sigma_Z^2\right)$$

and is achieved when

$$X \sim \mathcal{N}\left(0,P\right).$$

(e) Based on the discussions in the previous part, we have

$$C = \frac{1}{2}\log\left(1+\frac{P}{\sigma_Z^2}\right).$$

7.3.3 Further practice on channel coding

1. Consider a channel with input symbol $X \in \{0, A\}$ where $A \geq 0$ is a real number. The input of this channel is related to the output symbol Y as

$$Y = X + Z$$

where Z is uniformly distributed over $[0, 1]$. This means that the probability density function of Z is

$$f(z) = \begin{cases} 1 & 0 \leq z \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

(a) Determine the channel capacity.

(b) Find the input distribution with which the capacity of this channel is achieved.

(c) Plot the channel capacity in terms of A.

♠ **Solution:** This exercise describes a case in which the input and output of the channel are of different type: one is *discrete* and one is *continuous*. In the sequel, we see how all our calculations explained before extend to this case.

(a) To calculate the channel capacity, we start with our classic approach illustrated in Tutorial 5:

i. First we write the definition of the mutual information:

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

We hence need to calculate $h(Y)$ and $h(Y|X)$.

⇝ REMARK:

Note that Y is continuous random variable. Thus, we calculate its *differential* entropy.

ii. We now assume a *general* distribution for the input. Here, the input is a discrete random variable, hence we should consider a *probability distribution*. For the binary input, the most general case is the Bernoulli distribution. We hence set the input distribution as

$$\Pr\{X = 0\} = 1 - \Pr\{X = A\} = q$$

where $0 \leq q \leq 1$.

iii. We now calculate the conditional differential entropy $h(Y|X)$. To this end, we use the classic trick: writing $h(Y|X)$ in terms of $h(Y|X = x)$ for all possible outcomes $X = x$.

When input is *discrete* and output is *continuous*, the expansion is of the following form:

$$h(Y|X) = \sum_{x \in \mathcal{A}_X} h(Y|X = x) \Pr\{X = x\}$$

⇝ REMINDER:

We can summarize all expansion cases as bellow:

- *Discrete* input and *discrete* output

$$H\left(Y|X\right)=\sum_{x\in\mathcal{A}_X}H\left(Y|X=x\right)\Pr\left\{X=x\right\}$$

- *Continuous* input and *continuous* output

$$h\left(Y|X\right)=\int_{-\infty}^{+\infty}h\left(Y|X=x\right)f\left(x\right)\mathrm{d}x$$

- *Discrete* input and *continuous* output

$$h\left(Y|X\right)=\sum_{x\in\mathcal{A}_X}h\left(Y|X=x\right)\Pr\left\{X=x\right\}$$

⇝ REMARK:

We have always the following counterparts for discrete and continuous random variables

$$h\left(\cdot\right)\Longleftrightarrow H\left(\cdot\right)$$
$$\int_{-\infty}^{+\infty}\left(\cdot\right)f\left(x\right)\mathrm{d}x\Longleftrightarrow\sum_{x\in\mathcal{A}_X}\left(\cdot\right)\Pr\left\{X=x\right\}$$

Since we have only two outcomes for X, i.e., $\mathcal{A}_X=\{0,A\}$, we have

$$h\left(Y|X\right)=h\left(Y|X=0\right)q+h\left(Y|X=A\right)\left(1-q\right).$$

each term on the right hand side is calculated as follows:

- $h\left(Y|X=0\right)$ is the differential entropy of output Y, when X is set to zero. This means that

$$h\left(Y|X=0\right)=\int_{-\infty}^{+\infty}f\left(y|x=0\right)\log\frac{1}{f\left(y|x=0\right)}\mathrm{d}y$$

  As we set $X=0$, $Y=0+Z=Z$. Hence, $f\left(y|x=0\right)$ is the PDF of a uniform random variable in interval $[0,1]$. This means that

$$f\left(y|x=0\right)=\begin{cases}1 & 0\leq y\leq 1\\ 0 & \text{otherwise}\end{cases}.$$

As a result, we have

$$h\left(Y|X=0\right) = \int_0^1 \log 1 \mathrm{d}y = 0.$$

- $h\left(Y|X=A\right)$ is the differential entropy of output Y, when X is set to A. This means that

$$h\left(Y|X=A\right) = \int_{-\infty}^{+\infty} f\left(y|x=A\right) \log \frac{1}{f\left(y|x=A\right)} \mathrm{d}y$$

As we set $X = A$, $Y = A + Z$. Hence, in this case, Y is a continuous uniform random variable in interval $[A, 1+A]$. This means that

$$f\left(y|x=A\right) = \begin{cases} 1 & A \leq y \leq 1+A \\ 0 & \text{otherwise} \end{cases}.$$

As a result, we have

$$h\left(Y|X=A\right) = \int_A^{1+A} \log 1 \mathrm{d}y = 0.$$

Consequently, the conditional differential entropy reads

$$h\left(Y|X\right) = 0.$$

iv. As the next step, we calculate the distribution of Y. To this end, we use *marginalization* which in this case reads

> When X is *discrete* and Y is *continuous*, the marginalization rule is of the following form:
>
> $$f\left(y\right) = \sum_{x \in \mathcal{A}_X} f\left(y|x\right) \Pr\left\{X = x\right\}$$

We hence have

$$f\left(y\right) = f\left(y|x=0\right) q + f\left(y|x=A\right)\left(1-q\right).$$

Considering $f\left(y|x=0\right)$ and $f\left(y|x=A\right)$ being calculated in the previous step, we can conclude that

- When $A < 1$,

$$f\left(y\right) = \begin{cases} q & 0 \leq y \leq A \\ 1 & A \leq y \leq 1 \\ 1-q & 1 \leq y \leq 1+A \\ 0 & \text{otherwise} \end{cases}.$$

- When $A \geq 1$,

$$f(y) = \begin{cases} q & 0 \leq y \leq 1 \\ 1-q & A \leq y \leq 1+A \\ 0 & \text{otherwise} \end{cases}.$$

The differential entropy hence is calculated directly from $f(y)$ as

$$h(Y) = \int_{-\infty}^{+\infty} f(y) \log \frac{1}{f(y)} \mathrm{d}y$$

which for given $f(y)$ reads

- When $A < 1$,

$$\begin{aligned} h(Y) &= \underbrace{\int_0^A q \log \frac{1}{q} \mathrm{d}y}_{Aq \log \frac{1}{q}} + \underbrace{\int_A^1 \log 1 \mathrm{d}y}_{0} + \underbrace{\int_1^{1+A} (1-q) \log \frac{1}{1-q} \mathrm{d}y}_{A(1-q) \log \frac{1}{1-q}} \\ &= A \left(q \log \frac{1}{q} + (1-q) \log \frac{1}{1-q} \right) = AH_2(q). \end{aligned}$$

- When $A \geq 1$,

$$\begin{aligned} h(Y) &= \underbrace{\int_0^1 q \log \frac{1}{q} \mathrm{d}y}_{q \log \frac{1}{q}} + \underbrace{\int_A^{1+A} (1-q) \log \frac{1}{1-q} \mathrm{d}y}_{(1-q) \log \frac{1}{1-q}} \\ &= q \log \frac{1}{q} + (1-q) \log \frac{1}{1-q} = H_2(q). \end{aligned}$$

We hence conclude that

$$\begin{aligned} h(Y) &= \begin{cases} AH_2(q) & A < 1 \\ H_2(q) & A \geq 1 \end{cases} \\ &= H_2(q) \underbrace{\begin{cases} A & A < 1 \\ 1 & A \geq 1 \end{cases}}_{\min\{A,1\}} = \min\{A,1\} H_2(q). \end{aligned}$$

v. As te final step, we calculate the mutual information and maximize it over $q \in [0, 0.5]$. From previous steps, we have

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= \min\{A,1\} H_2(q) - 0 = \min\{A,1\} H_2(q) \end{aligned}$$

The capacity is hence determined as

$$\begin{aligned} C &= \max_{P(x)} I\left(X;Y\right) \\ &= \max_{q\in[0,0.5]} \min\left\{A,1\right\} H_2\left(q\right) \\ &= \min\left\{A,1\right\} \underbrace{\max_{q\in[0,0.5]} H_2\left(q\right)}_{=1 \text{ at } q=0.5} \\ &= \min\left\{A,1\right\}. \end{aligned}$$

(b) As it was shown in Part (a), the capacity is achieved for $q = 0.5$. This means that *uniform binary input* achieves the capacity of this channel.

(c) From Part (a), we know that $C = \min\left\{A,1\right\}$. Thus, the capacity for this channel in terms of A is plotted as below

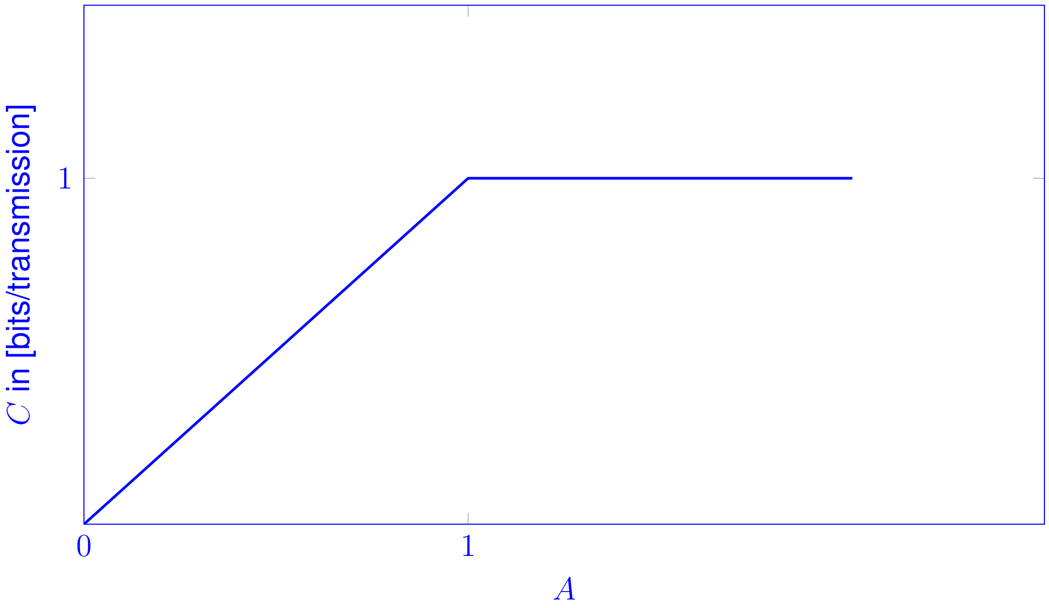


Figure 7.3.1: Capacity versus A.

7.4 Homework

The following exercises are suggested for further practice.

7.4.1 Primary exercises

1. Let X be a Gaussian random variable with mean μ and variance σ^2. This means that

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Calculate the differential entropy of X.

♠ **Solution:** $\boxed{h(X) = \frac{1}{2}\log 2\pi e\sigma^2}$ which is the same as a zero-mean Gaussian random variable with variance σ^2.

2. Determine the differential entropy of following random variables:

   (a) X which is uniformly distributed on $[0, 0.5)$.

   ♠ **Solution:** $h(X) = -1$

   (b) Y which is exponentially distributed with parameter λ, i.e.,

$$f(y) = \begin{cases} \lambda \exp\{-\lambda y\} & y \geq 0 \\ 0 & y < 0 \end{cases}$$

   ♠ **Solution:** Using the fact that

$$\mathcal{E}\{Y\} = \int_0^{+\infty} y f(y)\, \mathrm{d}y = \frac{1}{\lambda},$$

   we conclude that $h(Y) = \log\frac{e}{\lambda}$.

3. Starting from the definition of differential entropy, show that

$$h(X) = h(X + c)$$

where c is a constant real number.

♠ **Solution:** Simply google the solution!

4. Starting from the definition of differential entropy, show that

$$h(aX) = h(X) + \log|a|$$

where a is a constant real number.

♠ **Solution:** Simply google the solution!

> ⇝ REMARK:
>
> This property is different from the one we have for discrete random variables.
>
> Remember that for discrete random variables, we have
>
> $$H(aX) = H(X)$$
>
> for any a.

5. Starting from the definition of differential entropy, prove the *chain rule*, i.e., show that

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

♠ **Solution:** Simply google the solution!

7.4.2 Further exercises from the textbook

- Chapter 11: Exercise 11.1., Exercise 11.5.

7.5 Extra Notes on Channel Capacity

The following two sections are *not* typical exercises you would get in the exam, and are mainly designed for your better understanding of the channel coding theorem. You could hence ignore them, if you are preparing for the exam.

7.5.1 Nice toy example

1. Consider a BSC with flipping probability f. Assume we want to transmit message d_1 which consists of $\lfloor NR \rfloor$ bits with $0 \leq R \leq 1$. To this end, we encode d_1 as

$$X^N = f_N(d_1) = \underbrace{0 \ldots 0}_{N \text{ times}},$$

and send it over the channel in N consecutive intervals. Let Y^N be the sequence received by the receiver.

   (a) Given the fact that X^N is passed through a BSC, what is a *typical* Y^N when d_1 is sent.

   (b) Determine the number of typical realizations for Y^N, and approximate it using Stirling's approximation.

   Now, assume that we decode a receive sequence as follows:

   $\boxed{Y^N \text{ is decoded as } d_1, \text{ if it is a typical receive sequence for } X^N}$

   (c) For this decoding, find a necessary condition on R, such that the error probability converges to zero.

♠ **Solution:**

(a) If we transmit message d_1, X^N is an all-zero sequence. In this case, the distribution of Y^N reads

$$\Pr\left\{Y^N = y^N | X^N = 0 \dots 0\right\} = \prod_{n=1}^{N} \Pr\left\{Y_n = y_n | X_n = 0\right\}.$$

Since we transmit all the symbols X_n over the BSC, we can say that for all $n \in \{1, \dots, N\}$,

$$\Pr\left\{Y_n = y_n | X_n = 0\right\} = P\left(y_n | 0\right) = \begin{cases} 1-f & y_n = 0 \\ f & y_n = 1 \end{cases}.$$

Therefore, we have

$$\Pr\left\{Y^N = y^N | X^N = 0 \dots 0\right\} = \prod_{n=1}^{N} P\left(y_n | 0\right).$$

This indicate that, conditioned to transmission of d_1, Y^N is an i.i.d. Bernoulli sequence with:

$$\Pr\left\{Y_n = 1 | X_n = 0\right\} = 1 - \Pr\left\{Y_n = 0 | X_n = 0\right\} = f$$

A *typical* Y^N, as discussed in Tutorial 2, is a sequence with

$$N_1 \approx Nf$$
$$N_0 \approx N\left(1-f\right)$$

where N_0 and N_1 are the numbers of zeros and ones in Y^N, respectively.

(b) Using the result of the previous part, we can say that the number of typical Y^Ns, when message d_1 is sent over the channel, is

$$\text{\# of typical } Y^N\text{s when } d_1 \text{ is sent} = \mathcal{T}_Y\left(d_1\right) \approx \binom{N}{Nf}.$$

Using Stirling's approximation in Chapter 1 of the book, we have

$$\mathcal{T}_Y\left(d_1\right) \approx 2^{NH_2(f)}.$$

(c) In order to have small error probability, we should make sure that a given realization of Y^N is not decoded as two different messages. This means that two different messages *should not have a same typical* Y^N. We know that for message d_1, we have $\mathcal{T}_Y\left(d_1\right)$ typical sequences. Since the channel is symmetric, we can guess that for other messages we have approximately similar number of typical sequences. A necessary condition for these typical sequences to be distinct is that

$$\sum_{m=1}^{M} T_Y\left(d_m\right) \leq \text{\# of all possible realizations of } Y^N = 2^N.$$

This means that

$$MT_Y(d_1) \leq 2^N$$
$$2^{NR}2^{NH_2(f)} \leq 2^N \Rightarrow N(R + H_2(f)) \leq N$$

This concludes that a necessary condition to have a small error probability is that

$$\boxed{R \leq 1 - H_2(f)}.$$

This bound is actually the capacity of the channel.

7.5.2 Capacity of Gaussian channel: Operational meaning

1. Consider the AWGN channel over which we could transmit with maximum average power P. We intend to send B bits of information over it. To this end, we do the following:

   - We divide the interval $[-\sqrt{P}, +\sqrt{P}]$ into $M = 2^B$ *signal levels*, $x_0, \ldots, x_{M-1}$, where
   $$x_m = -\sqrt{P} + m\frac{2\sqrt{P}}{M-1}$$
   - To transmit $\mathtt{b}_1 \ldots \mathtt{b}_B$, we send x_m over the channel, where m is the decimal representation of $\mathtt{b}_1 \ldots \mathtt{b}_B$. For example, to transmit the sequence of all zeros, $\mathtt{0} \ldots \mathtt{0}$, we transmit $x_0 = -\sqrt{P}$.
   - Given received symbol Y, $x_{\hat{m}}$ is estimated as the transmitted symbol, where $x_{\hat{m}}$ is the closest level to Y.

   We intend to investigate the efficiency of this transmission scheme by means of the Shannon's theorem.

   (a) What is the transmission rate?

   ♠ **Solution:** In each symbol transmission, we are sending $M = 2^B$ information bits over the channel. Hence, the rate is

   $$R = B = \log M \,\frac{\text{bits}}{\text{transmissions}}.$$

   (b) Assuming that $\mathtt{b}_1 \ldots \mathtt{b}_B$ is an i.i.d. sequence with uniformly distributed bits, what is the distribution of the channel input X in this case.

   ♠ **Solution:** Since $\mathtt{b}_1 \ldots \mathtt{b}_B$ is an i.i.d. and uniform binary sequence, we can conclude each symbol level x_m occurs with probability

   $$\Pr\{X = x_m\} = \Pr\{\mathtt{b}_1 \ldots \mathtt{b}_B = \mathtt{binary}(x_m)\} = \left(\frac{1}{2}\right)^B = \frac{1}{M}.$$

   Hence, X is a uniform random variable.

(c) Determine the average transmit power.

> **Hint:** Use the following identity
>
> $$\sum_{m=0}^{M-1} (2m - M + 1)^2 = \frac{1}{3} M (M-1)(M+1).$$

♠ **Solution:** The average transmit power in this case is

$$\begin{aligned}\mathcal{E}\left(X^2\right) = \sum_{m=0}^{M-1} \Pr\left\{X = x_m\right\} x_m^2 &= \frac{1}{M} \sum_{m=0}^{M-1} \left(\frac{\sqrt{P}\,(2m - M + 1)}{M-1}\right)^2 \\ &= \frac{P}{M\,(M-1)^2} \sum_{m=0}^{M-1} (2m - M + 1)^2 .\end{aligned}$$

Using the identity in the hint, we can conclude that

$$\mathcal{E}\left(X^2\right) = \frac{P}{3M\,(M-1)^2} M(M-1)(M+1) = \frac{P}{3}\left(\frac{M+1}{M-1}\right).$$

(d) Write the capacity of this channel, assuming that the average transmit power is restricted by the result of Part (b), and approximate it when M and P are both very large.

♠ **Solution:** Following the result of the previous exercise, we know that

$$C = \frac{1}{2}\log\left(1 + \frac{\mathcal{E}\left(X^2\right)}{\sigma_Z^2}\right) = \frac{1}{2}\log\left(1 + \frac{P}{3}\left(\frac{M+1}{M-1}\right)\right).$$

When $M \to \infty$, we have

$$\lim_{M\to\infty} \frac{M+1}{M-1} = 1,$$

and hence,

$$C = \frac{1}{2}\log\left(1 + \frac{P}{3}\right).$$

When P is significantly large, we have $P/3 \gg 1$ and thus

$$C \approx \frac{1}{2}\log\frac{P}{3} = \frac{1}{2}\log P - \frac{1}{2}\log 3.$$

Therefore, for this amount of transmit power, we could say

$$\boxed{C \approx \frac{1}{2}\log P - 0.8}$$

(e) Determine the error probability $P_{\mathrm{E}}(x_1)$, when $X = x_1$ is transmitted. Show $P_{\mathrm{E}}(x_1)$ in terms of Q-function.

$$\mathrm{Q}\left(\frac{t}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_t^\infty \exp\left\{-\frac{u^2}{2\sigma^2}\right\} \mathrm{d}u$$

♠ **Solution:** When we transmit $X = x_1$, the received symbol is $Y = x_1 + Z$. The receiver then estimates the transmitted symbol correctly, if the closest symbol to Y is x_1. This means that the detected symbol is correct, if Y lies in the red interval in the following figure, i.e. $\ell_1 \leq Y \leq u_1$ where

$$u_1 = -\sqrt{P} + \frac{3\sqrt{P}}{M-1}, \qquad \text{and} \qquad \ell_1 = -\sqrt{P} + \frac{\sqrt{P}}{M-1}$$

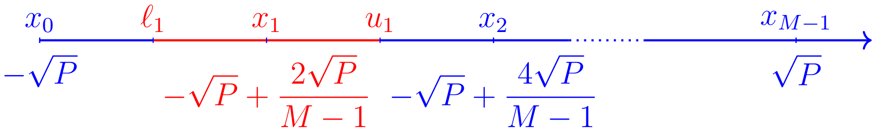

x_0
ℓ_1
x_1
u_1
x_2
x_{M-1}
$-\sqrt{P}$
$-\sqrt{P} + \frac{2\sqrt{P}}{M-1}$
$-\sqrt{P} + \frac{4\sqrt{P}}{M-1}$
$\sqrt{P}$

Figure 7.5.1: Diagram of the transmit constellation.

This means that we have error in detection, if

$$Y = x_1 + Z \geq u_1 \Rightarrow Z \geq \frac{\sqrt{P}}{M-1}, \qquad \text{or} \qquad Y = x_1 + Z \leq \ell_1 \Rightarrow Z \leq -\frac{\sqrt{P}}{M-1}.$$

As the result, the error probability reads

$$\begin{aligned} P_{\mathrm{E}}(x_1) &= \Pr\left\{Z \geq \frac{\sqrt{P}}{M-1}\right\} + \Pr\left\{Z \leq -\frac{\sqrt{P}}{M-1}\right\} \\ &= 2\Pr\left\{Z \geq \frac{\sqrt{P}}{M-1}\right\} = \frac{2}{\sqrt{2\pi\sigma_Z^2}} \int_{\frac{\sqrt{P}}{M-1}}^{\infty} \exp\left\{-\frac{z^2}{2\sigma_Z^2}\right\} \mathrm{d}z \\ &= 2\mathrm{Q}\left(\frac{\sqrt{P}}{\sigma_Z (M-1)}\right). \end{aligned}$$

Hence, we have

$$\boxed{P_{\mathrm{E}}(x_1) = 2\mathrm{Q}\left(\frac{\sqrt{P}}{M-1}\right)}$$

(f) For a given small ϵ we wish to have

$$P_{\mathrm{E}}(x_1) \leq \epsilon.$$

Determine the maximum transmission rate, achieved by our simple method, for which this condition is fulfilled. Approximate it for large values of P and M and compare it to the result of Part (d).

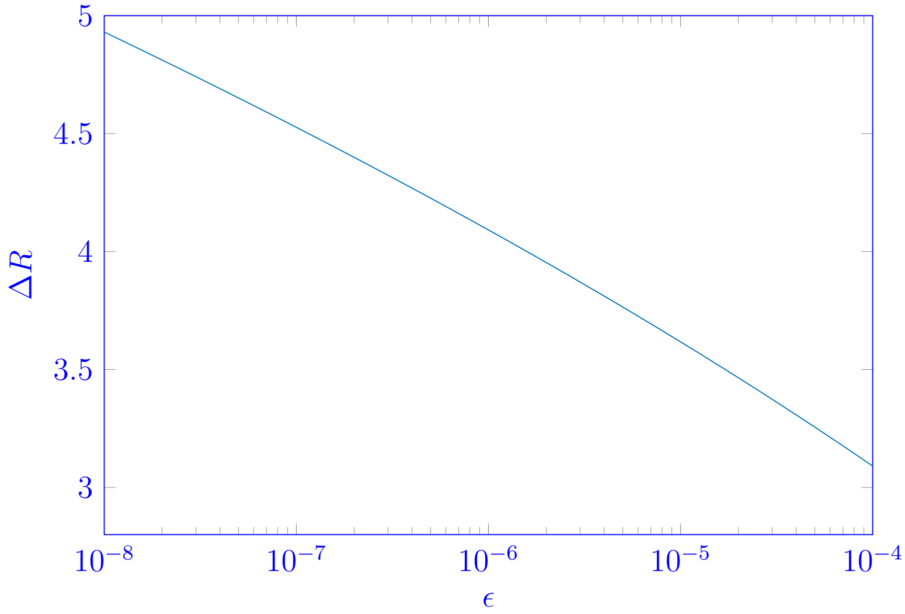


Figure 7.5.2: Rate difference versus ϵ.

♠ **Solution:** To have $P_{\mathrm{E}}(x_1) \leq \epsilon$, we need to

$$\mathrm{Q}\left(\frac{\sqrt{P}}{M-1}\right) \leq \frac{\epsilon}{2}.$$

Noting that $f(x) = \mathrm{Q}(x)$ is a decreasing function, we can write that

$$\frac{\sqrt{P}}{M-1} \geq \mathrm{Q}^{-1}\left(\frac{\epsilon}{2}\right),$$

or equivalently, we need to have

$$M \leq 1 + \frac{\sqrt{P}}{\mathrm{Q}^{-1}\left(\frac{\epsilon}{2}\right)}.$$

This means that the maximum possible value for M in this case is

$$M_{\max} = 1 + \frac{\sqrt{P}}{\mathrm{Q}^{-1}\left(\frac{\epsilon}{2}\right)},$$

and thus, a rate which achieves such an error rate satisfies

$$R \leq R_{\max} = \log M_{\max} = \log\left(1 + \frac{\sqrt{P}}{\mathrm{Q}^{-1}\left(\frac{\epsilon}{2}\right)}\right).$$

When P is large enough, we can write

$$R_{\max} \approx \log \frac{\sqrt{P}}{\mathrm{Q}^{-1}\left(\frac{\epsilon}{2}\right)} = \frac{1}{2}\log P - \log \mathrm{Q}^{-1}\left(\frac{\epsilon}{2}\right).$$

Comparing the results of Prats (d) and (f), we can can write

$$\Delta R = C - R_{\max} \approx \log \mathrm{Q}^{-1}\left(\frac{\epsilon}{2}\right) - 0.8\,\frac{\text{bits}}{\text{transmissions}}.$$

This parameter define the difference between the rate we can achieve for error probability of ϵ and the channel capacity derived by Shannon. This rate difference has been plotted in Figure 7.5.2.

To have a reference point, note that a very huge value for P is to have $P = 10000\sigma_Z^2$ which is not often possible in practice. In this case,

$$C \approx \frac{1}{2}\log 10000 = 6.64\,\frac{\text{bits}}{\text{transmissions}}.$$

This indicates that we transmit almost nothing, when we use the proposed naive approach, if we want to have small error probability. Shannon said that using a good code, we can reduce this difference close to zero. This is why we learn Shannon's theorem. We want to know how effective a proposed approach for transmission is, and how can we improve it?

Chapter 8

Linear Channel Codes

Up to this point, we have finished with the fundamental results in information theory, i.e., the source coding theorem and the channel coding theorem. We now start with some special topics in information theory. The first topic to discuss is the construction of practical channel code. In this chapter, we go through the basics of channel codes by focusing on the specific class of linear parity check codes. We then go through an specific form of parity check codes known as low density parity check (LDPC) codes. The contents of this chapter are consistent with Chapters 13 and 47 of the textbook.

8.1 Brief Review of Main Concepts

In channel coding, we map K information bits into a vector of $N > K$ bits. This longer bit contains some redundant bits which help us combat the channel noise and distortion.

To understand the concept of channel coding, consider the simple example of repetition code with three repetitions $\mathbb{R}_3$. In this code an information vector of length $K = 1$ is mapped into a vector of $N = 3$ bits by repeating the information bit three times. This means that

- When we want to send $b = 0$, we transmit over the channel

$$\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

- When we want to send $b = 1$, we transmit over the channel

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Depending on the value of b, we transmit one of $\mathbf{x}_0$ and $\mathbf{x}_1$. The receiver however observes the vector $\mathbf{y}$ which is a distorted version of the transmitted vector. For example, in $\mathbb{R}_3$, if $b = 0$ is to be sent, we transmit $\mathbf{x}_0$. This vector contains of $N = 3$ bits, hence we should transmit by three distinct transmissions. After transmitting all three bits in $\mathbf{x}_0$, we receive $\mathbf{y}$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}.$$

Here, y_1. y_2 and y_3 denote the received symbols in the first, second and the third transmission, respectively. These received symbols are not in general binary. For example, if the channel is an AWGN channel, then $y_n = \mathrm{x}_{0n} + z_n$ where z_n is a Gaussian random variable. Hence, in this case, y_n is a real number.

In channel coding, the following definitions are considered:

1. The *code rate* is defined as

$$R_{\mathrm{C}} = \frac{K}{N}$$

and describes the number of information bits per each transmitted symbol.

2. A given realization of the vector of N information bits is called an *information word*. For example, in $\mathbb{R}_3$, we have two information words, one is $b = 0$ and the other is $b = 1$.

3. The vector of N encoded bits corresponding to an information word is called a *codeword*. For example, in $\mathbb{R}_3$, we have two codewords, one is $\mathbf{x}_0$ and the other is $\mathbf{x}_1$.

In general, for a code whose information words comprise K bits, we can say

# of information words = # of code words = 2^K

4. The table which contains all the information words and their corresponding codewords is called the *codebook*. For example, for $\mathbb{R}_3$, the codebook is

information word	codeword
0	$\mathbf{x}_0$
1	$\mathbf{x}_1$

5. The function which generates a codeword from its information word is called the *encoder*. For example, in $\mathbb{R}_3$, the encoder is function $f_{\mathrm{ENC}}(b)$ which is

$$f_{\mathrm{ENC}}(b) = b \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

6. To recover the sent message from the received noisy signal, we further need to map the receive vector $\mathbf{y}$ back to an information word. This procedure consists of two phases

   (a) **Detection:** As indicated before received symbols are in general real numbers. We hence map these real numbers to binary symbols. This procedure is called *detection* in which we *detect* which binary symbol has been transmitted.
   There are several algorithms for detection. The well-known example of these algorithms is *maximum likelihood (ML) detection*.

(b) **Decoding:** The detected symbol might be a binary sequence which is not in the codebook. We hence should design a function which maps each detected binary sequence into either an information symbol or an error. This function is called *decoder*.

As an example assume that we have sent $b = 0$ over an AWGN channel via $\mathbb{R}_3$. In this case, we transmit $\mathbf{x}_0$ over the channel. Thus, the received vector is

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} .$$

where z_1, z_2 and z_3 are Gaussian noise. Assume that for the given realizations of the channel these noise terms are $z_1 = -0.01$, $z_2 = -0.51$ and $z_3 = 0.33$. Thus, we receive

$$\mathbf{y} = \begin{bmatrix} -0.01 \\ -0.51 \\ 0.33 \end{bmatrix} .$$

We use a *hard thresholding* algorithm for detection, such that we detect the transmitted symbol as 0, if the received symbol is negative. The transmitted symbol is detected as 1 otherwise. Therefore, based on the received vector $\mathbf{y}$, we detect

$$\hat{\mathbf{x}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} .$$

As it is seen $\hat{\mathbf{x}}$ is not in the codebook. Hence, we need to have a function which decode each binary sequence of length $N = 3$ to 0 or 1. In this example, we use a *majority vote* decoder, which decodes $\hat{\mathbf{x}}$ into the information symbol 0, if the number of zeros in $\hat{\mathbf{x}}$ is more than ones, and otherwise $\hat{\mathbf{x}}$ is decoded as 1. In this case $\hat{\mathbf{x}}$ is decoded as $\hat{b} = 0$.

The term *code* often refers to the set of these whole objects.

The *code* is the set of *codebook*, *encoder* and *decoder*.

8.1.1 Parity Check Codes

Parity check codes are special form of channel codes in which parity check equations are used for encoding and decoding. As it gets clear through the exercises, in parity check codes, one can always find a *parity check matrix* $\mathbf{H}$ for which we have

$$\mathbf{H}\mathbf{x} = 0 \qquad \text{Mod 2.}$$

The codeword further can be written in terms of the information word $\mathbf{i}$ as

$$\mathbf{x} = \mathbf{G}\mathbf{i} \text{Mod } 2.$$

where $\mathbf{G} \in \{0,1\}^{N \times K}$ is the generator matrix constructed from $\mathbf{H}$. Since the information word and the codeword are linearly related, these codes are *linear*.

LDPC codes are special form of parity check codes in which the parity check matrix is *sparse*, i.e., it has significantly less 1s than 0s. *Regular LDPC* codes are special forms of LDPC codes in which

- the numbers of ones in all rows of the parity check matrix $\mathbf{H}$ are the same, and
- the numbers of ones in all columns of the parity check matrix $\mathbf{H}$ are the same.

8.2 Exercises

8.2.1 Parity check codes

1. Let $\mathbf{x}_{7\times 1}$ be a codeword of a parity check code with the following parity check equations:

$$\begin{aligned} \mathrm{x}_1 \oplus \mathrm{x}_2 \oplus \mathrm{x}_3 \oplus \mathrm{x}_5 &= 0 \\ \mathrm{x}_1 \oplus \mathrm{x}_2 \oplus \mathrm{x}_4 \oplus \mathrm{x}_6 &= 0 \\ \mathrm{x}_1 \oplus \mathrm{x}_3 \oplus \mathrm{x}_4 \oplus \mathrm{x}_7 &= 0. \end{aligned}$$

(a) What is the code rate?

(b) Write the parity check matrix of this code.

(c) Plot the factor graph corresponding to this code.

(d) Is this parity check code a regular LDPC code? Give a reason.

(e) Could the vector

$$\mathbf{x}_o = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

be a codeword of this parity check code?

(f) A closest neighbor of $\mathbf{x}_o$ is

$$\mathbf{x}_c = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Determine the *minimum* distance of this code.

(g) How many bits can be corrected by this code?

8.2.2 LDPC codes

1. Let $\mathbf{x}_{12\times 1}$ be a codeword of a regular LDPC code with the following parity check equations:

$$\begin{aligned}
\mathrm{x}_3 \oplus \mathrm{x}_6 \oplus \mathrm{x}_7 \oplus \mathrm{x}_8 &= 0\\
\mathrm{x}_1 \oplus \mathrm{x}_2 \oplus \mathrm{x}_5 \oplus \mathrm{x}_{12} &= 0\\
\mathrm{x}_4 \oplus \mathrm{x}_9 \oplus \mathrm{x}_{10} \oplus \mathrm{x}_{11} &= 0\\
\mathrm{x}_2 \oplus \mathrm{x}_6 \oplus \mathrm{x}_7 \oplus \mathrm{x}_{10} &= 0\\
\mathrm{x}_1 \oplus \mathrm{x}_3 \oplus \mathrm{x}_8 \oplus \mathrm{x}_{11} &= 0\\
\mathrm{x}_4 \oplus \mathrm{x}_5 \oplus \mathrm{x}_9 \oplus \mathrm{x}_{12} &= 0\\
\mathrm{x}_1 \oplus \mathrm{x}_4 \oplus \mathrm{x}_5 \oplus \mathrm{x}_7 &= 0\\
\mathrm{x}_6 \oplus \mathrm{x}_8 \oplus \mathrm{x}_{11} \oplus \mathrm{x}_{12} &= 0\\
\mathrm{x}_2 \oplus \mathrm{x}_3 \oplus \mathrm{x}_9 \oplus \mathrm{x}_{10} &= 0.
\end{aligned}$$

   (a) Write the parity check matrix of this code. Explain why it is a *regular LDPC code*.

   (b) Plot the factor graph corresponding to this code.

   (c) Write the *maximum likelihood* decoder for this code.

2. In WiFi standard IEEE 802.11n, LDPC code is utilized. In this system, the codeword length is $N = 648$ and the code rate is $R_\mathrm{C} = 1/2$.

   (a) What is the computational complexity of *maximum likelihood* decoder for this system?

   (b) Assume that you have a computer which does a *macro-operation* within $T = 1$ psec $= 10^{-12}$ sec. Determine approximately the time required to detect a received signal with this computer.

8.3 Solutions to Exercises

8.3.1 Parity check codes

1. Let $\mathbf{x}_{7\times 1}$ be a codeword of a parity check code with the following parity check equations:

$$\begin{aligned}
\mathrm{x}_1 \oplus \mathrm{x}_2 \oplus \mathrm{x}_3 \oplus \mathrm{x}_5 &= 0\\
\mathrm{x}_1 \oplus \mathrm{x}_2 \oplus \mathrm{x}_4 \oplus \mathrm{x}_6 &= 0\\
\mathrm{x}_1 \oplus \mathrm{x}_3 \oplus \mathrm{x}_4 \oplus \mathrm{x}_7 &= 0.
\end{aligned}$$

   (a) What is the code rate?

(b) Write the parity check matrix of this code.

(c) Plot the factor graph corresponding to this code.

(d) Is this parity check code a regular LDPC code? Give a reason.

(e) Could the vector

$$\mathbf{x}_o = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

be a codeword of this parity check code?

(f) A closest neighbor of $\mathbf{x}_o$ is

$$\mathbf{x}_c = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Determine the *minimum* distance of this code.

(g) How many bits can be corrected by this code?

♠ **Solution:**

(a) This is a parity check code. In parity check codes, we encode an information sequence of K bits into a sequence of N bits. These N bits satisfy M parity check equations. The relation between K, N and M is

$$\boxed{N = K + M}$$

As we have codewords of length $N = 7$, and $M = 3$ parity check equations, we can write

$$K = N - M = 7 - 3 = 4.$$

Thus, the code rate reads

$$R_{\mathrm{C}} = \frac{K}{N} = \frac{4}{7}.$$

(b) The parity check equations are a set of linear equations which can be represented by

$$\mathbf{Hx} = 0 \qquad \text{Mod 2.}$$

Here, "Mod 2" means that all the summations are operated in the binary field. In other words, all sums are equivalent to an XOR operator: $1 \oplus 0 = 1$ and $0 \oplus 0 = 1 \oplus 1 = 0$.

Matrix $\mathbf{H}$ is called the *parity check matrix*. Considering the given parity check equations, we can write

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathrm{x}_1 \\ \vdots \\ \mathrm{x}_7 \end{bmatrix} = 0 \qquad \text{Mod 2.}$$

Thus, the parity check matrix is

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

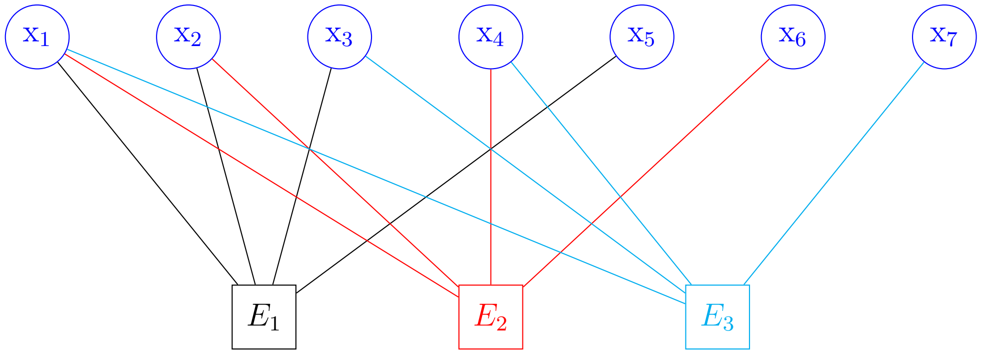
(c) The parity check equations can be also represented as follows:

$$E_m\left(\mathrm{x}_1, \ldots, \mathrm{x}_7\right) = 0 \qquad \text{for } m = 1, \ldots, M.$$

In our example, $E_m\left(\mathrm{x}_1, \ldots, \mathrm{x}_7\right)$ are

$$\begin{aligned} E_1\left(\mathrm{x}_1, \ldots, \mathrm{x}_7\right) &= \mathrm{x}_1 \oplus \mathrm{x}_2 \oplus \mathrm{x}_3 \oplus \mathrm{x}_5 \\ E_2\left(\mathrm{x}_1, \ldots, \mathrm{x}_7\right) &= \mathrm{x}_1 \oplus \mathrm{x}_2 \oplus \mathrm{x}_4 \oplus \mathrm{x}_6 \\ E_3\left(\mathrm{x}_1, \ldots, \mathrm{x}_7\right) &= \mathrm{x}_1 \oplus \mathrm{x}_3 \oplus \mathrm{x}_4 \oplus \mathrm{x}_7. \end{aligned}$$

These functions can be represented by a factor graph with $M = 3$ factor nodes and $N = 7$ variable nodes as shown below:



(d) As it is observed, this is not a regular LDPC code. For example, the first column of $\mathbf{H}$ has three ones whereas the second column has two ones. Hence, it is **not** a regular LDPC code.

(e) Any binary vector that satisfies the parity check equations could be a codeword. Since

$$\mathbf{H}\mathbf{x}_o = 0,$$

we can conclude that the parity check equations are all satisfied, and hence x_o could be a codeword.

(f) The minimum distance of a code is defined as the minimum possible Hamming distance between two distinct codewords of the code. In our example, we have $K = 4$. This means that we have in total $2^4 = 16$ codewords. If we denote these codewords with $\mathbf{x}_1, \ldots, \mathbf{x}_{16}$, then

$$d_{\text{min}} = \min_{\substack{k=1,\ldots,16 \\ j=1,\ldots,16 \\ k\neq j}} d_{\text{H}}\left(\mathbf{x}_k, \mathbf{x}_j\right)$$

where $d_{\text{H}}\left(\mathbf{x}_k, \mathbf{x}_j\right)$ is the Hamming distance between $\mathbf{x}_k$ and $\mathbf{x}_j$ which counts the number of bits these two codewords differ in. The derivation of this value requires to consider all possible distinct pairs of codewords, calculate their Hamming distance, and then find the minimum of them. This is task is computationally complex, and for very large values of K is intractable. Parity check codes however have a nice property[1] which says that the distance property of different codewords are the same. This means that if you concentrate in one codeword, and find the neighboring codeword which has the minimum distance, this value would be the same for the other codewords. This means that by fixing k, we can say

$$d_{\text{min}} = \min_{\substack{j=1,\ldots,16 \\ k\neq j}} d_{\text{H}}\left(\mathbf{x}_k, \mathbf{x}_j\right).$$

In our example, we focus on $\mathbf{x}_o$. As the closest neighboring codeword to $\mathbf{x}_o$ is $\mathbf{x}_c$, we could conclude that

$$d_{\text{min}} = \min_{\substack{\mathbf{x}_1,\ldots,\mathbf{x}_{16} \\ \mathbf{x}_j\neq \mathbf{x}_o}} d_{\text{H}}\left(\mathbf{x}_o, \mathbf{x}_j\right) = d_{\text{H}}\left(\mathbf{x}_o, \mathbf{x}_c\right) = 3.$$

Thus, the minimum distance of this code is $d_{\text{min}} = 3$.

(g) As given in Chapter 13 page 206 of the textbook, the numner of bits being corrected with a code with minimum distance d_{min} is

$$\boxed{t = \left\lfloor \frac{d_{\text{min}} - 1}{2} \right\rfloor}$$

Hence, for this code, we have

$$t = \left\lfloor \frac{3 - 1}{2} \right\rfloor = 1.$$

[1] Actually, this is a generic property of *linear* codes.

8.3.2 LDPC codes

1. Let $\mathrm{x}_{12\times 1}$ be a codeword of a regular LDPC code with the following parity check equations:

$$\begin{aligned}
\mathrm{x}_3 \oplus \mathrm{x}_6 \oplus \mathrm{x}_7 \oplus \mathrm{x}_8 &= 0\\
\mathrm{x}_1 \oplus \mathrm{x}_2 \oplus \mathrm{x}_5 \oplus \mathrm{x}_{12} &= 0\\
\mathrm{x}_4 \oplus \mathrm{x}_9 \oplus \mathrm{x}_{10} \oplus \mathrm{x}_{11} &= 0\\
\mathrm{x}_2 \oplus \mathrm{x}_6 \oplus \mathrm{x}_7 \oplus \mathrm{x}_{10} &= 0\\
\mathrm{x}_1 \oplus \mathrm{x}_3 \oplus \mathrm{x}_8 \oplus \mathrm{x}_{11} &= 0\\
\mathrm{x}_4 \oplus \mathrm{x}_5 \oplus \mathrm{x}_9 \oplus \mathrm{x}_{12} &= 0\\
\mathrm{x}_1 \oplus \mathrm{x}_4 \oplus \mathrm{x}_5 \oplus \mathrm{x}_7 &= 0\\
\mathrm{x}_6 \oplus \mathrm{x}_8 \oplus \mathrm{x}_{11} \oplus \mathrm{x}_{12} &= 0\\
\mathrm{x}_2 \oplus \mathrm{x}_3 \oplus \mathrm{x}_9 \oplus \mathrm{x}_{10} &= 0.
\end{aligned}$$

   (a) Write the parity check matrix of this code. Explain why it is a *regular LDPC code*.

   (b) Plot the factor graph corresponding to this code.

   (c) Write the *maximum likelihood* decoder for this code.

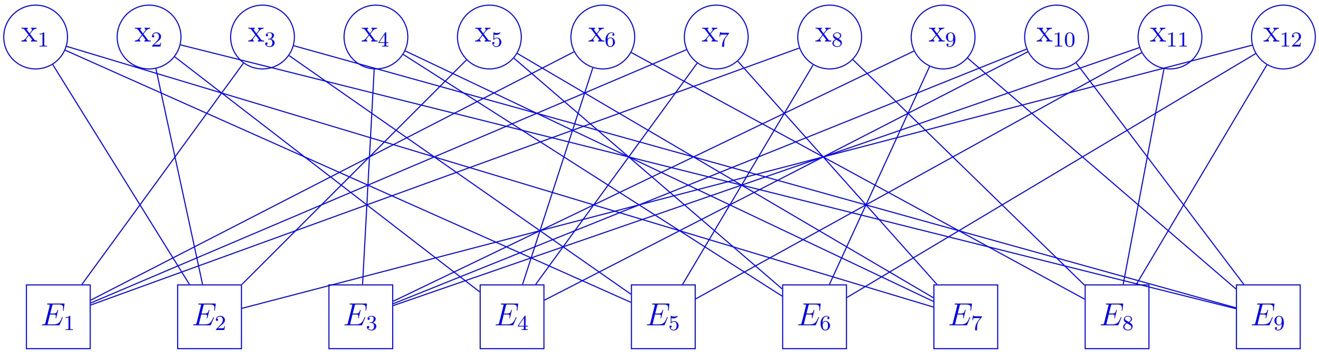
♠ **Solution:**

   (a) Considering the given parity check equations, we have

$$\mathbf{H} = \begin{bmatrix}
0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0\\
1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1\\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0\\
0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0\\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0\\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1\\
1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1\\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0
\end{bmatrix}.$$

   As you can see, in each row of $\mathbf{H}$, there are only 4 ones, and in each column, there are only 3 ones. Thus, it is a regular LDPC code.

(b) The factor graph of this code has $M = 9$ factor nodes and $N = 12$ variable nodes, and is shown below: Here, $E_1, \ldots, E_9$ correspond to the 1^{st} till 9^{th} parity check equations.



(c) In ML decoding, we do the detection and decoding task together. In this respect, let $\mathbf{y}$ denote the vector of received symbols. This vector is of length $N = 12$ and is received after transmitting $\mathrm{x}_1, \ldots, \mathrm{x}_{12}$. Noting that $M = 9$, we could conclude that $K = 3$, and hence the codebook consists of $2^3 = 8$ pairs of information word and codeword. Let us show the set of these 8 codewords with $\mathcal{C}$. This means that $\mathcal{C}$ contains 8 different binary vectors, each of size $N = 12$. The ML decoding is performed as following

- The ML decoder first detect the transmitted codeword $\mathbf{x}$ by finding the codeword which is closest to $\mathbf{y}$. This means that it detects $\hat{\mathbf{x}}$ which is

$$\hat{\mathbf{x}} = \underset{\mathbf{v} \in \mathcal{C}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{v}\|^2$$

  Noting that the ML detector minimizes the distance by searching over the codewords, we are sure that $\hat{\mathbf{x}}$ is in the codebook.

- In the next step, the information word $\mathbf{i}$ is decoded from the codebook, such that $\hat{\mathbf{x}}$ be the codeword of $\mathbf{i}$.

2. In WiFi standard IEEE 802.11n, LDPC code is utilized. In this system, the codeword length is $N = 648$ and the code rate is $R_{\mathrm{C}} = 1/2$.

   (a) What is the computational complexity of *maximum likelihood* decoder for this system?

   (b) Assume that you have a computer which does a *macro-operation* within $T = 1$ psec $= 10^{-12}$ sec. Determine approximately the time required to detect a received signal with this computer.

♠ **Solution:**

(a) The codeword length is $N = 648$ and the code rate is half. This means that

$$K = R_{\mathrm{C}} N = 324.$$

From the formulation in the previous exercise, we know that for ML decoding we need to search over all codewords. Since we have 2^K codewords, we need to do

$$N_{\mathrm{Operation}} = 2^{324}$$

macro-operations in general.

(b) Using this computer, we need

$$T_{\mathrm{Total}} = N_{\mathrm{Operation}} T = 2^{324} \times 10^{-12} \approx 3.4 \times 10^{85} \text{ sec.}$$

Noting that each year is 3.1536×10^7 sec, we could conclude that

$$T_{\mathrm{Total}} > 10^{76} \text{ centuries.}$$

where each century is 100 years! Although in practice this time can be reduced by some tricks, the resulting time is still impossible to handle. It is often said that the ML decoding is *computationally intractable*. This is why we use algorithms such as message passing in practice.

8.4 Homework

The following exercises from the textbook are suggested for further practice.

- Chapter 13: Exercise 13.7., Exercise 13.11., Exercise 13.12., Exercise 13.15., Exercise 13.22.
- Chapter 47: Exercise 47.2.,

8.5 Fun Facts

LDPC codes were initially invented by *Robert Gray Gallager* in his PhD dissertation in 1963. Back to then, he couldn't provide that much of simulation results, since it was computationally very complex to simulate these codes with the available computers. It hence believed to be *impractical* and only a nice toy example. David JC. MacKay, who authored your textbook, was the one who rediscovered these codes. He showed that these codes can be also *practically efficient*, if one uses the *message passing algorithm* for implementation. LDPC codes are hence known as codes which were *dormant* of 35 years! Or as best said

A bit of 21st-century coding that happened to fall in the 20th century

Chapter 9

Message Passing and Sum-Product

In this chapter, we go through the concept of message passing and the sum-product algorithm. These are two very recent topics in information theory which connect information theory to many topics in inference and machine learning. The contents of this chapter are consistent with Chapters 16 and 26 of the textbook.

9.1 Brief Review of Main Concepts

The concept of message passing follows a simple idea: when you want to cont all possible cases in a problem with multiple variables, you don't need to shoot in dark and consider all possible combinations of the variables. In many problems, you can start with one variable and the gradually omit a lot of impossible cases. Doing so, you could save a lot in terms of computational complexity. In such problems, many calculations can be effectively done by the so-called *sum-product* algorithm. We briefly go through the sum-product algorithm in the sequel.

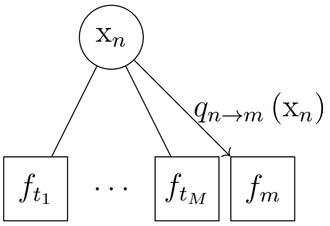
9.1.1 Sum-Product Algorithm

Following the discussions in Chapter 26 of the textbook, there are two types of messages being sent in a factor graph by the sum-product algorithm. The explicit expression for these messages are given in page 336 in a box at the middle of the page. To understand these formulas clearly, consider a factor graph which represents the function $F(\mathrm{x}_1, \dots, \mathrm{x}_V)$ with V variables. Assuming that this function could be written as

$$F(\mathrm{x}_1, \dots, \mathrm{x}_V) = \prod_{u=1}^{U} f_u\left(\mathrm{x}_{v_1}, \dots, \mathrm{x}_{v_{N_u}}\right),$$

the factor graph has V variable nodes and U factor nodes.

Let us first focus on a variable node x_n which is connected to some factor nodes. We show these nodes with $f_{t_1}, \dots, f_{t_M}, f_m$. This is shown in the following figure.



The variable node x_n sends the message $q_{n\to m}(\mathrm{x}_n)$ to the factor node f_m. This message is constructed as

$$q_{n\to m}(\mathrm{x}_n) = \prod_{i=1}^{M} r_{t_i\to n}(\mathrm{x}_n)$$

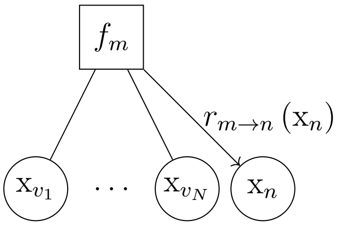
where $r_{t_i\to n}(\mathrm{x}_n)$ is the message from factor node f_{t_i} to the variable node x_n. This update rule can be simply written as

$q_{n\to m}(\mathrm{x}_n)$ = Product of all incoming messages except the message from the destination

Now, we focus on a factor node. Let the factor node f_m represent the function $f_m(\cdot)$. Assume that $f_m(\cdot)$ is a function of variables $\mathrm{x}_{v_1}, \ldots, \mathrm{x}_{v_N}$ and x_n, i.e.

$$f_m(\mathrm{x}_{v_1}, \ldots, \mathrm{x}_{v_N}, \mathrm{x}_n)$$

A schematic view of this node is given below.



The factor node f_m sends the message $r_{m\to n}(\mathrm{x}_n)$ to the variable node x_n. This message reads

$$r_{m\to n}(\mathrm{x}_n) = \sum_{\mathrm{x}_{v_1},\ldots,\mathrm{x}_{v_N}} f_m(\mathrm{x}_{v_1}, \ldots, \mathrm{x}_{v_N}, \mathrm{x}_n) \prod_{i=1}^{N} q_{v_i\to m}(\mathrm{x}_{v_i})$$

where $q_{v_i\to m}(\mathrm{x}_n)$ is the message from variable node x_{v_i} to the factor node f_m. This update rule can be simply written as

$r_{m\to n}(\mathrm{x}_n)$ = $\sum_{\text{all variables except } \mathrm{x}_n}$ { Function f_m × Product of all incoming messages except message from destination}

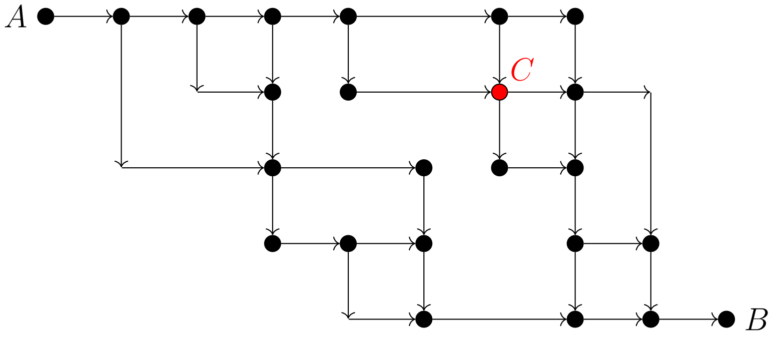
The sum-product algorithm starts from the leaves of the graph. After running all the update rule, the algorithm stops. At this point, we have several incoming messages at each variable node. In this case, we can write

$Z_{\mathrm{x}_n}(\mathrm{x}_n) = \sum_{\mathrm{x}_v, v\neq n} F(\mathrm{x}_1, \ldots, \mathrm{x}_V) =$ Product of all incoming messages to node x_n

9.2 Exercises

9.2.1 Message passing algorithm

1. Consider the following graph. In this graph, we are allowed to move either to the right or down.



   (a) Count the number of paths from node A to node B.

   (b) You move from node A to node B randomly as follows:

   > At each node, if you face two ways, a uniform coin is tossed. You move right, if it is tail; you go down, if it is head.

   Find the probability of passing through node C in your way.

   (c) You again move from node A to node B randomly; however, this time you do the following:

   > You consider all possible paths and choose one of them at random.

   What is the probability of passing through node C, in this case.

9.2.2 Sum-product algorithm

1. The functions $h(x,y)$ and $g(x)$ are defined as

$$h(x,y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases} \quad \text{and} \quad g(x) = \begin{cases} 0.25 & x = 1 \\ 0.75 & x = 0 \end{cases}$$

where $x, y \in \{0, 1\}$.

Consider the binary variables x, y, z and let the function $f(x,y,z)$ be

$$f(x,y,z) = g(x)\, h(x,y)\, h(y,z)\,.$$

   (a) Draw the factor graph corresponding to $f(x,y,z)$ and specify the factor and variable nodes.

(b) Determine the marginal function

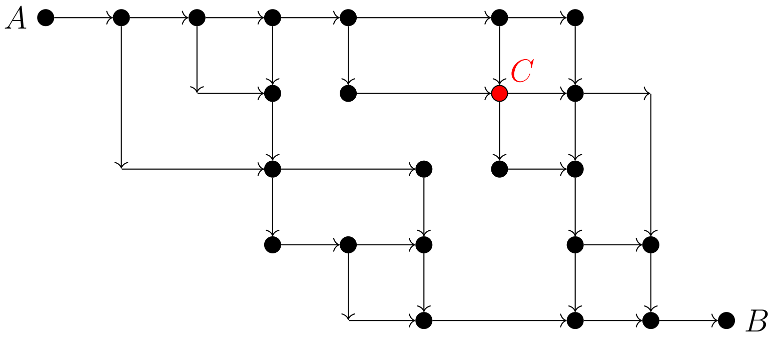
$$Z_X(x) = \sum_{y,z \in \{0,1\}} f(x,y,z)$$

via the *sum-product* algorithm.

9.3 Solutions to Exercises

9.3.1 Message passing algorithm

1. Consider the following graph. In this graph, we are allowed to move either to the right or down.



(a) Count the number of paths from node A to node B.

(b) You move from node A to node B randomly as follows:

> At each node, if you face two ways, a uniform coin is tossed. You move right, if it is tail; you go down, if it is head.

Find the probability of passing through node C in your way.

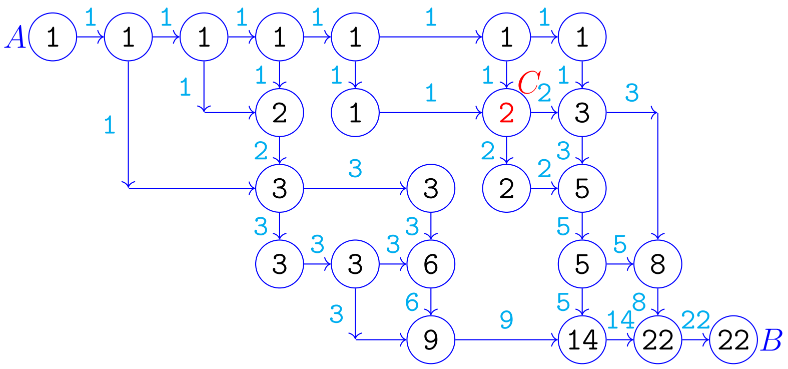
(c) You again move from node A to node B randomly; however, this time you do the following:

> You consider all possible paths and choose one of them at random.

What is the probability of passing through node C, in this case.

♠ **Solution:**

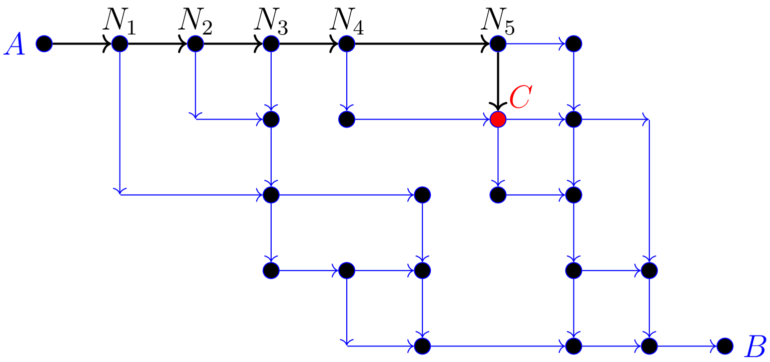
(a) To find the number of paths, we start from node A with message $M_A = 1$. At each following node, the nodes message is constructed by summing all *incoming messages*. This message is then broadcasted over all *outgoing links*. The message of node B then gives the total number of paths. The following figure shows the message passing process in this graph. The broadcasted message of each node is shown by cyan on the outgoing links. The message of each node is shown by black inside the node. As you can see, the message of each node is the sum of its incoming messages.

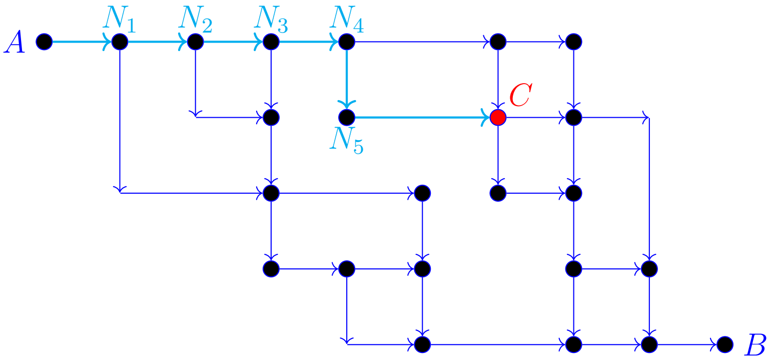


From the figure, we can conclude that

$$\boxed{\text{\# of paths } A \to B = 22}$$

(b) To pass through C by this random process, you should consider all possible events which take you from A to C. Considering Part (a), we have seen that the message of node C is $M_C = 2$. This means that there are two path which take us from A to C. There two paths are shown in the following figures with black and cyan where we have labeled the nodes in each path with N_1 till N_5.





The probability of passing through C with coin tossing is hence given by

$$\begin{aligned}\Pr\{A \to C \to B \text{ with coin tossing }\} = &\Pr\{A \to N_1 \to N_2 \to N_3 \to N_4 \to N_5 \to C \text{ with coin tossing }\}\\ &+ \Pr\{A \to N_1 \to N_2 \to N_3 \to N_4 \to N_5 \to C \text{ with coin tossing }\}.\end{aligned}$$

Noting that the tosses at nodes are independently done, we can write for the black path

$$\begin{aligned}
&\Pr\{A \to N_1 \to N_2 \to N_3 \to N_4 \to N_5 \to C \text{ with coin tossing }\}\\
&= \Pr\{A \to N_1 \text{ with coin tossing }\} \Pr\{N_1 \to N_2 \text{ with coin tossing }\}\\
&\quad \Pr\{N_2 \to N_3 \text{ with coin tossing }\} \Pr\{N_3 \to N_4 \text{ with coin tossing }\}\\
&\quad \Pr\{N_4 \to N_5 \text{ with coin tossing }\} \Pr\{N_5 \to C \text{ with coin tossing }\}\\
&=1 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{32}.
\end{aligned}$$

Similarly, for the cyan path, we have

$$\begin{aligned}
&\Pr\{A \to N_1 \to N_2 \to N_3 \to N_4 \to N_5 \to C \text{ with coin tossing }\}\\
&= \Pr\{A \to N_1 \text{ with coin tossing }\} \Pr\{N_1 \to N_2 \text{ with coin tossing }\}\\
&\quad \Pr\{N_2 \to N_3 \text{ with coin tossing }\} \Pr\{N_3 \to N_4 \text{ with coin tossing }\}\\
&\quad \Pr\{N_4 \to N_5 \text{ with coin tossing }\} \Pr\{N_5 \to C \text{ with coin tossing }\}\\
&=1 \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times 1 = \frac{1}{16}.
\end{aligned}$$

Hence, we can conclude that

$$\boxed{\Pr\{A \to C \to B \text{ with coin tossing }\} = \frac{3}{32}}$$

(c) In this second form of random walk from A to B, we choose our path from beginning. Hence, the probability of passing through C is equal to the probability of the event that the chosen paths contains node C. As a result, we can write

$$\Pr\{A \to C \to B \text{ with random path choosing }\} = \frac{\text{\# of paths } A \to C \to B}{\text{\# of paths } A \to B}$$

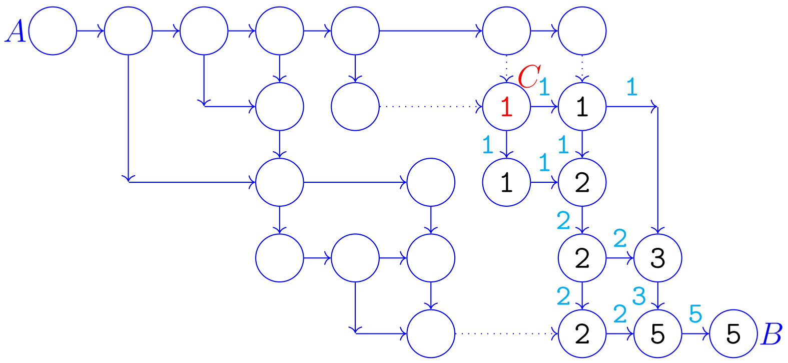
To count the number of path in $A \to C \to B$, we note that

$$\boxed{\text{\# of paths } A \to C \to B = (\text{ \# of paths } A \to C\ ) \times (\text{ \# of paths } C \to B\ )}$$

From Part (a), we know that

$$\boxed{\text{\# of paths } A \to C = 2}$$

To further count the number of paths from C to B, we use the message passing algorithm starting from node C. The details are shown in the following figure. Node C starts with message $M_C = 1$. It broadcasts this message on its *outgoing nodes*. The message at each node is calculated as the sum of *incoming messages*. The number of paths is then given by the message of node B. As it is shown, there



are 5 paths from C to B. Thus, we can conclude that

$$\text{\# of paths } A \to C \to B = 2 \times 5 = 10.$$

Consequently, we have

$$\Pr\{A \to C \to B \text{ with random path choosing }\} = \frac{10}{22}.$$

9.3.2 Sum-product algorithm

1. The functions $h(x,y)$ and $g(x)$ are defined as

$$h(x,y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases} \quad \text{and} \quad g(x) = \begin{cases} 0.25 & x = 1 \\ 0.75 & x = 0 \end{cases}$$

where $x, y \in \{0,1\}$.

Consider the binary variables x, y, z and let the function $f(x,y,z)$ be

$$f(x,y,z) = g(x)\, h(x,y)\, h(y,z).$$

(a) Draw the factor graph corresponding to $f(x,y,z)$ and specify the factor and variable nodes.

(b) Determine the marginal function

$$Z_X(x) = \sum_{y,z \in \{0,1\}} f(x,y,z)$$

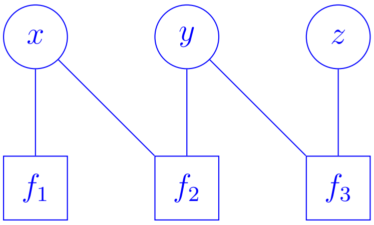
via the *sum-product* algorithm.

♠ **Solution:**

(a) Noting that $f(x, y, z)$ has there factor functions, the factor graph has three variable and three factor nodes. Defining the factors as

$$\begin{aligned} f_1(x) &= g(x) \\ f_2(x) &= h(x, y) \\ f_3(x) &= h(y, z), \end{aligned}$$

the factor graph is given by



(b) Using the sum-product algorithm, we note that the variable node x has only two neighbors f_1 and f_2. Thus, the marginal $Z_X(x)$ is given by

$$Z_X(x) = r_{1\to x}(x)\, r_{2\to x}(2)$$

where $r_{m\to x}(x)$ is the message from factor node f_m to the variable node x. To find $r_{1\to x}(x)$ and $r_{2\to x}(2)$, we write the update rules step by step:

i. Node f_1 has only one neighbor. This means that there is only one incoming message from x which is the destination. Hence, $r_{1\to x}(x)$ reads

$$r_{1\to x}(x) = \sum_{y,z} f_1(x) \times 1 = f_1(x) = g(x)$$

ii. Node f_2 has two neighbors, namely x and y. Thus, there are two incoming messages: one from x , and one from y. Since x is the destination, we should ignore it in the update rule. As a result, $r_{1\to x}(x)$ reads

$$r_{2\to x}(x) = \sum_{y,z} f_2(x, y) \times q_{y\to 2}(y) = \sum_{y} h(x, y)\, q_{y\to 2}(y).$$

For further calculations, we need to determine $q_{y\to 2}(y)$.

iii. Node y has two neighbors, namely f_2 and f_3. Thus, there are two incoming messages: one from f_2 , and one from f_3. Since f_2 is the destination, we should ignore it in the update rule. Hence,

$$q_{y\to 2}(y) = r_{3\to y}(y).$$

We now need to determine $r_{3\to y}(y)$.

iv. Node f_3 has two neighbors, namely y and z. Thus, there are two incoming messages: one from y , and one from z. Since y is the destination, we should ignore it in the update rule. Hence,

$$r_{3\to y}(y) = \sum_{y,z} f_3(x,z) \times q_{z\to 3}(z) = \sum_{z} h(y,z)\, q_{z\to 3}(z) \,.$$

We now need to determine $q_{z\to 3}(z)$.

v. Node z has only one neighbor f_3 which is the destination and should be ignored in the update rule. Hence,

$$q_{z\to 3}(z) = 1.$$

Now, we can successively replace the messages in each step as follows:

i. As $q_{z\to 3}(z) = 1$,

$$\begin{aligned} r_{3\to y}(y) &= \sum_{z} h(y,z) = h(y,0) + h(y,1) \\ &= \begin{cases} 1 & y=0 \\ 0 & y\neq 0 \end{cases} + \begin{cases} 1 & y=1 \\ 0 & y\neq 1 \end{cases} = \begin{cases} 1 & y=0 \\ 0 & y=1 \end{cases} + \begin{cases} 0 & y=0 \\ 1 & y=1 \end{cases} = 1 \end{aligned}$$

ii. As $r_{3\to y}(y) = 1$, we have

$$q_{y\to 2}(y) = r_{3\to y}(y) = 1$$

iii. As $q_{y\to 2}(y) = 1$, we have

$$\begin{aligned} r_{2\to x}(x) &= \sum_{y} h(x,y) = h(x,0) + h(x,1) \\ &= \begin{cases} 1 & x=0 \\ 0 & x\neq 0 \end{cases} + \begin{cases} 1 & x=1 \\ 0 & x\neq 1 \end{cases} = \begin{cases} 1 & x=0 \\ 0 & x=1 \end{cases} + \begin{cases} 0 & x=0 \\ 1 & x=1 \end{cases} = 1 \end{aligned}$$

Thus, we have

$$\boxed{r_{2\to x}(x) = 1}$$

which concludes that

$$Z_X(x) = g(x) \,.$$

⇝ REMARK:

Always start with what you are asked for and write it in terms of the messages in the graph. Then, start to calculate each message step by step. By doing so, you avoid calculating messages which are not required. For instance, in this example we did not calculate $r_{3\to z}(z)$, $q_{y\to 3}(y)$, $r_{2\to y}(y)$, and $q_{x\to 2}(x)$ as they are not required for determination of $Z_X(x)$.

9.4 Homework

The following exercises from the textbook are suggested for further practice.

- Chapter 16: Exercise 16.1., Exercise 16.2., Exercise 16.4.
- Chapter 26: Exercise 26.6., Exercise 26.7.

Chapter 10

Sample Exams With Solutions

In this chapter, some sample final exams of the Information Theory and Coding course in Friedrich-Alexander University of Erlangen are given. The solutions to the exams are further provided to help you preparing for the final exam.

10.1 Winter Semester 2017-2018 Exam

Exam Date: February 13th, 2018
8 Questions with total of 100 Points.
Exam Duration: 90 Minutes

10.1.1 Inference (10 Points)

Your friend is pregnant. She does not know whether she is pregnant with a single baby or twins; however, you wish to guess it. The history of her family indicates that 30% of women in this family have given birth to twins and 70% of them had single babies. You are moreover provided by the following information:
Pregnant women with twins suffer from morning sickness in 80% of the cases while those with single babies experience the morning sickness in 40% of the cases.
Assume that your friend experiences morning sickness; then, what is the probability that she is pregnant with twins?

♠ **Solution:**

$$\text{Twins} = T, \qquad \text{Single Baby} = S, \qquad \text{Morning Sickness} = M$$

$$\text{Pr}\{T|M\} = \frac{\text{Pr}\{M|T\}\text{Pr}\{T\}}{\text{Pr}\{M|T\}\text{Pr}\{T\} + \text{Pr}\{M|S\}\text{Pr}\{S\}}$$

$$= \frac{0.8 \times 0.3}{0.8 \times 0.3 + 0.4 \times 0.7} = 0.462$$

		Y		
		A	B	C
X	0	1/8	1/16	1/32
	1	1/32	1/2	1/4

10.1.2 Shannon Inequalities (8 Points)

Consider three discrete random variables X, Y and Z. Assume that

$$X \rightarrow Y \rightarrow Z$$

form a Markov chain. This means that

$$P(x,y,z) = P(x)P(y|x)P(z|y)$$

Show that the following inequality holds

$$I(X;Y) \geq I(X;Y|Z)$$

Hint: You may start by expanding $I(X;Y,Z)$.

♠ **Solution:**

$$I(X;Y,Z) = I(X;Y) + \underbrace{I(X;Z|Y)}_{\substack{=0 \\ \text{Markov Property}}}$$

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$$

$$I(X;Z) \geq 0 \implies I(X;Y) \geq I(X;Y|Z)$$

10.1.3 Entropy and Mutual Information (12 Points)

Consider the discrete dependent random variables $X \in \{0,1\}$ and $Y \in \{A,B,C\}$ which are jointly distributed with the following distribution.

(a) Determine

(a-1) the joint entropy $H(X,Y)$

♠ **Solution:**

$$H(X,Y) = \sum p(x,y) \log \frac{1}{p(x,y)} = \frac{31}{16}$$

(a-2) the conditional entropy $H(X|Y)$

♠ **Solution:**

$$H(X|Y) \Longrightarrow = \begin{cases} P(Y=A) = \frac{5}{32} \\ P(Y=B) = \frac{18}{32} \\ P(Y=C) = \frac{9}{32} \end{cases} \Longrightarrow H(X|Y) = H(X,Y) - H(Y) = 0.538$$

(a-3) the mutual information $I(X;Y)$

♠ **Solution:**

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

$$H(X) = H_2(\frac{7}{32}) \qquad \Longrightarrow I(X;Y) = 0.192$$

(b) Consider the function

$$f(y) = \begin{cases} 0 & y = A \text{ or } B \\ 1 & y = C \end{cases}$$

The random variable Z is defined as

$$Z = f(Y).$$

Determine the following terms

(b-1) $H(X,Y,Z)$

♠ **Solution:**

$$X \longrightarrow Y \longrightarrow Z$$

$$H(X,Y,Z) = H(X,Y) + \underbrace{H(Z|X,Y)}_{\substack{=0 \\ \text{Markov Property}}} = H(X,Y)$$

(b-2) $I(X;Z|Y)$

♠ **Solution:**

$$I(X;Z|Y) = 0 \text{ (Markov)}$$

Hint: You may use the chain rule.

10.1.4 Source Coding (20 Points)

Let the discrete source X be distributed as

X	A	B	C	D	E	F
Pr	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$

(a) Consider the binary source $\mathcal{C}(X)$

X	A	B	C	D	E	F
$\mathcal{C}(X)$	100	101	011	10	01	00

Is $\mathcal{C}(X)$ a uniquely decodable code? Justify your answer.

♠ **Solution:** No.

$$\text{Use Kraft's Inequality} \implies \frac{1}{2^3} + \frac{1}{2^3} + +\frac{1}{2^3} + \frac{1}{2^2} + \frac{1}{2^2} + \frac{1}{2^2} = \frac{9}{8}$$

$$\frac{9}{8} > 1 \implies \text{Kraft does NOT hold} \implies \text{It could not be uniquely decodable!}$$

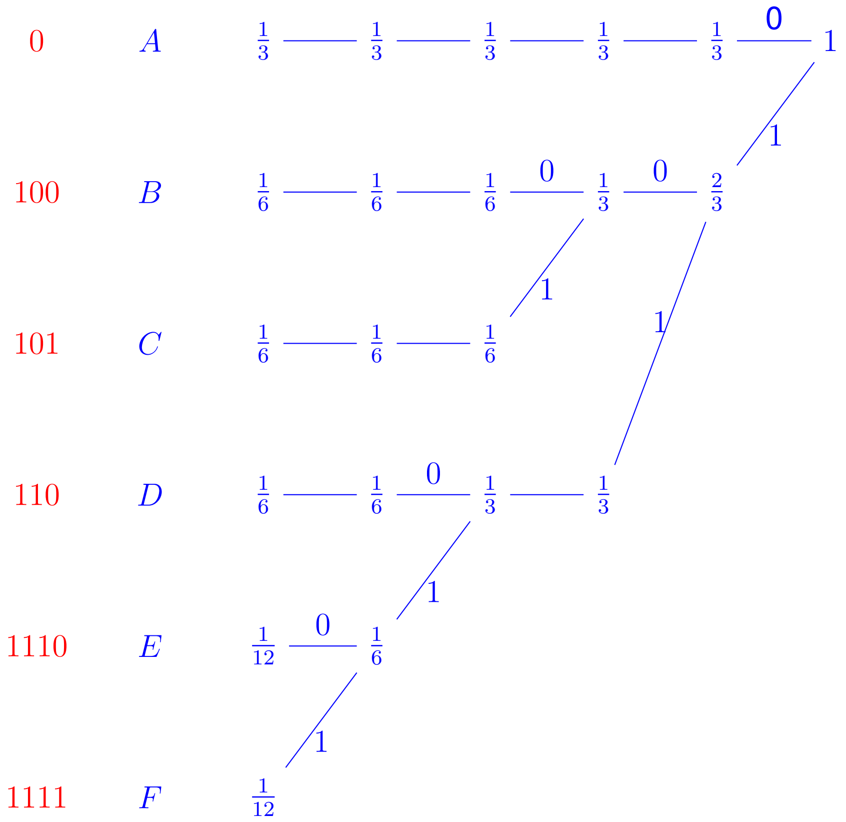
Another possible solution is to give a counter example.

$$\text{e.g} \quad AB \xrightarrow{\text{ENC}} 100, 101 \xrightarrow{\text{DEC}} AB$$

$$10, 01, 01 \xrightarrow{\text{DEC}} DEE$$

(b) Find a binary Huffman code for the source X and calculate its average length.

♠ **Solution:**



$$l(A) = 1,\ l(B) = l(C) = l(D) = 3,\ l(E) = l(F) = 4$$

$$L_{avg} = 1 \times \frac{1}{3} + \frac{1}{6} \times 3 \times 3 + \frac{1}{12} \times 4 \times 2 = \frac{2+9+4}{6} = \frac{5}{2}$$

$$L_{avg} = 2.5 \text{ bits}$$

(c) Consider the following function

$$g(x) = \begin{cases} 0 & x \in \{u_1, u_2\} \\ 1 & x \in \{u_3, u_4\} \\ 2 & x \in \{u_5, u_6\} \end{cases}$$

where $u_1, \ldots, u_6$ are distinct and chosen from $\{A, B, C, D, E, F\}$.

Find $u_1, \ldots, u_6$ such that there exist a binary code for the source $Z = g(X)$ whose expected length equals to the entropy $H(Z)$.

Hint: There is a condition on the distribution of a source under which the expected length of a binary Huffman code is equal to the entropy of the source. Remember this condition and try to find $u_1, \ldots, u_6$ such that is is fulfilled.

♠ **Solution:** To make Huffman code optimal for Z:

$$\text{Pr}(Z = z_i) = \frac{1}{2^{n_i}}, \quad \text{for some integer } n_i$$

Let:

$$u_1 = A, u_2 = B, u_3 = C, u_4 = D, u_5 = D, u_6 = F$$

$$\text{Pr}(z = 0) = \text{Pr}(X = A) + \text{Pr}(X = B) = \frac{1}{2}$$

$$\text{Pr}(z = 1) = \frac{1}{4}$$

$$\text{Pr}(z = 2) = \frac{1}{4}$$

This makes the average code length

$$L_{avg} = H(Z)$$

Other combinations which result in same probability are also true.

10.1.5 Lempel-Ziv Coding (7 Points)

A binary sequence has been encoded by the standard Lempel-Ziv algorithm, i.e., the simplest version of the algorithm with variable pointer length without any removal of codewords from the codebook nor removal of redundant data bits.
The encoded stream is

$$0110101100101$$

Find the binary sequence.

♠ **Solution:**

	0	11	010	110	0101
p	1	2	3	4	5
$\lceil \log x \rceil$	0	1	2	2	3
l	1	2	3	3	4
decode	0	0,1	0,0	00,0	01,1

The code book is as follows:

substream	index	
λ	0	0
0	1	01
01	2	10
00	3	11
000	4	100
011	5	101

Hence the decoded stream:

$$00100000011$$

10.1.6 Channel Capacity (18 Points)

Consider a discrete memoryless channel with input $X \in \{0, A\}$ where $A \geq 0$ is a real number. The input of the channel X is related to the output of the channel Y as

$$Y = X + Z$$

where noise Z is uniformly distributed over $[0, 1]$. This means that the probability density function of Z is

$$P(z) = \begin{cases} 1 & 0 \leq z \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) For which input distribution the capacity of this channel is achieved?

♠ **Solution:**

$$\left(\frac{1}{2}, \frac{1}{2}\right)$$

with standard approach.

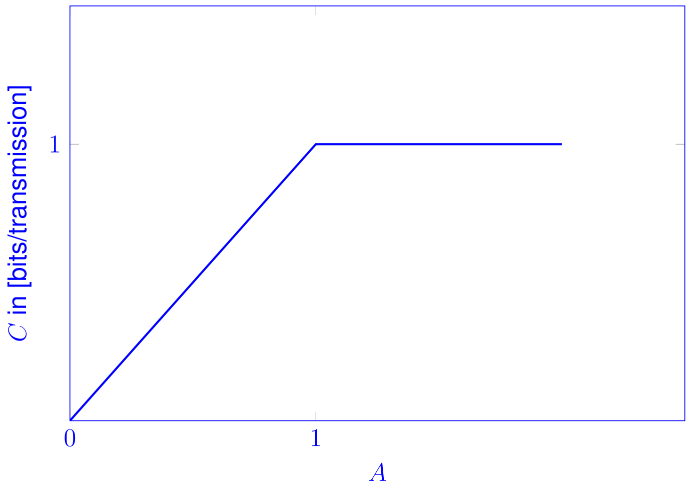
(b) Determine the channel capacity.

♠ **Solution:**

$$C = \min\{1, A\}$$

(c) Plot the channel capacity in terms of A.

♠ **Solution:**



10.1.7 Binary Codes (16 Points)

Consider a binary code whose codewords are of length $N = 12$. The parity-check matrix of this code is given by

$$\mathbf{H} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

(a) Is this binary code linear?

♠ **Solution:** Yes.

(b) Is the vector

$$\mathbf{t} = [0\;1\;1\;1\;0\;1\;1\;1\;1\;0\;1\;0]^T$$

a codeword of this binary code?

♠ **Solution:**

$$\mathbf{Ht} \neq 0 \implies \text{No}$$

(c) Is this code a regular Low-Density Parity Check (LDPC) code? Justify your answer by giving reasons.

♠ **Solution:**

Yes $\implies$ Number of zeros(ones) in rows and columns are the same

(d) Draw the factor graph of this binary code.

♠ **Solution:**

$\implies$ Standard

10.1.8 Sum-Product Algoritm (9 Points)

For binary inputs $x, y \in \{0, 1\}$, the functions $h(x, y)$ and $g(x)$ are defined as

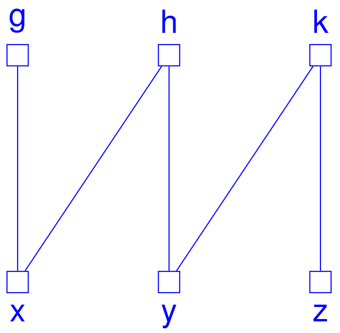
$$h(x,y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases}, \qquad g(x) = \begin{cases} 0.25 & x = 1 \\ 0.75 & x = 0 \end{cases}$$

Consider the binary variable x, y, z and let the function $f(x, y, z)$ be

$$f(x,y,z) = g(x)h(x,y)h(y,z).$$

(a) Draw the factor graph corresponding to $f(x, y, z)$ and specify the factor and variable nodes.

♠ **Solution:**



(b) Determine the marginal distribution

$$Z_X(x) = \sum_{y,z \in \{0,1\}} f(x,y,z)$$

via the sum-product algorithm.

♠ **Solution:**

Write standard max-sum $\implies$ $Z_X(x) = g(x)$

10.2 Summer Semester 2018 Exam

Exam Date: July 17th, 2018
9 Questions with total of 90 Points.
Exam Duration: 90 Minutes

10.2.1 Short Answer Questions (9 Points)

(a) Which distribution maximizes the entropy of a discrete random variable? How large is the entropy in this case?

♠ **Solution:** Uniform distribution, $H(X) = \log |\mathcal{A}_X|$

(b) Is the code $C = \{0, 01, 011, 111\}$ prefix free? Give a reason for you answer.

♠ **Solution:** No, 0 is prefix of 01.

(c) Is the code uniquely decodable?

♠ **Solution:** Yes.

(d) The source words associated with the code from question **(b)** have probabilities $\mathcal{P}_X = \{0.5, 0.25, 0.125, 0.125\}$. Is the code optimal? Give a reason for your answer.

♠ **Solution:** Yes, the probability of each codeword is 2^{-l}, where l is the length of the codeword.

For the following questions, assume the code is optimal.

(e) If we pick a random bit from the output of the encoder, what is the probability of the bit being a $"1"$?

♠ **Solution:** Since the encoder is optimal, both 0 and 1 appear with the same probability $\rightarrow p_1 = \frac{1}{2}$

(f) What is the probability of observing the string $"011"$ in the output of the decoder?

♠ **Solution:** As above, the probability for any symbol string should be equiprobable $\rightarrow p = \frac{1}{8}$

10.2.2 Probabilities and Inference (11 Points)

You meet another student, S. The student S has two siblings, A and B.

(a) What is the probability that the student S is the oldest?

♠ **Solution:** There are six possibilities: ABS,ASB,SAB,BAS,BSA,SBA. In two of the six cases, S is the oldest. $p = \frac{1}{3}$

(b) The student tells you that he is older than A. What is now the probability that S is the oldest? Show your thoughts!

♠ **Solution:** Now, only three possibilities remans: SAB,BSA,SBA. In two of those three, S is the oldest. $p = \frac{2}{3}$

In the following, X and Y are arbitrary discrete random variables.

(c) Show that $I(X;Y) \leq H(X,Y)$ holds.

♠ **Solution:**

$$\begin{aligned}
I(X;Y) &\leq H(X,Y)\\
I(X;Y) &\leq H(X) + H(Y|X) + \underbrace{H(X|Y) - H(X|Y)}_{=0}\\
I(X;Y) &\leq \underbrace{H(X) - H(X|Y)}_{=I(X;Y)} + H(X|Y) + H(Y|X)\\
0 &\leq \underbrace{H(X|Y)}_{\geq 0} + \underbrace{H(Y|X)}_{\geq 0}
\end{aligned}$$

(d) Show that $H(X,Z) - I(X;Z) \geq 0$ holds,
for $Z = f(X)$ with $f \in \mathbb{R}$ on X

♠ **Solution:**

$$\begin{aligned}
H(X,Z) - I(X;Z) &\geq 0\\
H(X,f(X)) - (H(X) - H(X|f(X))) &\leq 0\\
H(X|f(X)) &\leq 0
\end{aligned}$$

10.2.3 Asymptotic Equipartition Principle (10 Points)

(a) Explain in your own words, the meaning of the *smallest δ-sufficient subset* (S_δ).

♠ **Solution:** The smallest δ-sufficient subset contains the most probable elements of a source X^N, such that the sum of probabilities in the set is at least $1 - \delta$, while the cardinality of the subset is as small as possible.

(b) Explain, in your own words, the meaning of the *typical set* $(T_{N\beta})$

♠ **Solution:** The typical set of a source X^N contains elements, whose probabilities are close to the entropy of the source.

(c) Why cannot we use the smallest δ-sufficient subset for the proof of the source coding theorem?

♠ **Solution:** We do not know the cardinality of the smallest δ-sufficient subset.

(d) Show that the sequence 0011001110 is contained in the smallest $\delta = 0.01$-sufficient subset of the source X^{10} with $\mathcal{A}_X = \{0, 1\}$ and $\mathcal{P}_X = \{0.25, 0.75\}$.

♠ **Solution:** We will prove that at least one less probable element with $n_0 = 6$ and $n_1 = 4$ zeros and ones is part of the subset, which means that all elements with $n_0 = 5$ and $n_1 = 5$ zeros and ones must be part of it.

$$p_{6,4} = \binom{10}{4} \cdot 0.25^6 \cdot 0.75^4 \approx 0.016$$

This is larger than δ, therefore at least one of the elements must be in the δ-sufficient subset, therefore our sequence must be.

10.2.4 Burrows-Wheeler-Transform (8 Points)

(a) Calculate the inverse Burrows-Wheeler-Transform of

$$EFFEVCO$$

The index of the source word is 0.
Hint: The resulting string is not necessarily a real word

♠ **Solution:** $L = [5, 0, 3, 1, 2, 6, 4]$

$$\begin{array}{ccccccc}
C & O & V & F & E & F & E \\
E & C & O & V & F & E & F \\
E & F & E & C & O & V & F \\
F & E & C & O & V & F & E \\
F & E & F & E & C & O & V \\
O & V & F & E & F & E & C \\
V & F & E & F & E & C & O
\end{array}$$

$S = COVFEFE$

(b) Why would we use the Burrows-Wheeler-Transform? It obviously does not compress the source string.

♠ **Solution:** It exploits memory of the source to group same symbols together, making compression easier.

(c) Name a source which will be easier to compress when first transformed using the Burrows-Wheeler-Transform. Give a reason why.

♠ **Solution:** English language, some character combinations are very likely, BWT will exploit this.

10.2.5 Lempel-Ziv (Points)

(a) Encode the sequence 001000111010 using the optimized Lempel-Ziv algorithm, i.e. with replacement of no longer needed source words in the source word table. Underline symbols that may be omitted.

♠ **Solution:**

index	source word
00	1
01	00
10	010
11	011

$$(\underline{0},0)(1,1)(01,\underline{0})(10,1)(00,\underline{1})(10,\underline{0}) \rightarrow 011011010010$$

(b) Is there any information the receiver needs, apart from the encoded string? If yes, which?

♠ **Solution:** No.

(c) Name one advantage of Lempel-Ziv over both Huffman and Arithmetic Coding!

♠ **Solution:** No information about the source needed.

(d) Will a flipped bit in the encoded string have influence on more than the source word associated with the flipped bit? If yes, how so? If no, why not?

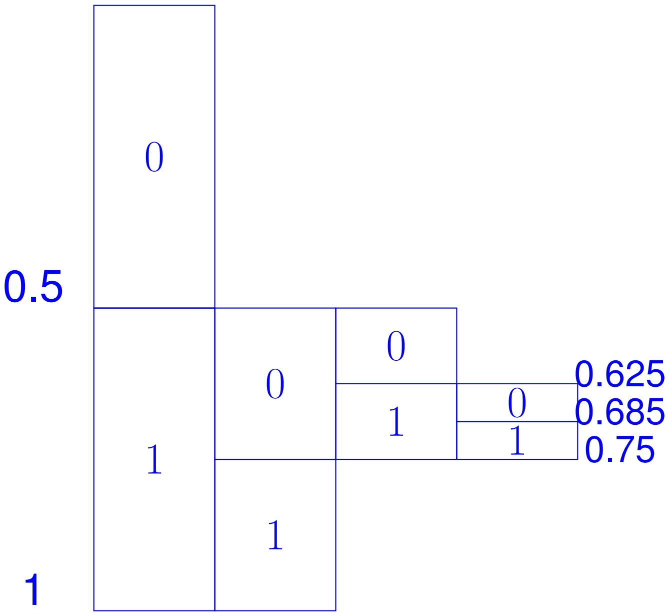
♠ **Solution:** Yes, it might have affected the table.

10.2.6 Arithmetic Coding (12 Points)

A source X emits symbols $\mathcal{A}_X = \{A, B\}$ with probabilities $\mathcal{P}_X = \{P_A = \frac{N-N_A}{N}, P_B = 1 - P_A\}$, where N is the number of symbols emitted so far, and N_A is the number of symbols A emitted so far. For the first symbol emitted, $P_A = P_B = 0.5$. The source words are of length $L = 5$.

(a) Find the interval for the code word 1011.

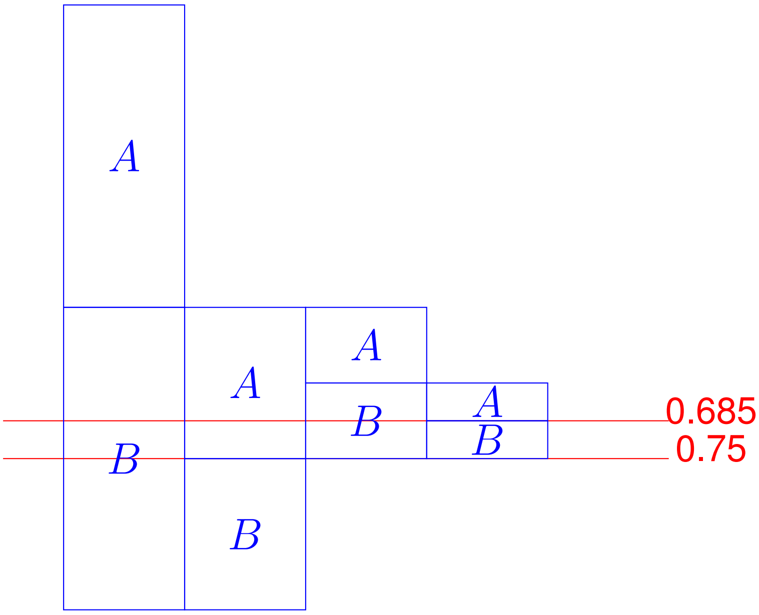
♠ **Solution:**



→ Interval is $[0.6875, 0.75]$

(b) Find the source word for the code word.

♠ **Solution:**



→ $BAABB$

10.2.7 Channel Capacity (14 Points)

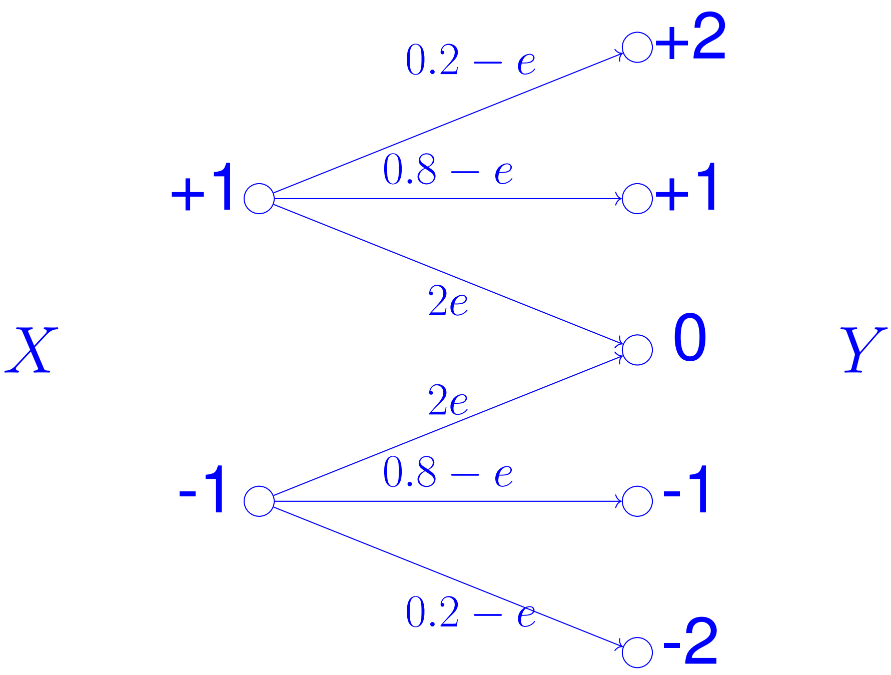
Consider a discrete, memoryless channel in which the channel input $X \in \{+1, -1\}$ is related to the output Y as follows:

$$Y = XZ,$$

where $Z \in \{0, 1, 2\}$ is a multiplicative noise with probabilities $\mathcal{P}_Z = \{2a, 0.8 - a, 0.2 - a\}$ for $0 \leq a \leq 0.2$.

(a) Draw the transition probability diagram for this channel.

♠ **Solution:**

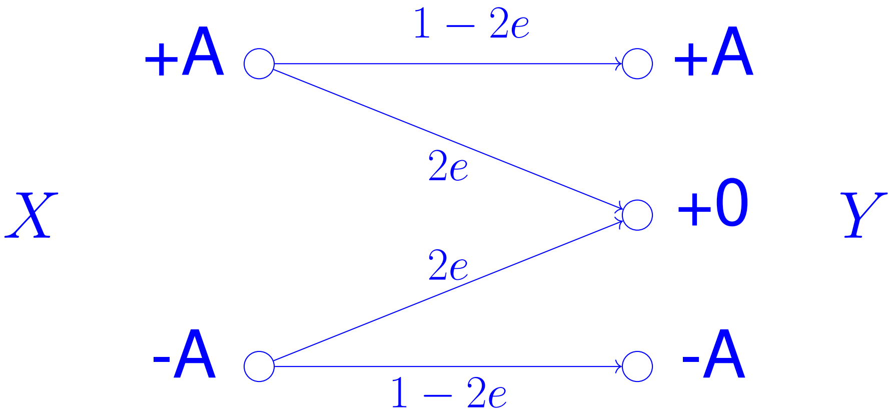


(b) Determine the input distribution $\mathcal{P}_X$ for which the channel capacity is achieved.

♠ **Solution:** There are two approaches to solve the problem. The first approach is quite simple and short but needs a bit of attention. The second approach is the standard approach which may take some more time.

Approach 1

By defining hyper-symbols $+A = \{+1, +2\}$ and $-A = \{-1, -2\}$,the channel reduces to:



which is a Binary Erasure Channel (BEC) with error probability $2e$. As the result, the channel capacity is achieved by the uniform input distribution, i.e, $\mathcal{P}_X = \{0.5, 0.5\}$.

Approach 2

The standard approach is to directly determine the capacity term. By this approach, we first assume an arbitrary input distribution $\mathcal{P}_X = \{p, 1-p\}$. Consequently, the capacity of the channel reads

$$\begin{aligned} C &= \max_{p\in[0,1]} I(X;Y) \\ &= \max_{p\in[0,1]} H(Y) - H(Y|X) \\ &= \max_{p\in[0,1]} H(Y) - pH(Y|X=+1) - (1-p)H(Y|X=-1) \end{aligned}$$

Considering the input-output relation, we have that

$$\text{Pr}\{Y = y|X = \pm 1\} \begin{cases} 2e & y = 0 \\ 0.8 - e & y = \pm 1 \\ 0.2 - e & y = \pm 2 \end{cases}$$

Therefore, we have

$$\begin{aligned} H(Y|X=\pm 1) &= 2e \log \frac{1}{2e} + (0.8-e)\log\frac{1}{0.8-e} + (0.2-e)\log\frac{1}{0.2-e} \\ &\overset{\text{Def.}}{=} F(e). \end{aligned}$$

which is independent of p. From the conditional distributions we have

$$\text{Pr}\{Y = y\} = \begin{cases} p(0.2-e) & y = +2 \\ p(0.8-e) & y = +1 \\ 2e & y = 0 \\ (1-p)(0.8-e) & y = -1 \\ (1-p)(0.2-e) & y = -2 \end{cases}$$

which concludes that

$$\begin{aligned} H(Y) =& 2e\log\frac{1}{2e} + p(0.8-e)\log\frac{1}{p(0.8-e)} + (1-p)(0.8-e)\log\frac{1}{(1-p)(0.8-e)} \\ &+ p(0.2-e)\log\frac{1}{p(0.2-e)} + (1-p)(0.2-e)\log\frac{1}{(1-p)(0.2-e)} \\ &= (1-2e)H_2(p) + F(e). \end{aligned}$$

Since $H(Y|X)$ is independent of p, maximizing $H(Y)$ maximizes the mutual information. As the result, the capacity is achieved by setting $p = 0.5$.

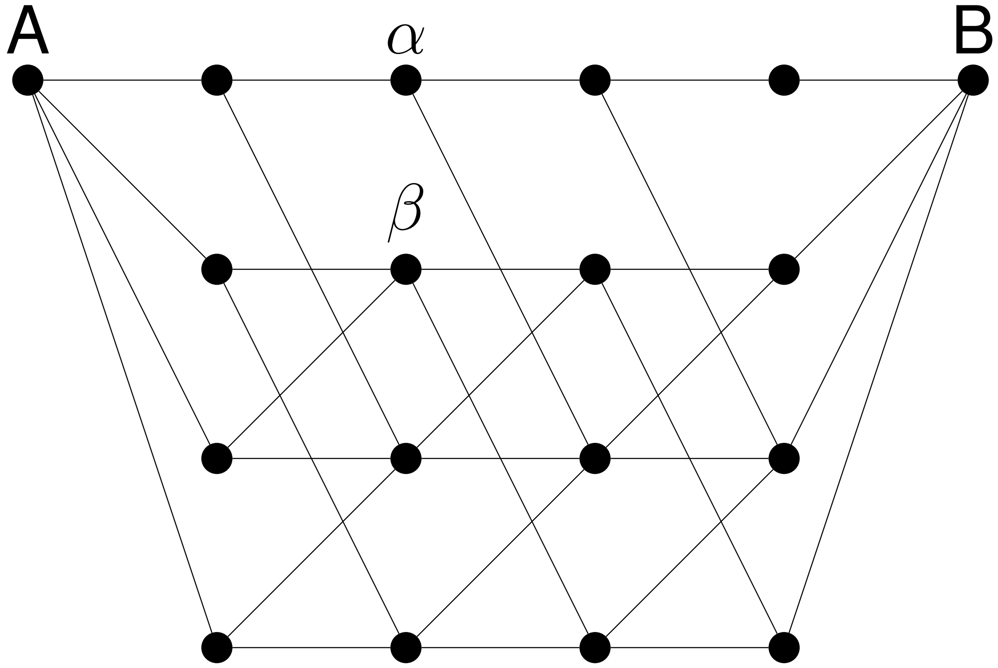
(c) Calculate the channel capacity.

♠ **Solution:**

$$\begin{aligned} C &= H(Y) - pH(Y|X=+1) - (1-p)H(Y|X=-1) \\ &= (1-2e) + F(e) - pF(e) - (1-p)F(e) \\ &= 1 - 2e \text{ bit/transmission} \end{aligned}$$

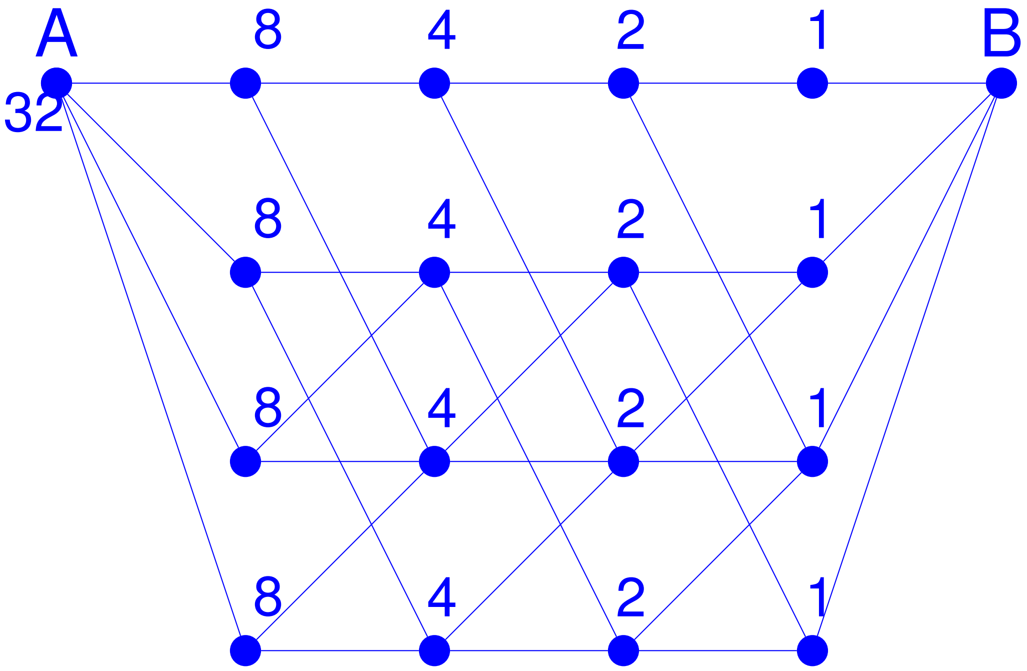
10.2.8 Message Passing (7 Points)

The following graph is given. Each edge may only be traversed from left to right.



(a) Determine the number of different paths from A to B.

♠ **Solution:**



(b) What is the probability that an inner node (neither A nor B) is part of a path chosen uniformly at random?

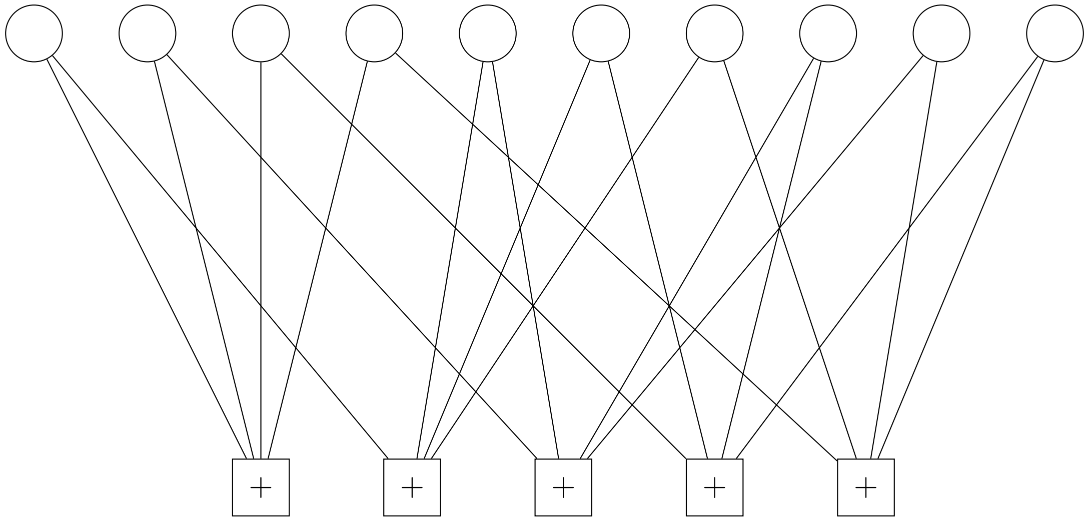
♠ **Solution:** Since the path is chosen uniformly at random, the number of paths that pass through that node divided by the number of total paths.

(c) What is the probability, if at every node an outgoing edge is chosen uniformly at random, to pass node α? What is the probability to pass node β?

♠ **Solution:** This is not the same as above: Choosing a path random is not the same as choosing each edge at random (edges are part of varying number of paths) Pr$\{\alpha\}$=$\frac{1}{8}$, Pr$\{\beta\}$=$\frac{1}{4}$.

10.2.9 Low-Density Parity-Check Codes (10 Points)

A binary Low-Density Parity-Check code is given by the factor graph below:



The variable nodes are numbered ascending from left to right.

(a) What is the rate of the code?

♠ **Solution:** $R = \frac{1}{2}$

(b) Write down the parity-check matrix corresponding to the factor graph.

♠ **Solution:**

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

(c) Is the word 1101010101 a valid codeword?

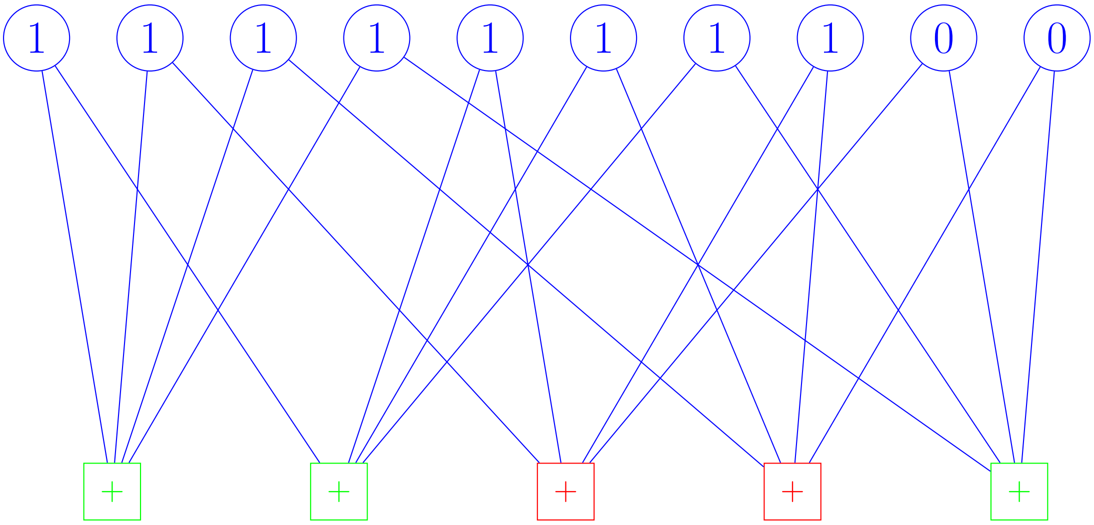
♠ **Solution:** No.

(d) Is the code regular? Give a reason for your answer.

♠ **Solution:** Yes, the same number of '1's in every row and column.

(e) Decode the received word 1111111100.

♠ **Solution:**



The only variable node involved in both red checks is the third to last.
$\rightarrow\ c = 1111111000$

10.3 Winter Semester 2018-2019 Exam

Exam Date: February 12th, 2019
6 Questions with total of 100 Points.
Exam Duration: 90 Minutes

10.3.1 Information Theoretic Identities and Inequalities (17 Points)

X is a **uniform** discrete random variable with alphabet

$$\mathcal{A}_X = \{0, 1, 2, 3, 4, 5, 6, 7\}.$$

Y is another discrete random variable whose alphabet is

$$\mathcal{A}_Y = \{A, B, C, D\}$$

The distribution of Y is **not** known. X and Y are **independent.**

Let discrete random variable Z be a deterministic function of Y; this means that

$$Z = f(Y)$$

Replace $\square$ in the following items with either $=$ or $\geq$ or $\leq$.

(a) $H(X|Y) \square H(Y)$

♠ **Solution:** Since X and Y are independent, we have

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

Moreover, as X is uniformly distributed, we can write

$$H(X|Y) = H(X) = \log 8 = 3$$

For Y, we know that $|\mathcal{A}_Y| = 4$, and hence

$$H(Y) \leq \log |\mathcal{A}_Y| = 2.$$

Therefore,

$$H(X|Y) = 3 \geq 2 \geq H(Y) \implies H(X|Y) \boxed{\geq} H(Y)$$

(b) $I(Z;Y) \square H(Z)$

♠ **Solution:** From the definition of mutual information, we have

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

Since $Z = f(Y)$, we have

$$H(Z|Y) = 0.$$

Therefore,

$$I(Z|Y) \boxed{=} H(Z)$$

(c) $I(Z;Y) \square H(Y)$

♠ **Solution:** By the alternative definition of mutual information, we have

$$I(Z;Y) = H(Y) - H(Y|Z).$$

In general, we have

$$H(Y|Z) \geq 0.$$

Thus,

$$I(Z|Y)\boxed{\leq}H(Y)$$

(d) $I(Z;Y) \square H(X)$

♠ **Solution:** From Part (c), we know that

$$I(Z;Y) \leq H(Y)$$

From Part (a), we know that $H(X) = 3$ and $H(Y) \leq 2$. Thus,

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

As a result,

$$I(Z|Y)\boxed{\leq}H(X)$$

(e) $I(Z;X) \square H(Z)$

♠ **Solution:** Since X and Y are independent, and Z is deterministic function of Y, we can conclude that X and Z are independent, as well. Therefore,

$$I(Z;X) = 0.$$

In general, we have

$$H(Z) \geq 0.$$

Thus,

$$I(Z|X)\boxed{\leq}H(Z)$$

10.3.2 Bayesian Inference (13 Points)

A sender can transmit four different signals, namely

$$s_1(t), s_2(t), s_3(t) \text{ and } s_4(t).$$

Each signal is a superposition of 100 distinct monotone signals. This means that $s_k(t)$ is written as

$$s_k(t) = \sum_{n=1}^{100} A_{k,n} \text{exp}\{2\pi j f_n t\}$$

where

- $f_1, \ldots, f_{100}$ are the *frequencies* of the monotone signals.
- $A_{k,1}, \ldots, A_{k,100}$ are **either zero or one** and called the *signal coefficients.*

The numbers of zero and non-zero signal coefficients, for each of the signal, are given in the following table:

Value of the signal coefficient	Zero	One
# of signal coefficients in $s_1(t)$	0	100
# of signal coefficients in $s_2(t)$	5	95
# of signal coefficients in $s_3(t)$	36	64
# of signal coefficients in $s_4(t)$	82	18

The sender transmits one of these signals to a receiver. The receiver has a simple hardware which can measure the value of the signal coefficients only in 5 frequencies. Therefore, it chooses 20 frequencies from $f_1, \ldots, f_{100}$ at random and measures the values of the signal coefficients in the chosen frequencies.
It then calculates the numbers of zero and non-zero coefficients and shows it on its monitor.

The following data is shown on the monitor of the receiver.

Value of the signal coefficient	Zero	One
# of measured signal coefficients	2	3

(a) Calculate the probability that the sender has transmitted $s_3(t)$.

♠ **Solution:** Let the random variable $S \in \{1, 2, 3, 4\}$ be the index of the transmitted signal. Define random vector $D = [D_0, D_1]$ to be the data observed on the monitor of the receiver, where D_0 is the number of zero-entries, and D_1 is the number of one-entries. We intend to determine,

$$\text{Pr}\{S = 3 | D = [2, 3]\}.$$

To do so, we use the Bayes rule which indicates

$$\Pr\{S=3|D=[2,3]\} = \frac{\Pr\{S=3, D=[2,3]\}}{\Pr\{D=[2,3]\}} = \frac{\Pr\{D=[2,3]|S=3\}\Pr\{S=3\}}{\Pr\{D=[2,3]\}}$$

Since we have no prior information about the transmitter, we assume that

$$\Pr\{S=1\} = \Pr\{S=2\} = \Pr\{S=3\} = \Pr\{S=4\} = \frac{1}{4}$$

For the term in the denominator, we have

$$\Pr\{D=[2,3]\} = \sum_{s=1}^{4} \Pr\{D=[2,3]|S=s\}\Pr\{S=s\} = \frac{1}{4}\sum_{s=1}^{4} \Pr\{D=[2,3]|S=s\}.$$

Therefore, we can write

$$\Pr\{S=3|D=[2,3]\} = \frac{\Pr\{D=[2,3]|S=3\}}{\sum_{s=1}^{4} \Pr\{D=[2,3]|S=s\}}$$

To determine $\Pr\{D=[2,3]|S=s\}$, we note that

$$\Pr\{D=[2,3]|S=s\} = (\Pr\{A_k=0|S=s\})^2(\Pr\{A_k=1|S=s\})^3.$$

Hence, we have

$$\begin{aligned}
\Pr\{D=[2,3]|S=1\} &= (0)^2(1)^3 = 0\\
\Pr\{D=[2,3]|S=2\} &= (0.05)^2(0.95)^3 = 0.0021\\
\Pr\{D=[2,3]|S=3\} &= (0.36)^2(0.64)^3 = 0.0340\\
\Pr\{D=[2,3]|S=4\} &= (0.82)^2(0.18)^3 = 0.0039
\end{aligned}$$

which concludes that

$$\boxed{\Pr\{S=3|D=[2,3]\} = \frac{0.0340}{0+0.0021+0.0340+0.0039} = 0.85}$$

(b) **Without** doing any further calculation, find the signal which has most likely been transmitted.

♠ **Solution:** Noting that $\Pr\{S=3|D=[2,3]\} = 0.85$, we have for all $s \neq 3$

$$\Pr\{S=s|D=[2,3]\} \leq \Pr\{S\neq 3|D=[2,3]\} = 1 - \Pr\{S=3|D=[2,3]\} = 0.15.$$

Therefore, for $s \neq 3$, we have

$$\Pr\{S=3|D=[2,3]\} \geq \Pr\{S=s|D=[2,3]\}.$$

This means that $s_3(t)$ has most likely been transmitted.

10.3.3 Source Coding (26 Points)

Consider the random sequence of length $2N$

$$X^{2N} = X_1, \ldots, X_{2N}$$

in which $X_n \in \{A, B, C, D\}$ for $n = 1, \ldots, 2N$. Let N be a power of 2. This means that $N = 2^U$ for some integer U.

The symbol X_1 in this sequence is generated by tossing a **uniform** four-faced die whose faces are labeled by A, B, C and D. This means that

$$\text{Pr}\{X_1 = A\} = \text{Pr}\{X_1 = B\} = \text{Pr}\{X_1 = C\} = \text{Pr}\{X_1 = D\} = 0.25$$

The symbol X_2 is then generated base on X_1 as follows:

- If $X_1 = A$, X_2 is generated by tossing a **uniform** coin whose faces are labeled by C and D. This means that

$$\text{Pr}\{X_2 = C|X_1 = A\} = \text{Pr}\{X_2 = D|X_1 = A\} = 0.5$$

- If $X_1 = B$ or C, X_2 is set to $X_2 = X_1$. This means that

$$\text{Pr}\{X_2 = B|X_1 = B\} = \text{Pr}\{X_2 = C|X_1 = C\} = 1$$

- If $X_1 = D$, X_2 is generated by tossing a **uniform** coin whose faces are labeled by A and B. This means that

$$\text{Pr}\{X_2 = A|X_1 = D\} = \text{Pr}\{X_2 = B|X_1 = D\} = 0.5$$

The symbols X_{2k-1} and X_{2k} for $k = 2, \ldots, N$ are then generated by repeating this procedure **independently** for $N - 1$ other times.

(a) Write all possible sequences for N=1.

♠ **Solution:** For $N = 1$, we have the following *six* possible sequences:

Sequence X^2	Probability
AC	1/8
AD	1/8
BB	1/4
CC	1/4
DA	1/8
DB	1/8

(b) Write two possible sequences for N=3.

♠ **Solution:** For $N = 3$, we have in general $6^2 = 36$ sequences. These sequences are given by cascading two of the sequences in Part (a). For example,

$$X^4 = ACAC \quad \text{and} \quad X^4 = ADBB$$

are two possibilities.

(c) Calculate the average entropy of this sequence which is defined as

$$\bar{H} = \frac{H(X^{2N})}{2N}$$

♠ **Solution:** Noting that (X_{2k-1}, X_{2k}) are independent for different values of k, we can write

$$H(X^{2N}) = \sum_{k=1}^{N} H(X_{2k-1}, X_{2k}).$$

(X_{2k-1}, X_{2k}) for $k = 1, \dots, N$ are identically distributed with the distribution given in the table in Part (a). Hence,

$$H(X^{2N}) = NH(X_1, X_2) \implies \bar{H} = \frac{H(X_1, X_2)}{2}$$

$H(X_1, X_2)$ is moreover calculated as

$$H(X_1, X_2) = 4 \times \frac{1}{8} \log 8 + 2 \times \frac{1}{4} \log 4 = 2.5 \text{ bits}$$

Therefore, $\boxed{\bar{H} = 1.25 \text{ bits}}$

(d) Consider the sequence V^N which is constructed from X^{2N} by grouping the symbols in X^{2N} into blocks of length $B = 2$. This means that

$$V^N = V_1, \dots, V_N$$

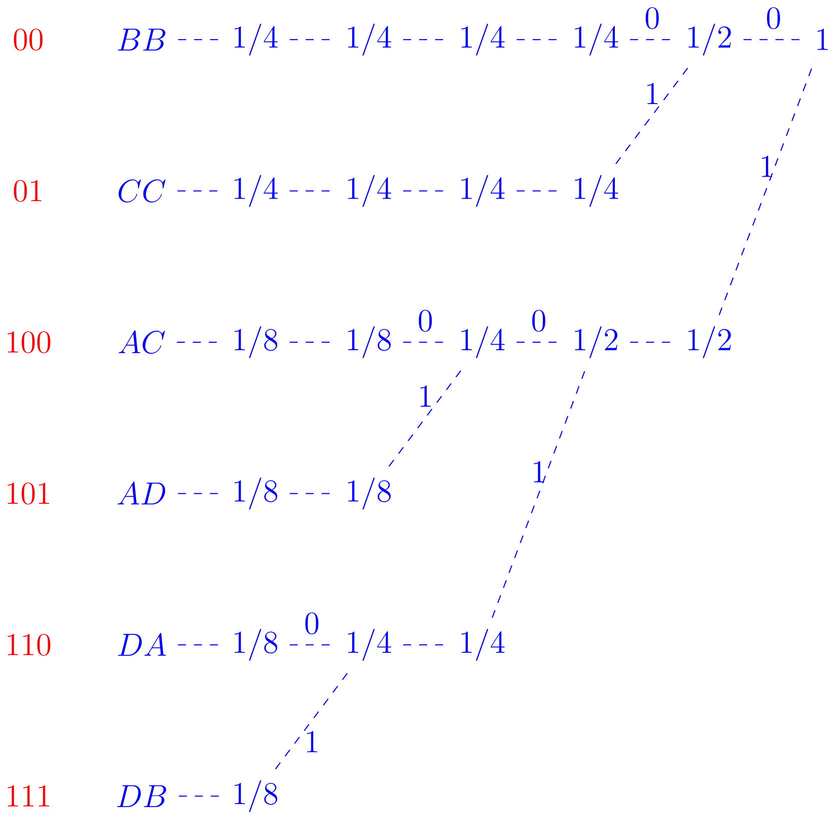
where

$$V_j = (X_{2j-1}, X_{2j})$$

for $j = 1, \dots, N$.

Give a **Huffman code** for the sequence V^N and calculate the expected length of this code $\bar{L}$.

♠ **Solution:** Following the discussions in the previous parts, V^N is an i.i.d sequence whose entries are distributed as (X_1, X_2). As a result, a Huffman code is



The average length is hence calculated as

$$\bar{L} = \sum_{i=1}^{6} p_i \ell_i = \frac{1}{4} \times 2 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = 2.5 \text{ bits}$$

(e) Consider the expected length $\bar{L}$ determined in Part (d) and the average entropy $\bar{H}$ derived in Part (c). Calculate the fraction

$$F = \frac{\bar{L}}{\bar{H}}$$

and compare it to the block length $B = 2$. Are the fraction F and the block length B equal?

Justify the result by giving a reason.

♠ **Solution:** The fraction F reads

$$F = \frac{\bar{L}}{\bar{H}} = \frac{2.5}{1.25} = 2.$$

Thus, we have

$$F = B.$$

This observation follows the fact that V^N is an independent and identically distributed (i.i.d) source, whose entries have probabilities $p_i = 2^{-l_i}$ for some integer l_i. From Chapter 5 of the textbook, we know that for such a source, the expected length of a Huffman code reads $\bar{L} = H(V_1) = H(X_1, X_2)$. On the other hand, we showed in Part (c) that $H(X_1, X_2) = 2\bar{H} = B\bar{H}$. This hence concludes that $F = B$.

(f) Now assume that we repeat Part (d) with some block size $B > 2$. For which choices of B the relation between F and B, derived in Part (e), holds? Give a reason for your answer.

♠ **Solution:** As each two adjusting entries of X^{2N} are independently generated, for any *even integer* B, we have

$$H(X_1, \ldots, X_B) = B\bar{H}.$$

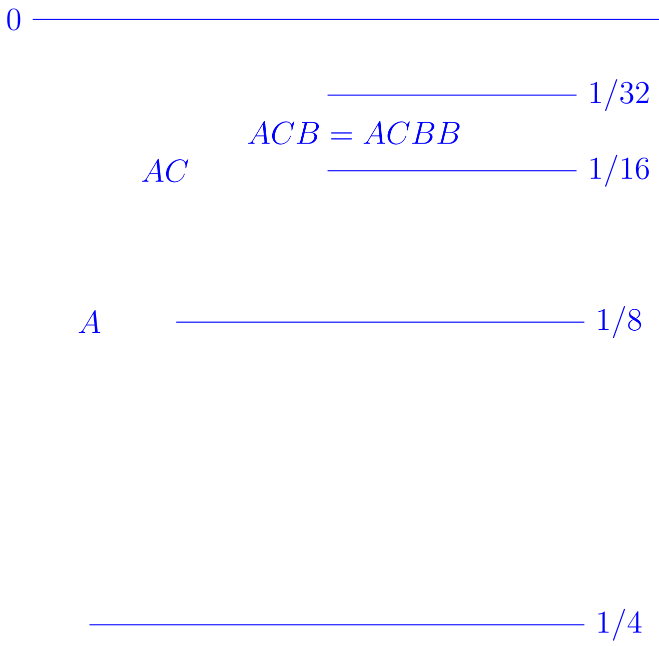
Noting that N is an integer power of two, we conclude that for any $B \leq N$ which is *integer power of two*, we have $F = B$.

(g) Give a binary codeword for the sequence

$$X^4 = ACBB$$

using the **arithmetic coding** algorithm.

♠ **Solution:** Following the standard approach, we have



Hence, $[1/32, 1/16)$ is the interval corresponding to $ACBB$. By same partitioning approach, one can find that this is also the interval for 00001. Therefore, $\boxed{00001}$ is an arithmetic code for $ACBB$.

10.3.4 Channel Capacity (14 Points)

The input-output relationship of a *nonlinear* discrete memoryless channel is given by

$$Y = X^2 + Z$$

Here, X is the input symbol chosen from alphabet

$$\mathcal{A}_X = \{\pm\sqrt{P}, \pm\sqrt{3P}, \pm\sqrt{5P}\}$$

Z is noise which is uniformly distributed over $\{0, 2P\}$. This means that

$$\text{Pr}\{Z = 0\} = \text{Pr}\{Z = 2P\} = 0.5$$

Y is the output symbol.

(a) Determine the transition probability matrix of this channel.

♠ **Solution:** The output symbol can take in general four different outcomes, i.e. $\mathcal{A}_Y = \{P, 3P, 5P, 7P\}$. Hence, the transmit matrix of the channel reads

$$\mathbf{P} = \begin{array}{c} \\ P \\ 3P \\ 5P \\ 7P \end{array} \begin{array}{cccccc} -\sqrt{5P} & -\sqrt{3P} & -\sqrt{P} & \sqrt{P} & \sqrt{3P} & \sqrt{5P} \\ \end{array}$$

$$\mathbf{P} = \begin{array}{c|cccccc} & -\sqrt{5P} & -\sqrt{3P} & -\sqrt{P} & \sqrt{P} & \sqrt{3P} & \sqrt{5P} \\ \hline P & 0 & 0 & 0.5 & 0.5 & 0 & 0 \\ 3P & 0 & 0.5 & 0.5 & 0.5 & 0.5 & 0 \\ 5P & 0.5 & 0.5 & 0 & 0 & 0.5 & 0.5 \\ 7P & 0.5 & 0 & 0 & 0 & 0 & 0.5 \end{array}$$

(b) Calculate the capacity of this channel.

♠ **Solution:** The solution is either given by the classic approach or by some tricks. Note that the channel in this case is *not* symmetric.
Following the classic approach, we note that for any input outcome x_i, the output symbol Y is a uniform Bernoulli random variable. Hence,

$$H(Y|X = x_i) = H_2(0.5) = 1 \text{ bit.}$$

As a result, regardless of the input distribution, we can write

$$H(Y|X) = \sum_{i=1}^{6} \text{Pr}\{X = x_i\} H(Y|X = x_i) = H_2(0.5) = 1 \text{ bit}$$

Hence, the capacity reads

$$C = \max_{P(x)} I(X;Y) = \max_{P(x)} H(Y) - 1$$

Noting that

$$\max H(Y) = \log |\mathcal{A}_Y| = 2 \text{ bits}$$

we can conclude that

$$C = 2 - 1 = 1 \quad \text{bit/channel use}$$

which is achieved when the Y is uniformly distributed.

(c) Find an input distribution which achieves the channel capacity. Is this distribution unique?

♠ **Solution:** To achieve the capacity, we should find the input distribution, such that Y is uniform, i.e.

$$\text{Pr}\{Y = y_i\} = 0.25 \quad \forall y_i \in \mathcal{A}_Y.$$

Noting that Y is function of X^2, one could choose

$$\text{Pr}\{X = -\sqrt{P}\} = \text{Pr}\{X = -\sqrt{3P}\} = \text{Pr}\{X = -\sqrt{5P}\} = 0$$

and

$$\text{Pr}\{X = \sqrt{P}\} = p_1, \ \text{Pr}\{X = \sqrt{3P}\} = p_2, \ \text{Pr}\{X = \sqrt{5P}\} = p_3$$

In this case, to have uniform Y, we need

$$\begin{aligned} 0.5p_1 &= 0.25 \\ 0.5p_1 + 0.5p_2 &= 0.25 \\ 0.5p_2 + 0.5p_3 &= 0.25 \\ 0.5p_3 &= 0.25 \end{aligned}$$

which is satisfied when $p_1 = p_3 = 0.5$ and $p_2 = 0$. Thus, a capacity-achieving input distribution is

$$\text{Pr}\{X = \sqrt{P}\} = \text{Pr}\{X = \sqrt{5P}\} = 0.5 \text{ and}$$

$$\text{Pr}\{X = x_i\} = 0 \quad \forall x_i \neq \sqrt{P}, \sqrt{5P}$$

Clearly,another capacity-achieving input distribution is

$$\text{Pr}\{X = -\sqrt{P}\} = \text{Pr}\{X = -\sqrt{5P}\} = 0.5 \text{ and}$$

$$\text{Pr}\{X = x_i\} = 0 \quad \forall x_i \neq -\sqrt{P}, -\sqrt{5P}$$

and hence, it is *not* unique.

10.3.5 Channel Codes (17 Points)

A channel code encodes binary information words of length $K = 2$ to binary codewords of length $N = 4$ by repeating the information words once again. For example, the information word

$$\mathbf{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

is encoded as

$$\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

(a) What is the code rate of this channel code?

♠ **Solution:** The code rate for this code is

$$R_C = \frac{K}{N} = 0.5.$$

(b) Find the *generator matrix* **G**, such that for each information word $\mathbf{b}_{2\times 1}$, the codeword $\mathbf{c}_{4\times 1}$ be determined as

$$\mathbf{c} = \mathbf{G}\mathbf{b}$$

♠ **Solution:** To find the generator matrix, we note that for

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

we have

$$\mathbf{c} = \begin{bmatrix} b_1 \\ b_2 \\ b_1 \\ b_2 \end{bmatrix}.$$

Thus, we have

$$\mathbf{G} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(c) Write **two** parity check equations for this code and determine the corresponding parity check matrix.

♠ **Solution:** For any codeword

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix},$$

we know that $c_1 = c_3$ and $c_2 = c_4$. Hence, we have

$$c_1 \oplus c_3 = 0$$
$$c_2 \oplus c_4 = 0.$$

The parity check matrix is therefore

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} = \mathbf{G}^T.$$

(d) How many codewords does this code have?

♠ **Solution:** Since $K = 2$, we have in total $2^K = 4$ codewords.

(e) Determine the minimum Hamming weight of the codewords.

♠ **Solution:** The codewords of this code are

$$\mathbf{c}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \ \mathbf{c}_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \ \mathbf{c}_3 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \text{ and } \mathbf{c}_4 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Hence, the minimum Hamming weight of this code,

- when we consider the all zero codeword is 0
- when we ignore the all zero codeword is 2.

(f) Determine the minimum Hamming distance of this code.

♠ **Solution:** In general, we have 6 combinations of two codewords which have the Hamming distance of either 4 or 2. As a result, the minimum Hamming distance of the code is 2.

(g) Are the values in Parts (f) and (e) equal? Explain your answers by giving a reason.

♠ **Solution:** If we ignore the all zero codeword, then the minimum Hamming weight of the code is equal to the minimum Hamming distance.
This observation follows the fact that the code is *linear*. From Chapter 13 of the textbook, we know that linear codes have symmetric distance property which means that the distance of neighbouring codewords are the same for each codeword. As the result, the minimum distance from the all zero codeword equals to the minimum Hamming distance of the code. The former is the minimum Hamming weight, when we ignore all zero codeword.

10.3.6 Factor Graphs (13 Points)

Consider **binary** random variables X_1, X_2, X_3, and X_4 whose joint distribution is

$$P(x_1, x_2, x_3, x_4) = Q_A(x_1, x_2) Q_B(x_2, x_3) Q_C(x_3, x_4)$$

for $x_1, x_2, x_3, x_4 \in \{0, 1\}$. The functions $Q_A(x_1, x_2)$, $Q_B(x_2, x_3)$ and $Q_C(x_3, x_4)$ are given by

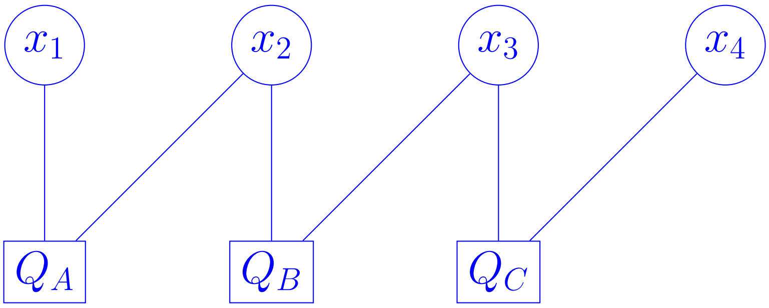
$$Q_A(x_1, x_2) = \begin{cases} 0.4 & \text{if } x_1 = x_2 \\ 0.6 & \text{if } x_1 \neq x_2 \end{cases},$$

$$Q_B(x_2, x_3) = \begin{cases} 0.2 & \text{if } x_2 = x_3 \\ 0.8 & \text{if } x_2 \neq x_3 \end{cases},$$

$$Q_C(x_3,x_4)=\begin{cases}0 & \text{if } x_3=x_4\\ 0.5 & \text{if } x_3\neq x_4\end{cases}$$

(a) Sketch the factor graph corresponding to the joint distribution $P(x_1,x_2,x_3,x_4)$.

♠ **Solution:** For this function, we have four variables and three factor. Hence, the factor graph is



(b) Calculate the marginal distribution of X_2 **via the sum-product algorithm**.

♠ **Solution:** Following Chapter 26 of the textbook, we know that

$$P(x_2)=\sum_{x_1,x_3,x_4} P(x_1,x_2,x_3,x_4)=r_{A\to 2}(x_2)r_{B\to 2}(x_2)$$

where $r_{A\to 2}(x_2)$ and $r_{B\to 2}(x_2)$ are the messages sent by the factor nodes Q_A and Q_B to the variable node x_2,in the sum-product algorithm, respectively. We hence need to calculate $r_{A\to 2}(x_2)$ and $r_{B\to 2}(x_2)$. For $r_{A\to 2}(x_2)$, we have

$$r_{A\to 2}(x_2)=\sum_{x_1,x_3,x_4} Q_A(x_1,x_2)\prod_{i=1,3,4} q_{i\to A}(x_i)$$

where $q_{i\to A}(x_i)$ is the message from the variable node x_i to the factor node Q_A. Thus,

$$r_{A\to 2}(x_2)=\sum_{x_1} Q_A(x_1,x_2)q_{1\to A}(x_1)\overset{\star}{=}\sum_{x_1} Q_A(x_1,x_2)=Q_A(0,x_2)+Q_A(1,x_2)$$
$$=\begin{cases}0.4 & \text{if } x_2=0\\ 0.6 & \text{if } x_2=1\end{cases}+\begin{cases}0.6 & \text{if } x_2=0\\ 0.4 & \text{if } x_2=1\end{cases}=1$$

where $\star$ follows the fact that x_1 is a leaf node, and has only $r_{A\to 1}(x_1)$ as the incoming message.
Similarly, for $r_{B\to 2}(x_2)$, we have

$$r_{B\to 2}(x_2)=\sum_{x_3} Q_B(x_2,x_3)q_{3\to B}(x_3).$$

We hence need to find $q_{3\to B}(x_3)$ which reads

$$q_{3\to B}(x_3)=r_{C\to 3}(x_3).$$

Continuing further with the algorithm we have

$$\begin{aligned} r_{C\to 3}(x_3) = \sum_{x_4} Q_C(x_3, x_4) q_{4\to C}(x_4) &\overset{\dagger}{=} \sum_{x_4} Q_C(x_3, x_4) = Q_C(x_3, 0) + Q_C(x_3, 1) \\ &= \begin{cases} 0 & \text{if } x_3 = 0 \\ 0.5 & \text{if } x_3 = 1 \end{cases} + \begin{cases} 0.5 & \text{if } x_3 = 0 \\ 0 & \text{if } x_3 = 1 \end{cases} = 0.5 \end{aligned}$$

where † again follows the fact that x_4 is a leaf node. Consequently, $q_{3\to B}(x_3) = 0.5$, and

$$\begin{aligned} r_{B\to 2}(x_2) = 0.5[Q_B(x_2, 0) + Q_B(x_2, 1)] &= 0.5 \left[\begin{cases} 0.2 & \text{if } x_2 = 0 \\ 0.8 & \text{if } x_2 = 1 \end{cases} + \begin{cases} 0.8 & \text{if } x_2 = 0 \\ 0.2 & \text{if } x_2 = 1 \end{cases} \right] \\ &= 0.5 \end{aligned}$$

Therefore, we have

$$P(x_2) = r_{A\to 2}(x_2) r_{B\to 2}(x_2) = 1 \times 0.5 = 0.5 = \begin{cases} 0.5 & \text{if } x_2 = 0 \\ 0.5 & \text{if } x_2 = 1 \end{cases}.$$

10.4 Summer Semester 2019 Exam

Exam Date: July 30th, 2019
6 Questions with total of 100 Points.
Exam Duration: 90 Minutes

10.4.1 Information Theoretic Identities (15 Points)

Four discrete random variables X, Y, U and Z are given. For these random variables, the following set of quantities have been calculated:

$$\mathcal{Q} = \{I(X;Y), I(X;U), I(U;Y), I(Y;Z), I(Y;Z|X), I(Y;X|Z)\}$$

For each of the following items, name the quantities in $\mathcal{Q}$ which you need to answer the given question. Moreover, explain how you answer the question, using the selected quantities.

(a) We know that among $\{X, Y, U\}$, there exists *exactly one* pair of *independent* random variables. How do you find this pair?

♠ **Solution:** We need $I(X;Y)$, $I(X;U)$, $I(U;Y)$.
The one which is zero determines the independent pair.
Eg. $I(X;U) = 0 \implies X, U$ independent.

(b) Assume that in Part (a), you conclude that the independent pair in $\{X, Y, U\}$ is (X, U). You are now informed that X and Z have a *deterministic* relation. This means that either $Z = f(X)$ or $X = f(Z)$ for some function f. Nevertheless, we do *not* know which relation is correct. How do you find the correct relation?

♠ **Solution:** We need $I(Y;Z|X)$ and $I(Y;X|Z)$.

- If $I(Y;Z|X) = 0 \implies Y \to X \to Z \implies Z = f(X)$
- If $I(Y;X|Z) = 0 \implies Y \to Z \to X \implies X = f(Z)$

(c) Assume that in Part (b), you find our that the correct relation is $Z = f(X)$. How do you calculate the new quantity $I(Y;X,Z)$?

♠ **Solution:** We need $I(Y;X)$.

$$\text{When } Z = f(X) \implies Y \to X \to Z \implies I(Y;Z|X) = 0$$

$$I(Y;X,Z) = I(Y;X) + \underbrace{I(Y;Z|X)}_{=0} = \boxed{I(Y;X)}$$

10.4.2 Entropy (12 Points)

Consider K *continuous* random variables

$$X_1, \ldots, X_K.$$

These random variables are *jointly independent*, however their distributions are *different*.

The *binary* random variables $U_1, \ldots, U_K$ are constructed from $X_1, \ldots, X_K$ as follows

$$U_k = \begin{cases} 0 & \text{if } |X_k| \geq T \\ 1 & \text{if } |X_k| < T \end{cases}$$

for some constant $T > 0$.

(a) Use the following fact:

$$\boxed{\text{For a Bernoulli random variable } B \text{, we have } H(B) \leq 1},$$

and give an upper-bound on $H(U_1, \ldots, U_K)$.

♠ **Solution:** Since U_k is a Bernoulli random variable, we have

$$H(U_k) \leq 1$$

Moreover, $X_1, \ldots, X_K$ are independent. Thus, $U_1, \ldots, U_K$ are also independent. This leads to

$$H(U_1, \ldots, U_K) = \sum_{k=1}^{K} H(U_k) \leq K$$

$$\Longrightarrow \boxed{\frac{1}{K} H(U_1, \ldots, U_K) \leq 1}$$

(b) Now assume that you are provided with this information: For $k = 1, \ldots, K$, we have

$$\mathcal{E}\{|X_k|\} = \frac{T}{4}$$

where $\mathcal{E}\{\cdot\}$ denotes the mathematical expectation. Show that

$$\frac{1}{K} H(U_1, \ldots, U_K) \leq H_2(0.25)$$

where $H_2(x)$ is the binary entropy function defined for $0 \leq x \leq 1$ as

$$H_2(x) = x \log \frac{1}{x} + (1 - x) \log \frac{1}{1 - x}$$

Hint: You can use *Chebyshev's inequality*.

♠ **Solution:** We know that:

$$\text{Pr}\{U_k = 1\} = \text{Pr}\{|X_k| \geq T\}$$

By Chebyshev's inequality, we can conclude that;

$$\text{Pr}\{|X_k| \geq T\} \leq \frac{\mathcal{E}\{|X_k|\}}{T} = \frac{1}{4}$$

Therefore,

$$\boxed{\text{Pr}\{U_k = 1\} \leq \frac{1}{4}}$$

Since $H_2(x)$ is an increasing function for $0 \leq x \leq \frac{1}{2}$, we have

$$H(U_k) = H_2(\text{Pr}\{U_k = 1\}) \leq H_2\left(\frac{1}{4}\right)$$

Thus,

$$H(U_1, \ldots, U_K) \leq KH_2\left(\frac{1}{4}\right) \implies \boxed{\frac{1}{K}H(U_1, \ldots, U_K) \leq H_2\left(\frac{1}{4}\right)}$$

10.4.3 Bayesian Inference (16 Points)

A network of computers consists of three computer clusters, namely *Cluster A*, *Cluster B* and *Cluster C*. Each cluster contains 80 computers. These computers use one of the following three operating systems: *Ubuntu*, *macOS*, or *Windows*. The numbers of computers in each cluster using each of these operating systems are given in the following table:

Cluster	Ubuntu	macOS	Windows
Cluster A	12	27	41
Cluster B	8	30	42
Cluster C	37	22	21

The control unit of the network has detected an error in the network caused by *one* of the clusters.
To resolve this error, the control unit tries to find the cluster in which the error has occurred. Before getting any query from the network, the control unit has the following *prior belief*:

Cluster	Chance of causing the error
Cluster A	19%
Cluster B	38%
Cluster C	43%

The control unit further gets a query from the network. The query has the following result:

It is known that in the cluster in which the error has occurred, the operating systems of the first 10 computers are as follows:

Windows, Ubuntu, Windows, macOS, Ubuntu, Windows, Windows, macOS, macOS, Windows

(a) Given the *query result* and the *prior belief* of the control unit, calculate the probability that the error has occurred in Cluster B.

♠ **Solution:** Let E be the index of the erroneous cluster. Then, we have

$$\text{Pr}\{E = B|\text{Query}\} = \frac{\text{Pr}\{\text{Query}|E = B\}\text{Pr}\{E = B\}}{\text{Pr}\{\text{Query}\}}$$

where

$$\text{Pr}\{\text{Query}\} = \sum_{e\in\{A,B,C\}} \text{Pr}\{\text{Query}|E = e\}\text{Pr}\{E = e\}$$

For the likelihood terms, we have:

$$\text{Pr}\{\text{Query}|E = A\} = \left(\frac{12}{80}\right)^2 \left(\frac{27}{80}\right)^3 \left(\frac{41}{80}\right)^5 = 0.306 \times 10^{-4}$$

$$\text{Pr}\{\text{Query}|E = B\} = \left(\frac{8}{80}\right)^2 \left(\frac{30}{80}\right)^3 \left(\frac{42}{80}\right)^5 = 0.21 \times 10^{-4}$$

$$\text{Pr}\{\text{Query}|E = C\} = \left(\frac{37}{80}\right)^2 \left(\frac{22}{80}\right)^3 \left(\frac{21}{80}\right)^5 = 0.055 \times 10^{-4}$$

Given the prior belief, we have

$$\boxed{\text{Pr}\{\text{Query}\} = 1.62 \times 10^{-5}}$$

The posterior probability is hence given by

$$\text{Pr}\{E = B|\text{Query}\} = \frac{0.21 \times 10^{-4} \times 0.38}{0.162 \times 10^{-4}} = \boxed{0.493}$$

(b) What is the best guess for the erroneous cluster basen on the *query result* and the *prior belief*?

♠ **Solution:** The other two posteriors read:

$$\text{Pr}\{E = A|\text{Query}\} = \frac{0.306 \times 10^{-4} \times 0.19}{0.162 \times 10^{-4}} = 0.359$$

$$\text{Pr}\{E = C|\text{Query}\} = \frac{0.55 \times 10^{-4} \times 0.43}{0.162 \times 10^{-4}} = 0.148$$

$\Longrightarrow$ $\text{Pr}\{E = B|\text{Query}\}$ is maximal and hence the best choice.

10.4.4 Source Coding (24 Points)

A picture consists of N pixels. Each pixel can be colored with one of eight different colors. These colors are denoted by $C_1, \dots, C_8$. A computer colors each pixel of this picture *independently* via the following procedure:

> It tosses a *uniform* coin *independently* for 7 times.
>
> - If the *first* head occurs in the *i*-th toss; then, it colors the pixel with color C_i.
> - If it observes *only tails*; then, it colors the pixel with color C_8.

After coloring all the pixels,the computer encodes this picture into M bits using a *binary source code*.

(a) The compression rate for the given source code is defined as

$$R = \frac{M}{N}$$

Calculate the minimum value of R given by the source coding theorem.

♠ **Solution:** The picture can be observed as sequence

$$X^N = X_1, \dots, X_N$$

where X_n for $n = 1, \dots, N$ are i.i.d with respect to

$$\text{Pr}(X_n = C_i) = \begin{cases} (\frac{1}{2})^i & \text{for } i = 1, \dots, 7 \\ (\frac{1}{2})^7 & \text{for } i = 8 \end{cases}$$

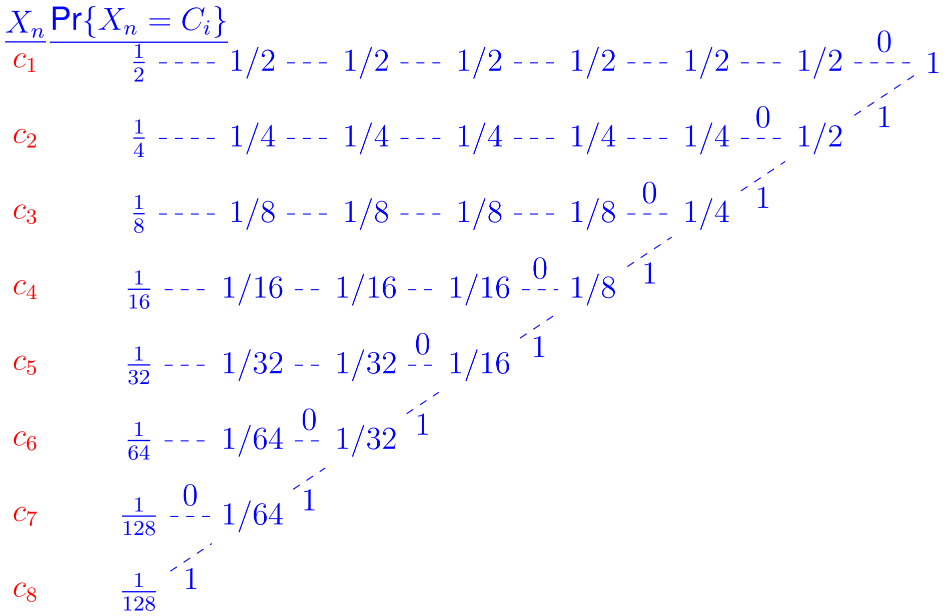
Thus,

$$\Longrightarrow \; H(X_n) = \sum_{i=1}^{7} \left(\frac{1}{2}\right)^i i + \frac{7}{2^7} = \boxed{1.9844 \text{ bits}}$$

which gives minimum compression rate.

(b) Give a binary symbol code whose average length equals to the minimum value of R.

♠ **Solution:** Since all probabilities are of the form $\frac{1}{2^i}$, Huffman code achieves the minimum rate. We hence have;


X_n Pr$\{X_n = C_i\}$
c_1 $\frac{1}{2}$ 1/2 1/2 1/2 1/2 1/2 1/2 0 1
c_2 $\frac{1}{4}$ 1/4 1/4 1/4 1/4 1/4 0 1/2 1
c_3 $\frac{1}{8}$ 1/8 1/8 1/8 1/8 0 1/4 1
c_4 $\frac{1}{16}$ 1/16 1/16 1/16 0 1/8 1
c_5 $\frac{1}{32}$ 1/32 1/32 0 1/16 1
c_6 $\frac{1}{64}$ 1/64 0 1/32 1
c_7 $\frac{1}{128}$ 0 1/64 1
c_8 $\frac{1}{128}$ 1

Codewords:

$$
\begin{aligned}
c_1 &\to 0\\
c_2 &\to 10\\
c_3 &\to 110\\
c_4 &\to 1110\\
c_5 &\to 11110\\
c_6 &\to 111110\\
c_7 &\to 1111110\\
c_8 &\to 1111111
\end{aligned}
$$

(c) To encode the picture, the computer does the following steps:

- First, it represents the color C_i with *three bits* which are the binary representation of $i - 1$. For instance, C_2 is represented by 001.
- The binary representation of the colors is then encoded via the *basic Lempel-Ziv algorithm*.

Consider a picture with $N = 4$ pixels whose colors are

$$C_8 C_1 C_1 C_2$$

Find the *encoded* binary sequence of this picture.

♠ **Solution:** The binary representation reads:

$$c_8c_1c_1c_2 = 111000000001$$

The Lempel-Ziv algorithm hence reads:

$$\begin{tikzcd}
\lambda & 1 \arrow[d] & 11 \arrow[d] & 0 \arrow[d] & 00 \arrow[d] & 000 \arrow[d] & 001 \arrow[d] \\
 & (0,1) \arrow[d, "p=1"'] & (1,1) \arrow[d, "p=2"'] & (0,0) \arrow[d, "p=3"'] & (3,0) \arrow[d, "p=4"'] & (4,0) \arrow[d, "p=5"'] & (4,1) \arrow[d, "p=6"'] \\
 & (1) & (1,1) & (00,0) & (11,0) & (100,0) & (100,1)
\end{tikzcd}$$

Codebook

Substream	Index	$\lceil \log p \rceil$
λ	0	
1	1	0
11	2	1
0	3	2
00	4	2
000	5	3
001	6	3

$\Longrightarrow$ Encoded sequence:

$$11100011010001001 \qquad \Longrightarrow \boxed{M = 17}$$

(d) Calculate the compression rate for the source code in Part (c) and compare it with the minimum value determined in Part (a).

♠ **Solution:**

$$M = 17, N = 4 \implies R_c = \frac{17}{4} = 4.25 \gg 1.98$$

This shows that the Lempel-Ziv is not efficient for small sequences.

10.4.5 Channel Capacity (14 Points)

For an integer I, the input-output relationship of a discrete memoryless channel is given by

$$Y = X + Z \qquad (\text{mod } I)$$

where $I \geq 2$, and

- X is an integer chosen from the alphabet

$$\mathcal{A}_X = \{1, \ldots, 2I\}$$

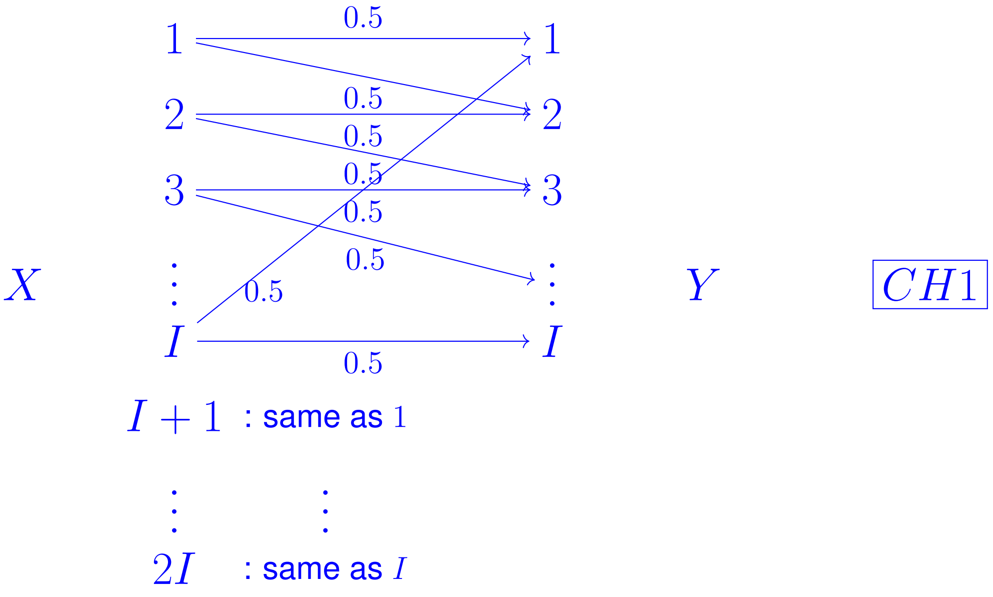
- Z is noise which is a uniform Bernoulli random variable. This means that $\mathcal{A}_Z = \{0, 1\}$,

$$\text{Pr}\{Z = 0\} = \text{Pr}\{Z = 1\} = 0.5$$

- The sum indicates a *modular* addition. For example, $5 + 2 \pmod 6 = 1$

(a) Calculate the capacity of this channel.

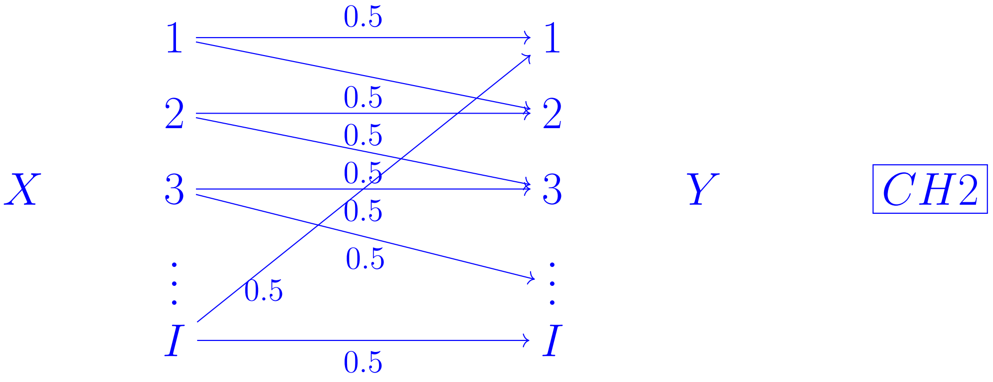
♠ **Solution:** The diagram of this channel is as follows:



From the diagram, it is clear that $I + j$ and j for $j = 1, \ldots, I$ are not distinguishable. Hence, we can conclude that in capacity-achieving distribution:

$$\text{Pr}\{I + j\} = 0, \quad \text{for } j = 1, \ldots, I.$$

In this case, the channel reduces to



which is the erroneous type-writer in the textbook. For this channel, we have

$$H(Y|X) = H_2(0.5) = 1 \text{ bit}$$

$$\Longrightarrow \boxed{C = \log I - 1 = \log \frac{I}{2}}$$

(b) Find an input distribution which achieves the channel capacity.

♠ **Solution:** The channel in $\boxed{CH2}$ is a symmetric channel whose capacity is given by uniform input. Thus,

$$\begin{cases} \text{Pr}\{X = j\} = \frac{1}{I} & \text{for } j = 1, \ldots, I \\ \text{and} & \\ \text{Pr}\{X = I + j\} = 0 & \text{otherwise.} \end{cases}$$

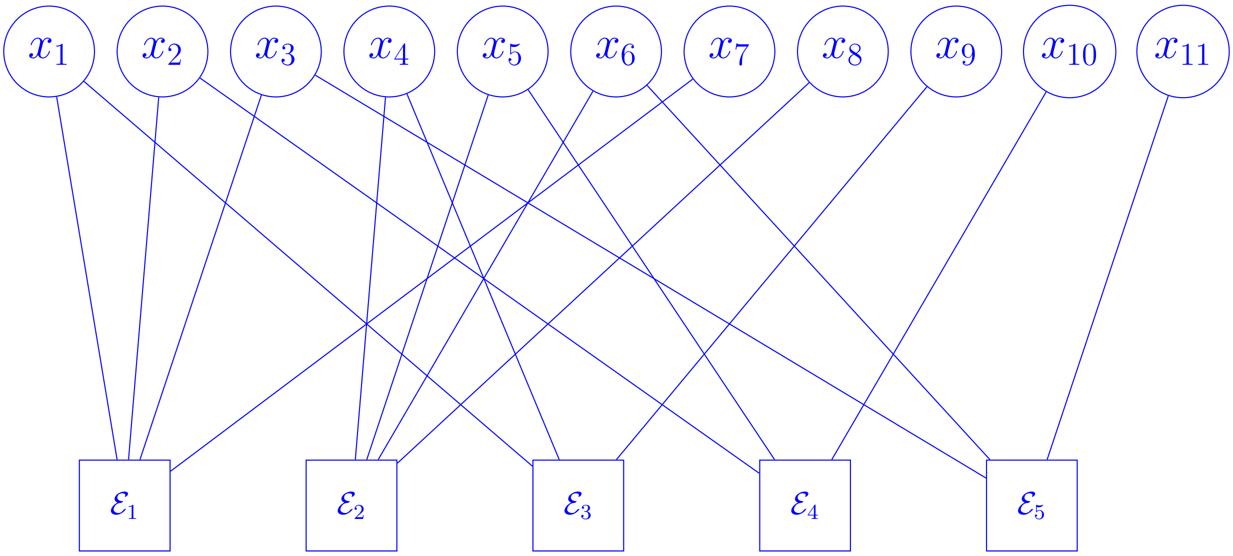
10.4.6 Channel Codes (19 Points)

The parity check matrix of an LDPC code is given below

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

(a) Sketch the factor graph which corresponds to the parity check equations of this code. Explicitly specify the code symbol corresponding to each variable node and the parity check equation corresponding to each factor node.

♠ **Solution:** $\mathbf{H}_{5\times11}$ corresponds to a parity-check code with 11 code symbols and 5 parity equations. Thus, the factor graph is:



(b) Calculate the *total number* of codewords in this code.

♠ **Solution:** The code length is $N = 11$ and the number of parity check equations is $P = 5$. Thus, the number of information symbols is

$$K = N - P = 11 - 5 = 6$$

$$\Longrightarrow \ \# \text{ of codewords} = 2^6 = \boxed{64}$$

(c) The codeword **c** is passed through a binary erasure channel (BEC). The output of the channel is

$$\mathbf{y} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ \epsilon \\ \epsilon \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

where ϵ is the erasure symbol. Find the codeword **c** sent over this channel.

♠ **Solution:** We need to correct x_4 and x_5:

- From parity check equation $\mathcal{E}_3$,we have

$$x_1 = 1, \ x_9 = 0$$

$$x_1 \oplus x_4 \oplus x_9 = 0 \implies 1 \oplus x_4 = 0 \implies \boxed{x_4 = 1}$$

- From parity check $\mathcal{E}_4$, we have

$$x_2 = 0,\ x_{10} = 0$$

$$x_2 \oplus x_5 \oplus x_{10} = 0 \implies x_5 \oplus 0 \implies \boxed{x_5 = 0}$$

(d) The generator matrix of this code is given by

$$\mathbf{G}^T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

This means that codeword **c** is constructed from its corresponding information word **b** as

$$\mathbf{c} = \mathbf{G}^T\mathbf{b}.$$

Decode the codeword **c** detected from **y** in Part (c); namely, find its corresponding information word **b**.

♠ **Solution:** From $\mathbf{G}^T$, it is seen that:

$$\boxed{c_1 = b_1, \ldots, c_6 = b_6}$$

and

$$\begin{aligned} c_7 &= b_1 \oplus b_2 \oplus b_3 \qquad & c_{10} &= b_2 \oplus b_5 \\ c_8 &= b_4 \oplus b_5 \oplus b_6 \qquad & c_{11} &= b_3 \oplus b_6 \\ c_9 &= b_1 \oplus b_4 \end{aligned}$$

Hence,

$$\mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

10.5 Winter Semester 2019-2020 Exam

Exam Date: February 11th, 2020
6 Questions with total of 100 Points.
Exam Duration: 90 Minutes

10.5.1 Short Questions (13 Points)

Answer the following questions:

(a) Three *dependent* random variables X, Y and Z make a *Markov chain*. It is known that these three random variables satisfy the following identity

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

Find the *Markov chain*.

♠ **Solution:** We know by the chain rule

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y)$$

The identity hence concludes that

$$I(X;Z|Y) = 0$$

Which means

$$X \to Y \to Z$$

(b) Let X be a *continuous* random variable with probability density function (PDF) $f(x)$ whose differential entropy is denoted by $h(X)$. Show that for any real *constant* μ, we have

$$h(X+\mu) = h(X)$$

♠ **Solution:** Let $Y = X + \mu$. For Y, we have

$$F_Y(y) = \text{Pr}\{Y \leq y\} = \text{Pr}\{X + \mu \leq y\} = F_X(y-\mu)$$

$$\implies f_Y(y) = \frac{d}{dy} F_X(y-\mu) = f(y-\mu)$$

$$\implies h(Y) = h(X+\mu) = \int_{-\infty}^{\infty} f_Y(y) \log f_Y(y)\, dy = -\int_{-\infty}^{\infty} f(y-\mu) \log f(y-\mu)\, dy$$

By exchanging $z = y - \mu$, we have

$$h(X+\mu) = -\int_{-\infty}^{\infty} f(z) \log f(z)\, dz = h(x).$$

(c) Consider a discrete memoryless channel (DMC). Let Y_1 and Y_2 be the output symbols corresponding to input symbols X_1 and X_2, respectively. Given that X_1 and X_2 are *independent*, what is the value of $I(Y_1; Y_2)$?

♠ **Solution:**

$$\left.\begin{array}{l} Y_1 \text{ only depends on } X_1 \\ Y_2 \text{ only depends on } X_2 \end{array}\right\} \Longrightarrow (X_1 \text{ and } X_2 \text{ are independent})$$

$$(X_1 \text{ and } X_2 \text{ are independent}) \Longrightarrow Y_1 \text{ and } Y_2 \text{ are also independent.}$$

$$\Longrightarrow \boxed{I(Y_1; Y_2) = 0}$$

10.5.2 Typicality (11 Points)

Consider an *independent and identically distributed (i.i.d)* binary sequence

$$X^N = X_1, \ldots, X_N$$

which is distributed according to a Bernoulli distribution with

$$\text{Pr}\{X_n = 1\} = 1 - \text{Pr}\{X_n = 0\} = 0.3$$

for $n = 1, \ldots, N$.

Let N_0 and N_1 denote the number of symbols in X^N which are 0 and 1, respectively. This means that

$$N_0 + N_1 = N$$

(a) Write a definition of the β*-typical set* T_{N_β} *only* in terms of N_0, N_1 and N.

♠ **Solution:** From Eq. (4.29) in page 80, we have

$$T_{N_\beta} = \{X^N \in \{0,1\}^N : \left| \frac{1}{N} \log_2 \frac{1}{p(X^N)} - H \right| < \beta\}$$

Following the fact that X^N is i.i.d, we have

$$\begin{aligned} p(X^N) = \prod_{n=1}^{N} p(X_n) &= (\text{Pr}\{X_n = 0\})^{N_0} (\text{Pr}\{X_n = 1\})^{N_1} \\ &= 0.7^{N_0}\, 0.3^{N_1} \end{aligned}$$

Thus,

$$\begin{aligned} \Longrightarrow \frac{1}{N} \log_2 \frac{1}{p(X^N)} &= \frac{1}{N} \log 0.7^{-N_0} 0.3^{-N_1} \\ &= \frac{1}{N} \left(\log 0.7^{-N_0} + \log 0.3^{-N_1} \right) \\ &= \frac{-N_0}{N} \log 0.7 + \frac{-N_1}{N} \log 0.3 \end{aligned}$$

On the other hand:

$$H = -0.3\log 0.3 - 0.7\log 0.7$$

$$\implies \left|\frac{1}{N}\log_2\frac{1}{p(X^N)} - H\right| = \left|\left(\frac{N_0}{N} - 0.7\right)\log_2 0.7 + \left(\frac{N_1}{N} - 0.3\right)\log_2 0.3\right|$$

Therefore:

$$\boxed{T_{N_\beta} = \{X^N \in \{0,1\}^N : \left|\left(\frac{N_0}{N} - 0.7\right)\log_2 0.7 + \left(\frac{N_1}{N} - 0.3\right)\log_2 0.3\right| < \beta\}}$$

(b) Write two different sequences which belong to the β-typical set T_{N_β} with $N = 10$ and $\beta = 0.001$.

♠ **Solution:** Since $\beta = 0.001$, we write

$$X_1^N = 0000000111$$
$$X_2^N = 0101010000$$

for both sequences:

$$\left|\frac{1}{N}\log_2\frac{1}{p(X^N)} - H\right| = 0 < 0.001$$

Any other sequences with 7 zeros and 3 ones are also solutions.

(c) Consider on of the sequences that you wrote in Part (b). Assume that this sequence is passed through a binary symmetric channel (BSC) with flipping probability $f = 0.2$. Let

$$Y^{10} = Y_1, \ldots, Y_{10}$$

denote the output sequence. Write a realization of Y^{10} which is *jointly typical* with the input sequence to tolerance $\beta = 0.001$.

♠ **Solution:** A jointly typical sequence would have only <u>two</u> bits flipped. Thus, considering

$$X_1^N = 0000000111$$

The sequence

$$Y_1^10 = 1001000111$$

is jointly typical with X_1^N with tolerance $0 < 0.001$. Any other Y^N which have only two bits different with X_1^N is also a solution.

Part (c) has a technical issue. Therefore, it was considered as a bonus point in the final correction.

10.5.3 Bayesian Inference (16 Points)

You are asked to design a *classifier machine*. This machine gets an image as input and decides whether this image contains *Object A*, *Object B* or *Object C*. Each image contains exactly one of these three objects. For classification, the machine performs the following processes on an input image:

- It divides the image into 4 pixels, namely *Pixel 1*, *Pixel 2*, *Pixel 3* and *Pixel 4*.
- Based on the light intensity in each pixel, the machine quantifies the pixel with 0 or 1.
- Given the values of all pixels, the machine identifies *only one* of the objects in the input image.

To make this machine perform well, we first train it with 100 sample images. The statistics of these sample images are as follows:

- Among this 100 sample images, 48 images contain *Object A*, 27 images contain *Object B* and 25 images contain *Object C*.
- In the following table, the number of images whose pixel value is 1 is shown for each object and each pixel.

	Pixel 1	Pixel 2	Pixel 3	Pixel 4
Object A	18	22	39	31
Object B	20	22	2	7
Object C	1	16	2	21

For example,the table indicates that among 48 sample images containing *Object A*, *Pixel 1* is quantified with 1 in 18 images, and in the remaining 30 images this pixel is quantified with 0.

The classifier now operates on a new image whose pixel values are given in the following table:

Pixel 1	Pixel 2	Pixel 3	Pixel 4
1	1	0	0

(a) Interpret the classification of the new image as a *Bayesian inference* problem. Specify the *prior*, *likelihood* and *posterior*.

♠ **Solution:** Let X denote the object in the new image. We want to find

$$\Pr\{X = A|Data\}, \ \Pr\{X = B|Data\}, \ \Pr\{X = C|Data\}$$

where "Data" is the given data.
We then select the object whose probability is maximum. It is hence a Bayesian inference problem.

Posteriors $= \{\text{Pr}\{X = A|Data\},\ \text{Pr}\{X = B|Data\},\ \text{Pr}\{X = C|Data\}\}$

Priors $= \{\text{Pr}\{X = A\},\ \text{Pr}\{X = B\},\ \text{Pr}\{X = C\}\}$

Likelihoods $= \{\text{Pr}\{Data|X = A\},\ \text{Pr}\{Data|X = B\},$
$\text{Pr}\{Data|X = C\}\}$

(b) Given the training data, find the object which is identified in the new image, when the classification is done by the Bayesian inference problem specified in Part (a).

♠ **Solution:** We start with calculating priors and likelihoods:

$$\begin{aligned}\text{Pr}\{X = A\} &= \frac{\#\text{ of pics with object } A}{\#\text{ of total training pictures}}\\ &= \frac{48}{100} = 0.48\end{aligned}$$

Similarly,

$$\implies \text{Pr}\{X = B\} = 0.27 \qquad \text{Pr}\{X = C\} = 0.25.$$

For likelihoods, we also have:

$$\begin{aligned}\text{Pr}\{Data|X = A\} = \text{Pr}\{Pixel\ 1 = 1|X = A\}.\text{Pr}\{Pixel\ 2 = 1|X = A\}\\ .\text{Pr}\{Pixel\ 3 = 0|X = A\}.\text{Pr}\{Pixel\ 4 = 0|X = A\}\end{aligned}$$

From the training data we have:

$$\begin{aligned}\text{Pr}\{Pixel\ 1 = 1|X = A\} &= \frac{18}{48}\\ \text{Pr}\{Pixel\ 2 = 1|X = A\} &= \frac{22}{48}\\ \text{Pr}\{Pixel\ 3 = 0|X = A\} &= 1 - \frac{39}{48} = \frac{9}{48}\\ \text{Pr}\{Pixel\ 4 = 0|X = A\} &= 1 - \frac{31}{48} = \frac{17}{48}\end{aligned}$$

$$\implies \text{Pr}\{Data|X = A\} = \frac{18 \times 22 \times 9 \times 17}{48^4} = 0.0114$$

Similarly,

$$\text{Pr}\{Data|X = B\} = \frac{20 \times 22 \times 25 \times 20}{27^4} = 0.4140$$

$$\text{Pr}\{Data|X = C\} = \frac{1 \times 16 \times 23 \times 4}{25^4} = 0.0038$$

Thus, marginals read:

$$\begin{aligned}\text{Pr}\{Data\} &= \sum_{x\in\{A,B,C\}} \text{Pr}\{X=x\}\text{Pr}\{Data|X=x\}\\ &= 0.48\times 0.0114 + 0.27\times 0.414 + 0.25\times 0.0038\\ &= 0.1182\end{aligned}$$

Thus, posteriors are,

$$\begin{aligned}\text{Pr}\{X=A|Data\} &= \frac{\text{Pr}\{X=A\}\text{Pr}\{Data|X=A\}}{\text{Pr}\{Data\}} = 0.0463\\ \text{Pr}\{X=B|Data\} &= \frac{0.4140\times 0.27}{0.1182} = 0.9457\\ \text{Pr}\{X=C|Data\} &= 0.0082\end{aligned}$$

Thus, Bayesian inference chooses $\boxed{X=B}$.

10.5.4 Source Coding (20 Points)

The source sequence of length $2N$

$$X^{2N} = X_1, \dots, X_{2N}$$

is generated as below:

X_1 and X_2 are specified as follows:

- A *uniform* coin is tossed
  - If a head occurs; then, we set $X_1 = A$ and $X_2 = A$.
  - If we observe a tail; then, we set $X_1 = B$ and specify X_2 by tossing the uniform coin *independently* once again:
    - If a head occurs, we set $X_2 = A$.
    - If a tail occurs, we set $X_2 = B$.

The next pairs, i.e., $(X_3, X_4), (X_5, X_6), \dots, (X_{2N-1}, X_{2N})$ are generated by repeating the above procedure *independently*.

(a) Determine the normalized entropy $\bar{H}$ which is defined as

$$\bar{H} = \frac{1}{2N} H(X_1, \dots, X_{2N})$$

♠ **Solution:**

$$H(X_1,\dots,H_{2N}) = \sum_{n=1}^{N} H(X_{2n-1},X_{2n}) \\ = NH(X_1,X_2)$$

$$X_1,X_2 = \begin{cases} A,A & Pr=\frac{1}{2} \\ B,A & Pr=\frac{1}{4} \\ B,B & Pr=\frac{1}{4} \end{cases} \implies H(X_1,X_2) = \frac{1}{2}+\frac{1}{4}\times 2+\frac{1}{4}\times 2 = \boxed{\frac{3}{2}}$$

$$\implies H(X_1,\dots,H_{2N}) = \frac{3}{2}N \implies \bar{H} = \frac{1}{2N}\cdot\frac{3N}{2} = \boxed{\frac{3}{4}}$$

(b) Assume $C(X^{2N})$ denotes the binary codeword of sequence X^{2N} determined via the code C. Let $L(C,X^{2N})$ denote the length of this codeword, i.e., the number of bits in $C(X^{2N})$. The average length of C is defined as

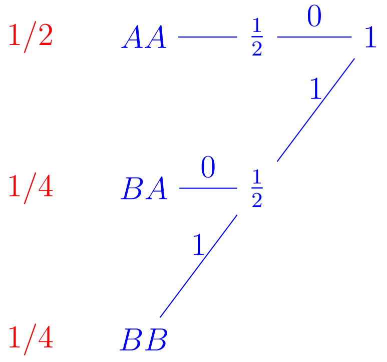
$$\bar{L} = \frac{1}{2N}\mathcal{E}\{L(C,X^{2N})\}$$

Give a binary encoding for which we have $\bar{L} = \bar{H}$ for any N.

♠ **Solution:** Such a binary code is simply given using Huffman algorithm. Define:

$$V_n = (X_{2n-1},X_{2n}) \implies V_n = \begin{cases} AA & p=\frac{1}{2} \\ BA & p=\frac{1}{4} \\ BB & p=\frac{1}{4} \end{cases}$$

Thus, a Huffman code for V_n is



$$C(AA) = 0 \\ C(BA) = 10 \\ C(BB) = 11$$

The code for X^{2N} is simply given by repeating same thing. It means:

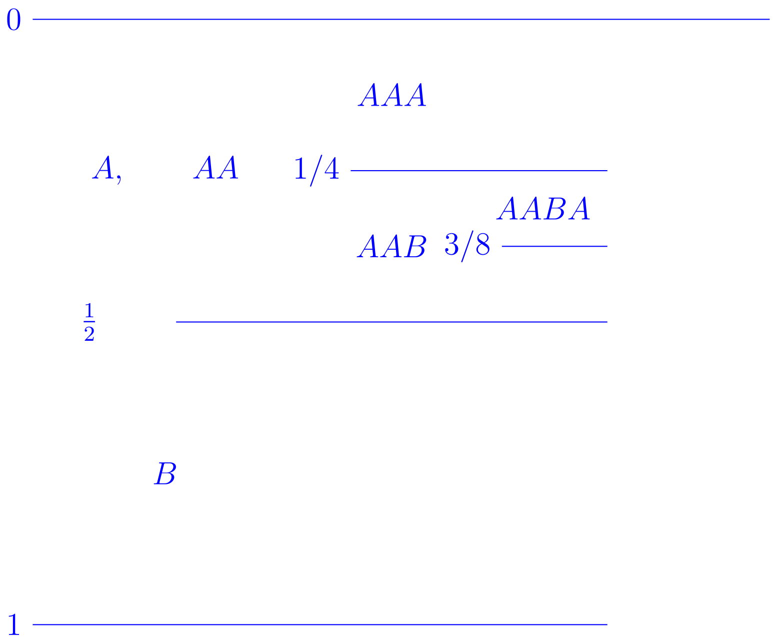
$$C(X^{2N}) = C(X_1,X_2)C(X_3,X_4),\dots,C(X_{2N-1},X_{2N})$$

(c) Give a binary codeword for sequence

$$AABA$$

using the *arithmetic coding* algorithm.

♠ **Solution:** We first find the interval:



$$\text{Pr}\{A\}\frac{1}{2}, \quad \text{Pr}\{A|A\} = 1$$

$$\text{Pr}\{B|AA\} = \frac{1}{2}, \quad \text{Pr}\{A|AAB\} = \frac{1}{2}$$

The solution is then 010 for this code. Given the agreement in the book, 0100 or 0101 is also correct.

Same as Huffman.

10.5.5 Channel Capacity (16 Points)

In a channel, the input symbol X is related to the output symbol Y as follows

$$Y = f(X) + Z$$

Here,

- X is a *real* input.

- Z is uniformly distributed on the interval $[0, 1.5)$,
- The function $f(\cdot)$ is defined as

$$f(x) = \begin{cases} 1 & \text{if } |x| \geq 2 \\ 0 & \text{if } |x| < 2 \end{cases}.$$

(a) Calculate the capacity of this channel.

♠ **Solution:** Let us define: $V = f(x)$.
V is a binary variable. The capacity of the channel is hence given by

$$C = \max_{p(v)} I(Y;V)$$

Noting that $Y = V + Z$, we have

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

where

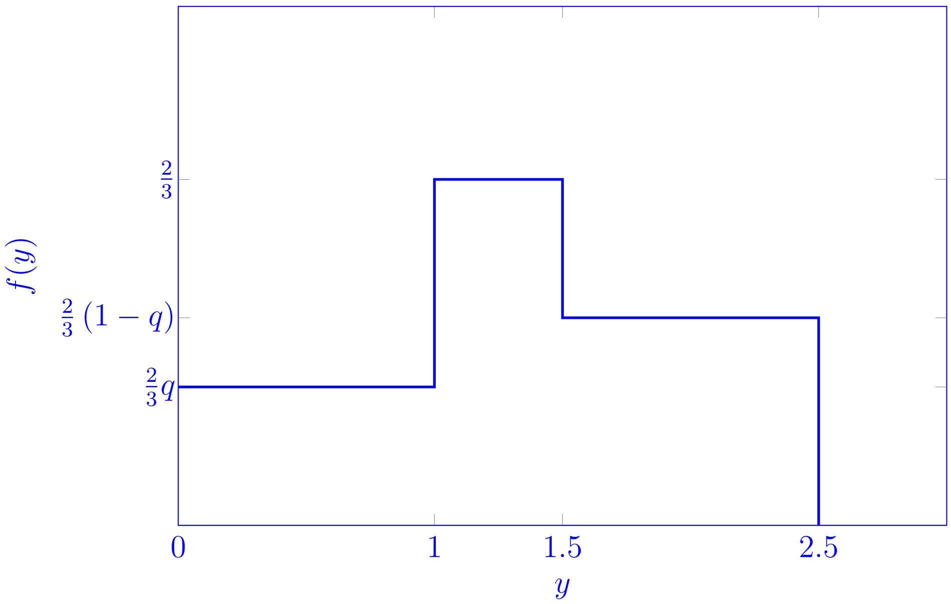
$$h(Y|V) = h(Z) = \log 1.5$$

Moreover, assuming an arbitrary distribution for V, we have

$$V = \begin{cases} 0 & q \\ 1 & 1-q \end{cases}.$$

Hence, $f(y)$ reads:

$$\begin{aligned} f(y) &= \sum_v f(y|v)p(v) \\ &= qf(y|v=0) + (1-q)f(y|v=1) \end{aligned}$$



$$\begin{aligned} h(Y) = -\int_{\infty}^{\infty} f(y)\log f(y)\, dy &= q\log q \times 1 + (1-q)\log(\frac{1}{1-q}) \times 1 \\ &= \tfrac{2}{3}H_2(q) + \log 1.5 \end{aligned}$$

$$\Longrightarrow\ I(V;Y) = \tfrac{2}{3}H_2(q)$$

And thus,

$$\Longrightarrow\ C = \max_{q\in[0,0.5]} I(V;Y) = \boxed{\tfrac{2}{3}}$$

(b) Find two input distributions which achieve the channel capacity.

♠ **Solution:** To achieve the capacity, V should be uniform. Thus, any distribution of X which leads to uniform V is capacity achieving. We hence have;

$$\begin{aligned}\text{Pr}\{V=0\} &= \text{Pr}\{|X|<2\} = \int_{-2}^{2} f(x)\,dx\\ \text{Pr}\{V=1\} &= 1-\text{Pr}\{V=0\}.\end{aligned}$$

Therefore, any $f(x)$ which satisfy

$$\boxed{\int_{-2}^{2} f(x)\,dx = \frac{1}{2}}$$

is an answer. For example:

$$\begin{aligned}X &\sim \text{Uniform}[-4,4)\\ X &\sim \text{Uniform}[-2,6)\end{aligned}$$

10.5.6 Channel Codes (24 Points)

A channel code with code-length $N = 18$ has the following codewords:

$$\mathbf{c}_1 = \begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}^T$$

$$\mathbf{c}_2 = \begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1\end{bmatrix}^T$$

$$\mathbf{c}_3 = \begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}^T$$

$$\mathbf{c}_4 = \begin{bmatrix}0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1\end{bmatrix}^T$$

$$\mathbf{c}_5 = \begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}^T$$

$$\mathbf{c}_6 = \begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1\end{bmatrix}^T$$

$$\mathbf{c}_7 = \begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0\end{bmatrix}^T$$

$$\mathbf{c}_8 = \begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1\end{bmatrix}^T$$

This code encodes *binary* information words.

(a) What is the length of information words in this code?

♠ **Solution:** We have 8 codewords. Thus;

$$2^K = 8 \implies \boxed{K = 3}$$

(b) Find a *generator matrix* **G** for this code, such that for each information word $\mathbf{b}_i$ and its corresponding codeword $\mathbf{c}_i$, we have

$$\mathbf{c}_i = \mathbf{G}^T \mathbf{b}_i.$$

♠ **Solution:** First we can infer that:

$$\mathbf{b}_1 = \begin{bmatrix}0 & 0 & 0\end{bmatrix}^T$$

$$\mathbf{b}_2 = \begin{bmatrix}0 & 0 & 1\end{bmatrix}^T$$

$$\mathbf{b}_3 = \begin{bmatrix}0 & 1 & 0\end{bmatrix}^T$$

$$\mathbf{b}_4 = \begin{bmatrix}0 & 1 & 1\end{bmatrix}^T$$

$$\mathbf{b}_5 = \begin{bmatrix}1 & 0 & 0\end{bmatrix}^T$$

$$\mathbf{b}_6 = \begin{bmatrix}1 & 0 & 1\end{bmatrix}^T$$

$$\mathbf{b}_1 = \begin{bmatrix}1 & 1 & 0\end{bmatrix}^T$$

$$\mathbf{b}_1 = \begin{bmatrix}1 & 1 & 1\end{bmatrix}^T$$

And then the generator matrix $\mathbf{G}^T$ reads:

$$\mathbf{G}^T = \left[\begin{array}{ccc}
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
\hdashline
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
\hdashline
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1
\end{array}\right]_{18\times 3}$$

$$\mathbf{G} = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}$$

(c) Find a parity-check matrix **H** for this code, such that

$$\mathbf{Hc}_i = 0$$

for $i = 1, \ldots, 8$.

♠ **Solution:** We first make the code symmetric:

$$\begin{bmatrix} c_1 \\ \vdots \\ \vdots \\ c_{18} \end{bmatrix} = \mathbf{G}^T\mathbf{b} \Longrightarrow \tilde{c} = \begin{bmatrix} c_1 \\ c_7 \\ c_{13} \\ c_2 \\ \vdots \\ c_6 \\ c_8 \\ \vdots \\ c_{12} \\ c_{14} \\ \vdots \\ c_{18} \end{bmatrix} = \tilde{\mathbf{G}}^T\mathbf{b}$$

$$\tilde{\mathbf{G}}^T = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hdashline 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{array}\right]$$

$$\tilde{\mathbf{G}}^T = \left[\begin{array}{c} I_3 \\ \hdashline P \end{array}\right]$$

$$\tilde{\mathbf{G}} = \left[\begin{array}{c:c} I_3 & P^T \end{array}\right]$$

$$\tilde{\mathbf{H}} = \left[\begin{array}{c:c} P & I_{15} \end{array}\right]$$

Then **H** is constructed by shifting row 2 to row 7 and row 3 to row 13.

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & e_1 \\ 1 & 0 & 0 & e_4 \\ 1 & 0 & 0 & e_5 \\ 1 & 0 & 0 & e_2 \\ 0 & 1 & 0 & e_6 \\ 0 & 1 & 0 & e_7 \\ 0 & 1 & 0 & e_8 \\ 0 & 1 & 0 & e_9 \\ 0 & 1 & 0 & e_{10} \\ 1 & 0 & 0 & e_3 \\ 0 & 0 & 1 & e_{11} \\ 0 & 0 & 1 & e_{12} \\ 0 & 0 & 1 & e_{13} \\ 0 & 0 & 1 & e_{14} \\ 0 & 0 & 1 & e_{15} \end{bmatrix}$$

(d) Assume that this code is used to communicate over a binary symmetric channel (BSC) with flipping probability

$$f = \frac{2}{9}$$

The receiver, receives the following word

$$\mathbf{y} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix}^T$$

It then uses the maximum likelihood (ML) decoder to recover the transmitter codeword. Find the decoded codeword.

♠ **Solution:** By maximum likelihood, we find

$$\hat{\mathbf{c}} = \operatorname*{argmax}_{i=1,\ldots,8} \ p(\mathbf{y}|\mathbf{c}_i)$$

for BSC, we have

$$p(\mathbf{y}|\mathbf{c}_i) = \prod_{j=1}^{18} p(y_j|c_{i,j}) = \left(\frac{2}{9}\right)^{N_{f,i}} \left(\frac{7}{9}\right)^{N-N_{f,i}}$$

where

$$N_{f,i} = \text{ \# of bits of } \mathbf{y} \text{ differs from } \mathbf{c}_i$$

$$\Longrightarrow \ \hat{\mathbf{c}} = \operatorname*{argmin}_{i=1,\ldots,8} \ N_{f,i} \ \Longrightarrow \ \boxed{\hat{\mathbf{c}} = \mathbf{c}_1}$$

(e) Assume that in Part (d), the receiver exchanges the ML decoder with a *typical set* decoder.
Find the decoded codeword in this case.

♠ **Solution:** We should find $\hat{\mathbf{c}}$, such that $(\hat{\mathbf{c}}, \mathbf{y})$ are jointly typical. Since, $f = \frac{2}{9}$, we have

$$N \cdot f = 18 \cdot \frac{2}{9} = \boxed{4}$$

So, $\hat{\mathbf{c}}$ are jointly typical if they are different only in 4 bits. This concludes that

$$\boxed{\hat{\mathbf{c}} = \mathbf{c}_1}$$

(f) Repeat Part (e) for the case in which the receiver exchanges the ML detector with a *bounded distance* decoder. Find the decoded codeword in this case.

♠ **Solution:** The bounded distance decoder finds $\hat{\mathbf{c}}$ such that

$$d_H(\hat{\mathbf{c}}, \mathbf{y})$$

is minimum. This means that

$$\hat{\mathbf{c}} = \underset{i=1,\ldots,8}{\operatorname{argmin}}\ d_H(\mathbf{c}_i, \mathbf{y}) = \operatorname{argmin}\ N_{f,i}$$

which is same as ML. Thus,

$$\boxed{\hat{\mathbf{c}} = \mathbf{c}_1}$$

10.6 Summer Semester 2020 Exam

Exam Date: August 11th, 2020
5 Questions with total of 100 Points.
Exam Duration: 90 Minutes

10.6.1 Basic Questions (18 Points)

Consider discrete random variables X and Y with alphabets $\mathcal{A}_X$ and $\mathcal{A}_Y$, respectively. Let $P_{X,Y}$ denote the joint distribution, and P_X and P_Y represent the marginal distribution of X and Y, respectively. Moreover, $P_{X,u}$ is the uniform distribution over $\mathcal{A}_X$.

Show the validity of the following identities:

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

♠ **Solution:**

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} P_{X,Y}(x,y)\log P_{X,Y}(x,y) \\ &= -\sum_{x,y} P_{Y|X}(y|x)P_X(x)\left[\log P_{Y|X}(y|x) + \log P_X(x)\right] \\ &= -\sum_{x} P_X(x)\log P_X(x)\sum_{Y} P_{Y|X}(y|x) - \sum_{x,y} P_{X,Y}(x,y)\log P_{Y|X}(y|x) \\ &= H(X) + H(Y|X) \end{aligned}$$

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

♠ **Solution:**

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

(c) $I(X;X) = H(X)$

♠ **Solution:**

$$I(X;X) = H(X) - \underbrace{H(X|X)}_{=0} = H(X)$$

(d) $D_{KL}(P_X||P_{X,u}) = \log|\mathcal{A}_X| - H(X)$

♠ **Solution:**

$$\begin{aligned}
P_{X,u}(x) &= \frac{1}{|\mathcal{A}_X|} \\
D_{KL}(P_X||P_{X,u}) &= \sum_{x\in\mathcal{A}_X} P_X(x)\log\frac{P_X(x)}{P_{X,u}(x)} \\
&= \sum_{x\in\mathcal{A}_X} P_X(x)\log P_X(x) + \sum_{x\in\mathcal{A}_X} P_X(x)\log|\mathcal{A}_X| \\
&= -H(X) + \log\mathcal{A}_\mathcal{X}
\end{aligned}$$

10.6.2 Information Content and Entropy (22 Points)

A murder case is being investigated. The investigators have found DNA traces of the murderer at the crime scene. There is for sure only a single murderer and you are in charge of identifying him among 16 available suspects. The investigators are sure that the murderer is one of these suspects. Due to budget limitations, you are asked to use only four DNA tests.

The DNA tests are performed by means of saliva samples. This means that you have the possibility of *group-wise* testing. A group-wise test is positive, if the murderer is in the tested group.

Let us denote the results of the four DNA tests by T_i for $i \in \{1,2,3,4\}$, where $T_i \in \{$positiv, negativ$\}$, and the probability distribution of T_i by $P_{T_i}(t_i)$, i.e., $P_{T_i} = \Pr\{T_i = t_i\}$.

We further define the random variable $M \in \{1,2,\ldots,16\}$ whose probability distribution $P_{M|\mathcal{H}}(i) = \Pr\{M = i|\mathcal{H}\}$, for a given hypothesis $\mathcal{H}$, gives the probability of suspect i being the murderer when the prior information provided by the hypothesis $\mathcal{H}$ is considered.

Now assume that you have no prior information on the suspects and their connection to the murder case. We refer to this hypothesis as hypothesis $\mathcal{H}_1$.

(a) Give probability distribution $P_{M|\mathcal{H}}(i)$. Describe a testing strategy by which you could surely identify the murderer using only four DNA tests.

♠ **Solution:**

$$P_{M|H_1}(i) = \frac{1}{16}\,, \qquad i = 1,\ldots,16$$

Test strategy: Binary search; divide group of suspects into two halves and test one of them; if positive, divide this group further into two halves; if negative, divide other group into two halves and test one of them;...

(b) For hypothesis $\mathcal{H}_1$, determine the probability distribution P_{T_i} considering your proposed testing strategy in Part (a).

♠ **Solution:**

$$P_{T_i}(t_i) = \begin{cases} \frac{1}{2}, & t_i = \text{positive} \\ \frac{1}{2}, & t_i = \text{negative} \end{cases}, \qquad i = 1, \ldots, 4$$

(c) Consider you proposed strategy in Part (a), and assume that only the first and third tests are positive, i.e., $[t_1, t_2, t_3, t_4] = [\text{positiv}, \text{negativ}, \text{positiv}, \text{negativ}]$. Specify the *information content* of each test result in bits. Determine how many bits of information you obtain in total by performing the four DNA tests. Compare your answer to the entropy of M when it is distributed with distribution $P_{M|\mathcal{H}_1}(i)$.

♠ **Solution:**

$$h(t_i) = -\log P_{T_1}(t_1) = 1 \text{ Bit} = h_2(t_2) = h_3(t_3) = h_4(t_4)$$

$\Longrightarrow$ In total, we obtain 4 bits of information

$$H(M) = \log |\{1, \ldots, 16\}| = 4 \text{ bits} = \sum_{i=1}^{4} h(t_i)$$

Now consider an alternative condition: Instead of having only four DNA tests, you are permitted to use an *arbitrary number* of DNA tests. Before you start testing the suspects, you are further informed that on the crime night, four of the suspects have been seen close to the crime scene. You hence postulate that the murderer is one of these four suspects with probability 88% and the probability of the murderer being among the other suspects is only 12%. We refer to this hypothesis as hypothesis $\mathcal{H}_2$.

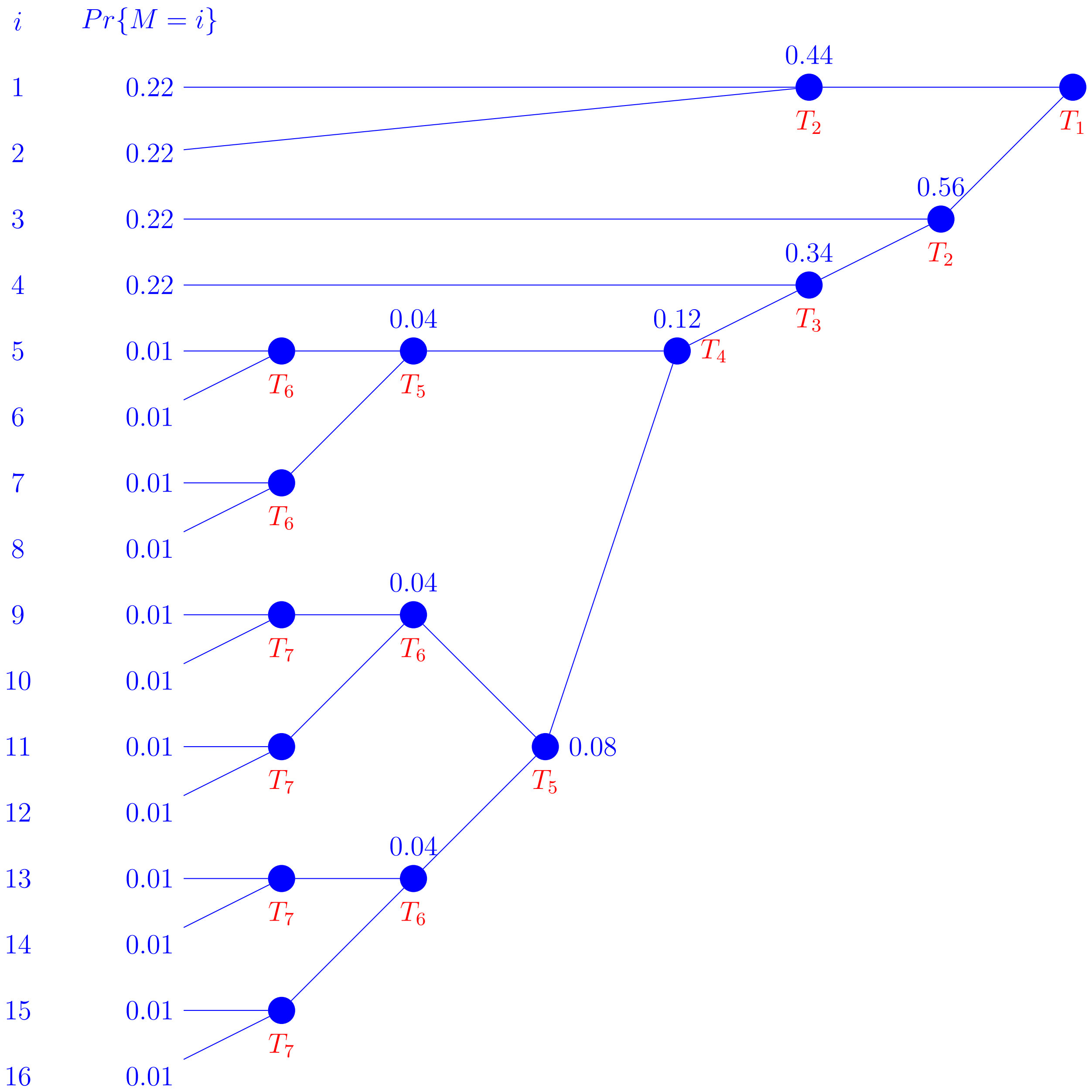
(d) What is the entropy of random variable M under distribution $P_{M|\mathcal{H}_2}(i)$?

♠ **Solution:**

$$H(M) = -4 \cdot 0.22 \cdot \log 0.22 - 12 \cdot 0.01 \cdot \log 0.01 \approx 2.72 \quad \text{bits}$$

(e) We define an *optimal* testing strategy as a strategy which requires minimum average number of DNA tests to identify the murderer. Specify an *optimal* testing strategy under hypothesis $\mathcal{H}_2$. What is the minimum number of DNA tests required by this strategy?

♠ **Solution:** Assume without loss of generality $\{1, \ldots, 4\}$ are the main suspects. The optimal testing strategy is given by the Huffman tree.



$\Longrightarrow$ Minimum number of tests is 2.

(f) Under which conditions the murderer is identified with the minimum number of DNA tests? Specify further the maximum number of required test by the given strategy in Part (e).

♠ **Solution:** Two tests are sufficient if the murderer is one of the main suspects tested in the first test against the rest or the main suspect testes in the second test against the rest.

The maximum number of required tests is 7.

(g) Assume that a series of murders with the same number of suspects, i.e., 16 suspects each, are happening in the city. You use your strategy in Part (e) to find the murderer in each case.
How many tests do you expect to need to identify the murderer in 30 cases?
Hint: You may assume that each murder has been committed by only one person who is different from the murderers of the other cases.

♠ **Solution:**

$$\mathbb{E}\{\#\text{ of Tests}\} = 8 \cdot 0.01 \cdot 7 + 4 \cdot 0.01 \cdot 6 + 0.22 \cdot 3 + 3 \cdot 0.22 \cdot 2 = 2.78$$

(h) Compare the results of Parts (d) and (g). Explain why the results are different?

♠ **Solution:** The probabilities $P_{M|H_2}(i)$ are no powers of $\frac{1}{2}$, hence the entropy is not reached.

10.6.3 Source Coding (23 Points)

Consider the ternary source $X \in \{A, B, C\} = \mathcal{A}_X$. The following table define the source code $\mathcal{C}$:

X	$\mathcal{C}(X)$
A	0
B	01
C	1

(a) Is the source code $\mathcal{C}$ uniquely decodable? Explain your answer by giving a reason.

♠ **Solution:**

$$C('AC') = 01 = C('B')$$

$$\Longrightarrow\ C \text{ is not uniquely decodable}$$

(b) For the given source, give a criterion for existence of a uniquely decodable source code. Does the source code $\mathcal{C}$ fulfils this criterion?

♠ **Solution:**

$$\text{Kraft Inequality : } \sum_i 2^{-\ell_i} \leq 1$$

$$\text{Here, we have } \ell_A = \ell_C = 1, \ell_B = 2$$

$$\implies \sum_i 2^{-\ell_i} = \frac{1}{2} + \frac{1}{4} + \frac{1}{2} = \frac{5}{4} > 4$$

$$\implies \text{Kraft inqeuality is not fulfilled for } \mathcal{C}.$$

Now, consider the following probability distribution

$$P_X(x) = \Pr\{X = x\} = \begin{cases} \frac{2}{3} & x = A \\ \frac{1}{6} & x = B \\ \frac{1}{6} & x = C \end{cases}.$$

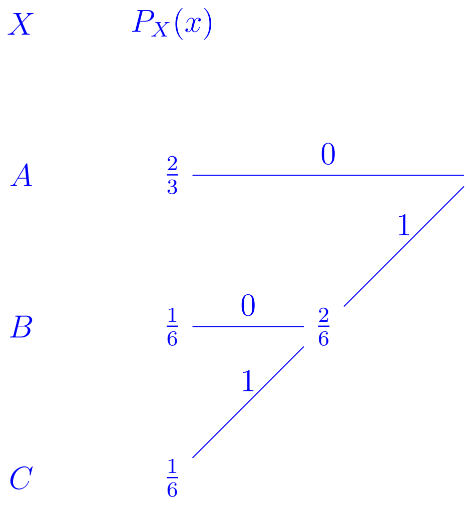
(c) How many bits per symbol is required, such that a long independent and identically distributed (iid) sequence of X with probability distribution P_X is compressed without any information loss?

♠ **Solution:**

$$\begin{aligned} H(X) &= -\sum_i P_X(x) \log P_X(x) \\ &= \frac{2}{3} \log \frac{2}{3} + \frac{2}{6} \log 6 \\ &\approx 1.25 \text{ Bits} \end{aligned}$$

(d) For the given source X with probability distribution P_X, give a prefix-free source code with minimum average codelength. Determine the average codelength of this source code.

♠ **Solution:** Huffman code:



C(x)	ℓ_i
0	1
10	2
11	2

$$\Longrightarrow \ \mathbb{E}\{L(C(X))\} = \frac{2}{3} + \frac{1}{6}\cdot 2 = \frac{4}{3}$$

(e) Compare your answers to Parts (c) and (d). Why are they different? Name a strategy by which the difference between these two answers can be reduced (at the expense of a larger codebook).

♠ **Solution:**

$$\mathbb{E}\{L(C(X))\} > H(X) \ \text{ because } 2^{-\ell_i} \neq P_X(x_i)$$

Strategy to reduce difference: Encode in blocks X^N

(f) Determine the probability distribution P_I which is induced on $\mathcal{A}_X$ by your given code in Part (d).

♠ **Solution:**

$$P_I(x) = \begin{cases} \frac{1}{2} & x = A \\ \frac{1}{4} & x = B \\ \frac{1}{4} & x = C \end{cases}.$$

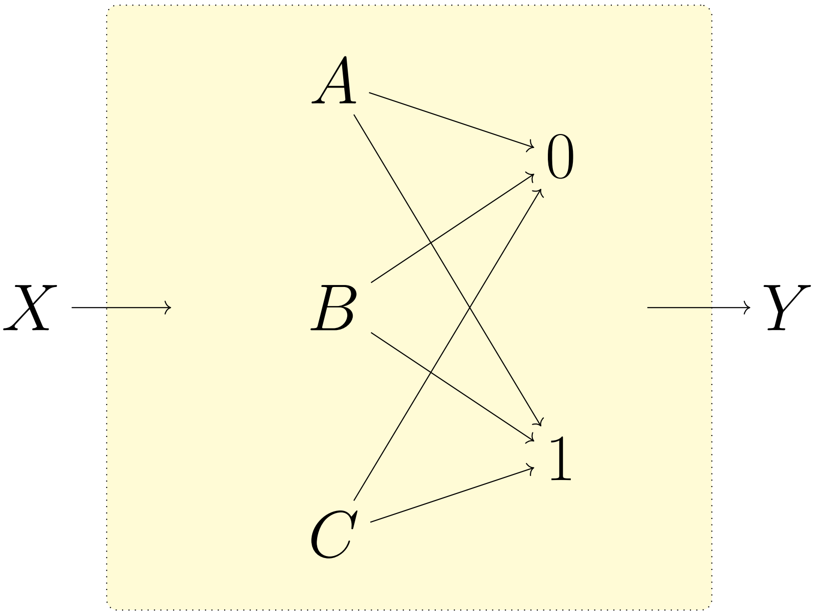
(g) Determine the Kullback-Leibler divergence between the induced distribution P_I, determined in Part (f), and the distribution P_X, i.e., $D_{KL}(P_I||P_X)$.

♠ **Solution:**

$$\begin{aligned} D_{KL}(P_I||P_X) &= \sum_{x \in \mathcal{A}_X} P_I(x) \log \frac{P_I(x)}{P_X(x)} \\ &= \frac{1}{2} \log \frac{3}{4} + 2 \log \frac{1}{4} \log \frac{6}{4} \\ &\approx 0.085 \end{aligned}$$

10.6.4 Channel Capacity (17 Points)

Consider the following channel



with input symbol $X \in \{A, B, C\}$, output symbol $Y \in \{0, 1\}$ and transition rule

$$\text{Pr}\{Y = 1|X = A\} = 1 - \text{Pr}\{Y = 0|X = A\} = f$$
$$\text{Pr}\{Y = 0|X = B\} = 1 - \text{Pr}\{Y = 1|X = B\} = f$$
$$\text{Pr}\{Y = 0|X = C\} = 1 - \text{Pr}\{Y = 1|X = C\} = f$$

(a) Is this channel symmetric? Explain your answer by giving a reason.

♠ **Solution:** Transition matrix:

$$\begin{bmatrix} 1-f & f & f \\ f & 1-f & 1-f \end{bmatrix}$$

There is no rowwise partitioning of the transition matrix such that each row and each column are permutations of each other row and each other column, respectively.

$\Longrightarrow$ The channel is not symmetric.

(b) For a generic input distribution, calculate the marginal output distribution P_Y. What does the result say about the capacity achieving distribution? Explain your answer.

♠ **Solution:**

$$P_X(x) = \begin{cases} q_1 & x = A \\ q_2 & x = B \\ q_3 & x = C \end{cases}.$$

$$P_{Y|X}(Y|X = A) = \begin{cases} 1-f & ,y = 0 \\ f & ,y = 1 \end{cases}.$$

$$P_{Y|X}(Y|X = B) = P_{Y|X}(Y|X = C) = \begin{cases} f & ,y = 0 \\ 1-f & ,y = 1 \end{cases}.$$

$$\begin{aligned} P_Y(0) &= q_1(1-f) + q_2 f + (1 - q_1 - q_2)f \\ &= q_1(1-f) + (1-q_1)f \\ P_Y(1) &= 1 - P_Y(0) = q_1 f_1 + (1-q_1)(1-f) \end{aligned}$$

$\Longrightarrow$ Output distribution is independent of $q_2 = P_X(B)$

Justification: As $P_{Y|X}(Y|X = B) = P_{Y|X}(Y|X = C)$, it is irrelevant in terms of the channel output whether B or C is sent.

(c) Determine the capacity of this channel. Give *two* different distributions with which the channel capacity is achieved.

♠ **Solution:**

$$\begin{aligned} H(Y|X) &= P_X(A)H(Y|X = A) + P_X(B)H(Y|X = B) + P_X(C)H(Y|X = C) \\ &= q_1 H_2(f) + q_2 H_2(f) + (1 - q_1 - q_2)H_2(f) \\ &= H_2(f) \end{aligned}$$

$$C = \max_{P_X} I(X;Y) = \max_{q_1} \; H(Y) - H_2(f) = 1 - H_2(f)$$

$$q_1^* = \operatorname*{argmax}_{q_1} H(Y)$$

$$q_1^*(1-f) + (1-q_1^*)f \overset{!}{=} \frac{1}{2} \quad \left(\text{Because } P_Y(0) \overset{!}{=} \frac{1}{2}\right)$$

$$q_1^* = \frac{\frac{1}{2} - f}{1 - 2f} = \frac{1}{2}$$

Capacity achieving distributions are e.g

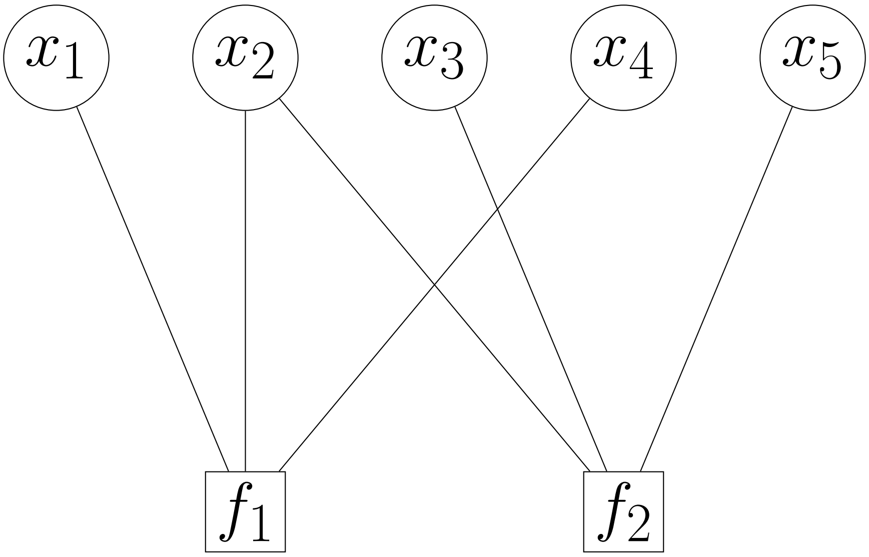
$$P_{X,1} = \left(\frac{1}{2}, 0, \frac{1}{2}\right) = (P_{X,1}(A), P_{X,1}(B), P_{X,1}(C))$$

and

$$P_{X,2} = \left(\frac{1}{2}, \frac{1}{2}, 0\right).$$

10.6.5 Channel Coding and Decoding (20 Points)

Consider the following factor graph with variable nodes $x_1, \ldots, x_5$ and factor nodes f_1 and f_2:



The functions $f_1(\cdot)$ and $f_2(\cdot)$ are defined as:

$$f_1(x_1, x_2, x_4) = \begin{cases} 1, & \text{if} \quad x_1 + x_2 + x_4 = 0 \quad \text{mod } 2 \\ 0, & \text{otherwise} \end{cases},$$

$$f_2(x_2, x_3, x_5) = \begin{cases} 1, & \text{if} \quad x_2 + x_3 + x_5 = 0 \quad \text{mod } 2 \\ 0, & \text{otherwise} \end{cases}.$$

(a) Determine the parity check matrix **H** defined by this factor graph.

♠ **Solution:**

$$H = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \end{bmatrix}$$

(b) Determine the code rate for the channel code $\mathcal{C}$ defined by **H**.

♠ **Solution:**

$$R = \frac{3}{5}$$

(c) Show that $[x_1, x_2, x_3, x_4, x_5] = [1, 0, 0, 1, 0]$ is a valid codeword of channel code $\mathcal{C}$. Determine the number of errors (bit flips) in a transmitted codeword which can be corrected by the channel code $\mathcal{C}$.

♠ **Solution:**

$$H \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1+1 \\ 0+0 \end{bmatrix} = \underline{0}$$

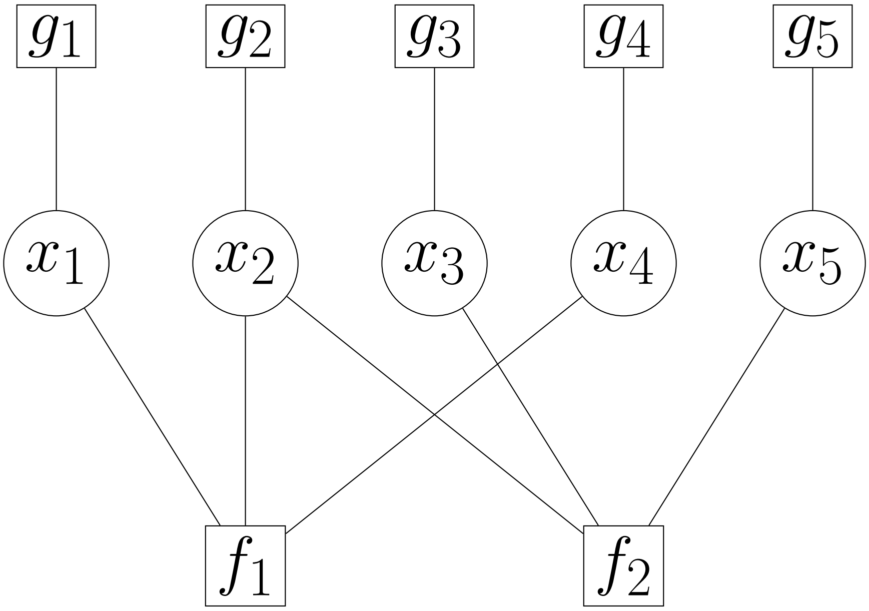
$\implies d_min \leq 2$ (as $\underline{0}$ is also a valid codewoed)

$\#$ of errors that can be corrected:

$$\lfloor \frac{d_{\text{min}} - 1}{2} \rfloor \leq \lfloor \frac{2-1}{2} \rfloor = 0$$

No error can be corrected.

Now assume that a codeword of channel code $\mathcal{C}$ has been transmitted over a *binary erasure channel* (BEC) with error probability $\text{Pr}\{Y = \text{Error}|X = 0\} = \text{Pr}\{Y = \text{Error}|X = 1\} = \epsilon$. The codeword is received as $\mathbf{r} = [r_1, r_2, r_3, r_4, r_5] = [1, 0, 0, \text{Error}, 0]$. We intend to detect **r** by means of the following factor graphs:



(d) Define functions $g_i(x_i)$, for $i \in \{1, \ldots, 5\}$, in the factor graph such that $g_i(x_i) = \text{Pr}\{r_i|X_i = x_i\}$.

♠ **Solution:**

$$g_1(x_1) = \begin{cases} 0, & x_1 = 0 \\ 1-\epsilon, & x_1 = 1 \end{cases}$$

$$g_2(x_2) = \begin{cases} 1-\epsilon, & x_2 = 0 \\ 0, & x_2 = 1 \end{cases}$$

$$g_3(x_3) = \begin{cases} 1-\epsilon, & x_3 = 0 \\ 0, & x_3 = 1 \end{cases}$$

$$g_4(x_4) = \begin{cases} \epsilon, & x_4 = 0 \\ \epsilon, & x_4 = 1 \end{cases}$$

$$g_5(x_5) = \begin{cases} 1-\epsilon, & x_5 = 0 \\ 0, & x_5 = 1 \end{cases}$$

(e) Using the sum-product algorithm, determine the *posterior* distribution of x_4, i.e., Pr$\{X_4 = x_4|\mathbf{r}\}$. Determine further the detected codeword.

♠ **Solution:**

$$\text{Pr}\{X_4 = x_4|r\} = \frac{Z_4(x_4)}{Z}$$

$$Z_4(x_4) = r_{f_1\to 4}(x_4) \cdot r_{g_4\to 4}(x_4)$$

$$\begin{aligned} r_{g_4\to 4}(x_4) &= g_4(x_4) \\ r_{f_1\to 4}(x_4) &= \sum_{x_1,x_2} f_1(x_1,x_2,x_4) q_{2\to f_1}(x_2) q_{1\to f_1}(x_1) \end{aligned}$$

$$\begin{aligned} q_{2\to f_1}(x_2) &= r_{g_2\to 2}(x_2) r_{f_2\to 2}(x_2) = g_2(x_2) r_{f_2\to 2}(x_2) \\ q_{1\to f_1}(x_1) &= r_{g_1\to 1}(x_1) = g_1(x_1) \\ r_{f_2\to 2}(x_2) &= \sum_{x_3,x_5} f_2(x_2,x_3,x_5) q_{3\to f_2}(x_3) q_{5\to f_2}(x_5) \end{aligned}$$

$$\begin{aligned} q_{3\to f_2}(x_3) &= r_{g_3\to 3}(x_3) = g_3(x_3) \\ q_{5\to f_2}(x_5) &= r_{g_5\to 5}(x_5) = g_5(x_5) \end{aligned}$$

$$\begin{aligned} r_{f_2\to 2} &= \sum_{x_3,x_5} f_2(x_2,x_3,x_5)g_3(x_3)g_5(x_5) = f_2(x_2,0,0)(1-\epsilon)^2 \\ r_{f_1\to 4}(x_4) &= \sum_{x_1,x_2} f_1(x_1,x_2,x_4)g_2(x_2)f_2(x_2,0,0)(1-\epsilon)^2 g_1(x_1) \\ &= f_1(1,0,x_4)(1-\epsilon)^4 \end{aligned}$$

$$Z_4(x_4) = \begin{cases} 0, & x_4 = 0 \\ (1-\epsilon)^4\epsilon, & x_4 = 1 \end{cases}$$

$$\text{Pr}\{X_4 = 0|r\} = 0 \qquad \text{Pr}\{X_4 = 1|r\} = 1$$

$$\implies \hat{x} = [1,0,0,1,0]$$

(f) Is there another approach to find your answer to Part (e)? If yes, explain this approach.

♠ **Solution:** We could come to the same result by considering that, given $\underline{r} = [1,0,0,\epsilon,0]$, the only two possibly sent words are:

$$\underline{x_1} = [1,0,0,0,0]$$

and

$$\underline{x_2} = [1,0,0,1,0]$$

As $H\underline{x_1} \neq 0$, $\underline{x_1}$ is not a valid codeword, and

$$\underline{\hat{x}} = [1,0,0,1,0]$$

10.7 Winter Semester 2020-2021 Exam

Exam Date: February 16th, 2021
5 Questions with total of 100 Points.
Exam Duration: 90 Minutes

10.7.1 Information Theoretic Identities and Inequalities (20 Points)

Assume that X is a *continuous* random variable with the support $[a, a+I)$. This means that the probability density function (PDF) of X is of the following form

$$f_X(x) = \begin{cases} f_0(x) & a \leq x \leq a+I \\ 0 & \text{otherwise} \end{cases}$$

where $f_0(x)$ satisfies $\displaystyle\int_a^{a+I} f_0(x)\,\mathrm{d}x = 1$.

(a) Show that the *differential* entropy of X, i.e., $h(X)$, satisfies the following inequality:

$$h(X) \leq \log I$$

Hint: You could use *Gibbs' inequality.*

> ⇝ REMINDER:
>
> The Kullback-Leibler divergence between two PDFs f_X and g_X is defined as
>
> $$D_{\mathsf{KL}}(f_X || g_X) = \int_{-\infty}^{+\infty} f_X(x) \log \frac{f_X(x)}{g_X(x)}\,\mathrm{d}x,$$
>
> and has the same properties as those given for the Kullback-Leibler divergence between two probability mass functions (PMFs).

♠ **Solution:** Let

$$g_X(x) = \begin{cases} \frac{1}{I} & a \leq x < a+I \\ 0 & \text{otherwise} \end{cases}.$$

Then,

$$\begin{aligned} D_{\mathsf{KL}}(f_X || g_X) &= \int_a^{a+I} f_X(x) \log \frac{f_X(x)}{\frac{1}{I}} \mathrm{d}x = \\ &= \int_a^{a+I} f_X(x) \left[\log f_X(x) + \log I\right] \mathrm{d}x \\ &= \underbrace{\int_a^{a+I} f_X(x) \log f_X(x) \mathrm{d}x}_{=-h(X)} + \log I \underbrace{\int_a^{a+I} f_X(x) \mathrm{d}x}_{=1} = -h(X) + \log I \end{aligned}$$

Gibb's inequality states that : $D_{\text{KL}}(f_X||g_X) \geq 0$. Thus,

$$-h(X) + \log I \geq 0 \implies \boxed{h(X) \leq \log I}$$

(b) Explain why Part (a) does *not* contradict with the following statement:

> The *maximum* differential entropy of a real continuous random variable whose variance is *bounded from above* by σ^2 is given by a Gaussian distribution with variance σ^2.

♠ **Solution:** Two major constraints are not fulfilled:

1. X is not defined over the whole real axis.
2. The variance of X is not restricted.

Now consider two *independent discrete* random variables X and Y with non-zero entropies, i.e., $H(X) \neq 0$ and $H(Y) \neq 0$. The *discrete* random variable Z is determined as

$$Z = X + 2Y$$

Fill the empty fields, i.e., □, in the following items with *identity sign*, i.e, $=$, *greater than*, i.e., $>$, or *less than*, i.e., $<$.

(c) $I(X;Z|Y) \quad \square \quad H(X)$

♠ **Solution:**

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

$$\text{Ⓐ} \quad X, Y \text{ independent.} \implies H(X|Y) = H(X)$$

$$\text{Ⓑ} \quad Z = X + 2Y \implies X = Z - 2Y \implies H(X|Z,Y) = 0$$

$$\text{Ⓐ} + \text{Ⓑ} \implies \boxed{I(X;Z|Y) = H(X)}$$

Answer: $\boxed{=}$

(d) $I(X;Y|Z) \quad \square \quad 0$

♠ **Solution:**

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

From last part, we know:

$$H(X|Y,Z) = 0$$

Thus,

$$I(X;Y|Z) = H(X|Z) \geq 0$$

Since,

$$H(Y) \neq 0 \implies H(X|Z) \neq 0 \implies H(X|Z) > 0$$

Answer: $\boxed{>}$

(e) $I(X;Z|Y) \quad \square \quad I(X;Z) + I(X;Y|Z)$

♠ **Solution:**

$$I(X;Z|Y) = \underbrace{H(X|Y)}_{=H(X)} - \underbrace{H(X|Y,Z)}_{=0} = H(X)$$

$$\underbrace{I(X;Z)}_{\textcircled{1}} + \underbrace{I(X;Y|Z)}_{\textcircled{2}} = \underbrace{H(X) - \cancel{H(X|Z)}}_{\textcircled{1}} + \underbrace{\cancel{H(X|Z)} - H(X|Y,Z)}_{\textcircled{2}}$$
$$= H(X) - \underbrace{H(X|Y,Z)}_{=0} = H(X)$$

Answer: $\boxed{=}$

Also possible to solve using the fact that $I(X;Y) = 0$, and that the two sides are decompositions of $I(X;Y,Z)$.

10.7.2 Bayesian Inference (26 Points)

In Nuremberg airport, four passengers have arrived from four different departure points, namely A, B, C and D. Based on the statistical records, the following *prior beliefs* are available for each person:

Departure Point	Probability of being infected with COVID-19
Point A	0.10
Point B	0.05
Point C	0.40
Point D	0.25

Three passengers are now testes for COVID-19. The results of the tests are as follows:

Departure Point	Test Result
Point A	negative
Point B	positive
Point C	negative
Point D	positive

The COVID-19 tests which have been used for these passengers are not completely accurate.

> For an *infected* person, the result of the test could wrongly be *negative* with probability 0.2, and for an *uninfected* person, it could be wrongly *positive* with probability 0.05.

The *true infection probability* for each person is defined as the *posterior* probability of the person being infected with COVID-19, *conditioned* on the test result, when the *prior infection probability* of the person is set according to the *prior belief* on the departure point.

(a) Formulate the calculation of the *true infection probability* as a Bayesian inference problem. Specify the *prior probability* and *likelihood* for each passenger.

♠ **Solution:** Status of a person arrived from point i: X_i

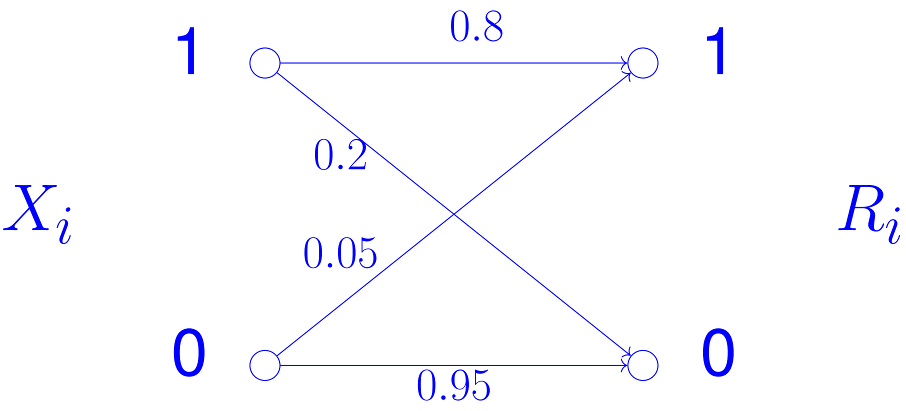
$$X_i = \begin{cases} 1 & \text{infected} \\ 0 & \text{uninfected} \end{cases}$$

Result of the test for person i:

$$R_i = \begin{cases} 1 & \text{positive} \\ 0 & \text{negative} \end{cases}$$

$$\text{Pr}\{X_i = x_i\} : \text{ Prior Belief (Table 1)}$$

$$\text{Pr}\{R_i = r_i | X_i = x_i\} : \text{ Likelihood :}$$



$$\text{Pr}\{X_i = x_i | R_i = r_i\} : \text{ Posterior}$$

(b) Determine the *true infection probability* for each person.

♠ **Solution:** The true infection probability is defined as the posterior.

Let us show it with P_i^T for person arrived from point i. Then, we have

$$P_i^T = \text{Pr}\{X_i = 1|R_i = r_i\} = \frac{\text{Pr}\{R_i = r_i|X_i = 1\}\text{Pr}\{X_i = 1\}}{\sum_{x_i \in \{0,1\}} \text{Pr}\{R_i = r_i|X_i = x_i\}\text{Pr}\{X_i = x_i\}}$$

We get the priors from Table 1, and r_i from Table 2.

Point A:

$$\text{Pr}\{X_A = 1\} = 0.1, \qquad \text{Pr}\{X_A = 0\} = 0.9$$

$$\boxed{r_A = 0} \implies \begin{cases} \text{Pr}\{R_A = 0|X_A = 1\} = 0.2 \\ \text{Pr}\{R_A = 0|X_A = 0\} = 0.95 \end{cases} \implies \text{Pr}\{R_A = 0\} = \frac{7}{8} = 0.875$$

$$\boxed{\text{Thus: } P_A^T = \frac{0.02}{0.875} = 0.023}$$

Point B:

$$\text{Pr}\{X_B = 1\} = 0.05, \qquad \text{Pr}\{X_B = 0\} = 0.95$$

$$\boxed{r_B = 1} \implies \begin{cases} \text{Pr}\{R_B = 1|X_B = 1\} = 0.8 \\ \text{Pr}\{R_B = 1|X_B = 0\} = 0.05 \end{cases} \implies \text{Pr}\{R_B = 1\} = \frac{7}{80} = 0.0875$$

$$\boxed{\text{Thus: } P_B^T = \frac{0.04}{0.0875} = 0.457}$$

Point C:

$$\text{Pr}\{X_C = 1\} = 0.4, \qquad \text{Pr}\{X_C = 0\} = 0.6$$

$$\boxed{r_C = 0} \implies \begin{cases} \text{Pr}\{R_C = 0|X_C = 1\} = 0.2 \\ \text{Pr}\{R_C = 0|X_C = 0\} = 0.95 \end{cases} \implies \text{Pr}\{R_C = 0\} = 0.65$$

$$\boxed{\text{Thus: } P_C^T = \frac{0.08}{0.65} = 0.123}$$

Point D:

$$\text{Pr}\{X_D = 1\} = 0.25, \qquad \text{Pr}\{X_D = 0\} = 0.75$$

$$\boxed{r_D = 1} \implies \begin{cases} \text{Pr}\{R_D = 1|X_D = 1\} = 0.8 \\ \text{Pr}\{R_D = 1|X_D = 0\} = 0.05 \end{cases} \implies \text{Pr}\{R_D = 1\} = 0.2375$$

$$\boxed{\text{Thus: } P_D^T = \frac{0.2}{0.2375} = 0.842}$$

The airport has received the following instruction:

> A person should stay in quarantine, if his/her *true infection probability* is more than 0.2

(c) Specify the departure points of the passengers who should stay in quarantine.

♠ **Solution:** B and D.

A person has arrived from another departure point E which is different from points A, B, C and D. Based on the statistical records, the *prior belief* is that the person is infected with COVID-19 with probability p, where $p \in [0, 1]$.

(d) Assume that *the same threshold* of true infection probability, i.e., 0.2, is used to decide, whether a passenger should stay in quarantine or not. Find the minimum value of p for which COVID-19 test is useless, i.e, the person should stay in quarantine regardless of the test result.

♠ **Solution:** We need

$$(P_E^T \text{ for } r_E = 0) \quad \geq 0.2$$

When $r_E = 0$: $P_E^T = \Pr\{X_E = 1 | R_E = 0\}$

$$P_E^T = \frac{\overbrace{\Pr\{R_E = 0 | X_E = 1\}}^{0.2} \overbrace{\Pr\{X_E = 1\}}^{p}}{\sum_{x \in \{0,1\}} \Pr\{R_E = 0 | X_E = x\} \Pr\{X_E = x\}}$$

Also,

$$\begin{aligned}\Pr\{R_E = 0\} &= \overbrace{\Pr\{R_E = 0 | X_E = 1\}}^{0.2} \overbrace{\Pr\{X_E = 1\}}^{p} \\ &+ \underbrace{\Pr\{R_E = 0 | X_E = 0\}}_{0.8} \underbrace{\Pr\{X_E = 0\}}_{1-p} \qquad = 0.8 - 0.6p\end{aligned}$$

Thus:

$$\begin{aligned} P_E^T = \frac{0.2p}{0.8 - 0.6p} &\geq 0.2 \\ &\implies \boxed{p \geq 0.5} \\ &\implies \text{For } p \geq 0.5\text{, the person goes to quarantine anyway.}\end{aligned}$$

- Important: Note that for $r_E = 1$:

$$P_E^T = \frac{0.8}{0.8 - 0.6p}$$

which is always better than P_E^T for $r_E = 0$. Thus, it does not need to be checked.

10.7.3 Source Coding (22 Points)

A company produces surfaces with N pieces. In these surfaces, each piece is a reflective material with an *integer* index between 1 and 7. This means that each piece has an index which is in set $\{1, \dots, 7\}$. The surfaces are made by *independent* collection of pieces. This means that each piece is chosen from the available 7 choices *independently*.

A statistical study has tested several surfaces with large number of elements. The result of this study has revealed that

> For a given surface, let N_i for $i \in \{1, \dots, 7\}$ denote the number of pieces whose indices are i. In a *typical* surface, we have
>
> $$N_i = \frac{N_1}{i}.$$

You are asked to design a labeling strategy for surfaces with a *large number of pieces*, i.e., surfaces with N pieces where $N \to \infty$. Your strategy should label a given surface with a binary sequence, such that this surface is identified uniquely from its label with *probability one*.

(a) Model a surface as an independent and identically distributed (i.i.d) sequence of pieces. Use the result of the statistical study to find the probability of each *piece index*.

♠ **Solution:** Let X_n be label of piece $n \in \{1, \dots, N\}$. Then the pieces are denoted by

$$X_1, X_2, \dots, X_N; \qquad X_n \in \{1, \dots, 7\}$$

$$\Pr\{X_n = i\} = \frac{N_i}{N} = \frac{N_1}{i \times N} \implies \boxed{\Pr\{X_n = i\} = \frac{p_1}{i}}$$

To find N_i, we note that

$$\sum_{i=1}^{7} N_i = N \implies N_1 + \frac{N_1}{2} + \frac{N_1}{3} + \frac{N_1}{4} + \cdots + \frac{N_1}{7} = N$$

$$\implies \frac{N}{N_1} = 1 + \sum_{i=2}^{7} \frac{1}{i} = 2.59$$

$$P_1 = \frac{N_1}{N} = \frac{1}{2.59} = \boxed{0.386}, \qquad \boxed{p_i = \frac{0.386}{i}}$$

(b) Assume that you have a *typical* surface with $N = 10^6$ pieces. Use the source doing theorem to bound *approximately the minimum* length of its label from below. Justify your answer.

> **Hint:** Pay attention to the fact that this surface is a *typical* surface.

♠ **Solution:** Let length of the label be M. The first Shannon's Theorem would suggest that:

$$\frac{M}{N} \geq H(X) \implies \boxed{M \geq N\,H(X)}$$

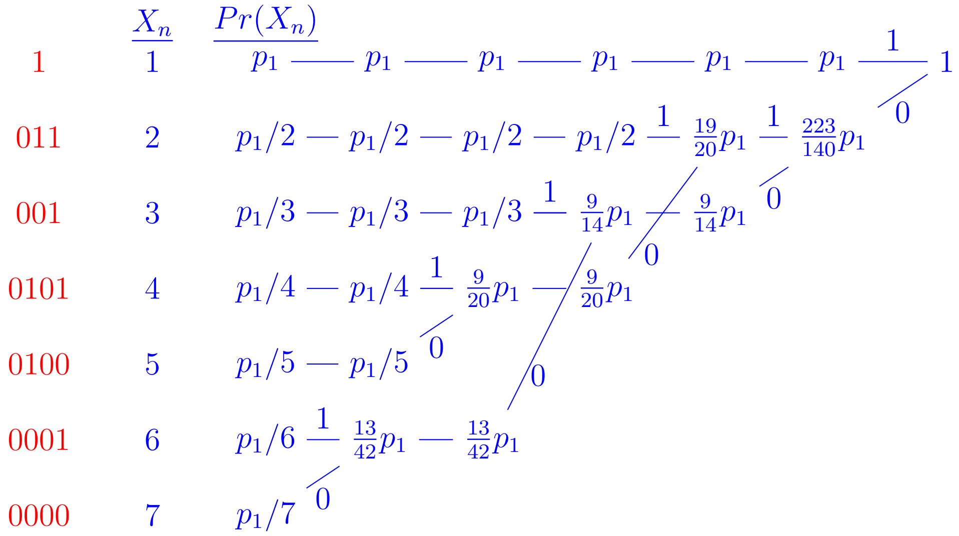
This is a good bound, since the sequence is typical. We know that:

$$H(X) = \sum p_i \log \frac{1}{p_i} = p_1 \underbrace{\sum \frac{1}{i} \log i}_{2.82} - \log p_1$$

$$\implies \boxed{H(X) = 2.46} \implies \boxed{M \geq 10^6 \times 2.46 = 2460000}$$

(c) Design a labeling strategy via Huffman coding which labels a surface *piece-wise*, i.e., each piece is labeled individually. Determine the *average length of labels per piece* for this strategy.

♠ **Solution:** For Huffman Coding, we have



$$\overline{L}(X,C) = \sum_{i=1}^{7} p_i \log L(X_n = i) = p_1 \sum_{i=1}^{7} \frac{\ell_i}{i} \boxed{= 2.5237}$$

(d) Consider once again the surface in Part (b). Determine *approximately* the length of its label, when the strategy in Part (c) is used for labeling.
Hint: Pay attention to the fact that this surface is a *typical* surface.

♠ **Solution:** Now, we have for a typical surface:

$$\frac{M}{N} \geq \overline{L} \implies M \geq N \times 2.5237 = 2523700$$

Now assume that your labeling should also include surfaces with a *small number* of pieces.

(e) Name a *universal* dictionary-based labeling strategy which *neither* uses the piece index probabilities *nor* constructs prior beliefs on piece indices for labeling.

♠ **Solution:** Lempel-Ziv algorithm.

(f) Assume you have a small surface with $N = 6$ pieces. These 6 pieces have the following indices:

$$1, 2, 1, 4, 6, 7.$$

Label this surface via the universal strategy named in Part (e), taking the following steps:

i. Replace the index of each piece with its binary representation, i.e., 1 is represented by 001 and 7 is represented by 111.

ii. Use the universal strategy, names in Part (e), to find the label.

♠ **Solution:** The binary representation is:

$$1, 2, 1, 4, 6, 7$$

$$001, 010, 001, 100, 110, 111$$

$$\begin{tikzcd}
0 \arrow[d] & 01 \arrow[d] & 010 \arrow[d] & 00 \arrow[d] & 1 \arrow[d] & 10 \arrow[d] & 011 \arrow[d] & 0111 \arrow[d] \\
(\lambda, 0) \arrow[d] & (1, 1) \arrow[d] & (2, 0) \arrow[d] & (1, 0) \arrow[d] & (0, 1) \arrow[d] & (5, 0) \arrow[d] & (2, 1) \arrow[d] & (7, 1) \arrow[d] \\
(-, 0) & (1, 1) & (10, 0) & (01, 0) & (000, 1) & (101, 0) & (010, 1) & (111, 1)
\end{tikzcd}$$

Codebook

Substream	Index
λ	0
0	1
01	2
010	3
00	4
1	5
10	6
011	7
0111	8

$\Longrightarrow$ Encoded sequence (Label) :

$$0111000100001101001011111$$

10.7.4 Channel Coding (17 Points)

The *complex-valued* input of a communication channel is denoted by $X = I + \mathrm{j}Q$, where $\mathrm{j} = \sqrt{-1}$ is the imaginary unit. X is related to the output of the channel as follows:

$$Y = \mathrm{sgn}(Z_I)\,\mathrm{sgn}(I) + \mathrm{j}\,\mathrm{sgn}(Z_Q)\,\mathrm{sgn}(Q).$$

Here,

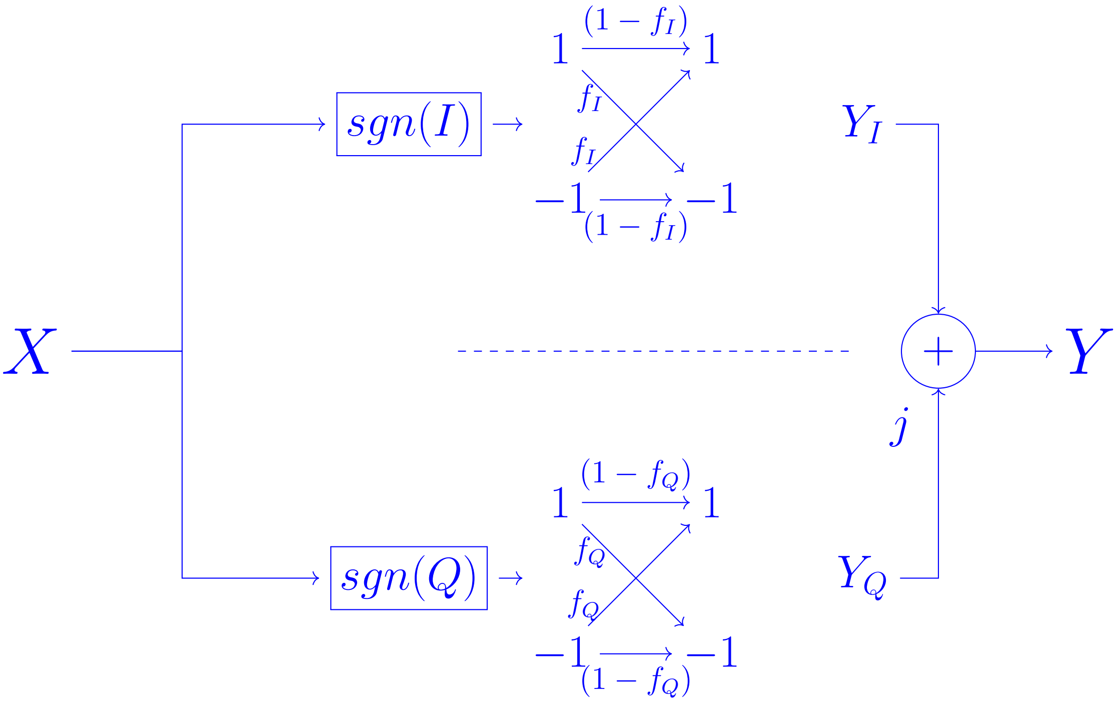
- $\mathrm{sgn}(\cdot)$ is the *sign function* defined as

$$\mathrm{sgn}(x) = \begin{cases} +1 & x \geq 0 \\ -1 & x < 0 \end{cases}.$$

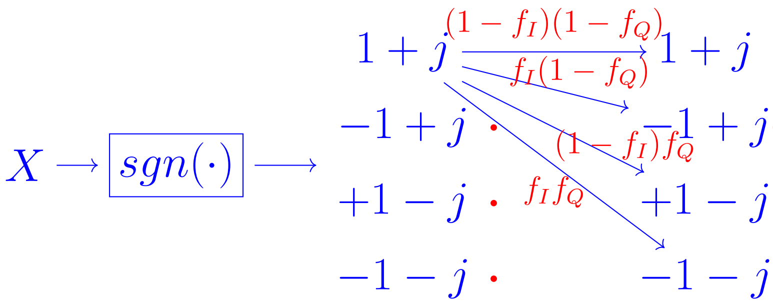
- Z_I and Z_Q are two *independent* continuous real-valued random variables which model the noise process. Z_I and Z_Q are distributed uniformly in intervals $[-f_I, 1 - f_I]$ and $[-f_Q, 1 - f_Q]$, respectively, where f_I, $f_Q \in [0, 0.5]$.

(a) Draw the input-output diagram of this channel.

♠ **Solution:** Let $Y = Y_I + j\,Y_Q$



Thus, we could say



Other branches are derived similarly.

(b) Determine the capacity of this channel.

♠ **Solution:** To determine the capacity:

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

→ For $H(Y|X)$, we note that:

$$H(Y|X = x_i) = H_2(f_I) + H_2(f_Q).$$

Thus,

$$\boxed{H(Y|X) = H_2(f_I) + H_2(f_Q)}$$

→ For $H(Y)$: We set the distribution of X after operating the sgn$(\cdot)$ function to be

$$X = \begin{cases} 1+j, & p_1 \\ 1-j, & p_2 \\ -1+j, & p_3 \\ -1-j, & 1-p_1-p_2-p_3 = p_4 \end{cases}$$

The output distribution is then

$$Y = \begin{cases} 1+j, & (1-f_I)(1-f_Q)p_1 + (1-f_I)f_Qp_2 + f_I(1-f_Q)p_3 + f_If_Qp_4 \\ 1-j, & (1-f_I)(1-f_Q)p_2 + (1-f_I)f_Qp_1 + f_If_Qp_3 + f_I(1-f_Q)p_4 \\ -1+j, & (1-f_Q)f_Ip_1 + f_If_Qp_2 + (1-f_I)(1-f_Q)p_3 + f_Q(1-f_I)p_4 \\ -1-j, & f_If_Qp_1 + f_I(1-f_Q)p_2 + (1-f_I)f_Qp_3 + (1-f_I)(1-f_Q)p_4 \end{cases}$$

$$\implies I(X;Y) = H(Y) - H(Y|X) = H(Y) - H_2(f_I) - H_2(f_Q)$$

$$C = \max_{p_1,p_2,p_3,p_4} \underbrace{H(Y)}_{H(Y)\leq \log 4 = 2} - H(Y|X) = 2 - H_2(f_I) - H_2(f_Q)$$

Alternatively, one could say that

The channel consists of TWO PARALLEL BSCs.

Thus, capacity is

$$C = \underbrace{C_I}_{Ⓐ} + \underbrace{C_Q}_{Ⓑ} = 1 - H_2(f_I) + (1 - H_2(f_Q)) = \boxed{2 - H_2(f_I) - H_2(f_Q)}$$

Ⓐ is the channel from the real value of X to real value of Y and Ⓑ is the channel from the imaginary of X to imaginary of Y.

(c) Find an input distribution which achieves the channel capacity.

♠ **Solution:** Given the distribution of Y: We have the maximum mutual information when Y is uniform.

Using the distribution of Y, one can see:

$$\text{Pr}\{Y = y\} = \frac{1}{4} \text{ iff } \boxed{p_i = \frac{1}{4}} \text{ for } i = 1, \dots, 4$$

Thus, X should be distributed such that the output of sgn$(\cdot)$ is uniform. There are various examples:

① $X \sim \mathcal{CN}(0, 1)$

② $X \sim$ Uniform on/in unit circle

③ $X \sim$ Uniform on/in a symmetric square with center at $(0 + j\,0)$

Any other example with Pr$\{I + j\,Q = i + j\,q\} = \frac{1}{4}$ is correct.

10.7.5 Factor Graphs (15 Points)

The four binary random variables X_1, X_2, X_3 and X_4 are statistically *dependent*. For the random variable X_i with $i \in \{2, 3, 4\}$, we have the following property

$$P(x_i|x_{i-1}, \dots, x_1) = P(x_i|x_{i-1})$$

Assume that Pr$\{X_1 = 0\} = 1 - $ Pr$\{X_1 = 1\} = 0.4$, and

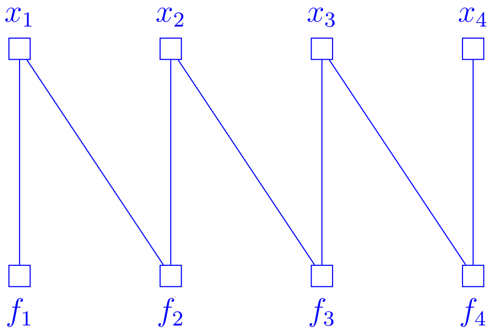
$$P(x_i|x_{i-1}) = \frac{1}{8} + \frac{x_{i-1}}{8} + \frac{3x_i}{4} - \frac{x_{i-1}x_i}{4}.$$

(a) Draw the factor graph of the joint distribution $P(x_1, x_2, x_3, x_4)$ with the *maximum* number of factor and variable nodes. Specify the factor and variable nodes in the graph.

♠ **Solution:**

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)\underbrace{p(x_3|x_2, x_1)}_{p(x_3|x_2)}\underbrace{p(x_4|x_3, x_2, x_1)}_{p(x_4|x_3)}$$

$$= \underbrace{p(x_1)}_{f_1}\underbrace{p(x_2|x_1)}_{f_2}\underbrace{p(x_3|x_2)}_{f_3}\underbrace{p(x_4|x_3)}_{f_4}$$



(b) Determine the marginal distribution of X_2 via the *sum-product* algorithm.

♠ **Solution:**

$$p(x_2) = \sum_{x_1,x_3,x_4} p(x_1,x_2,x_3,x_4) = Z(x_2)$$

Sum-product says:

$$Z(x_2) = r_{f_2\to x_2}(x_2) r_{f_3\to x_2}(x_2)$$

$$r_{f_2\to x_2}(x_2) = \sum_{x_1} f_2(x_1,x_2) q_{x_1\to f_2}(x_1)$$

And we have

$$q_{x_1\to f_2}(x_1) = r_{f_1\to x_1}(x_1) = \sum_{\emptyset} f_1(x_1) = p(x_1)$$

Thus:

$$\begin{aligned} r_{f_2\to x_2}(x_2) &= \sum_{x_1} p(x_1)p(x_2|x_1) \\ &= 0.4\,p(x_2|0) + 0.6\,p(x_2|1) \\ &= 0.4\left(\frac{1}{8}+\frac{3x_2}{4}\right) + 0.6\left(\frac{1}{4}+\frac{x_2}{2}\right) \\ &= \boxed{\frac{1}{5}+\frac{3}{5}x_2} \end{aligned}$$

$$
\begin{aligned}
r_{f_3\to x_2}(x_2) &= \sum_{x_3} f_3(x_3,x_2)q_{x_3\to f_3}(x_3)\\
&= \sum_{x_3} f_3(x_3,x_2)r_{f_4\to x_3}(x_3)\\
&= \sum_{x_3} f_3(x_3,x_2)\left(\sum_{x_4} f_4(x_3,x_4)\,\underbrace{q_{x_4\to f_4}(x_4)}_{1}\right)\\
&= \sum_{x_3} p(x_3|x_2)\left(\underbrace{\sum_{x_4} p(x_4|x_3)}_{=1,\text{ Probability dist.}}\right)\\
&= \sum_{x_3} p(x_3|x_2) = 1\\
&= 1 \implies \boxed{Z_2(x_2) = \frac{1}{5}+\frac{3}{5}x_2 = \begin{cases}\frac{1}{5}, & x_2=0\\ \frac{4}{5}, & x_2=1\end{cases}}
\end{aligned}
$$

Important: Direct calculation of $p(x_2)$ is NOT accepted!

10.8 Summer Semester 2021 Exam

Exam Date: July 20th, 2021
6 Questions with total of 100 Points.
Exam Duration: 90 Minutes

10.8.1 Entropy and Mutual Information (18 Points)

Consider the two *independent* random variables X and $Y \in \{0, 1\}$ which are distributed as $X \sim$ Bernoulli(q_X) and $Y \sim$ Bernoulli(q_Y). The parameters q_X and q_Y are $q_X = q_Y = 0.5$.

(a) Determine $H(X)$.

♠ **Solution:**

$$H(X) = H_2\left(\frac{1}{2}\right) = 1$$

(b) Determine $I(X;Y)$.

♠ **Solution:**

$$I(X;Y) = H(X) - H(X|Y) \underbrace{=}_{\text{independence}} H(X) - H(X) = 0$$

We now define the random variable Z as $Z = X + Y$.

(c) Determine the probability distribution $P_Z(z)$, i.e., calculate $P_Z(z) = \Pr\{Z = z\}$ for $z \in \{0, 1, 2\}$.

♠ **Solution:**

$$P_Z(z) = \begin{cases} \Pr\{X=0\}\Pr\{Y=0\} = \frac{1}{4}, & \text{if } z = 0 \\ \Pr\{X=0\}\Pr\{Y=1\} + \Pr\{X=1\}\Pr\{Y=0\} = \frac{1}{2}, & \text{if } z = 1 \\ \Pr\{X=1\}\Pr\{Y=1\} = \frac{1}{4}, & \text{if } z = 2 \end{cases}$$

(d) Determine $H(X|Y, Z)$.

♠ **Solution:** Since $X = Z - Y$, we have

$$H(X|Y, Z) = 0$$

(e) Determine $I(X;Y|Z = 0)$ and $I(X;Y|Z = 1)$.

♠ **Solution:**

$$I(X;Y|Z=0) = H(X|Z=0) - \underbrace{H(X|Y,Z=0)}_{=0} = H(X|Z=0) = 0$$
$$I(X;Y|Z=1) = H(X|Z=1) - \underbrace{H(X|Y,Z=1)}_{=0} = H(X|Z=1) = 1$$

(f) Determine $I(X;Y|Z)$.

♠ **Solution:**

$$I(X;Y|Z=2) = 0 \qquad \text{(as in } (e))$$
$$I(X;Y|Z) = \sum_z P_Z(z) I(X;Y|Z=z) = \frac{1}{4} \times 0 + \frac{1}{2} \times 1 + \frac{1}{4} \times 0 = \frac{1}{2}$$

10.8.2 Bayesian Inference (22 Points)

You apply for a job in a network monitoring company. In the job interview, you are given with the following data stream.

$$D = 0110101100010110$$

You are further informed that D is generated by a data source S. Your task is to find a statistical model for the source S from the data stream D based on some given information and assumptions.

(a) Which kind of probability problem is to be solved in this task?

♠ **Solution:** It is a backward/inverse problem.

First, you are informed that there is *only* the source S in the network which can be described as an independent and identically distributed (i.i.d) Bernoulli source with parameter $q \in [0,1]$. This means that S generates an i.i.d sequence $S_0, S_1, \ldots,$ where $S_i \in \{0,1\}$ and Pr$\{S_i = 1\} = 1 - \{S_i = 0\} = q$. Assume that you have no prior information on the value of q.

(b) Considering the given assumptions, determine the *prior* probability density function (PDF) of q and the *likelihood* of q conditioned on the observation of the data stream D. Using the Bayes rule, formulate the *posterior* PDF of q conditioned on the data stream D, i.e., $P(q|D)$.

♠ **Solution:**

Prior:

$$f_q(q) = 1, \quad q \in [0,1]$$

Likelihood:

$$f_{D|q}(D|q) = q^{N_1(D)}(1-q)^{N_0(D)}, \qquad N_1(D) = \#\text{1's in } D, N_0(D) = \#\text{0's in } D$$

Posterior:

$$P(q|D) = \frac{f_{D|q}(D|q)f_q(q)}{\int_0^1 f_q(q')f_{D|q}(D|q')\,\mathrm{d}q'} = \frac{q^{N_1(D)}(1-q)^{N_0(D)}}{\int_0^1 (q')^{N_1(D)}(1-q')^{N_0(D)}\,\mathrm{d}q'}$$

(c) Assume D^{2N} is a binary sequence of length $2N$ whose N entries are 1 and N entries are 0. For this data stream, determine the limiting distribution to which $P(q|D^{2N})$ converges as $N \to \infty$.

♠ **Solution:**

As N goes to ∞, $P(q|D^{2N})$ approaches the Dirac delta at $q = \frac{1}{2}$, i.e.,

$$\lim_{N\to\infty} P(q|D^{2N}) = \delta\left(q - \frac{1}{2}\right)$$

You are now asked to consider the following alternative assumptions: It is assumed that there are *two* sources in the network, namely the sources $S^{(1)}$ and $S^{(2)}$. The source $S^{(1)}$ generates an i.i.d sequence of random variables $S_0^{(1)}, S_1^{(1)}, \ldots$ with parameter $p = 0.3$, i.e., $\Pr\{S_i^{(1)} = 1\} = 1 - \Pr\{S_i^{(1)} = 0\} = p$ for any i. $S^{(2)}$ generates a binary sequence $S_0^{(2)}, S_1^{(2)}, \ldots,$ whose distribution of is as follows:

$$\Pr\{S_i^{(2)} = 1\} = 1 - \Pr\{S_i^{(2)} = 0\} = \begin{cases} 0.5, & \text{if } i = 0, \\ 0.3, & \text{if } i \geq 1 \text{ and } S_{(i-1)}^{(2)} = 1, \\ 0.7, & \text{if } i \geq 1 \text{ and } S_{(i-1)}^{(2)} = 0. \end{cases}$$

You know that the data stream D is generated by one of these two sources. It is further believed that in this network, the chance that a data stream is generated by source $S^{(1)}$ is *twice higher* than the chance that the data stream is generated by the source $S^{(2)}$.

(d) Specify the source which has *most probably* generated the data stream D.

♠ **Solution:**

$$P(S = S^{(1)}|D) = \frac{P(D|S^{(1)})P(S^{(1)})}{Z}, \qquad \text{where } Z \text{ is the normalization constant.}$$

$$P(S^{(1)}) = \frac{2}{3}$$
$$P(D|S^{(1)}) = p^8(1-p)^8 = (0.3)^8 \times (0.7)^8 \approx 3.78 \times 10^{-6}$$

$$
\begin{aligned}
P(S = S^{(2)}|D) &= \frac{P(D|S^{(2)})P(S^{(2)})}{Z} \\
P(S^{(2)}) &= \frac{1}{3} \\
P(D|S^{(2)}) &= 0.5 \times 0.7 \times 0.3 \times 0.7 \times 0.7 \times 0.7 \times 0.7 \times 0.3 \times 0.7 \\
&\quad 0.3 \times 0.3 \times 0.7 \times 0.7 \times 0.7 \times 0.3 \times 0.7 \\
&= 0.5 \times 0.3^5 \times 0.7^10 \approx 3.43 \times 10^{-5}
\end{aligned}
$$

$$\Longrightarrow \frac{P(S = S^{(1)}|D)}{P(S = S^{(2)}|D)} < 1$$

$\Longrightarrow$ D was most likely generated by $S^{(2)}$.

10.8.3 Source Coding (13 Points)

Consider the following binary string

$$00111110010011$$

(a) Encode this string via the standard (basic) Lempel-Ziv algorithm, i.e., *no* further codebook optimization is needed.

♠ **Solution:**

Dictionary

Word	Pointer
λ	000
0	001
01	010
1	011
11	100
10	101
010	110
011	111

Encoded string:

$$(0)0 \mid 11 \mid 001 \mid 111 \mid 0110 \mid 0100 \mid 0101$$

(b) *Name* two possible approaches by which the standard Lempel-Ziv algorithm can be improved. Explain each approach briefly.

♠ **Solution:**

- Obsolete words can be deleted from the dictionary to keep the dictionary size small.
- Redundant bits can be omitted from the encoded string, i.e., the second time the same word from the dictionary is referenced, the additional bit is redundant.

(c) Name an advantage of the Lempel-Ziv algorithm against the arithmetic code. Also, name an advantage of the arithmetic code against the Lempel-Ziv algorithm.

♠ **Solution:**

- Lempel-Ziv can be used without knowing the source statistics, it is a general purpose compression algorithm.
- Arithmetic coding allows one to exploit knowledge of the source statistics, i.e., it is preferable over Lempel-Ziv if a (good) statistical model for the source is available.

10.8.4 Burrows-Wheeler Transform (10 Points)

Consider the following string

ABAC_BAC

(a) Determine the Burrows-Wheeler transform of this string. To this end, add a *termination character* first.

♠ **Solution:**

Termination character: $

ABAC_BAC$

Rotation table:

$ABAC_BAC
_BAC$ABAC
ABAC_BAC$
AC$ABAC_B
AC_BAC$AB
BAC$ABAC_
BAC_BAC$A
C$ABAC_BA
C_BAC$ABA

Transformed word:

CC$BB_AAA

(b) Explain why the Burrows-Wheeler transform is used along with source coding?

♠ **Solution:**

The BWT groups same characters together. This makes stream codes, such as the Lempel-Ziv, more effective.

10.8.5 Channel Coding (15 Points)

Consider the following discrete memoryless channel (DMC):

$$Y = X + Z \text{ (mod } I), \; I \geq 2$$

where mod I describes the *modulo* I addition, $X \in \{1, \ldots, I\}$ is the channel input, and $Z \in \{-1, 0, 1\}$ models random noise whose probability distribution is

$$\text{Pr}\{Z = z\} = \begin{cases} 0.1 & \text{if } z = -1 \\ 0.1 & \text{if } z = 1 \\ 0.8 & \text{if } z = 0 \end{cases} .$$

(a) Determine the capacity of this channel and an optimal input distribution $P_X(x)$ by which the channel capacity is achieved.

♠ **Solution:**

The channel is symmetric, hence

$$P_X(x) = \frac{1}{I} \quad \forall x \in \{1, \ldots, I\} \quad \text{is an optimal input distribution.}$$

The capacity of the channel is given by

$$\begin{aligned} C &= \max_{\tilde{P}_X} I(X;Y) \\ &= \max_{\tilde{P}_X} H(X) - H(X|Y) \\ &= \log_2(I) - 2 \cdot \frac{1}{10} \log_2(10) \cdot \frac{8}{10} \log_2\left(\frac{10}{8}\right) \\ &= \log_2\left(\frac{I}{10}\right) + \frac{12}{5} \end{aligned}$$

(b) For the distribution $P_X(x)$ determined in Part (a), define the *typical set* $T_{N\beta}$ which contains typical sequences $\boldsymbol{x} = [x_1, \ldots, x_N]$ of length N with tolerance level $\beta > 0$.

♠ **Solution:**

$$T_{N\beta} = \left\{ \boldsymbol{x} \in \{1,,\dots,I\}^N : \left| \frac{1}{N} \log_2 \left(\frac{1}{P_X(\boldsymbol{x})} \right) - H(X) < \beta \right| \right\}$$

where

$$\begin{aligned} H(X) &= \log_2(I), \\ P_X(\boldsymbol{x}) &= I^{-N} \end{aligned}$$

Hence,

$$\begin{aligned} T_{N\beta} &= \left\{ \boldsymbol{x} \in \{1,,\dots,I\}^N, : \quad \left| \frac{1}{N} \log_2 \left(I^N \right) - \log_2 I < \beta \right| \right\} \\ &= \{1,\dots,I\}^N \end{aligned}$$

We now set $I = 10$ and use the channel for N times.

(c) Give two different sequences of length $N = 10$, which for any tolerance level β are in typical set $T_{N\beta}$ (regardless of the value of β).

♠ **Solution:**

$$1, 1, 1, 1, 1, 1, 1, 1, 1, 1$$

$$2, 2, 3, 3, 4, 5, 6, 7, 1, 2$$

(d) For an arbitrary joint distribution $P_{X,Y}(x,y)$, define the *jointly typical set* $J_{N\beta}$ containing the sequence pairs $(\boldsymbol{x}, \boldsymbol{y}) = ([x_1,\dots,x_N],[y_1,\dots,y_N])$ of length N which are jointly typical with tolerance level $\beta > 0$.

♠ **Solution:**

$$J_{N\beta} = \left\{ (\boldsymbol{x},\boldsymbol{y}) \in \{1,\dots,I\}^N \times \{0,\dots,I-1\}^N : \left| \frac{1}{N} \log_2 \left(\frac{1}{P_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})} \right) - H(X,Y) \right| < \beta \right\}$$

Then,

$$J_{N\beta} = \left\{ (\boldsymbol{x},\boldsymbol{y}) \in \{1,\dots,10\}^N \times \{0,\dots,9\}^N : \left| \frac{1}{N} \log_2 \left(\frac{1}{P_{\boldsymbol{x},\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{y})} \right) - H(X,Y) \right| < \beta \right\}$$

With

$$P_{\boldsymbol{x},\boldsymbol{y}} = \prod_{i=1}^{N} P_{X,Y}(x_i, y_i)$$

(e) Now, let $P_{X,Y}(x, y)$ be given by the channel when we set the input distribution to be $P_X(x)$ calculated in Part (a). Give two pairs of typical sequences $(\boldsymbol{x}_1, \boldsymbol{y}_1)$ and $(\boldsymbol{x}_2, \boldsymbol{y}_2)$ of length $N = 10$, which are always in the jointly typical set $J_{N\beta}$ (regardless of the value of β).

♠ **Solution:**

$$\begin{aligned}\boldsymbol{x}_1 &= 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\\ \boldsymbol{y}_1 &= 2, 0, 1, 1, 1, 1, 1, 1, 1, 1\end{aligned}$$

$$\begin{aligned}\boldsymbol{x}_2 &= 2, 2, 2, 2, 2, 2, 2, 2, 2, 2\\ \boldsymbol{y}_2 &= 2, 2, 2, 2, 2, 2, 1, 3, 2, 2\end{aligned}$$

10.8.6 Sum-Product Algorithm (18 Points)

Consider the sum-product algorithm which operates on a factor graph with N variable nodes and M factor nodes. Let

- $q_{n\to m}(x_n)$ denote the message from variable node n to factor node m, and
- $r_{m\to n}(x_n)$ denote the message from factor node m to variable node n.

(a) Simplify $q_{n\to m}(x_n)$ for the case in which all variable nodes of the factor graph have degree 2, i.e., each variable node is connected only to 2 factor nodes.

♠ **Solution:**

$$q_{n\to m}(x_n) = r_{m'\to n}(x_n),$$

where m and m' are the indices of the two factor nodes connected to variable node n.

(b) Simplify $r_{m\to n}(x_n)$ for the case in which all factor nodes of the factor graph have degree 2, i.e., each factor node is connected only to 2 variable nodes.

♠ **Solution:**

$$r_{m\to n}(x_n) = \sum_{x_{n'}} f_m(x_{n'}, x_n)\, q_{n'\to m}(x_{n'}),$$

where n and n' are the indices of the two variable nodes connected to factor node m.

(c) What is the purpose of the sum-product algorithm, in general?

♠ **Solution:**

Compute the marginal of a factorizable function, e.g., the marginal of a A-Posteriori probability distribution.

(d) Under which condition is the algorithm *guaranteed* to serve its purpose?

♠ **Solution:**

If the factor graph does not contain cycles, i.e., if the factor graph is a tree.

(e) Under which condition is the algorithm *likely* to serve its purpose?

♠ **Solution:**

If the factor graph is sparse.

(f) Name a problem in communications where the sum-product algorithm is used.

♠ **Solution:**

Bitwise decoding / Decoding of LDPC codes.

10.9 Winter Semester 2021-2022 Exam

Exam Date: February 15th, 2022
5 Questions with total of 100 Points.
Exam Duration: 90 Minutes

10.9.1 Information Theoretic Metrics and Inequalities (22 Points)

Consider the discrete random variables X, Y, and Z which are defined as follows:

- X is uniformly distributed over alphabet $\mathcal{A}_X$ = {0, 1, 5, 9}.
- Y is the minimum number of required bits for binary representation of X.

> EXAMPLE: The binary representation of X = 5 is 101. Therefore, the minimum number of required bits to represent X = 5 is Y = 3.

- Z is the minimum number of required bits for binary representation of Y.

(a) Determine the entropy of Y; that is $H(Y)$.

♠ **Solution:** X and Y and Z are distributed as follows:

X	Pr{ X = x }
0	$1/4$
1	$1/4$
5	$1/4$
9	$1/4$

→

Y	Pr{ Y = y }
1	$1/2$
3	$1/4$
4	$1/4$

→

Z	Pr{ Z = z }
1	$1/2$
2	$1/4$
3	$1/4$

$\rightarrow H(Y) = \frac{1}{2} + \frac{1}{4} \times 2 + \frac{1}{4} \times 2$ = 1.5 bits

(b) Determine the conditional entropy of X given Y; that is $H(X|Y)$.

♠ **Solution:**

$$\begin{aligned} H(X|Y) &= \sum_{y\in\{1,3,4\}} H(X|Y=y)Pr\{Y=y\} \\ &= \frac{1}{2} \times H(X|Y=1) + \frac{1}{4} \times 0 + \frac{1}{4} \times 0 \end{aligned}$$

We know that

$$p(X|Y=1) = \begin{cases} \frac{1}{2} & x = 0 \\ \frac{1}{2} & x = 1 \\ 0 & x = 5, 9 \end{cases}$$

$\rightarrow H(X|Y=1) = H_2(\frac{1}{2}) = 1 \rightarrow H(X|Y) = \frac{1}{2}$ bit

(c) Determine the conditional entropy of X given Z; that is $H(X|Z)$.

♠ **Solution:** Two solutions: Either repeat the same approach as (b)

$$\rightarrow H(X|Z) = \frac{1}{2}$$

Or, note that $Z = f(Y)$ and $Y = g(Z)$

$$\rightarrow H(X|Z) = H(X|Y) = \frac{1}{2} \text{ bit}$$

(d) Compare $H(X|Z)$ with $H(X|Y)$, and justify your observation.

♠ **Solution:** $H(X|Z) = H(X|Y)$, because Z is a one-to-one function of Y: $Y = f(Z)$ and $Z = f(Y)$

(e) Determine the mutual information between X and Y; that is $I(X;Y)$.

♠ **Solution:**

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \log 4 - 0.5 \\ &= 1.5 \end{aligned}$$

Or

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

Since

$$Y = f(X) \rightarrow H(Y|X) = 0 \rightarrow I(X;Y) = H(Y) = 1.5 \text{ bits}$$

(f) Determine the conditional information between X and Y given Z; that is $I(X;Y|Z)$.

♠ **Solution:**

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

Since Y is a one-to-one function of Z:

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

Thus:

$$I(X;Y|Z) = 0$$

(g) Compare $I(X;Y)$ with $I(X;Y|Z)$, and justify your observation.

♠ **Solution:**

$$I(X;Y) > I(X;Y|Z)$$

This is intuitive, since knowing Z reveals all information in Y; thus, knowing Y would not add any information about X.
One could also use data processing inequality.

10.9.2 Bayesian Inference (20 Points)

The opinion of a user about a purchased product is shown by the parameter $U \in \{-1, 0, +1\}$:

- U = -1 means that the user is not satisfied.
- U = 0 means that the user is indifferent.
- U = +1 means that the user is satisfied.

A computer program reads the comment of a user about a purchased product and gives the comment a label $D \in \{-1, +1\}$:

- D = -1 means that the comment is negative.
- D = + means that the comment is positive.

This label is related to the satisfaction parameter U of the user via the following generative model:

$$D = \text{sign}(U + W)$$

where W is a continuous random variable uniformly distributed on [-2, 2); that is

$$f_W(w) = \begin{cases} 0.25 & -2 \leq w < 2 \\ 0 & \text{otherwise} \end{cases}$$

and sign(x) determines the sign of x; that is

$$\text{sign}(x) = \begin{cases} -1 & x \leq 0 \\ +1 & x > 0 \end{cases}$$

You are provided with the label D of the comment by a user about a particular product and asked to infer the satisfaction parameter of the user about this product from the given data. You are further informed about the following prior belief:

> Only 5% of the users are not satisfied about the product while 80% of the users seem to be indifferent after purchasing

(a) Formulate your task as a Bayesian inference problem. Specify the prior probability distribution and the likelihood.

♠ **Solution:**
Bayesian inference:
Parameter = U
Data = D
Prior: Pr{U = 0} = 0.8, Pr{U = -1} = 0.05, Pr{U = +1} = 0.15

Likelihood: Pr{D = d | U = u}

u = 0 → Pr{D = ± 1 | U = 0} = 0.5
u = 1 → Pr{D = 1 | U = 1} = 0.75
u = -1 → Pr{D = -1 | U = -1} = 0.25

(b) Assume that the label is $D = \pm 1$. Determine the likelihood for all outcomes of the satisfaction parameter U.

♠ **Solution:** As indicated is part(a);

$$\begin{aligned} Pr\{D = +1 | U = +1\} &= 0.75 \\ Pr\{D = +1 | U = 0\} &= 0.5 \\ Pr\{D = +1 | U = -1\} &= 0.25 \end{aligned}$$

(c) What do you infer, if you use maximum likelihood approach?

♠ **Solution:**

$$\hat{U} = +1 = \underset{u}{\operatorname{argmax}}\, Pr\{D = +1 | U = u\}$$

(d) For D = +1, determine the posterior probability distribution of the satisfaction parameter U.

♠ **Solution:**

$$Pr\{U = u | D = +1\} = \frac{Pr\{D = +1 | U = u\} Pr\{U = u\}}{\sum_u Pr\{D = +1 | U = u\} Pr\{U = u\}}$$

Given the prior distribution, we have

$$\begin{aligned} Pr\{D = 1\} &= 0.75 Pr\{U = 1\} + 0.5 Pr\{U = 0\} + 0.25 Pr\{U = -1\} \\ &= 0.75 \times 0.15 + 0.5 \times 0.8 + 0.25 \times 0.05 \\ &= 0.525 \end{aligned}$$

$$\begin{aligned} Pr\{U=1|D=1\} &= \frac{0.15}{0.525} Pr\{D=1|U=1\} \\ &= 0.215 \end{aligned}$$

$$\begin{aligned} Pr\{U=0|D=1\} &= \frac{0.8}{0.525} Pr\{D=1|U=0\} \\ &= 0.76 \end{aligned}$$

$$\begin{aligned} Pr\{U=-1|D=1\} &= \frac{0.05}{0.525} Pr\{D=1|U=-1\} \\ &= 0.025 \end{aligned}$$

(e) What do you infer, if you use maximum-a-posteriori approach?

♠ **Solution:**

$$\hat{U} = 0 = \operatorname*{argmax}_{u} Pr\{U=u|D=+1\}$$

(f) Compare the inferred satisfaction parameters in Parts (c) and (e). Which one is more reliable from a Bayesian viewpoint? Explain briefly your answer.

♠ **Solution:** (e) is more reliable since it takes into account the prior belief. The ML detection assumes a uniform prior which is wrong.

10.9.3 Source Coding (25 Points)

A source X^N is generated independent and identically distributed (i.i.d.) from the discrete random variable X which is distributed on the alphabet $\mathcal{A}_X = \{A, B, C, D\}$ with the following probability distribution

$$P_X(x) = \begin{cases} \frac{1}{3} & x = A, B \\ \frac{1}{6} & x = C, D \end{cases}$$

We intend to design a uniquely-decodable source code to compress X^N into a sequence of M bits. The compression rate is defined as

$$R = \frac{M}{N}.$$

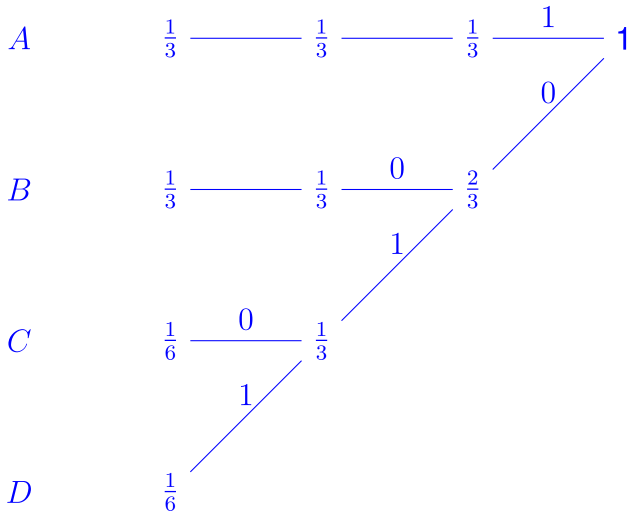
(a) Determine the minimum compression rate given by Shannon's theorem.

♠ **Solution:** The minimum compression rate is

$$
\begin{aligned}
R_{min} = H(X) &= 2 \times \frac{1}{3} \times \log 3 + 2 \times \frac{1}{6} \times \log 6 \\
&= \frac{2}{3} \log 3 + \frac{1}{3}(\log 3 + 1) \\
&= \log 3 + \frac{1}{3} \\
&= 1.9183
\end{aligned}
$$

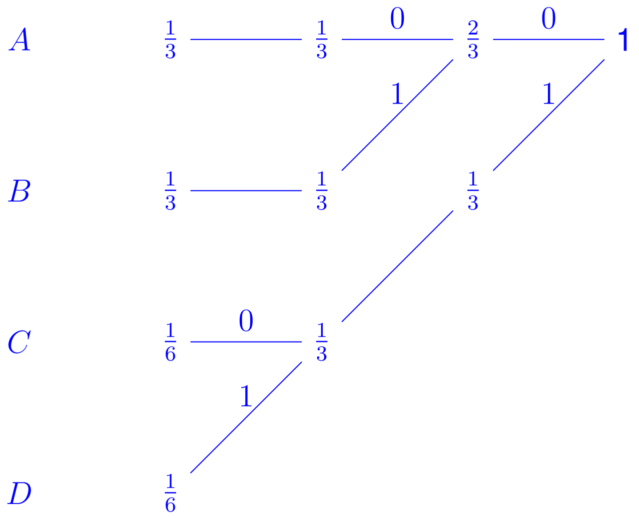
(b) Give a Huffman code for this source.

♠ **Solution:** There are various possible solutions. Here, we give two solutions:
Code 1:



$$
\begin{aligned}
A &\rightarrow 1 \\
B &\rightarrow 00 \\
C &\rightarrow 010 \\
D &\rightarrow 011
\end{aligned}
$$

Code 2:



$$
\begin{aligned}
A &\rightarrow 00\\
B &\rightarrow 01\\
C &\rightarrow 10\\
D &\rightarrow 11
\end{aligned}
$$

Other possible solutions are also correct.

(c) Determine the average length of the Huffman code given is Part (b).

♠ **Solution:** For any variant of Huffman Code, we get $L(X, C) = 2$.

(d) Compare the average length of Part (c) with the minimum compression rate which you determined in Part (a). Explain your observation shortly.

♠ **Solution:**

$$L(X, C) > H(X) = R_{min}$$

This is due to the fact that $\log \frac{1}{P_X(x)}$ are not integers.
We now use the Huffman code of Part (b) to encode the following sequence of length N = 6.

$$x^6 = DDDDDD$$

(e) Write down the encoded binary sequence.

♠ **Solution:**

If we use code 1 → C(x^6) = 011 011 011 011 011 011
If we use code 2 → C(x^6) = 11 11 11 11 11 11

(f) Determine the compression rate which you achieve by encoding this particular sequence and compare it to the average length of Part(c). Give a reason for your observation.

♠ **Solution:**
For code 1:

$$R = \frac{M}{N} = \frac{18}{6} = 3 > L(X, C)$$

This is due to the fact that X^6 is not typical.
For code 2:

$$R = \frac{M}{N} = 2 = L(X, C)$$

This is due to the fact that code 2 has a fixed length.

(g) Give a sequence of length N = 6, such that using Huffman code of Part (b , a compression rate equal to the average length of Part (c) is achieved.

♠ **Solution:** For Code 1: It must be typical. For example,

$$x^6 = AABBCD$$

For Code 2: Any sequence is correct.

(h) What property does your suggested sequence in Part (g) have? Is the compression rate equal to the average length for every sequence with this property? Justify your answer briefly.

♠ **Solution:** For Code 1: Typicality, and yes. A typical sequence has the same empirical distribution as the true distribution and hence, we have always R = L(X; C)
For Code 2: Any sequence and yes. Because the code has a fixed length.

10.9.4 Channel Coding (15 Points)

The discrete input of a communication channel X is selected from the input alphabet $\mathcal{A}_X = \{1, 2\}$ and is related to the output of the channel as follows:

$$Y = (Z^X)^{\frac{1}{X}}$$

Here, Z is a continuous random noise term which is distributed uniformly on the interval $[-0.5, 0.5)$. This means that the probability density function (PDF) of Z is given by

$$f_Z(z) = \begin{cases} 1 & -0.5 \leq z < 0.5 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the input distribution which achieves the channel capacity.

Hint: The following items might be useful:

- The derivative of the binary entropy function $H_2(x)$ with respect to x is

$$\frac{\partial}{\partial x} H_2(x) = \log_2(\frac{1-x}{x})$$

- Note that $\sqrt{x^2} = |x|$

♠ **Solution:**

$$\star I(X;Y) = h(Y) - h(Y|X)$$
$$\star h(Y|X) = \sum p_X(x) h(Y|X=x)$$

Assume a general input distribution:

$$x = \begin{cases} 1 & q \\ 2 & 1-q \end{cases}$$

Then, for x=1 $\rightarrow Y = Z$

$$h(Y|X=1) = h(Z) = 0$$

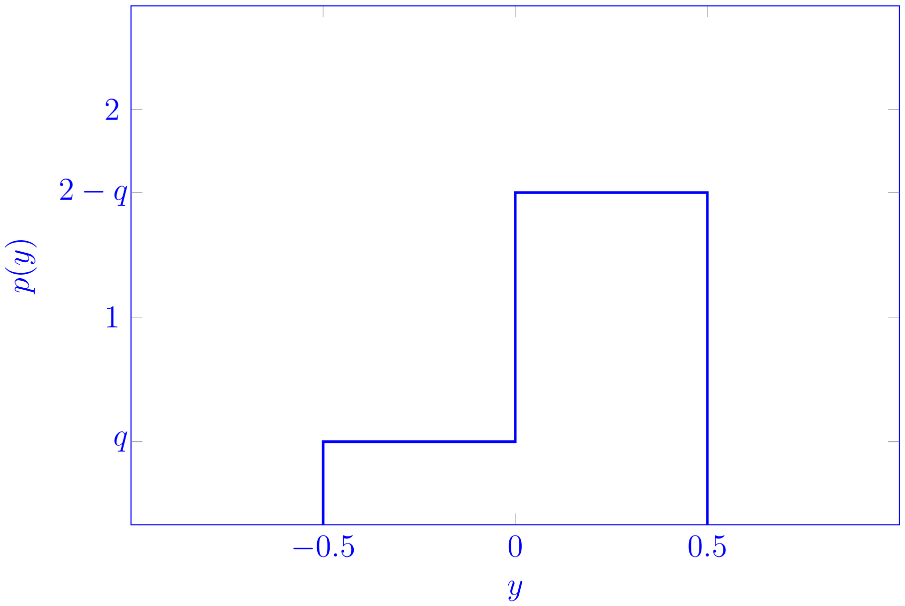
for x = 2 $\rightarrow Y = \sqrt{Z^2} = |Z|$,

$$h(Y|X=2) = h(|Z|)$$

if Z ~Uniform[-0.5, 0.5) $\rightarrow |Z|$ ~Uniform[0, 0.5]

$$\rightarrow h(Y|X=2) = h(|Z|) = \log\frac{1}{2} = -1$$
$$\rightarrow h(Y|X) = q \times 0 + (1-q) \times (-1) = q - 1$$

$$p(y) = \sum p_X(x) p(y|x)$$
$$= q \times Unif[-0.5, 0.5) + (1-q) \times Unif[0, 0.5)$$



$$\begin{aligned}
\rightarrow h(Y) &= \int_{-0.5}^{0} q\log\frac{1}{q}dy + \int_{0}^{0.5}(2-q)\log\frac{1}{2-q}dy \\
&= \frac{1}{2}q\log\frac{1}{q} + \frac{1}{2}(2-q)\log\frac{1}{2-q} \\
&= \frac{q}{2}[\log\frac{2}{q}-1] + (1-\frac{q}{2})[\log\frac{1}{1-\frac{q}{2}}-1] \\
&= \frac{q}{2}\log\frac{1}{\frac{q}{2}} + (1-\frac{q}{2})\log\frac{1}{1-\frac{q}{2}} - \frac{q}{2} + \frac{q}{2} - 1
\end{aligned}$$

$$h(Y) = H_2(\tfrac{q}{2}) - 1 \rightarrow I(X;Y) = H_2(\tfrac{q}{2}) - q$$

(b) Determine the capacity of this channel.

♠ **Solution:** To find the Capacity, we write:

$$C = \max_q I(X;Y) = I(X;Y)|_{q=q^*}$$

To find q^*, we have:

$$\begin{aligned}
\frac{\partial}{\partial q}I(X;Y) &= 0 \\
\frac{1}{2}\log(\frac{1-\frac{q^*}{2}}{\frac{q^*}{2}}) - 1 &= 0 \\
q^* = \frac{2}{5} &= 0.4
\end{aligned}$$

$$C = H_2(0.2) - 0.4 = 0.3219$$

Clearly, the capacity achieving distribution is

$$X = \begin{cases} 1 & 0.2 \\ 2 & 0.8 \end{cases}$$

10.9.5 Question 5: Parity Check Codes and Factor Graphs

A parity check code encodes a sequence of K = 6 information bits to a binary codeword of length N = 11. The parity check equations of this code are given below:

$$\begin{aligned} \mathrm{x}_2 \oplus \mathrm{x}_3 \oplus \mathrm{x}_4 \oplus \mathrm{x}_7 \oplus \mathrm{x}_9 &= 0 \\ \mathrm{x}_2 \oplus \mathrm{x}_4 \oplus \mathrm{x}_6 \oplus \mathrm{x}_8 \oplus \mathrm{x}_{10} &= 0 \\ \mathrm{x}_1 \oplus \mathrm{x}_2 \oplus \mathrm{x}_4 \oplus \mathrm{x}_5 \oplus \mathrm{x}_9 \oplus \mathrm{x}_{10} &= 0 \\ \mathrm{x}_2 \oplus \mathrm{x}_3 \oplus \mathrm{x}_5 \oplus \mathrm{x}_6 \oplus \mathrm{x}_{10} \oplus \mathrm{x}_{11} &= 0 \\ \mathrm{x}_1 \oplus \mathrm{x}_2 \oplus \mathrm{x}_4 \oplus \mathrm{x}_5 \oplus \mathrm{x}_7 \oplus \mathrm{x}_8 \oplus \mathrm{x}_{11} &= 0 \end{aligned}$$

(a) Write down the number of codewords and the code rate for this code.

♠ **Solution:**

$$K = 6 \rightarrow \#\text{of codewords} = 2^6 = 64$$
$$R = \frac{K}{N} = \frac{6}{11} = 0.5455$$

(b) Write the parity check matrix of this code.

♠ **Solution:**

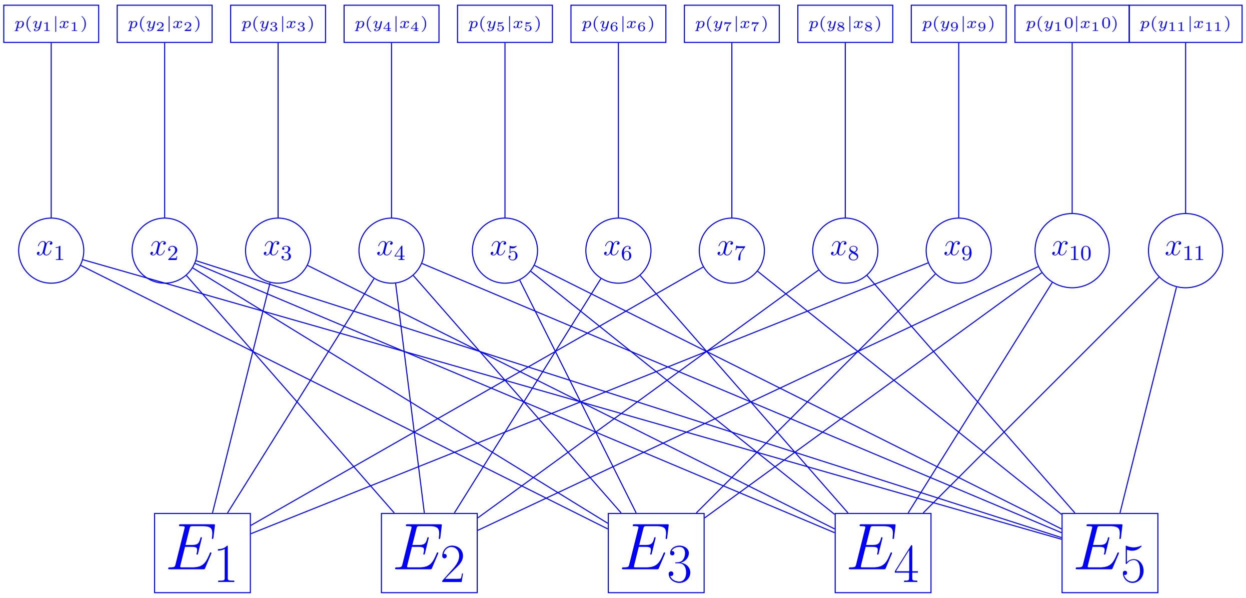
$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

(c) Is this code a regular low-density parity check (LDPC) code? Give a reason for your answer.

♠ **Solution:** No, $\#$ of 1 s in the rows / columns are not the same.

(d) Draw the factor graph for this code and specify all the variable and factor nodes.

♠ **Solution:**



Now assume that the factor graph of Part (d) represents the function F(x), where

$$\mathbf{x} = \left[\begin{array}{ccccccccccc} \mathrm{x}_1 & \mathrm{x}_2 & \mathrm{x}_3 & \mathrm{x}_4 & \mathrm{x}_5 & \mathrm{x}_6 & \mathrm{x}_7 & \mathrm{x}_8 & \mathrm{x}_9 & \mathrm{x}_{10} & \mathrm{x}_{11} \end{array} \right]^T.$$

We intend to calculate the marginalization function

$$Z(x_1) = \sum_{\mathbf{x} \setminus x_1} F(\mathbf{x})$$

Answer the following items:

(e) Can we necessarily calculate $Z(x_1)$ exactly by applying the sum product algorithm on the factor graph of Part (d)? Give a reason for your answer.

♠ **Solution:** No, the factor graph has loops.

(f) Can we calculate $Z(x_1)$ approximately by applying the sum product algorithm on the factor graph of Part (d)?

- If yes, explain briefly how we can do it. You need not give any derivations.
- If no, give a reason for your answer.

♠ **Solution:** Yes, we can do it iteratively. We start with some initial values for priors. We update posteriors and take them as priors for next iterations. We keep on updating, till the updated probabilities are almost the same as the ones in previous steps.

Chapter 11

Sample Exams Without Solutions

To have further practice, some older sample exams are given in this chapter. These exams have no solutions and hence can be used to test your self.

11.1 Summer Semester 2016 Exam

Exam Date: July 19, 2016
8 Short Problems with total of 78 Credits.
Exam Duration: 90 Minutes

11.1.1 Mixed Questions (14 Credits)

(a) Which distribution maximizes the entropy of a discrete, finite and memoryless random variable? How large is the entropy in this case?

(b) When is a code *prefixfree*? Why is this property important?

(c) What is the statement of the source coding theorem?

(d) What is the statement of the sampling theorem?

(e) You want to determine the entropy of the English language. Why is it not sufficient to approximate each letter's probability by the relative frequency in a sufficiently long text and then use these values for the entropy formula?

(f) Given are three discrete random variables X,Y and Z. It holds that $I(Y;Z|X) = 0$. Simplify

$$H(X,Y,Z) + H(X) - H(X,Y) - H(X,Z)$$

as much as possible.

11.1.2 Probability Calculation and Inference (5 Credits)

In a shared apartment, Annette, Bernd and Christian take turns in doing the dishes. 60% of the time, Bernd does the dishes, Annette does it 30% of the time and Christian does the rest. The probability that a piece of dishes breaks is 0.05 if Annette does the dishes, 0.1 if Bernd does it and 0.2 if Christian does it.

(a) You can hear a clanking noise from the kitchen. What is the probability that Bernd is doing the dishes?

(b) What is the probability that Annette does the dishes and that she breaks something?

11.1.3 Huffman Code (11 Credits)

A memoryless source emits symbols from the Alphabet $\mathcal{A}_X = \{A, B, C\}$ with $Pr(A) = 0.6,\ Pr(B) = 0.3,\ Pr(C) = 0.1$.

(a) Calculate the entropy of the source.

(b) Create a binary Huffman code for this source. Use source words of length 2 for your code. What is the average code word length of your code?

(c) We want to construct a Huffman code for source words of length 2 again. This time, the code alphabet is quaternary (Codesymbols 0, 1, 2, 3). Construct the code. What is the average code word length?

(d) Compare the average code word lengths from parts (b) and (c). Does it make sense to compare these values? Explain your answer.

11.1.4 Arithmetic Coding (12 Credits)

(a) Name three significant differences between a Huffman code and an arithmetic code.

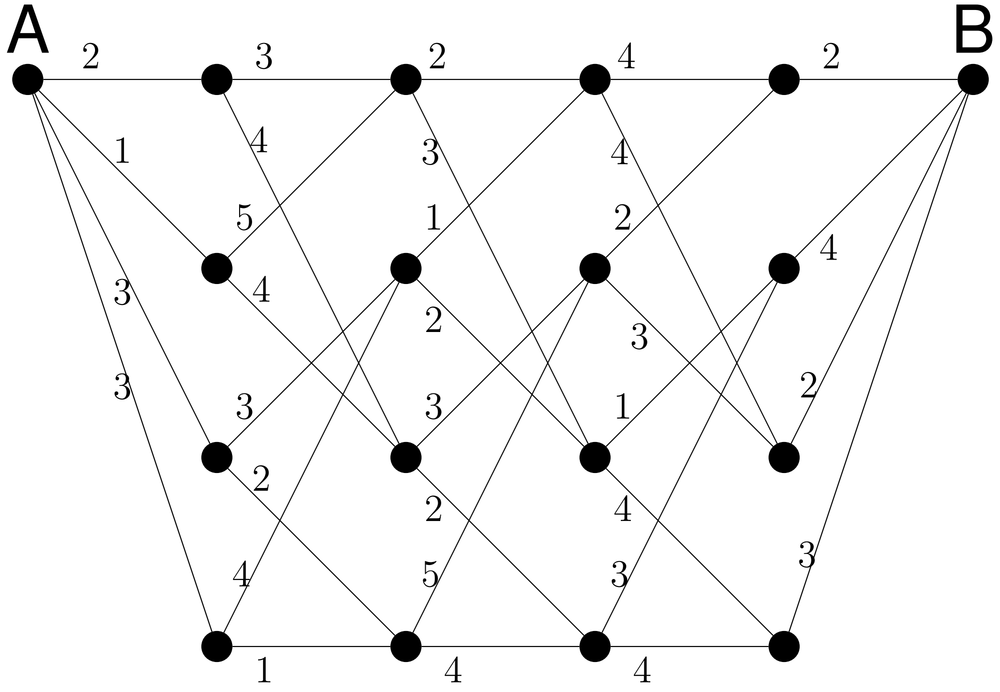
(b) Explain the encoding progress of any given data sequence by way of an arithmetic encoder.

(c) A binary source emits symbols from the alphabet $\{A, B\}$. It is known that no point in time, the number of emitted symbols A may differ by more than 1 from the number of emitted symbols B.
You want to encode a source sequence of length 5. Give the probability intervals of all possible sequences.

(d) How can you generate binary random numbers following an arbitrary distribution by means of an arithmetic code?

11.1.5 Message Passing (9 Credits)

The following graph is given. All connections may only be traversed from left to right. The numbers on each edge correspond to the costs of these connections.

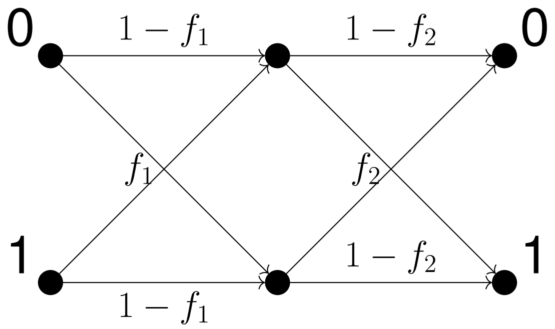


(a) How many different paths from A to B are there?

(b) What is the cheapest path from A to B? Draw the correct path and calculate its cost.

11.1.6 Channel Capacity (12 Credits)

(a) By concatenating two binary symmetric channels (BSCs) with error probabilities f_1 and f_2, the following new channel is generated.



Calculate the capacity of this new channel.

Given is a complex-valued, memoryless channel with multiplicative noise,

$$Y = e^{\psi i} X$$

with input X and output Y. ψ is continuous and uniformly distributed on $[0, 2\pi]$.

(b) At first, X is uniformly distributed over the 4-PSK-alphabet $\{+1, -1, +i, -i\}$. Calculate the mutual information $I(X;Y)$ for this input distribution.

(c) You can now freely choose the input distribution of X. Give the channel capacity as well as an input distribution which achieves capacity. Can you achieve this capacity in a practical communication system? Explain your answer.

11.1.7 Low Density Parity Check Codes (9 Credits)

The parity check matrix of an LDPC code is given as

$$H = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

(a) What is the rate of this code?

(b) Draw the factor graph for this code.

(c) A code word of this code is transmitted over the binary erasure channel (BEC). The receiver detects the sequence

$$1??0??101$$

Which code word was transmitted with highest probability?

11.1.8 Burrows-Wheeler-Transform (6 Credits)

The Burrows-Wheeler-Transformation of a sequence is

rfnzieen

with index 5 (counting starts at 1). What is the original sequence?

11.2 Winter Semester 2016-2017 Exam

Exam Date: February 14th, 2017
9 Short Problems with total of 86 Credits.
Exam Duration: 90 Minutes

11.2.1 Mixed Questions (16 Credits)

(a) Below are four (in-)equalities which lack the relation symbol ($\leq, =, \geq$). Insert the correct relation and explain why it holds. X, Y, Z are all *discrete* random variables.

(1) $H(X) \quad (?) \quad H(X|Y)$

(2) $I(XY; YZ) \quad (?) \quad I(X; YZ)$

(3) $I(X; Y) \quad (?) \quad I(f(X); Y)$ with arbitrary function $f \in \mathbb{R}$ on X

(4) $H(X) \quad (?) \quad H(aX)$ with real constant $a \in \mathbb{R}\ \{0\}$

(b) What is a code? Give the correct definition.

(c) Name one reason why additive noise is often modeled as Gaussian.

(d) The sequence 00001100100010100001 is given. Has this sequence already been source encoded? Explain your answer.

(e) Two discrete random variables X, Y with $|\mathcal{A}_X| = 3$ and $|\mathcal{A}_Y| = 4$ are given. Give the tightest upper and lower bound for $I(X; Y)$, that you can derive form this information.

11.2.2 Asymptotic Equipartition Principle / Source Coding (8 Credits)

(a) What is the difference between a *typical set* and a *smallest δ-sufficient subset*?

(b) Why is the typical set used for the proof of the source coding theorem?

(c) Write down a typical sequence of length $N = 10$ with parameter $\beta = 0.05$ for a binary source X with Pr$(X = A) = 0.2$ and Pr$(X = B) = 0.8$

(d) Imagine a yes/no game: You have to guess a certain person, but you may only ask questions which can be answered with yes or no. What is the optimal strategy for this game to find the person with as few questions as possible?

11.2.3 Inference (4 Credits)

Color blindness is a congenital impairment of the eyes. It occurs in 9% of boys, but only in 0.6% of girls. A newborn is male with a probability of 51% and with a probability of 49% it is female. A mother tells you that her child suffers from the color blindness. What is the probability that her child is a boy?

11.2.4 Arithmetic Coding (11 Credits)

A ternary source X with Pr$(X = A) = \frac{1}{2}$, Pr$(X = B) =$ Pr$(X = C) = \frac{1}{4}$ is given. It is known in advance that only sequences of length 4 will be encoded using an arithmetic code.

(a) Give a source interval for the sequence $ABAC$.

(b) Give a possible binary code word of minimal length for the sequence $ABAC$.

(c) Is it possible to encode arbitrarily close to entropy using arithmetic coding? Explain your answer.

11.2.5 Source Coding (17 Credits)

(a) Name one scenario where Huffman coding is superior to arithmetic coding and one where the opposite is the case. Also explain why each code is superior in that scenario.

(b) A binary source X with Pr$(X = A) = \frac{3}{4}$, Pr$(X = B) = \frac{1}{4}$ is given. Form source words of length 3 and construct a *ternary* Huffman code. Calculate the average code word length.

(c) The sequence 010001101000101100001001 is given. This sequence shall be encoded twice using the Lempel-Ziv algorithm. First, encode the sequence employing a Lempel-Ziv algorithm of your choice. Afterwards, perform a second encoding procedure. This time, the code sequence has to be shorter than the original one.

11.2.6 Factor Graphs (11 Credits)

You are given the function

$$f(a, b, c, d, e, f) = \frac{e}{a+b} \sin(a^2cd^2) \tanh(b+c) \frac{1}{de+f}$$

(a) Draw the factor graph for the function $f(a, b, c, d, e, f)$. Do not forget to label the nodes.

(b) The graph from question (a) is interpreted as the factor graph of the parity-check matrix **H** of an LDPC code. Write down **H**. What is the code rate of this code?

(c) We now compare two decoding procedures for the LDPC code from question (b). The code shall be decoded via iterative Belief Propagation decoding and Maximum A-Posteriori decoding. Which approach yields the better results for this code? Explain your answer.

11.2.7 Channel Capacity (9 Credits)

We consider a *continuous* modulo channel $Y = (X + N)$ mod 5 with input X, output Y and noise N. N is continuous and uniformly distributed on $[0, \frac{3}{2})$.

(a) Calculate the mutual information for a binary input distribution Pr$(X = 0) = \frac{1}{2} =$ Pr$(X = 2)$

(b) We now allow an arbitrary input alphabet with an arbitrary distribution, including continuous ones. Calculate the capacity of this channel. Give both a continuous and a discrete distribution, which achieve capacity.

11.2.8 Burrows-Wheeler-Transform (4 Credits)

Calculate the Burrows-Wheeler-Transform of

$$shannon$$

Indexing should start at 1.

11.2.9 Binary Codes (6 Credits)

(a) What is the advantage of *perfect* codes compared to any other error correcting codes?

(b) Someone asks you to create a *perfect* linear code with block length $N = 10$ and rate $R = \frac{2}{5}$ with a minimum distance of *at least* 5. Prove that such a code cannot exist.

11.3 Summer Semester 2017 Exam

Exam Date: August 1st, 2017
8 Short Problems with total of 81 Credits.
Exam Duration: 90 Minutes

11.3.1 Entropy and Mutual Information (15 Credits)

(a) Below are four (-in)equalities which lack the relation symbol $(\leq, =, \geq)$. Insert the correct relation and explain why it holds. X, Y, Z are all *discrete* random variables.

(1) $H(X) \quad (?) \quad H(X|Y)$

(2) $H(XY) \quad (?) \quad H(X)$

(3) $H(XY) \quad (?) \quad H(X) + H(Y)$

(4) $H(X) \quad (?) \quad H(aX)$ with constant value $a \in \mathbb{R} \setminus 0$

(b) State which condition(s) have to hold that the first three inequalities from part (a) are fulfilled with equality.

(c) The chain rule of mutual information is given as $I(XY; Z) = I(X; Z) + I(Y; Z|X)$. Derive the chain rule of mutual information by using the chain rule of entropy and the definition of mutual information.

(d) Assume that you receive a noise-free random sequence $X_k,\ k \in \mathbb{Z}$. You know that the elements have memory of length L, but L is unknown. How can you use mutual information to identify L?
Note: You can assume that you can measure the exact mutual information $I(X_k; X_l)$ for any values of k and l without error.

11.3.2 Inference (5 Credits)

Assume that an AIDS test has an error probability of 10^{-3}. Assume further that 1 out of 1000 people in Germany is infected with AIDS without showing visible symptoms.

(a) A randomly selected person, looking healthy, is tested positive for AIDS. What is the probability that this person is infected?

(b) Assume that the same test is used to qualify people for blood donations. What is the probability that a blood bag (which only contains blood from a single donor) contains the AIDS virus?

11.3.3 Prefix-free Codes (9 Credits)

A source X emits symbols from the alphabet $\mathcal{A}_X$ with $|\mathcal{A}_X| = 8$. You want to construct a prefix-free source code for this source.

(a) What is the advantage of a prefix-free code over a non-prefix-free code?

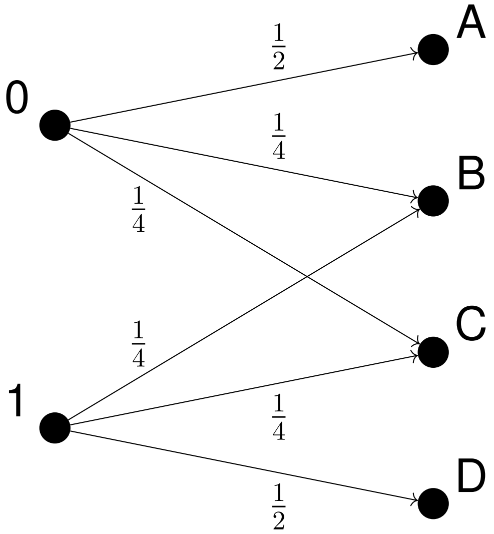
(b) We want to find a code with code word lengths $\{1, 3, 3, 3, 5, 5, 5, 5\}$. Can such a code exist? If yes, give such a code. If not, prove that.

(c) How should you map source symbols to code words to achieve the best compression?

(d) We now consider a code with code word lengths $\{1, 3, 3, 4, 5, 5, 5, 6\}$ for the source X. What condition of the input distribution has to be fulfilled that this particular code can compress down to entropy? Is this possible?

11.3.4 Channel Capacity (11 Credits)

The following channel transition diagram is given.



(a) Give the channel transition matrix.

(b) Is this channel symmetric or not? Explain your answer.
A simple "yes/no" answer without any explanation will give no points, even if it is correct.

(c) Calculate the capacity of this channel.

11.3.5 Huffman-Coding (8 Credits)

A source X with alphabet $\mathcal{A}_X = \{A, B, C, D, E, F\}$ and symbol probabilities according to the table below is given. We want to find a Huffman code for this source.

X	A	B	C	D	E	F
$\Pr(X=\cdot)$	0.3	0.1	0.05	0.4	0.075	0.075

(a) What is the entropy of this source?

(b) Create a Huffman code for this source and give the average code word length.

(c) Give the loss to entropy per source symbol. How can we reduce this loss?

11.3.6 Lempel-Ziv-Coding (11 Credits)

(a) Give one advantage and one disadvantage of Lempel-Ziv-coding compared to Huffman-coding.

(b) Using the standard Lempel-Ziv algorithm (i.e, without any removal of code words from the code book), the following code sequence was created:

$$0010100111001100011010110 00101$$

The source alphabet is binary. Give the original source sequence.

(c) We now consider a source with a ternary alphabet with elements $\{0,1,2\}$. Find a *binary* code sequence for the source sequence

$$012112121011022$$

using a Lempel-Ziv algorithm of your choice.

11.3.7 LDPC Codes (12 Credits)

The following parity-check matrix of an LDPC code is given:

$$H = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

(a) What is the code rate of the code defined by **H**?

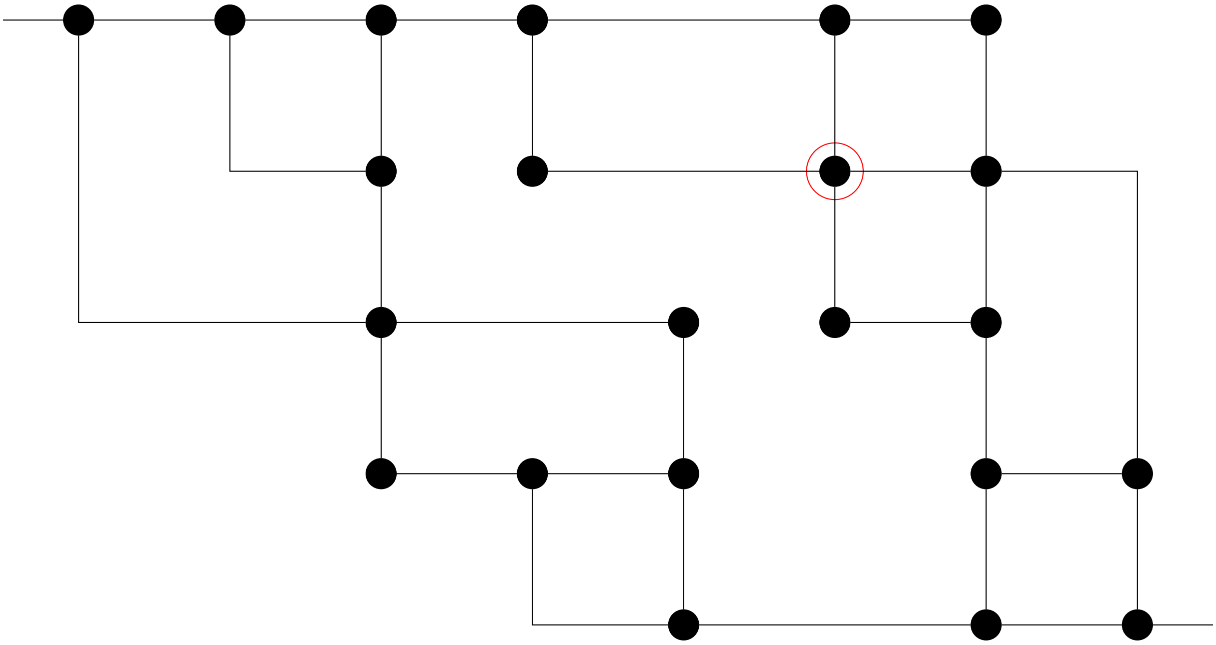
(b) Give one valid code word of the code defined by **H**.

(c) Draw the factor graph corresponding to **H**.

(d) Explain the idea of belief propagation decoding by explaining the different steps at variable and check nodes.
Note: You do not have to given any equations here. An intuitive explanation of the steps of this algorithm is sufficient.

11.3.8 Message Passing (10 Credits)

The following graph is given. In each node, you may only move either down or to the right.



(a) How many different paths from the node at the top left to the node at the bottom right are there?

(b) A path is chosen by uniformly deciding between going down or going right in each node (if possible). What is the probability of going through the encircled node?

(c) Now, a path is chosen by uniformly deciding between all possible paths. What is the probability of going through the encircled node in this case?

Abbreviations

BSC binary symmetric channel

AEP asymptotic equipartition property

i.i.d. identically and independently distributed

BEC binary erasure channel

CDF cumulative distribution function

PDF probability density function

AWGN additive white Gaussian noise

LDPC low density parity check

ML maximum likelihood