Elements of Information Theory
Second Edition

Solutions to Problems

Thomas M. Cover
Joy A. Thomas

October 17, 2006

Contents

Preface

Here we have the solutions to all the problems in the second edition of Elements of Information Theory. First a word about how the problems and solutions were generated.

The problems arose over the many years the authors taught this course. At first the homework problems and exam problems were generated each week. After a few years of this double duty, the homework problems were rolled forward from previous years and only the exam problems were fresh. So each year, the midterm and final exam problems became candidates for addition to the body of homework problems that you see in the text. The exam problems are necessarily brief, with a point, and reasonable free from time consuming calculation, so the problems in the text for the most part share these properties.

The solutions to the problems were generated by the teaching assistants and graders for the weekly homework assignments and handed back with the graded homeworks in the class immediately following the date the assignment was due. Homeworks were optional and did not enter into the course grade. Nonetheless most students did the homework. A list of the many students who contributed to the solutions is given in the book acknowledgment. In particular, we would like to thank Laura Ekroot, Will Equitz, Don Kimber, Mitchell Trott, Andrew Nobel, Jim Roche, Vittorio Castelli, Mitchell Oslick, Chien-Wen Tseng, Michael Morrell, Marc Goldberg, George Gemelos, Navid Hassanpour, Young-Han Kim, Charles Mathis, Styrmir Sigurjonsson, Jon Yard, Michael Baer, Mung Chiang, Suhas Diggavi, Elza Erkip, Paul Fahn, Garud Iyengar, David Julian, Yiannis Kontoyiannis, Amos Lapidoth, Erik Ordentlich, Sandeep Pombra, Arak Sutivong, Josh Sweetkind-Singer and Assaf Zeevi. We would like to thank Prof. John Gill and Prof. Abbas El Gamal for many interesting problems and solutions.

The solutions therefore show a wide range of personalities and styles, although some of them have been smoothed out over the years by the authors. The best way to look at the solutions is that they offer more than you need to solve the problems. And the solutions in some cases may be awkward or inefficient. We view that as a plus. An instructor can see the extent of the problem by examining the solution but can still improve his or her own version.

The solution manual comes to some 400 pages. We are making electronic copies available to course instructors in PDF. We hope that all the solutions are not put up on an insecure website—it will not be useful to use the problems in the book for homeworks and exams if the solutions can be obtained immediately with a quick Google search. Instead, we will put up a small selected subset of problem solutions on our website, http://www.elementsofinformationtheory.com, available to all. These will be problems that have particularly elegant or long solutions that would not be suitable homework or exam problems.

We have also seen some people trying to sell the solutions manual on Amazon or Ebay. Please note that the Solutions Manual for Elements of Information Theory is copyrighted and any sale or distribution without the permission of the authors is not permitted.

We would appreciate any comments, suggestions and corrections to this solutions manual.

Tom Cover
Durand 121, Information Systems Lab
Stanford University
Stanford, CA 94305.
Ph. 650-723-4505
FAX: 650-723-8473
Email: cover@stanford.edu

Joy Thomas
Stratify
701 N Shoreline Avenue
Mountain View, CA 94043.
Ph. 650-210-2722
FAX: 650-988-2159
Email: joythomas@stanfordalumni.org

Chapter 1

Introduction

Chapter 2

Entropy, Relative Entropy and Mutual Information

1. **Coin flips.** A fair coin is flipped until the first head occurs. Let X denote the number of flips required.

   (a) Find the entropy $H(X)$ in bits. The following expressions may be useful:

   $$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}, \qquad \sum_{n=0}^{\infty} nr^n = \frac{r}{(1-r)^2}.$$

   (b) A random variable X is drawn according to this distribution. Find an "efficient" sequence of yes-no questions of the form, "Is X contained in the set S?" Compare $H(X)$ to the expected number of questions required to determine X.

   Solution:

   (a) The number X of tosses till the first head appears has the geometric distribution with parameter $p = 1/2$, where $P(X = n) = pq^{n-1}$, $n \in \{1, 2, \ldots\}$. Hence the entropy of X is

   $$\begin{aligned} H(X) &= -\sum_{n=1}^{\infty} pq^{n-1} \log(pq^{n-1}) \\ &= -\left[\sum_{n=0}^{\infty} pq^n \log p + \sum_{n=0}^{\infty} npq^n \log q \right] \\ &= \frac{-p \log p}{1-q} - \frac{pq \log q}{p^2} \\ &= \frac{-p \log p - q \log q}{p} \\ &= H(p)/p \text{ bits.} \end{aligned}$$

   If $p = 1/2$, then $H(X) = 2$ bits.

(b) Intuitively, it seems clear that the best questions are those that have equally likely chances of receiving a yes or a no answer. Consequently, one possible guess is that the most "efficient" series of questions is: Is $X = 1$? If not, is $X = 2$? If not, is $X = 3$? ... with a resulting expected number of questions equal to $\sum_{n=1}^{\infty} n(1/2^n) = 2$. This should reinforce the intuition that $H(X)$ is a measure of the uncertainty of X. Indeed in this case, the entropy is exactly the same as the average number of questions needed to define X, and in general $E(\#\text{ of questions}) \geq H(X)$. This problem has an interpretation as a source coding problem. Let $0 = $no, $1 = $yes, $X = $Source, and $Y = $Encoded Source. Then the set of questions in the above procedure can be written as a collection of (X,Y) pairs: $(1,1)$, $(2,01)$, $(3,001)$, etc. . In fact, this intuitively derived code is the optimal (Huffman) code minimizing the expected number of questions.

2. **Entropy of functions.** Let X be a random variable taking on a finite number of values. What is the (general) inequality relationship of $H(X)$ and $H(Y)$ if

(a) $Y = 2^X$?

(b) $Y = \cos X$?

Solution: Let $y = g(x)$. Then

$$p(y) = \sum_{x:\, y=g(x)} p(x).$$

Consider any set of x's that map onto a single y. For this set

$$\sum_{x:\, y=g(x)} p(x) \log p(x) \leq \sum_{x:\, y=g(x)} p(x) \log p(y) = p(y) \log p(y),$$

since log is a monotone increasing function and $p(x) \leq \sum_{x:\, y=g(x)} p(x) = p(y)$. Extending this argument to the entire range of X (and Y), we obtain

$$\begin{aligned} H(X) &= -\sum_x p(x) \log p(x) \\ &= -\sum_y \sum_{x:\, y=g(x)} p(x) \log p(x) \\ &\geq -\sum_y p(y) \log p(y) \\ &= H(Y), \end{aligned}$$

with equality iff g is one-to-one with probability one.

(a) $Y = 2^X$ is one-to-one and hence the entropy, which is just a function of the probabilities (and not the values of a random variable) does not change, i.e., $H(X) = H(Y)$.

(b) $Y = \cos(X)$ is not necessarily one-to-one. Hence all that we can say is that $H(X) \geq H(Y)$, with equality if cosine is one-to-one on the range of X.

3. **Minimum entropy.** What is the minimum value of $H(p_1,...,p_n) = H(\mathbf{p})$ as $\mathbf{p}$ ranges over the set of n-dimensional probability vectors? Find all $\mathbf{p}$'s which achieve this minimum.

   Solution: We wish to find *all* probability vectors $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ which minimize

$$H(\mathbf{p}) = -\sum_i p_i \log p_i .$$

   Now $-p_i \log p_i \geq 0$, with equality iff $p_i = 0$ or 1. Hence the only possible probability vectors which minimize $H(\mathbf{p})$ are those with $p_i = 1$ for some i and $p_j = 0, j \neq i$. There are n such vectors, i.e., $(1,0,\ldots,0)$, $(0,1,0,\ldots,0)$, $\ldots$, $(0,\ldots,0,1)$, and the minimum value of $H(\mathbf{p})$ is 0.

4. **Entropy of functions of a random variable.** Let X be a discrete random variable. Show that the entropy of a function of X is less than or equal to the entropy of X by justifying the following steps:

$$\begin{aligned} H(X, g(X)) &\overset{(a)}{=} H(X) + H(g(X) \mid X) && (2.1)\\ &\overset{(b)}{=} H(X); && (2.2)\\ H(X, g(X)) &\overset{(c)}{=} H(g(X)) + H(X \mid g(X)) && (2.3)\\ &\overset{(d)}{\geq} H(g(X)). && (2.4) \end{aligned}$$

   Thus $H(g(X)) \leq H(X)$.

   Solution: *Entropy of functions of a random variable.*

   (a) $H(X, g(X)) = H(X) + H(g(X)|X)$ by the chain rule for entropies.

   (b) $H(g(X)|X) = 0$ since for any particular value of X, g(X) is fixed, and hence $H(g(X)|X) = \sum_x p(x) H(g(X)|X = x) = \sum_x 0 = 0$.

   (c) $H(X, g(X)) = H(g(X)) + H(X|g(X))$ again by the chain rule.

   (d) $H(X|g(X)) \geq 0$, with equality iff X is a function of $g(X)$, i.e., $g(.)$ is one-to-one. Hence $H(X, g(X)) \geq H(g(X))$.

   Combining parts (b) and (d), we obtain $H(X) \geq H(g(X))$.

5. **Zero conditional entropy.** Show that if $H(Y|X) = 0$, then Y is a function of X, i.e., for all x with $p(x) > 0$, there is only one possible value of y with $p(x,y) > 0$.

   Solution: *Zero Conditional Entropy.* Assume that there exists an x, say x_0 and two different values of y, say y_1 and y_2 such that $p(x_0, y_1) > 0$ and $p(x_0, y_2) > 0$. Then $p(x_0) \geq p(x_0, y_1) + p(x_0, y_2) > 0$, and $p(y_1|x_0)$ and $p(y_2|x_0)$ are not equal to 0 or 1. Thus

$$\begin{aligned} H(Y|X) &= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) && (2.5)\\ &\geq p(x_0)(-p(y_1|x_0) \log p(y_1|x_0) - p(y_2|x_0) \log p(y_2|x_0)) && (2.6)\\ > &> 0, && (2.7) \end{aligned}$$

since $-t \log t \geq 0$ for $0 \leq t \leq 1$, and is strictly positive for t not equal to 0 or 1. Therefore the conditional entropy $H(Y|X)$ is 0 if and only if Y is a function of X.

6. **Conditional mutual information vs. unconditional mutual information.** Give examples of joint random variables X, Y and Z such that

   (a) $I(X;Y \mid Z) < I(X;Y)$,

   (b) $I(X;Y \mid Z) > I(X;Y)$.

   Solution: *Conditional mutual information vs. unconditional mutual information.*

   (a) The last corollary to Theorem 2.8.1 in the text states that if $X \rightarrow Y \rightarrow Z$ that is, if $p(x,y \mid z) = p(x \mid z)p(y \mid z)$ then, $I(X;Y) \geq I(X;Y \mid Z)$. Equality holds if and only if $I(X;Z) = 0$ or X and Z are independent.

   A simple example of random variables satisfying the inequality conditions above is, X is a fair binary random variable and $Y = X$ and $Z = Y$. In this case,

$$I(X;Y) = H(X) - H(X \mid Y) = H(X) = 1$$

   and,

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

   So that $I(X;Y) > I(X;Y \mid Z)$.

   (b) This example is also given in the text. Let X, Y be independent fair binary random variables and let $Z = X + Y$. In this case we have that,

$$I(X;Y) = 0$$

   and,

$$I(X;Y \mid Z) = H(X \mid Z) = 1/2.$$

   So $I(X;Y) < I(X;Y \mid Z)$. Note that in this case X, Y, Z are not markov.

7. **Coin weighing.** Suppose one has n coins, among which there may or may not be one counterfeit coin. If there is a counterfeit coin, it may be either heavier or lighter than the other coins. The coins are to be weighed by a balance.

   (a) Find an upper bound on the number of coins n so that k weighings will find the counterfeit coin (if any) and correctly declare it to be heavier or lighter.

   (b) *(Difficult)* What is the coin weighing strategy for $k = 3$ weighings and 12 coins?

   Solution: *Coin weighing.*

   (a) For n coins, there are $2n + 1$ possible situations or “states”.

   - One of the n coins is heavier.
   - One of the n coins is lighter.
   - They are all of equal weight.

Each weighing has three possible outcomes - equal, left pan heavier or right pan heavier. Hence with k weighings, there are 3^k possible outcomes and hence we can distinguish between at most 3^k different "states". Hence $2n+1 \leq 3^k$ or $n \leq (3^k-1)/2$.

Looking at it from an information theoretic viewpoint, each weighing gives at most $\log_2 3$ bits of information. There are $2n+1$ possible "states", with a maximum entropy of $\log_2(2n+1)$ bits. Hence in this situation, one would require at least $\log_2(2n+1)/\log_2 3$ weighings to extract enough information for determination of the odd coin, which gives the same result as above.

(b) There are many solutions to this problem. We will give one which is based on the ternary number system.

We may express the numbers $\{-12,-11,\ldots,-1,0,1,\ldots,12\}$ in a ternary number system with alphabet $\{-1,0,1\}$. For example, the number 8 is (-1,0,1) where $-1\times 3^0+0\times 3^1+1\times 3^2=8$. We form the matrix with the representation of the positive numbers as its columns.

	1	2	3	4	5	6	7	8	9	10	11	12	
3^0	1	-1	0	1	-1	0	1	-1	0	1	-1	0	$\Sigma_1=0$
3^1	0	1	1	1	-1	-1	-1	0	0	0	1	1	$\Sigma_2=2$
3^2	0	0	0	0	1	1	1	1	1	1	1	1	$\Sigma_3=8$

Note that the row sums are not all zero. We can negate some columns to make the row sums zero. For example, negating columns 7,9,11 and 12, we obtain

	1	2	3	4	5	6	7	8	9	10	11	12	
3^0	1	-1	0	1	-1	0	-1	-1	0	1	1	0	$\Sigma_1=0$
3^1	0	1	1	1	-1	-1	1	0	0	0	-1	-1	$\Sigma_2=0$
3^2	0	0	0	0	1	1	-1	1	-1	1	-1	-1	$\Sigma_3=0$

Now place the coins on the balance according to the following rule: For weighing #i, place coin n

- On left pan, if $n_i=-1$.
- Aside, if $n_i=0$.
- On right pan, if $n_i=1$.

The outcome of the three weighings will find the odd coin if any and tell whether it is heavy or light. The result of each weighing is 0 if both pans are equal, -1 if the left pan is heavier, and 1 if the right pan is heavier. Then the three weighings give the ternary expansion of the index of the odd coin. If the expansion is the same as the expansion in the matrix, it indicates that the coin is heavier. If the expansion is of the opposite sign, the coin is lighter. For example, (0,-1,-1) indicates $(0)3^0+(-1)3+(-1)3^2=-12$, hence coin #12 is heavy, (1,0,-1) indicates #8 is light, (0,0,0) indicates no odd coin.

Why does this scheme work? It is a single error correcting Hamming code for the ternary alphabet (discussed in Section 8.11 in the book). Here are some details.

First note a few properties of the matrix above that was used for the scheme. All the columns are distinct and no two columns add to (0,0,0). Also if any coin is heavier, it will produce the sequence of weighings that matches its column in the matrix. If it is lighter, it produces the negative of its column as a sequence of weighings. Combining all these facts, we can see that any single odd coin will produce a unique sequence of weighings, and that the coin can be determined from the sequence.

One of the questions that many of you had whether the bound derived in part (a) was actually achievable. For example, can one distinguish 13 coins in 3 weighings? No, not with a scheme like the one above. Yes, under the assumptions under which the bound was derived. The bound did not prohibit the division of coins into halves, neither did it disallow the existence of another coin known to be normal. Under both these conditions, it is possible to find the odd coin of 13 coins in 3 weighings. You could try modifying the above scheme to these cases.

8. **Drawing with and without replacement.** An urn contains r red, w white, and b black balls. Which has higher entropy, drawing $k \geq 2$ balls from the urn with replacement or without replacement? Set it up and show why. (There is both a hard way and a relatively simple way to do this.)

Solution: *Drawing with and without replacement.* Intuitively, it is clear that if the balls are drawn with replacement, the number of possible choices for the i-th ball is larger, and therefore the conditional entropy is larger. But computing the conditional distributions is slightly involved. It is easier to compute the unconditional entropy.

- With replacement. In this case the conditional distribution of each draw is the same for every draw. Thus

$$X_i = \begin{cases} \text{red} & \text{with prob.} \frac{r}{r+w+b} \\ \text{white} & \text{with prob.} \frac{w}{r+w+b} \\ \text{black} & \text{with prob.} \frac{b}{r+w+b} \end{cases} \tag{2.8}$$

and therefore

$$H(X_i|X_{i-1},\ldots,X_1) = H(X_i) \tag{2.9}$$

$$= \quad \log(r+w+b) - \frac{r}{r+w+b}\log r - \frac{w}{r+w+b}\log w - \frac{b}{r+w+b}\log b \tag{2.10}$$

- Without replacement. The unconditional probability of the i-th ball being red is still $r/(r+w+b)$, etc. Thus the unconditional entropy $H(X_i)$ is still the same as with replacement. The conditional entropy $H(X_i|X_{i-1},\ldots,X_1)$ is less than the unconditional entropy, and therefore the entropy of drawing without replacement is lower.

9. **A metric.** A function $\rho(x,y)$ is a metric if for all x, y,

- $\rho(x,y) \geq 0$
- $\rho(x,y) = \rho(y,x)$

- $\rho(x,y) = 0$ if and only if $x = y$
- $\rho(x,y) + \rho(y,z) \geq \rho(x,z)$.

(a) Show that $\rho(X,Y) = H(X|Y) + H(Y|X)$ satisfies the first, second and fourth properties above. If we say that $X = Y$ if there is a one-to-one function mapping from X to Y, then the third property is also satisfied, and $\rho(X,Y)$ is a metric.

(b) Verify that $\rho(X,Y)$ can also be expressed as

$$\begin{aligned} \rho(X,Y) &= H(X) + H(Y) - 2I(X;Y) && (2.11)\\ &= H(X,Y) - I(X;Y) && (2.12)\\ &= 2H(X,Y) - H(X) - H(Y). && (2.13) \end{aligned}$$

Solution: *A metric*

(a) Let

$$\rho(X,Y) = H(X|Y) + H(Y|X). \tag{2.14}$$

Then

- Since conditional entropy is always ≥ 0, $\rho(X,Y) \geq 0$.
- The symmetry of the definition implies that $\rho(X,Y) = \rho(Y,X)$.
- By problem 2.6, it follows that $H(Y|X)$ is 0 iff Y is a function of X and $H(X|Y)$ is 0 iff X is a function of Y. Thus $\rho(X,Y)$ is 0 iff X and Y are functions of each other - and therefore are equivalent up to a reversible transformation.
- Consider three random variables X, Y and Z. Then

$$\begin{aligned} H(X|Y) + H(Y|Z) &\geq H(X|Y,Z) + H(Y|Z) && (2.15)\\ &= H(X,Y|Z) && (2.16)\\ &= H(X|Z) + H(Y|X,Z) && (2.17)\\ &\geq H(X|Z), && (2.18) \end{aligned}$$

from which it follows that

$$\rho(X,Y) + \rho(Y,Z) \geq \rho(X,Z). \tag{2.19}$$

Note that the inequality is strict unless $X \to Y \to Z$ forms a Markov Chain and Y is a function of X and Z.

(b) Since $H(X|Y) = H(X) - I(X;Y)$, the first equation follows. The second relation follows from the first equation and the fact that $H(X,Y) = H(X) + H(Y) - I(X;Y)$. The third follows on substituting $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

10. **Entropy of a disjoint mixture.** Let X_1 and X_2 be discrete random variables drawn according to probability mass functions $p_1(\cdot)$ and $p_2(\cdot)$ over the respective alphabets $\mathcal{X}_1 = \{1, 2, \ldots, m\}$ and $\mathcal{X}_2 = \{m+1, \ldots, n\}$. Let

$$X = \begin{cases} X_1, & \text{with probability } \alpha, \\ X_2, & \text{with probability } 1-\alpha. \end{cases}$$

(a) Find $H(X)$ in terms of $H(X_1)$ and $H(X_2)$ and α.

(b) Maximize over α to show that $2^{H(X)} \leq 2^{H(X_1)} + 2^{H(X_2)}$ and interpret using the notion that $2^{H(X)}$ is the effective alphabet size.

Solution: *Entropy.* We can do this problem by writing down the definition of entropy and expanding the various terms. Instead, we will use the algebra of entropies for a simpler proof.

Since X_1 and X_2 have disjoint support sets, we can write

$$X = \begin{cases} X_1 & \text{with probability} \quad \alpha \\ X_2 & \text{with probability} \quad 1-\alpha \end{cases}$$

Define a function of X,

$$\theta = f(X) = \begin{cases} 1 & \text{when } X = X_1 \\ 2 & \text{when } X = X_2 \end{cases}$$

Then as in problem 1, we have

$$\begin{aligned} H(X) &= H(X, f(X)) = H(\theta) + H(X|\theta) \\ &= H(\theta) + p(\theta=1)H(X|\theta=1) + p(\theta=2)H(X|\theta=2) \\ &= H(\alpha) + \alpha H(X_1) + (1-\alpha)H(X_2) \end{aligned}$$

where $H(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$.

11. **A measure of correlation**. Let X_1 and X_2 be identically distributed, but not necessarily independent. Let

$$\rho = 1 - \frac{H(X_2 \mid X_1)}{H(X_1)}.$$

(a) Show $\rho = \frac{I(X_1;X_2)}{H(X_1)}$.

(b) Show $0 \leq \rho \leq 1$.

(c) When is $\rho = 0$?

(d) When is $\rho = 1$?

Solution: *A measure of correlation.* X_1 and X_2 are identically distributed and

$$\rho = 1 - \frac{H(X_2|X_1)}{H(X_1)}$$

(a)

$$\begin{aligned} \rho &= \frac{H(X_1) - H(X_2|X_1)}{H(X_1)} \\ &= \frac{H(X_2) - H(X_2|X_1)}{H(X_1)} \text{ (since } H(X_1) = H(X_2)) \\ &= \frac{I(X_1;X_2)}{H(X_1)}. \end{aligned}$$

(b) Since $0 \le H(X_2|X_1) \le H(X_2) = H(X_1)$, we have

$$0 \le \frac{H(X_2|X_1)}{H(X_1)} \le 1$$

$$0 \le \rho \le 1.$$

(c) $\rho = 0$ iff $I(X_1;X_2) = 0$ iff X_1 and X_2 are independent.

(d) $\rho = 1$ iff $H(X_2|X_1) = 0$ iff X_2 is a function of X_1. By symmetry, X_1 is a function of X_2, i.e., X_1 and X_2 have a one-to-one relationship.

12. **Example of joint entropy.** Let $p(x,y)$ be given by

X \ Y	0	1
0	$\frac{1}{3}$	$\frac{1}{3}$
1	0	$\frac{1}{3}$

Find

(a) $H(X), H(Y)$.

(b) $H(X \mid Y), H(Y \mid X)$.

(c) $H(X,Y)$.

(d) $H(Y) - H(Y \mid X)$.

(e) $I(X;Y)$.

(f) Draw a Venn diagram for the quantities in (a) through (e).

Solution: *Example of joint entropy*

(a) $H(X) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3 = 0.918$ bits $= H(Y)$.

(b) $H(X|Y) = \frac{1}{3}H(X|Y=0) + \frac{2}{3}H(X|Y=1) = 0.667$ bits $= H(Y|X)$.

(c) $H(X,Y) = 3 \times \frac{1}{3}\log 3 = 1.585$ bits.

(d) $H(Y) - H(Y|X) = 0.251$ bits.

(e) $I(X;Y) = H(Y) - H(Y|X) = 0.251$ bits.

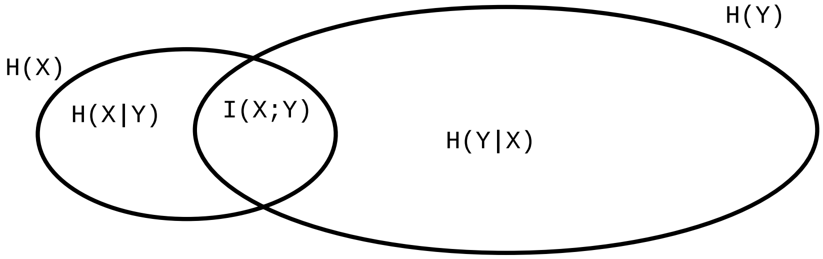
(f) See Figure 1.

13. **Inequality.** Show $\ln x \ge 1 - \frac{1}{x}$ for $x > 0$.

Solution: *Inequality.* Using the Remainder form of the Taylor expansion of $\ln(x)$ about $x = 1$, we have for some c between 1 and x

$$\ln(x) = \ln(1) + \left(\frac{1}{t}\right)_{t=1}(x-1) + \left(\frac{-1}{t^2}\right)_{t=c}\frac{(x-1)^2}{2} \le x - 1$$

Figure 2.1: Venn diagram to illustrate the relationships of entropy and relative entropy



since the second term is always negative. Hence letting $y = 1/x$, we obtain

$$-\ln y \leq \frac{1}{y} - 1$$

or

$$\ln y \geq 1 - \frac{1}{y}$$

with equality iff $y = 1$.

14. **Entropy of a sum.** Let X and Y be random variables that take on values $x_1, x_2, \ldots, x_r$ and $y_1, y_2, \ldots, y_s$, respectively. Let $Z = X + Y$.

(a) Show that $H(Z|X) = H(Y|X)$. Argue that if X, Y are independent, then $H(Y) \leq H(Z)$ and $H(X) \leq H(Z)$. Thus the addition of *independent* random variables adds uncertainty.

(b) Give an example of (necessarily dependent) random variables in which $H(X) > H(Z)$ and $H(Y) > H(Z)$.

(c) Under what conditions does $H(Z) = H(X) + H(Y)$?

Solution: *Entropy of a sum.*

(a) $Z = X + Y$. Hence $p(Z = z|X = x) = p(Y = z - x|X = x)$.

$$\begin{aligned}
H(Z|X) &= \sum p(x) H(Z|X = x) \\
&= -\sum_x p(x) \sum_z p(Z = z|X = x) \log p(Z = z|X = x) \\
&= \sum_x p(x) \sum_y p(Y = z - x|X = x) \log p(Y = z - x|X = x) \\
&= \sum p(x) H(Y|X = x) \\
&= H(Y|X).
\end{aligned}$$

If X and Y are independent, then $H(Y|X) = H(Y)$. Since $I(X;Z) \geq 0$, we have $H(Z) \geq H(Z|X) = H(Y|X) = H(Y)$. Similarly we can show that $H(Z) \geq H(X)$.

(b) Consider the following joint distribution for X and Y Let

$$X = -Y = \begin{cases} 1 & \text{with probability } 1/2 \\ 0 & \text{with probability } 1/2 \end{cases}$$

Then $H(X) = H(Y) = 1$, but $Z = 0$ with prob. 1 and hence $H(Z) = 0$.

(c) We have

$$H(Z) \leq H(X,Y) \leq H(X) + H(Y)$$

because Z is a function of (X,Y) and $H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$. We have equality iff (X,Y) is a function of Z and $H(Y) = H(Y|X)$, i.e., X and Y are independent.

15. **Data processing.** Let $X_1 \to X_2 \to X_3 \to \cdots \to X_n$ form a Markov chain in this order; i.e., let

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1) \cdots p(x_n|x_{n-1}).$$

Reduce $I(X_1; X_2, \ldots, X_n)$ to its simplest form.

Solution: *Data Processing.* By the chain rule for mutual information,

$$I(X_1; X_2, \ldots, X_n) = I(X_1; X_2) + I(X_1; X_3|X_2) + \cdots + I(X_1; X_n|X_2, \ldots, X_{n-2}). \quad (2.20)$$

By the Markov property, the past and the future are conditionally independent given the present and hence all terms except the first are zero. Therefore

$$I(X_1; X_2, \ldots, X_n) = I(X_1; X_2). \quad (2.21)$$

16. **Bottleneck.** Suppose a (non-stationary) Markov chain starts in one of n states, necks down to $k < n$ states, and then fans back to $m > k$ states. Thus $X_1 \to X_2 \to X_3$, i.e., $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$, for all $x_1 \in \{1, 2, \ldots, n\}$, $x_2 \in \{1, 2, \ldots, k\}$, $x_3 \in \{1, 2, \ldots, m\}$.

(a) Show that the dependence of X_1 and X_3 is limited by the bottleneck by proving that $I(X_1; X_3) \leq \log k$.

(b) Evaluate $I(X_1; X_3)$ for $k = 1$, and conclude that no dependence can survive such a bottleneck.

Solution:

Bottleneck.

(a) From the data processing inequality, and the fact that entropy is maximum for a uniform distribution, we get

$$\begin{aligned} I(X_1;X_3) &\leq I(X_1;X_2) \\ &= H(X_2) - H(X_2 \mid X_1) \\ &\leq H(X_2) \\ &\leq \log k. \end{aligned}$$

Thus, the dependence between X_1 and X_3 is limited by the size of the bottleneck. That is $I(X_1;X_3) \leq \log k$.

(b) For $k = 1$, $I(X_1;X_3) \leq \log 1 = 0$ and since $I(X_1,X_3) \geq 0$, $I(X_1,X_3) = 0$. Thus, for $k = 1$, X_1 and X_3 are independent.

17. **Pure randomness and bent coins.** Let $X_1, X_2, \ldots, X_n$ denote the outcomes of independent flips of a *bent* coin. Thus $\Pr\{X_i = 1\} = p$, $\Pr\{X_i = 0\} = 1-p$, where p is unknown. We wish to obtain a sequence $Z_1, Z_2, \ldots, Z_K$ of *fair* coin flips from $X_1, X_2, \ldots, X_n$. Toward this end let $f : \mathcal{X}^n \to \{0,1\}^*$, (where $\{0,1\}^* = \{\Lambda, 0, 1, 00, 01, \ldots\}$ is the set of all finite length binary sequences), be a mapping $f(X_1, X_2, \ldots, X_n) = (Z_1, Z_2, \ldots, Z_K)$, where $Z_i \sim$ Bernoulli $(\frac{1}{2})$, and K may depend on $(X_1, \ldots, X_n)$. In order that the sequence $Z_1, Z_2, \ldots$ appear to be fair coin flips, the map f from bent coin flips to fair flips must have the property that all 2^k sequences $(Z_1, Z_2, \ldots, Z_k)$ of a given length k have equal probability (possibly 0), for $k = 1, 2, \ldots$. For example, for $n = 2$, the map $f(01) = 0$, $f(10) = 1$, $f(00) = f(11) = \Lambda$ (the null string), has the property that $\Pr\{Z_1 = 1 | K = 1\} = \Pr\{Z_1 = 0 | K = 1\} = \frac{1}{2}$.

Give reasons for the following inequalities:

$$\begin{aligned} nH(p) &\overset{(a)}{=} H(X_1, \ldots, X_n) \\ &\overset{(b)}{\geq} H(Z_1, Z_2, \ldots, Z_K, K) \\ &\overset{(c)}{=} H(K) + H(Z_1, \ldots, Z_K | K) \\ &\overset{(d)}{=} H(K) + E(K) \\ &\overset{(e)}{\geq} EK. \end{aligned}$$

Thus no more than $nH(p)$ fair coin tosses can be derived from $(X_1, \ldots, X_n)$, on the average. Exhibit a good map f on sequences of length 4.

Solution: *Pure randomness and bent coins.*

$$\begin{aligned} nH(p) &\overset{(a)}{=} H(X_1, \ldots, X_n) \\ &\overset{(b)}{\geq} H(Z_1, Z_2, \ldots, Z_K) \end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{=} H(Z_1, Z_2, \ldots, Z_K, K) \\
&\overset{(d)}{=} H(K) + H(Z_1, \ldots, Z_K|K) \\
&\overset{(e)}{=} H(K) + E(K) \\
&\overset{(f)}{\geq} EK \ .
\end{aligned}
$$

(a) Since $X_1, X_2, \ldots, X_n$ are i.i.d. with probability of $X_i = 1$ being p, the entropy $H(X_1, X_2, \ldots, X_n)$ is $nH(p)$.

(b) $Z_1, \ldots, Z_K$ is a function of $X_1, X_2, \ldots, X_n$, and since the entropy of a function of a random variable is less than the entropy of the random variable, $H(Z_1, \ldots, Z_K) \leq H(X_1, X_2, \ldots, X_n)$.

(c) K is a function of $Z_1, Z_2, \ldots, Z_K$, so its conditional entropy given $Z_1, Z_2, \ldots, Z_K$ is 0. Hence $H(Z_1, Z_2, \ldots, Z_K, K) = H(Z_1, \ldots, Z_K) + H(K|Z_1, Z_2, \ldots, Z_K) = H(Z_1, Z_2, \ldots, Z_K)$.

(d) Follows from the chain rule for entropy.

(e) By assumption, $Z_1, Z_2, \ldots, Z_K$ are pure random bits (given K), with entropy 1 bit per symbol. Hence

$$
\begin{aligned}
H(Z_1, Z_2, \ldots, Z_K|K) &= \sum_k p(K=k) H(Z_1, Z_2, \ldots, Z_k|K=k) && (2.22) \\
&= \sum_k p(k)k && (2.23) \\
&= EK. && (2.24)
\end{aligned}
$$

(f) Follows from the non-negativity of discrete entropy.

(g) Since we do not know p, the only way to generate pure random bits is to use the fact that all sequences with the same number of ones are equally likely. For example, the sequences 0001,0010,0100 and 1000 are equally likely and can be used to generate 2 pure random bits. An example of a mapping to generate random bits is

$$
\begin{array}{llll}
0000 \rightarrow \Lambda & & & \\
0001 \rightarrow 00 & 0010 \rightarrow 01 & 0100 \rightarrow 10 & 1000 \rightarrow 11 \\
0011 \rightarrow 00 & 0110 \rightarrow 01 & 1100 \rightarrow 10 & 1001 \rightarrow 11 \\
1010 \rightarrow 0 & 0101 \rightarrow 1 & & \\
1110 \rightarrow 11 & 1101 \rightarrow 10 & 1011 \rightarrow 01 & 0111 \rightarrow 00 \\
1111 \rightarrow \Lambda & & &
\end{array}
\qquad (2.25)
$$

The resulting expected number of bits is

$$
\begin{aligned}
EK &= 4pq^3 \times 2 + 4p^2q^2 \times 2 + 2p^2q^2 \times 1 + 4p^3q \times 2 && (2.26) \\
&= 8pq^3 + 10p^2q^2 + 8p^3q. && (2.27)
\end{aligned}
$$

For example, for $p \approx \frac{1}{2}$, the expected number of pure random bits is close to 1.625. This is substantially less then the 4 pure random bits that could be generated if p were exactly $\frac{1}{2}$.

We will now analyze the efficiency of this scheme of generating random bits for long sequences of bent coin flips. Let n be the number of bent coin flips. The algorithm that we will use is the obvious extension of the above method of generating pure bits using the fact that all sequences with the same number of ones are equally likely.

Consider all sequences with k ones. There are $\binom{n}{k}$ such sequences, which are all equally likely. If $\binom{n}{k}$ were a power of 2, then we could generate $\log \binom{n}{k}$ pure random bits from such a set. However, in the general case, $\binom{n}{k}$ is not a power of 2 and the best we can to is the divide the set of $\binom{n}{k}$ elements into subset of sizes which are powers of 2. The largest set would have a size $2^{\lfloor \log \binom{n}{k} \rfloor}$ and could be used to generate $\lfloor \log \binom{n}{k} \rfloor$ random bits. We could divide the remaining elements into the largest set which is a power of 2, etc. The worst case would occur when $\binom{n}{k} = 2^{l+1} - 1$, in which case the subsets would be of sizes $2^l, 2^{l-1}, 2^{l-2}, \ldots, 1$.

Instead of analyzing the scheme exactly, we will just find a lower bound on number of random bits generated from a set of size $\binom{n}{k}$. Let $l = \lfloor \log \binom{n}{k} \rfloor$. Then at least half of the elements belong to a set of size 2^l and would generate l random bits, at least $\frac{1}{4}$th belong to a set of size 2^{l-1} and generate $l-1$ random bits, etc. On the average, the number of bits generated is

$$\begin{aligned} E[K|k \text{ 1's in sequence}] &\geq \frac{1}{2}l + \frac{1}{4}(l-1) + \cdots + \frac{1}{2^l}1 && (2.28)\\ &= l - \frac{1}{4}\left(1 + \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \cdots + \frac{l-1}{2^{l-2}}\right) && (2.29)\\ &\geq l-1, && (2.30) \end{aligned}$$

since the infinite series sums to 1.

Hence the fact that $\binom{n}{k}$ is not a power of 2 will cost at most 1 bit on the average in the number of random bits that are produced.

Hence, the expected number of pure random bits produced by this algorithm is

$$\begin{aligned} EK &\geq \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \lfloor \log \binom{n}{k} - 1 \rfloor && (2.31)\\ &\geq \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left(\log \binom{n}{k} - 2 \right) && (2.32)\\ &= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log \binom{n}{k} - 2 && (2.33)\\ &\geq \sum_{n(p-\epsilon) \leq k \leq n(p+\epsilon)} \binom{n}{k} p^k q^{n-k} \log \binom{n}{k} - 2. && (2.34) \end{aligned}$$

Now for sufficiently large n, the probability that the number of 1's in the sequence is close to np is near 1 (by the weak law of large numbers). For such sequences, $\frac{k}{n}$ is close to p and hence there exists a δ such that

$$\binom{n}{k} \geq 2^{n(H(\frac{k}{n})-\delta)} \geq 2^{n(H(p)-2\delta)} \tag{2.35}$$

using Stirling's approximation for the binomial coefficients and the continuity of the entropy function. If we assume that n is large enough so that the probability that $n(p-\epsilon) \leq k \leq n(p+\epsilon)$ is greater than $1-\epsilon$, then we see that $EK \geq (1-\epsilon)n(H(p)-2\delta)-2$, which is very good since $nH(p)$ is an upper bound on the number of pure random bits that can be produced from the bent coin sequence.

18. **World Series.** The World Series is a seven-game series that terminates as soon as either team wins four games. Let X be the random variable that represents the outcome of a World Series between teams A and B; possible values of X are AAAA, BABABAB, and BBBAAAA. Let Y be the number of games played, which ranges from 4 to 7. Assuming that A and B are equally matched and that the games are independent, calculate $H(X)$, $H(Y)$, $H(Y|X)$, and $H(X|Y)$.

Solution:

World Series. Two teams play until one of them has won 4 games.

There are 2 (AAAA, BBBB) World Series with 4 games. Each happens with probability $(1/2)^4$.

There are $8 = 2\binom{4}{3}$ World Series with 5 games. Each happens with probability $(1/2)^5$.

There are $20 = 2\binom{5}{3}$ World Series with 6 games. Each happens with probability $(1/2)^6$.

There are $40 = 2\binom{6}{3}$ World Series with 7 games. Each happens with probability $(1/2)^7$.

The probability of a 4 game series ($Y=4$) is $2(1/2)^4 = 1/8$.

The probability of a 5 game series ($Y=5$) is $8(1/2)^5 = 1/4$.

The probability of a 6 game series ($Y=6$) is $20(1/2)^6 = 5/16$.

The probability of a 7 game series ($Y=7$) is $40(1/2)^7 = 5/16$.

$$\begin{aligned} H(X) &= \sum p(x) log \frac{1}{p(x)} \\ &= 2(1/16)\log 16 + 8(1/32)\log 32 + 20(1/64)\log 64 + 40(1/128)\log 128 \\ &= 5.8125 \end{aligned}$$

$$\begin{aligned} H(Y) &= \sum p(y) log \frac{1}{p(y)} \\ &= 1/8\log 8 + 1/4\log 4 + 5/16\log(16/5) + 5/16\log(16/5) \\ &= 1.924 \end{aligned}$$

Y is a deterministic function of X, so if you know X there is no randomness in Y. Or, $H(Y|X)=0$.

Since $H(X)+H(Y|X)=H(X,Y)=H(Y)+H(X|Y)$, it is easy to determine $H(X|Y)=H(X)+H(Y|X)-H(Y)=3.889$

19. **Infinite entropy.** This problem shows that the entropy of a discrete random variable can be infinite. Let $A=\sum_{n=2}^{\infty}(n\log^2 n)^{-1}$. (It is easy to show that A is finite by bounding the infinite sum by the integral of $(x\log^2 x)^{-1}$.) Show that the integer-valued random variable X defined by $\Pr(X=n)=(An\log^2 n)^{-1}$ for $n=2,3,\ldots,$ has $H(X)=+\infty$.

    Solution: *Infinite entropy.* By definition, $p_n=\Pr(X=n)=1/An\log^2 n$ for $n\geq 2$. Therefore

$$\begin{aligned}
H(X) &= -\sum_{n=2}^{\infty} p(n)\log p(n)\\
&= -\sum_{n=2}^{\infty}\left(1/An\log^2 n\right)\log\left(1/An\log^2 n\right)\\
&= \sum_{n=2}^{\infty}\frac{\log(An\log^2 n)}{An\log^2 n}\\
&= \sum_{n=2}^{\infty}\frac{\log A+\log n+2\log\log n}{An\log^2 n}\\
&= \log A+\sum_{n=2}^{\infty}\frac{1}{An\log n}+\sum_{n=2}^{\infty}\frac{2\log\log n}{An\log^2 n}.
\end{aligned}$$

The first term is finite. For base 2 logarithms, all the elements in the sum in the last term are nonnegative. (For any other base, the terms of the last sum eventually all become positive.) So all we have to do is bound the middle sum, which we do by comparing with an integral.

$$\sum_{n=2}^{\infty}\frac{1}{An\log n}>\int_2^{\infty}\frac{1}{Ax\log x}dx=K\ln\ln x\Big|_2^{\infty}=+\infty.$$

We conclude that $H(X)=+\infty$.

20. **Run length coding.** Let $X_1,X_2,\ldots,X_n$ be (possibly dependent) binary random variables. Suppose one calculates the run lengths $\mathbf{R}=(R_1,R_2,\ldots)$ of this sequence (in order as they occur). For example, the sequence $\mathbf{X}=0001100100$ yields run lengths $\mathbf{R}=(3,2,2,1,2)$. Compare $H(X_1,X_2,\ldots,X_n)$, $H(\mathbf{R})$ and $H(X_n,\mathbf{R})$. Show all equalities and inequalities, and bound all the differences.

    Solution: *Run length coding.* Since the run lengths are a function of $X_1,X_2,\ldots,X_n$, $H(\mathbf{R})\leq H(\mathbf{X})$. Any X_i together with the run lengths determine the entire sequence

$X_1, X_2, \ldots, X_n$. Hence

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_i, \mathbf{R}) && (2.36)\\ &= H(\mathbf{R}) + H(X_i|\mathbf{R}) && (2.37)\\ &\le H(\mathbf{R}) + H(X_i) && (2.38)\\ &\le H(\mathbf{R}) + 1. && (2.39) \end{aligned}$$

21. **Markov's inequality for probabilities.** Let $p(x)$ be a probability mass function. Prove, for all $d \ge 0$,

$$\Pr\{p(X) \le d\} \log\left(\frac{1}{d}\right) \le H(X). \tag{2.40}$$

Solution: *Markov inequality applied to entropy.*

$$\begin{aligned} P(p(X) < d) \log\frac{1}{d} &= \sum_{x:p(x)<d} p(x)\log\frac{1}{d} && (2.41)\\ &\le \sum_{x:p(x)<d} p(x)\log\frac{1}{p(x)} && (2.42)\\ &\le \sum_x p(x)\log\frac{1}{p(x)} && (2.43)\\ &= H(X) && (2.44) \end{aligned}$$

22. **Logical order of ideas.** Ideas have been developed in order of need, and then generalized if necessary. Reorder the following ideas, strongest first, implications following:

(a) Chain rule for $I(X_1, \ldots, X_n; Y)$, chain rule for $D(p(x_1, \ldots, x_n)||q(x_1, x_2, \ldots, x_n))$, and chain rule for $H(X_1, X_2, \ldots, X_n)$.

(b) $D(f||g) \ge 0$, Jensen's inequality, $I(X;Y) \ge 0$.

Solution: *Logical ordering of ideas.*

(a) The following orderings are subjective. Since $I(X;Y) = D(p(x,y)||p(x)p(y))$ is a special case of relative entropy, it is possible to derive the chain rule for I from the chain rule for D.

Since $H(X) = I(X;X)$, it is possible to derive the chain rule for H from the chain rule for I.

It is also possible to derive the chain rule for I from the chain rule for H as was done in the notes.

(b) In class, Jensen's inequality was used to prove the non-negativity of D. The inequality $I(X;Y) \ge 0$ followed as a special case of the non-negativity of D.

23. **Conditional mutual information.** Consider a sequence of n binary random variables $X_1, X_2, \ldots, X_n$. Each sequence with an even number of 1's has probability $2^{-(n-1)}$ and each sequence with an odd number of 1's has probability 0. Find the mutual informations

$$I(X_1; X_2), \quad I(X_2; X_3|X_1), \ldots, \; I(X_{n-1}; X_n|X_1, \ldots, X_{n-2}).$$

Solution: *Conditional mutual information.*

Consider a sequence of n binary random variables $X_1, X_2, \ldots, X_n$. Each sequence of length n with an even number of 1's is equally likely and has probability $2^{-(n-1)}$.

Any $n-1$ or fewer of these are independent. Thus, for $k \leq n-1$,

$$I(X_{k-1}; X_k|X_1, X_2, \ldots, X_{k-2}) = 0.$$

However, given $X_1, X_2, \ldots, X_{n-2}$, we know that once we know either X_{n-1} or X_n we know the other.

$$\begin{aligned} I(X_{n-1}; X_n|X_1, X_2, \ldots, X_{n-2}) &= H(X_n|X_1, X_2, \ldots, X_{n-2}) - H(X_n|X_1, X_2, \ldots, X_{n-1}) \\ &= 1 - 0 = 1 \text{ bit.} \end{aligned}$$

24. **Average entropy.** Let $H(p) = -p \log_2 p - (1-p) \log_2(1-p)$ be the binary entropy function.

(a) Evaluate $H(1/4)$ using the fact that $\log_2 3 \approx 1.584$. *Hint:* You may wish to consider an experiment with four equally likely outcomes, one of which is more interesting than the others.

(b) Calculate the average entropy $H(p)$ when the probability p is chosen uniformly in the range $0 \leq p \leq 1$.

(c) *(Optional)* Calculate the average entropy $H(p_1, p_2, p_3)$ where (p_1, p_2, p_3) is a uniformly distributed probability vector. Generalize to dimension n.

Solution: *Average Entropy.*

(a) We can generate two bits of information by picking one of four equally likely alternatives. This selection can be made in two steps. First we decide whether the first outcome occurs. Since this has probability $1/4$, the information generated is $H(1/4)$. If not the first outcome, then we select one of the three remaining outcomes; with probability $3/4$, this produces $\log_2 3$ bits of information. Thus

$$H(1/4) + (3/4) \log_2 3 = 2$$

and so $H(1/4) = 2 - (3/4) \log_2 3 = 2 - (.75)(1.585) = 0.811$ bits.

(b) If p is chosen uniformly in the range $0 \leq p \leq 1$, then the average entropy (in nats) is

$$-\int_0^1 p \ln p + (1-p)\ln(1-p) dp = -2\int_0^1 x \ln x \, dx = -2\left(\frac{x^2}{2}\ln x + \frac{x^2}{4}\right)\Big|_0^1 = \tfrac{1}{2}\,.$$

Therefore the average entropy is $\frac{1}{2}\log_2 e = 1/(2\ln 2) = .721$ bits.

(c) Choosing a uniformly distributed probability vector (p_1, p_2, p_3) is equivalent to choosing a point (p_1, p_2) uniformly from the triangle $0 \leq p_1 \leq 1$, $p_1 \leq p_2 \leq 1$. The probability density function has the constant value 2 because the area of the triangle is 1/2. So the average entropy $H(p_1, p_2, p_3)$ is

$$-2\int_0^1\int_{p_1}^1 p_1 \ln p_1 + p_2 \ln p_2 + (1-p_1-p_2)\ln(1-p_1-p_2) dp_2 dp_1\,.$$

After some enjoyable calculus, we obtain the final result $5/(6\ln 2) = 1.202$ bits.

25. **Venn diagrams.** There isn't realy a notion of mutual information common to three random variables. Here is one attempt at a definition: Using Venn diagrams, we can see that the mutual information common to three random variables X, Y and Z can be defined by

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

This quantity is symmetric in X, Y and Z, despite the preceding asymmetric definition. Unfortunately, $I(X;Y;Z)$ is not necessarily nonnegative. Find X, Y and Z such that $I(X;Y;Z) < 0$, and prove the following two identities:

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

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

The first identity can be understood using the Venn diagram analogy for entropy and mutual information. The second identity follows easily from the first.

Solution: *Venn Diagrams.* To show the first identity,

$$\begin{aligned}
I(X;Y;Z) &= I(X;Y) - I(X;Y|Z) \quad \text{by definition} \\
&= I(X;Y) - (I(X;Y,Z) - I(X;Z)) \quad \text{by chain rule} \\
&= I(X;Y) + I(X;Z) - I(X;Y,Z) \\
&= I(X;Y) + I(X;Z) - (H(X) + H(Y,Z) - H(X,Y,Z)) \\
&= I(X;Y) + I(X;Z) - H(X) + H(X,Y,Z) - H(Y,Z) \\
&= I(X;Y) + I(X;Z) - H(X) + H(X,Y,Z) - (H(Y) + H(Z) - I(Y;Z)) \\
&= I(X;Y) + I(X;Z) + I(Y;Z) + H(X,Y,Z) - H(X) - H(Y) - H(Z).
\end{aligned}$$

To show the second identity, simply substitute for $I(X;Y)$, $I(X;Z)$, and $I(Y;Z)$ using equations like

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

These two identities show that $I(X;Y;Z)$ is a symmetric (but not necessarily nonnegative) function of three random variables.

26. **Another proof of non-negativity of relative entropy.** In view of the fundamental nature of the result $D(p||q) \geq 0$, we will give another proof.

    (a) Show that $\ln x \leq x - 1$ for $0 < x < \infty$.

    (b) Justify the following steps:

$$-D(p||q) = \sum_x p(x) \ln \frac{q(x)}{p(x)} \tag{2.45}$$

$$\leq \sum_x p(x) \left(\frac{q(x)}{p(x)} - 1 \right) \tag{2.46}$$

$$\leq 0 \tag{2.47}$$

    (c) What are the conditions for equality?

Solution: *Another proof of non-negativity of relative entropy.* In view of the fundamental nature of the result $D(p||q) \geq 0$, we will give another proof.

(a) Show that $\ln x \leq x - 1$ for $0 < x < \infty$.
There are many ways to prove this. The easiest is using calculus. Let

$$f(x) = x - 1 - \ln x \tag{2.48}$$

for $0 < x < \infty$. Then $f'(x) = 1 - \frac{1}{x}$ and $f''(x) = \frac{1}{x^2} > 0$, and therefore $f(x)$ is strictly convex. Therefore a local minimum of the function is also a global minimum. The function has a local minimum at the point where $f'(x) = 0$, i.e., when $x = 1$. Therefore $f(x) \geq f(1)$, i.e.,

$$x - 1 - \ln x \geq 1 - 1 - \ln 1 = 0 \tag{2.49}$$

which gives us the desired inequality. Equality occurs only if $x = 1$.

(b) We let A be the set of x such that $p(x) > 0$.

$$-D_e(p||q) = \sum_{x \in A} p(x) ln \frac{q(x)}{p(x)} \tag{2.50}$$

$$\leq \sum_{x \in A} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) \tag{2.51}$$

$$= \sum_{x \in A} q(x) - \sum_{x \in A} p(x) \tag{2.52}$$

$$\leq 0 \tag{2.53}$$

The first step follows from the definition of D, the second step follows from the inequality $\ln t \leq t - 1$, the third step from expanding the sum, and the last step from the fact that the $q(A) \leq 1$ and $p(A) = 1$.

(c) What are the conditions for equality?

We have equality in the inequality $\ln t \leq t-1$ if and only if $t=1$. Therefore we have equality in step 2 of the chain iff $q(x)/p(x) = 1$ for all $x \in A$. This implies that $p(x) = q(x)$ for all x, and we have equality in the last step as well. Thus the condition for equality is that $p(x) = q(x)$ for all x.

27. **Grouping rule for entropy:** Let $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ be a probability distribution on m elements, i.e, $p_i \geq 0$, and $\sum_{i=1}^m p_i = 1$. Define a new distribution $\mathbf{q}$ on $m-1$ elements as $q_1 = p_1$, $q_2 = p_2$,. . . , $q_{m-2} = p_{m-2}$, and $q_{m-1} = p_{m-1} + p_m$, i.e., the distribution $\mathbf{q}$ is the same as $\mathbf{p}$ on $\{1, 2, \ldots, m-2\}$, and the probability of the last element in $\mathbf{q}$ is the sum of the last two probabilities of $\mathbf{p}$. Show that

$$H(\mathbf{p}) = H(\mathbf{q}) + (p_{m-1} + p_m) H\left(\frac{p_{m-1}}{p_{m-1}+p_m}, \frac{p_m}{p_{m-1}+p_m}\right). \tag{2.54}$$

Solution:

$$\begin{aligned}
H(\mathbf{p}) &= -\sum_{i=1}^{m} p_i \log p_i && (2.55)\\
&= -\sum_{i=1}^{m-2} p_i \log p_i - p_{m-1} \log p_{m-1} - p_m \log p_m && (2.56)\\
&= -\sum_{i=1}^{m-2} p_i \log p_i - p_{m-1} \log \frac{p_{m-1}}{p_{m-1}+p_m} - p_m \log \frac{p_m}{p_{m-1}+p_m} && (2.57)\\
&\qquad -(p_{m-1}+p_m)\log(p_{m-1}+p_m) && (2.58)\\
&= H(\mathbf{q}) - p_{m-1} \log \frac{p_{m-1}}{p_{m-1}+p_m} - p_m \log \frac{p_m}{p_{m-1}+p_m} && (2.59)\\
&= H(\mathbf{q}) - (p_{m-1}+p_m)\left(\frac{p_{m-1}}{p_{m-1}+p_m} \log \frac{p_{m-1}}{p_{m-1}+p_m} - \frac{p_m}{p_{m-1}+p_m} \log \frac{p_m}{p_{m-1}+p_m}\right) && (2.60)\\
&= H(\mathbf{q}) + (p_{m-1}+p_m) H_2\left(\frac{p_{m-1}}{p_{m-1}+p_m}, \frac{p_m}{p_{m-1}+p_m}\right), && (2.61)
\end{aligned}$$

where $H_2(a,b) = -a \log a - b \log b$.

28. **Mixing increases entropy.** Show that the entropy of the probability distribution, $(p_1, \ldots, p_i, \ldots, p_j, \ldots, p_m)$, is less than the entropy of the distribution $(p_1, \ldots, \frac{p_i+p_j}{2}, \ldots, \frac{p_i+p_j}{2}, \ldots, p_m)$. Show that in general any transfer of probability that makes the distribution more uniform increases the entropy.

Solution:

Mixing increases entropy.

This problem depends on the convexity of the log function. Let

$$\begin{aligned}
P_1 &= (p_1, \ldots, p_i, \ldots, p_j, \ldots, p_m)\\
P_2 &= (p_1, \ldots, \frac{p_i+p_j}{2}, \ldots, \frac{p_j+p_i}{2}, \ldots, p_m)
\end{aligned}$$

Then, by the log sum inequality,

$$\begin{aligned} H(P_2)-H(P_1) &= -2(\frac{p_i+p_j}{2})\log(\frac{p_i+p_j}{2})+p_i\log p_i+p_j\log p_j \\ &= -(p_i+p_j)\log(\frac{p_i+p_j}{2})+p_i\log p_i+p_j\log p_j \\ &\geq 0. \end{aligned}$$

Thus,

$$H(P_2)\geq H(P_1).$$

29. **Inequalities.** Let X, Y and Z be joint random variables. Prove the following inequalities and find conditions for equality.

(a) $H(X,Y|Z)\geq H(X|Z)$.

(b) $I(X,Y;Z)\geq I(X;Z)$.

(c) $H(X,Y,Z)-H(X,Y)\leq H(X,Z)-H(X)$.

(d) $I(X;Z|Y)\geq I(Z;Y|X)-I(Z;Y)+I(X;Z)$.

Solution: *Inequalities.*

(a) Using the chain rule for conditional entropy,

$$H(X,Y|Z)=H(X|Z)+H(Y|X,Z)\geq H(X|Z),$$

with equality iff $H(Y|X,Z)=0$, that is, when Y is a function of X and Z.

(b) Using the chain rule for mutual information,

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

with equality iff $I(Y;Z|X)=0$, that is, when Y and Z are conditionally independent given X.

(c) Using first the chain rule for entropy and then the definition of conditional mutual information,

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

with equality iff $I(Y;Z|X)=0$, that is, when Y and Z are conditionally independent given X.

(d) Using the chain rule for mutual information,

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

and therefore

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

We see that this inequality is actually an equality in all cases.

30. **Maximum entropy.** Find the probability mass function $p(x)$ that maximizes the entropy $H(X)$ of a non-negative integer-valued random variable X subject to the constraint

$$EX = \sum_{n=0}^{\infty} np(n) = A$$

for a fixed value $A > 0$. Evaluate this maximum $H(X)$.

Solution: *Maximum entropy*

Recall that,

$$-\sum_{i=0}^{\infty} p_i \log p_i \le -\sum_{i=0}^{\infty} p_i \log q_i.$$

Let $q_i = \alpha(\beta)^i$. Then we have that,

$$\begin{aligned} -\sum_{i=0}^{\infty} p_i \log p_i &\le -\sum_{i=0}^{\infty} p_i \log q_i \\ &= -\left(\log(\alpha)\sum_{i=0}^{\infty} p_i + \log(\beta)\sum_{i=0}^{\infty} ip_i\right) \\ &= -\log\alpha - A\log\beta \end{aligned}$$

Notice that the final right hand side expression is independent of $\{p_i\}$, and that the inequality,

$$-\sum_{i=0}^{\infty} p_i \log p_i \le -\log\alpha - A\log\beta$$

holds for all α, β such that,

$$\sum_{i=0}^{\infty} \alpha\beta^i = 1 = \alpha\frac{1}{1-\beta}.$$

The constraint on the expected value also requires that,

$$\sum_{i=0}^{\infty} i\alpha\beta^i = A = \alpha\frac{\beta}{(1-\beta)^2}.$$

Combining the two constraints we have,

$$\begin{aligned} \alpha\frac{\beta}{(1-\beta)^2} &= \left(\frac{\alpha}{1-\beta}\right)\left(\frac{\beta}{1-\beta}\right) \\ &= \frac{\beta}{1-\beta} \\ &= A, \end{aligned}$$

which implies that,

$$\begin{aligned}\beta &= \frac{A}{A+1}\\ \alpha &= \frac{1}{A+1}.\end{aligned}$$

So the entropy maximizing distribution is,

$$p_i = \frac{1}{A+1}\left(\frac{A}{A+1}\right)^i .$$

Plugging these values into the expression for the maximum entropy,

$$-\log\alpha - A\log\beta = (A+1)\log(A+1) - A\log A.$$

The general form of the distribution,

$$p_i = \alpha\beta^i$$

can be obtained either by guessing or by Lagrange multipliers where,

$$F(p_i,\lambda_1,\lambda_2) = -\sum_{i=0}^{\infty} p_i \log p_i + \lambda_1(\sum_{i=0}^{\infty} p_i - 1) + \lambda_2(\sum_{i=0}^{\infty} i p_i - A)$$

is the function whose gradient we set to 0.

To complete the argument with Lagrange multipliers, it is necessary to show that the local maximum is the global maximum. One possible argument is based on the fact that $-H(p)$ is convex, it has only one local minima, no local maxima and therefore Lagrange multiplier actually gives the global maximum for $H(p)$.

31. **Conditional entropy.** Under what conditions does $H(X \mid g(Y)) = H(X \mid Y)$?

Solution: (*Conditional Entropy*). If $H(X|g(Y)) = H(X|Y)$, then $H(X) - H(X|g(Y)) = H(X) - H(X|Y)$, i.e., $I(X;g(Y)) = I(X;Y)$. This is the condition for equality in the data processing inequality. From the derivation of the inequality, we have equality iff $X \to g(Y) \to Y$ forms a Markov chain. Hence $H(X|g(Y)) = H(X|Y)$ iff $X \to g(Y) \to Y$. This condition includes many special cases, such as g being one-to-one, and X and Y being independent. However, these two special cases do not exhaust all the possibilities.

32. **Fano.** We are given the following joint distribution on (X,Y)

Y			
X	a	b	c
1	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$
2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$
3	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$

Let $\hat{X}(Y)$ be an estimator for X (based on Y) and let $P_e = \Pr\{\hat{X}(Y) \neq X\}$.

(a) Find the minimum probability of error estimator $\hat{X}(Y)$ and the associated P_e.

(b) Evaluate Fano's inequality for this problem and compare.

Solution:

(a) From inspection we see that

$$\hat{X}(y) = \begin{cases} 1 & y = a \\ 2 & y = b \\ 3 & y = c \end{cases}$$

Hence the associated P_e is the sum of $P(1,b)$, $P(1,c)$, $P(2,a)$, $P(2,c)$, $P(3,a)$ and $P(3,b)$. Therefore, $P_e = 1/2$.

(b) From Fano's inequality we know

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}.$$

Here,

$$\begin{aligned} H(X|Y) &= H(X|Y=a)\Pr\{y=a\} + H(X|Y=b)\Pr\{y=b\} + H(X|Y=c)\Pr\{y=c\} \\ &= H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)\Pr\{y=a\} + H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)\Pr\{y=b\} + H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)\Pr\{y=c\} \\ &= H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right)(\Pr\{y=a\} + \Pr\{y=b\} + \Pr\{y=c\}) \\ &= H\left(\frac{1}{2},\frac{1}{4},\frac{1}{4}\right) \\ &= 1.5 \text{ bits.} \end{aligned}$$

Hence

$$P_e \geq \frac{1.5 - 1}{\log 3} = .316.$$

Hence our estimator $\hat{X}(Y)$ is not very close to Fano's bound in this form. If $\hat{X} \in \mathcal{X}$, as it does here, we can use the stronger form of Fano's inequality to get

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}|\text{-}1)}.$$

and

$$P_e \geq \frac{1.5 - 1}{\log 2} = \frac{1}{2}.$$

Therefore our estimator $\hat{X}(Y)$ is actually quite good.

33. **Fano's inequality.** Let $\Pr(X = i) = p_i$, $i = 1, 2, \ldots, m$ and let $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_m$. The minimal probability of error predictor of X is $\hat{X} = 1$, with resulting probability of error $P_e = 1 - p_1$. Maximize $H(\mathbf{p})$ subject to the constraint $1 - p_1 = P_e$

to find a bound on P_e in terms of H. This is Fano's inequality in the absence of conditioning.

Solution: (*Fano's Inequality.*) The minimal probability of error predictor when there is no information is $\hat{X} = 1$, the most probable value of X. The probability of error in this case is $P_e = 1 - p_1$. Hence if we fix P_e, we fix p_1. We maximize the entropy of X for a given P_e to obtain an upper bound on the entropy for a given P_e. The entropy,

$$\begin{aligned} H(\mathbf{p}) &= -p_1 \log p_1 - \sum_{i=2}^{m} p_i \log p_i \qquad (2.62) \\ &= -p_1 \log p_1 - \sum_{i=2}^{m} P_e \frac{p_i}{P_e} \log \frac{p_i}{P_e} - P_e \log P_e \qquad (2.63) \\ &= H(P_e) + P_e H\left(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \ldots, \frac{p_m}{P_e}\right) \qquad (2.64) \\ &\leq H(P_e) + P_e \log(m-1), \qquad (2.65) \end{aligned}$$

since the maximum of $H\left(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \ldots, \frac{p_m}{P_e}\right)$ is attained by an uniform distribution. Hence any X that can be predicted with a probability of error P_e must satisfy

$$H(X) \leq H(P_e) + P_e \log(m-1), \qquad (2.66)$$

which is the unconditional form of Fano's inequality. We can weaken this inequality to obtain an explicit lower bound for P_e,

$$P_e \geq \frac{H(X) - 1}{\log(m-1)}. \qquad (2.67)$$

34. **Entropy of initial conditions.** Prove that $H(X_0|X_n)$ is non-decreasing with n for any Markov chain.

Solution: *Entropy of initial conditions.* For a Markov chain, by the data processing theorem, we have

$$I(X_0; X_{n-1}) \geq I(X_0; X_n). \qquad (2.68)$$

Therefore

$$H(X_0) - H(X_0|X_{n-1}) \geq H(X_0) - H(X_0|X_n) \qquad (2.69)$$

or $H(X_0|X_n)$ increases with n.

35. **Relative entropy is not symmetric:** Let the random variable X have three possible outcomes $\{a, b, c\}$. Consider two distributions on this random variable

Symbol	$p(x)$	$q(x)$
a	1/2	1/3
b	1/4	1/3
c	1/4	1/3

Calculate $H(p)$, $H(q)$, $D(p||q)$ and $D(q||p)$. Verify that in this case $D(p||q) \neq D(q||p)$.

Solution:

$$H(p) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = 1.5 \text{ bits.} \tag{2.70}$$

$$H(q) = \frac{1}{3}\log 3 + \frac{1}{3}\log 3 + \frac{1}{3}\log 3 = \log 3 = 1.58496 \text{ bits.} \tag{2.71}$$

$$D(p||q) = \frac{1}{2}\log\frac{3}{2} + \frac{1}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{3}{4} = \log(3) - 1.5 = 1.58496 - 1.5 = 0.08496 \tag{2.72}$$

$$D(q||p) = \frac{1}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{4}{3} + \frac{1}{3}\log\frac{4}{3} = \frac{5}{3} - \log(3) = 1.66666 - 1.58496 = 0.08170 \tag{2.73}$$

36. **Symmetric relative entropy:** Though, as the previous example shows, $D(p||q) \neq D(q||p)$ in general, there could be distributions for which equality holds. Give an example of two distributions p and q on a binary alphabet such that $D(p||q) = D(q||p)$ (other than the trivial case $p = q$).

    Solution:

    A simple case for $D((p, 1-p)||(q, 1-q)) = D((q, 1-q)||(p, 1-p))$, i.e., for

    $$p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p} \tag{2.74}$$

    is when $q = 1 - p$.

37. **Relative entropy:** Let X, Y, Z be three random variables with a joint probability mass function $p(x, y, z)$. The relative entropy between the joint distribution and the product of the marginals is

    $$D(p(x,y,z)||p(x)p(y)p(z)) = E\left[\log\frac{p(x,y,z)}{p(x)p(y)p(z)}\right] \tag{2.75}$$

    Expand this in terms of entropies. When is this quantity zero?

    Solution:

    $$\begin{aligned} D(p(x,y,z)||p(x)p(y)p(z)) &= E\left[\log\frac{p(x,y,z)}{p(x)p(y)p(z)}\right] & (2.76)\\ &= E[\log p(x,y,z)] - E[\log p(x)] - E[\log p(y)] - E[\log p(z)] & (2.77)\\ &= -H(X,Y,Z) + H(X) + H(Y) + H(Z) & (2.78) \end{aligned}$$

    We have $D(p(x,y,z)||p(x)p(y)p(z)) = 0$ if and only $p(x,y,z) = p(x)p(y)p(z)$ for all (x,y,z), i.e., if X and Y and Z are independent.

38. **The value of a question** Let $X \sim p(x)$, $x = 1, 2, \ldots, m$. We are given a set $S \subseteq \{1, 2, \ldots, m\}$. We ask whether $X \in S$ and receive the answer

    $$Y = \begin{cases} 1, & \text{if } X \in S \\ 0, & \text{if } X \notin S. \end{cases}$$

    Suppose $\Pr\{X \in S\} = \alpha$. Find the decrease in uncertainty $H(X) - H(X|Y)$.

Apparently any set S with a given α is as good as any other.

Solution: *The value of a question.*

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

since $H(Y|X) = 0$.

39. **Entropy and pairwise independence.**
Let X, Y, Z be three binary Bernoulli $(\frac{1}{2})$ random variables that are pairwise independent, that is, $I(X;Y) = I(X;Z) = I(Y;Z) = 0$.

    (a) Under this constraint, what is the minimum value for $H(X,Y,Z)$?

    (b) Give an example achieving this minimum.

    Solution:

    (a)

    $$\begin{aligned} H(X,Y,Z) &= H(X,Y) + H(Z|X,Y) && (2.79) \\ &\geq H(X,Y) && (2.80) \\ &= 2. && (2.81) \end{aligned}$$

    So the minimum value for $H(X,Y,Z)$ is at least 2. To show that is is actually equal to 2, we show in part (b) that this bound is attainable.

    (b) Let X and Y be iid Bernoulli($\frac{1}{2}$) and let $Z = X \oplus Y$, where $\oplus$ denotes addition mod 2 (xor).

40. **Discrete entropies**

    Let X and Y be two independent integer-valued random variables. Let X be uniformly distributed over $\{1, 2, \ldots, 8\}$, and let $\Pr\{Y = k\} = 2^{-k}$, $\quad k = 1, 2, 3, \ldots$

    (a) Find $H(X)$

    (b) Find $H(Y)$

    (c) Find $H(X + Y, X - Y)$.

    Solution:

    (a) For a uniform distribution, $H(X) = \log m = \log 8 = 3$.

    (b) For a geometric distribution, $H(Y) = \sum_k k2^{-k} = 2$. (See solution to problem 2.1

(c) Since $(X,Y) \to (X+Y, X-Y)$ is a one to one transformation, $H(X+Y,X-Y) = H(X,Y) = H(X) + H(Y) = 3 + 2 = 5$.

41. **Random questions**

One wishes to identify a random object $X \sim p(x)$. A question $Q \sim r(q)$ is asked at random according to $r(q)$. This results in a deterministic answer $A = A(x,q) \in \{a_1, a_2, \ldots\}$. Suppose X and Q are independent. Then $I(X;Q,A)$ is the uncertainty in X removed by the question-answer (Q,A).

(a) Show $I(X;Q,A) = H(A|Q)$. Interpret.

(b) Now suppose that two i.i.d. questions $Q_1, Q_2, \sim r(q)$ are asked, eliciting answers A_1 and A_2. Show that two questions are less valuable than twice a single question in the sense that $I(X;Q_1,A_1,Q_2,A_2) \le 2I(X;Q_1,A_1)$.

Solution: *Random questions.*

(a)

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

The interpretation is as follows. The uncertainty removed in X by (Q,A) is the same as the uncertainty in the answer given the question.

(b) Using the result from part a and the fact that questions are independent, we can easily obtain the desired relationship.

$$\begin{aligned} I(X;Q_1,A_1,Q_2,A_2) &\overset{(a)}{=} I(X;Q_1) + I(X;A_1|Q_1) + I(X;Q_2|A_1,Q_1) + I(X;A_2|A_1,Q_1,Q_2) \\ &\overset{(b)}{=} I(X;A_1|Q_1) + H(Q_2|A_1,Q_1) - H(Q_2|X,A_1,Q_1) + I(X;A_2|A_1,Q_1,Q_2 \\ &\overset{(c)}{=} I(X;A_1|Q_1) + I(X;A_2|A_1,Q_1,Q_2) \\ &= I(X;A_1|Q_1) + H(A_2|A_1,Q_1,Q_2) - H(A_2|X,A_1,Q_1,Q_2) \\ &\overset{(d)}{=} I(X;A_1|Q_1) + H(A_2|A_1,Q_1,Q_2) \\ &\overset{(e)}{\le} I(X;A_1|Q_1) + H(A_2|Q_2) \\ &\overset{(f)}{=} 2I(X;A_1|Q_1) \end{aligned}$$

(a) Chain rule.
(b) X and Q_1 are independent.

(c) Q_2 are independent of X, Q_1, and A_1.
(d) A_2 is completely determined given Q_2 and X.
(e) Conditioning decreases entropy.
(f) Result from part a.

42. **Inequalities**. Which of the following inequalities are generally $\geq, =, \leq$? Label each with $\geq, =,$ or $\leq$.

    (a) $H(5X)$ vs. $H(X)$
    (b) $I(g(X);Y)$ vs. $I(X;Y)$
    (c) $H(X_0|X_{-1})$ vs. $H(X_0|X_{-1},X_1)$
    (d) $H(X_1,X_2,\ldots,X_n)$ vs. $H(c(X_1,X_2,\ldots,X_n))$, where $c(x_1,x_2,\ldots,x_n)$ is the Huffman codeword assigned to $(x_1,x_2,\ldots,x_n)$.
    (e) $H(X,Y)/(H(X)+H(Y))$ vs. 1

    Solution:

    (a) $X \to 5X$ is a one to one mapping, and hence $H(X) = H(5X)$.
    (b) By data processing inequality, $I(g(X);Y) \leq I(X;Y)$.
    (c) Because conditioning reduces entropy, $H(X_0|X_{-1}) \geq H(X_0|X_{-1},X_1)$.
    (d) $H(X,Y) \leq H(X)+H(Y)$, so $H(X,Y)/(H(X)+H(Y)) \leq 1$.

43. **Mutual information of heads and tails.**

    (a) Consider a fair coin flip. What is the mutual information between the top side and the bottom side of the coin?
    (b) A 6-sided fair die is rolled. What is the mutual information between the top side and the front face (the side most facing you)?

    Solution:

    Mutual information of heads and tails.

    To prove (a) observe that

$$\begin{aligned} I(T;B) &= H(B) - H(B|T) \\ &= \log 2 = 1 \end{aligned}$$

since $B \sim Ber(1/2)$, and $B = f(T)$. Here B, T stand for Bottom and Top respectively.

To prove (b) note that having observed a side of the cube facing us F, there are four possibilities for the top T, which are equally probable. Thus,

$$\begin{aligned} I(T;F) &= H(T) - H(T|F) \\ &= \log 6 - \log 4 \\ &= \log 3 - 1 \end{aligned}$$

since T has uniform distribution on $\{1,2,\ldots,6\}$.

44. **Pure randomness**

We wish to use a 3-sided coin to generate a fair coin toss. Let the coin X have probability mass function

$$X = \begin{cases} A, & p_A \\ B, & p_B \\ C, & p_C \end{cases}$$

where p_A, p_B, p_C are unknown.

(a) How would you use 2 independent flips X_1, X_2 to generate (if possible) a Bernoulli($\frac{1}{2}$) random variable Z?

(b) What is the resulting maximum expected number of fair bits generated?

Solution:

(a) The trick here is to notice that for any two letters Y and Z produced by two independent tosses of our bent three-sided coin, YZ has the same probability as ZY. So we can produce $B \sim$Bernoulli($\frac{1}{2}$) coin flips by letting $B = 0$ when we get AB, BC or AC, and $B = 1$ when we get BA, CB or CA (if we get AA, BB or CC we don't assign a value to B.)

(b) The expected number of bits generated by the above scheme is as follows. We get one bit, except when the two flips of the 3-sided coin produce the same symbol. So the expected number of fair bits generated is

$$0 * [P(AA) + P(BB) + P(CC)] + 1 * [1 - P(AA) - P(BB) - P(CC)], \quad (2.82)$$

$$\text{or,} \quad 1 - p_A^2 - p_B^2 - p_C^2. \quad (2.83)$$

45. **Finite entropy.** Show that for a discrete random variable $X \in \{1, 2, \ldots\}$, if $E \log X < \infty$, then $H(X) < \infty$.

Solution: Let the distribution on the integers be $p_1, p_2, \ldots$. Then $H(p) = -\sum p_i logp_i$ and $E \log X = \sum p_i logi = c < \infty$.

We will now find the maximum entropy distribution subject to the constraint on the expected logarithm. Using Lagrange multipliers or the results of Chapter 12, we have the following functional to optimize

$$J(p) = -\sum p_i \log p_i - \lambda_1 \sum p_i - \lambda_2 \sum p_i \log i \quad (2.84)$$

Differentiating with respect to p_i and setting to zero, we find that the p_i that maximizes the entropy set $p_i = ai^\lambda$, where $a = 1/(\sum i^\lambda)$ and λ chosed to meet the expected log constraint, i.e.

$$\sum i^\lambda \log i = c \sum i^\lambda \quad (2.85)$$

Using this value of p_i, we can see that the entropy is finite.

46. **Axiomatic definition of entropy.** If we assume certain axioms for our measure of information, then we will be forced to use a logarithmic measure like entropy. Shannon used this to justify his initial definition of entropy. In this book, we will rely more on the other properties of entropy rather than its axiomatic derivation to justify its use. The following problem is considerably more difficult than the other problems in this section.

If a sequence of symmetric functions $H_m(p_1, p_2, \ldots, p_m)$ satisfies the following properties,

- Normalization: $H_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$,
- Continuity: $H_2(p, 1-p)$ is a continuous function of p,
- Grouping: $H_m(p_1, p_2, \ldots, p_m) = H_{m-1}(p_1+p_2, p_3, \ldots, p_m) + (p_1+p_2) H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$,

prove that H_m must be of the form

$$H_m(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i, \qquad m = 2, 3, \ldots. \tag{2.86}$$

There are various other axiomatic formulations which also result in the same definition of entropy. See, for example, the book by Csiszár and Körner[3].

Solution: *Axiomatic definition of entropy.* This is a long solution, so we will first outline what we plan to do. First we will extend the grouping axiom by induction and prove that

$$\begin{aligned} H_m(p_1, p_2, \ldots, p_m) &= H_{m-k}(p_1 + p_2 + \cdots + p_k, p_{k+1}, \ldots, p_m) \\ &\quad + (p_1 + p_2 + \cdots + p_k) H_k\left(\frac{p_1}{p_1 + p_2 + \cdots + p_k}, \ldots, \frac{p_k}{p_1 + p_2 + \cdots + p_k}\right) \end{aligned} \tag{2.87}$$

Let $f(m)$ be the entropy of a uniform distribution on m symbols, i.e.,

$$f(m) = H_m\left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right). \tag{2.88}$$

We will then show that for any two integers r and s, that $f(rs) = f(r) + f(s)$. We use this to show that $f(m) = \log m$. We then show for rational $p = r/s$, that $H_2(p, 1-p) = -p \log p - (1-p) \log(1-p)$. By continuity, we will extend it to irrational p and finally by induction and grouping, we will extend the result to H_m for $m \geq 2$.

To begin, we extend the grouping axiom. For convenience in notation, we will let

$$S_k = \sum_{i=1}^{k} p_i \tag{2.89}$$

and we will denote $H_2(q, 1-q)$ as $h(q)$. Then we can write the grouping axiom as

$$H_m(p_1, \ldots, p_m) = H_{m-1}(S_2, p_3, \ldots, p_m) + S_2 h\left(\frac{p_2}{S_2}\right). \tag{2.90}$$

Applying the grouping axiom again, we have

$$\begin{align}
H_m(p_1,\ldots,p_m) &= H_{m-1}(S_2,p_3,\ldots,p_m)+S_2h\left(\frac{p_2}{S_2}\right) \tag{2.91}\\
&= H_{m-2}(S_3,p_4,\ldots,p_m)+S_3h\left(\frac{p_3}{S_3}\right)+S_2h\left(\frac{p_2}{S_2}\right) \tag{2.92}\\
&\vdots \tag{2.93}\\
&= H_{m-(k-1)}(S_k,p_{k+1},\ldots,p_m)+\sum_{i=2}^{k}S_ih\left(\frac{p_i}{S_i}\right). \tag{2.94}
\end{align}$$

Now, we apply the same grouping axiom repeatedly to $H_k(p_1/S_k,\ldots,p_k/S_k)$, to obtain

$$\begin{align}
H_k\left(\frac{p_1}{S_k},\ldots,\frac{p_k}{S_k}\right) &= H_2\left(\frac{S_{k-1}}{S_k},\frac{p_k}{S_k}\right)+\sum_{i=2}^{k-1}\frac{S_i}{S_k}h\left(\frac{p_i/S_k}{S_i/S_k}\right) \tag{2.95}\\
&= \frac{1}{S_k}\sum_{i=2}^{k}S_ih\left(\frac{p_i}{S_i}\right). \tag{2.96}
\end{align}$$

From (2.94) and (2.96), it follows that

$$H_m(p_1,\ldots,p_m)=H_{m-k}(S_k,p_{k+1},\ldots,p_m)+S_kH_k\left(\frac{p_1}{S_k},\ldots,\frac{p_k}{S_k}\right), \tag{2.97}$$

which is the extended grouping axiom.

Now we need to use an axiom that is not explicitly stated in the text, namely that the function H_m is symmetric with respect to its arguments. Using this, we can combine any set of arguments of H_m using the extended grouping axiom.

Let $f(m)$ denote $H_m(\frac{1}{m},\frac{1}{m},\ldots,\frac{1}{m})$.

Consider

$$f(mn)=H_{mn}(\frac{1}{mn},\frac{1}{mn},\ldots,\frac{1}{mn}). \tag{2.98}$$

By repeatedly applying the extended grouping axiom, we have

$$\begin{align}
f(mn) &= H_{mn}(\frac{1}{mn},\frac{1}{mn},\ldots,\frac{1}{mn}) \tag{2.99}\\
&= H_{mn-n}(\frac{1}{m},\frac{1}{mn},\ldots,\frac{1}{mn})+\frac{1}{m}H_n(\frac{1}{n},\ldots,\frac{1}{n}) \tag{2.100}\\
&= H_{mn-2n}(\frac{1}{m},\frac{1}{m},\frac{1}{mn},\ldots,\frac{1}{mn})+\frac{2}{m}H_n(\frac{1}{n},\ldots,\frac{1}{n}) \tag{2.101}\\
&\vdots \tag{2.102}\\
&= H_m(\frac{1}{m},\ldots.\frac{1}{m})+H(\frac{1}{n},\ldots,\frac{1}{n}) \tag{2.103}\\
&= f(m)+f(n). \tag{2.104}
\end{align}$$

We can immediately use this to conclude that $f(m^k) = kf(m)$.

Now, we will argue that $H_2(1,0) = h(1) = 0$. We do this by expanding $H_3(p_1,p_2,0)$ ($p_1 + p_2 = 1$) in two different ways using the grouping axiom

$$\begin{aligned} H_3(p_1,p_2,0) &= H_2(p_1,p_2) + p_2H_2(1,0) && (2.105)\\ &= H_2(1,0) + (p_1+p_2)H_2(p_1,p_2) && (2.106)\end{aligned}$$

Thus $p_2H_2(1,0) = H_2(1,0)$ for all p_2, and therefore $H(1,0) = 0$.

We will also need to show that $f(m+1) - f(m) \to 0$ as $m \to \infty$. To prove this, we use the extended grouping axiom and write

$$\begin{aligned} f(m+1) &= H_{m+1}(\frac{1}{m+1},\ldots,\frac{1}{m+1}) && (2.107)\\ &= h(\frac{1}{m+1}) + \frac{m}{m+1}H_m(\frac{1}{m},\ldots,\frac{1}{m}) && (2.108)\\ &= h(\frac{1}{m+1}) + \frac{m}{m+1}f(m) && (2.109)\end{aligned}$$

and therefore

$$f(m+1) - \frac{m}{m+1}f(m) = h(\frac{1}{m+1}). \tag{2.110}$$

Thus $\lim f(m+1) - \frac{m}{m+1}f(m) = \lim h(\frac{1}{m+1})$. But by the continuity of H_2, it follows that the limit on the right is $h(0) = 0$. Thus $\lim h(\frac{1}{m+1}) = 0$.

Let us define

$$a_{n+1} = f(n+1) - f(n) \tag{2.111}$$

and

$$b_n = h(\frac{1}{n}). \tag{2.112}$$

Then

$$\begin{aligned} a_{n+1} &= -\frac{1}{n+1}f(n) + b_{n+1} && (2.113)\\ &= -\frac{1}{n+1}\sum_{i=2}^{n} a_i + b_{n+1} && (2.114)\end{aligned}$$

and therefore

$$(n+1)b_{n+1} = (n+1)a_{n+1} + \sum_{i=2}^{n} a_i. \tag{2.115}$$

Therefore summing over n, we have

$$\sum_{n=2}^{N} nb_n = \sum_{n=2}^{N}(na_n + a_{n-1} + \ldots + a_2) = N\sum_{n=2}^{N} a_i. \tag{2.116}$$

Dividing both sides by $\sum_{n=1}^N n = N(N+1)/2$, we obtain

$$\frac{2}{N+1}\sum_{n=2}^N a_n = \frac{\sum_{n=2}^N nb_n}{\sum_{n=2}^N n} \tag{2.117}$$

Now by continuity of H_2 and the definition of b_n, it follows that $b_n \to 0$ as $n \to \infty$. Since the right hand side is essentially an average of the b_n's, it also goes to 0 (This can be proved more precisely using ϵ's and δ's). Thus the left hand side goes to 0. We can then see that

$$a_{N+1} = b_{N+1} - \frac{1}{N+1}\sum_{n=2}^N a_n \tag{2.118}$$

also goes to 0 as $N \to \infty$. Thus

$$f(n+1) - f(n) \to 0 \quad \text{as} n \to \infty. \tag{2.119}$$

We will now prove the following lemma

Lemma 2.0.1 *Let the function $f(m)$ satisfy the following assumptions:*

- *$f(mn) = f(m) + f(n)$ for all integers m, n.*
- $\lim_{n\to\infty}(f(n+1) - f(n)) = 0$
- *$f(2) = 1$,*

then the function $f(m) = \log_2 m$.

Proof of the lemma: Let P be an arbitrary prime number and let

$$g(n) = f(n) - \frac{f(P)\log_2 n}{\log_2 P} \tag{2.120}$$

Then $g(n)$ satisfies the first assumption of the lemma. Also $g(P) = 0$.

Also if we let

$$\alpha_n = g(n+1) - g(n) = f(n+1) - f(n) + \frac{f(P)}{\log_2 P}\log_2\frac{n}{n+1} \tag{2.121}$$

then the second assumption in the lemma implies that $\lim \alpha_n = 0$.

For an integer n, define

$$n^{(1)} = \left\lfloor \frac{n}{P} \right\rfloor. \tag{2.122}$$

Then it follows that $n^{(1)} < n/P$, and

$$n = n^{(1)}P + l \tag{2.123}$$

where $0 \leq l < P$. From the fact that $g(P) = 0$, it follows that $g(Pn^{(1)}) = g(n^{(1)})$, and

$$g(n) = g(n^{(1)}) + g(n) - g(Pn^{(1)}) = g(n^{(1)}) + \sum_{i=Pn^{(1)}}^{n-1} \alpha_i \tag{2.124}$$

Just as we have defined $n^{(1)}$ from n, we can define $n^{(2)}$ from $n^{(1)}$. Continuing this process, we can then write

$$g(n) = g(n^{(k)}) + \sum_{j=1}^{k} \left(\sum_{i=Pn^{(i)}}^{n^{(i-1)}} \alpha_i \right). \tag{2.125}$$

Since $n^{(k)} \leq n/P^k$, after

$$k = \left\lfloor \frac{\log n}{\log P} \right\rfloor + 1 \tag{2.126}$$

terms, we have $n^{(k)} = 0$, and $g(0) = 0$ (this follows directly from the additive property of g). Thus we can write

$$g(n) = \sum_{i=1}^{t_n} \alpha_i \tag{2.127}$$

the sum of b_n terms, where

$$b_n \leq P \left(\frac{\log n}{\log P} + 1 \right). \tag{2.128}$$

Since $\alpha_n \to 0$, it follows that $\frac{g(n)}{\log_2 n} \to 0$, since $g(n)$ has at most $o(\log_2 n)$ terms α_i. Thus it follows that

$$\lim_{n\to\infty} \frac{f(n)}{\log_2 n} = \frac{f(P)}{\log_2 P} \tag{2.129}$$

Since P was arbitrary, it follows that $f(P)/\log_2 P = c$ for every prime number P. Applying the third axiom in the lemma, it follows that the constant is 1, and $f(P) = \log_2 P$.

For composite numbers $N = P_1 P_2 \ldots P_l$, we can apply the first property of f and the prime number factorization of N to show that

$$f(N) = \sum f(P_i) = \sum \log_2 P_i = \log_2 N. \tag{2.130}$$

Thus the lemma is proved.

The lemma can be simplified considerably, if instead of the second assumption, we replace it by the assumption that $f(n)$ is monotone in n. We will now argue that the only function $f(m)$ such that $f(mn) = f(m) + f(n)$ for all integers m, n is of the form $f(m) = \log_a m$ for some base a.

Let $c = f(2)$. Now $f(4) = f(2 \times 2) = f(2) + f(2) = 2c$. Similarly, it is easy to see that $f(2^k) = kc = c \log_2 2^k$. We will extend this to integers that are not powers of 2.

For any integer m, let $r > 0$, be another integer and let $2^k \leq m^r < 2^{k+1}$. Then by the monotonicity assumption on f, we have

$$kc \leq rf(m) < (k+1)c \tag{2.131}$$

or

$$c\frac{k}{r} \leq f(m) < c\frac{k+1}{r} \tag{2.132}$$

Now by the monotonicity of $\log$, we have

$$\frac{k}{r} \leq \log_2 m < \frac{k+1}{r} \tag{2.133}$$

Combining these two equations, we obtain

$$\left| f(m) - \frac{\log_2 m}{c} \right| < \frac{1}{r} \tag{2.134}$$

Since r was arbitrary, we must have

$$f(m) = \frac{\log_2 m}{c} \tag{2.135}$$

and we can identify $c = 1$ from the last assumption of the lemma.

Now we are almost done. We have shown that for any uniform distribution on m outcomes, $f(m) = H_m(1/m, \ldots, 1/m) = \log_2 m$.

We will now show that

$$H_2(p, 1-p) = -p\log p - (1-p)\log(1-p). \tag{2.136}$$

To begin, let p be a rational number, r/s, say. Consider the extended grouping axiom for H_s

$$\begin{aligned} f(s) = H_s(\frac{1}{s}, \ldots, \frac{1}{s}) &= H(\underbrace{\frac{1}{s}, \ldots, \frac{1}{s}}_{r}, \frac{s-r}{s}) + \frac{s-r}{s} f(s-r) && (2.137) \\ &= H_2(\frac{r}{s}, \frac{s-r}{s}) + \frac{s}{r} f(s) + \frac{s-r}{s} f(s-r) && (2.138) \end{aligned}$$

Substituting $f(s) = \log_2 s$, etc, we obtain

$$H_2(\frac{r}{s}, \frac{s-r}{s}) = -\frac{r}{s}\log_2\frac{r}{s} - \left(1 - \frac{s-r}{s}\right)\log_2\left(1 - \frac{s-r}{s}\right). \tag{2.139}$$

Thus (2.136) is true for rational p. By the continuity assumption, (2.136) is also true at irrational p.

To complete the proof, we have to extend the definition from H_2 to H_m, i.e., we have to show that

$$H_m(p_1, \ldots, p_m) = -\sum p_i \log p_i \tag{2.140}$$

for all m. This is a straightforward induction. We have just shown that this is true for $m = 2$. Now assume that it is true for $m = n-1$. By the grouping axiom,

$$\begin{aligned} H_n(p_1, \ldots, p_n) &= H_{n-1}(p_1+p_2, p_3, \ldots, p_n) & (2.141)\\ &\quad +(p_1+p_2)H_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right) & (2.142)\\ &= -(p_1+p_2)\log(p_1+p_2) - \sum_{i=3}^{n} p_i \log p_i & (2.143)\\ &\quad -\frac{p_1}{p_1+p_2}\log\frac{p_1}{p_1+p_2} - \frac{p_2}{p_1+p_2}\log\frac{p_2}{p_1+p_2} & (2.144)\\ &= -\sum_{i=1}^{n} p_i \log p_i. & (2.145) \end{aligned}$$

Thus the statement is true for $m = n$, and by induction, it is true for all m. Thus we have finally proved that the only symmetric function that satisfies the axioms is

$$H_m(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i. \tag{2.146}$$

The proof above is due to Rényi[10]

47. **The entropy of a missorted file.**
A deck of n cards in order $1, 2, \ldots, n$ is provided. One card is removed at random then replaced at random. What is the entropy of the resulting deck?

Solution: *The entropy of a missorted file.*

The heart of this problem is simply carefully counting the possible outcome states. There are n ways to choose which card gets mis-sorted, and, once the card is chosen, there are again n ways to choose where the card is replaced in the deck. Each of these shuffling actions has probability $1/n^2$. Unfortunately, not all of these n^2 actions results in a unique mis-sorted file. So we need to carefully count the number of distinguishable outcome states. The resulting deck can only take on one of the following three cases.

- The selected card is at its original location after a replacement.
- The selected card is at most one location away from its original location after a replacement.
- The selected card is at least two locations away from its original location after a replacement.

To compute the entropy of the resulting deck, we need to know the probability of each case.

Case 1 (resulting deck is the same as the original): There are n ways to achieve this outcome state, one for each of the n cards in the deck. Thus, the probability associated with case 1 is $n/n^2 = 1/n$.

Case 2 (adjacent pair swapping): There are $n-1$ adjacent pairs, each of which will have a probability of $2/n^2$, since for each pair, there are two ways to achieve the swap, either by selecting the left-hand card and moving it one to the right, or by selecting the right-hand card and moving it one to the left.

Case 3 (typical situation): None of the remaining actions "collapses". They all result in unique outcome states, each with probability $1/n^2$. Of the n^2 possible shuffling actions, $n^2-n-2(n-1)$ of them result in this third case (we've simply subtracted the case 1 and case 2 situations above).

The entropy of the resulting deck can be computed as follows.

$$\begin{aligned} H(X) &= \frac{1}{n}log(n) + (n-1)\frac{2}{n^2}log(\frac{n^2}{2}) + (n^2-3n+2)\frac{1}{n^2}log(n^2) \\ &= \frac{2n-1}{n}log(n) - \frac{2(n-1)}{n^2} \end{aligned}$$

48. **Sequence length.**
How much information does the length of a sequence give about the content of a sequence? Suppose we consider a Bernoulli $(1/2)$ process $\{X_i\}$.

Stop the process when the first 1 appears. Let N designate this stopping time. Thus X^N is an element of the set of all finite length binary sequences $\{0,1\}^* = \{0,1,00,01,10,11,000,\ldots\}$.

(a) Find $I(N;X^N)$.

(b) Find $H(X^N|N)$.

(c) Find $H(X^N)$.

Let's now consider a different stopping time. For this part, again assume $X_i \sim$ Bernoulli $(1/2)$ but stop at time $N=6$, with probability $1/3$ and stop at time $N=12$ with probability $2/3$. Let this stopping time be independent of the sequence $X_1X_2\ldots X_{12}$.

(d) Find $I(N;X^N)$.

(e) Find $H(X^N|N)$.

(f) Find $H(X^N)$.

Solution:

(a)

$$\begin{aligned} I(X^N;N) &= H(N) - H(N|X^N) \\ &= H(N) - 0 \end{aligned}$$

$$\begin{aligned} &\overset{(a)}{=} E(N) \\ I(X^N;N) &= 2 \end{aligned}$$

where (a) comes from the fact that the entropy of a geometric random variable is just the mean.

(b) Since given N we know that $X_i = 0$ for all $i < N$ and $X_N = 1$,

$$H(X^N|N) = 0.$$

(c)

$$\begin{aligned} H(X^N) &= I(X^N;N) + H(X^N|N) \\ &= I(X^N;N) + 0 \\ H(X^N) &= 2. \end{aligned}$$

(d)

$$\begin{aligned} I(X^N;N) &= H(N) - H(N|X^N) \\ &= H(N) - 0 \\ I(X^N;N) &= H_B(1/3) \end{aligned}$$

(e)

$$\begin{aligned} H(X^N|N) &= \frac{1}{3}H(X^6|N=6) + \frac{2}{3}H(X^{12}|N=12) \\ &= \frac{1}{3}H(X^6) + \frac{2}{3}H(X^{12}) \\ &= \frac{1}{3}6 + \frac{2}{3}12 \\ H(X^N|N) &= 10. \end{aligned}$$

(f)

$$\begin{aligned} H(X^N) &= I(X^N;N) + H(X^N|N) \\ &= I(X^N;N) + 10 \\ H(X^N) &= H(1/3) + 10. \end{aligned}$$

Chapter 3

The Asymptotic Equipartition Property

1. **Markov's inequality and Chebyshev's inequality.**

   (a) (Markov's inequality.) For any non-negative random variable X and any $t > 0$, show that
   $$\Pr\{X \geq t\} \leq \frac{EX}{t}. \tag{3.1}$$
   Exhibit a random variable that achieves this inequality with equality.

   (b) (Chebyshev's inequality.) Let Y be a random variable with mean μ and variance σ^2. By letting $X = (Y-\mu)^2$, show that for any $\epsilon > 0$,
   $$\Pr\{|Y-\mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}. \tag{3.2}$$

   (c) (The weak law of large numbers.) Let $Z_1, Z_2, \ldots, Z_n$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Let $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ be the sample mean. Show that
   $$\Pr\left\{\left|\overline{Z}_n - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}. \tag{3.3}$$
   Thus $\Pr\left\{\left|\overline{Z}_n - \mu\right| > \epsilon\right\} \to 0$ as $n \to \infty$. This is known as the weak law of large numbers.

Solution: *Markov's inequality and Chebyshev's inequality.*

(a) If X has distribution $F(x)$,
$$\begin{aligned} EX &= \int_0^\infty x \mathrm{d}F \\ &= \int_0^\delta x \mathrm{d}F + \int_\delta^\infty x \mathrm{d}F \end{aligned}$$

$$\begin{aligned}
&\geq \int_{\delta}^{\infty} x \mathrm{d}F \\
&\geq \int_{\delta}^{\infty} \delta \mathrm{d}F \\
&= \delta \Pr\{X \geq \delta\}.
\end{aligned}$$

Rearranging sides and dividing by δ we get,

$$\Pr\{X \geq \delta\} \leq \frac{EX}{\delta}. \tag{3.4}$$

One student gave a proof based on conditional expectations. It goes like

$$\begin{aligned}
EX &= E(X|X \leq \delta)\Pr\{X \geq \delta\} + E(X|X < \delta)\Pr\{X < \delta\} \\
&\geq E(X|X \leq \delta)\Pr\{X \geq \delta\} \\
&\geq \delta \Pr\{X \geq \delta\},
\end{aligned}$$

which leads to (3.4) as well.
Given δ, the distribution achieving

$$\Pr\{X \geq \delta\} = \frac{EX}{\delta},$$

is

$$X = \begin{cases} \delta & \text{with probability } \frac{\mu}{\delta} \\ 0 & \text{with probability } 1 - \frac{\mu}{\delta}, \end{cases}$$

where $\mu \leq \delta$.

(b) Letting $X = (Y-\mu)^2$ in Markov's inequality,

$$\begin{aligned}
\Pr\{(Y-\mu)^2 > \epsilon^2\} &\leq \Pr\{(Y-\mu)^2 \geq \epsilon^2\} \\
&\leq \frac{E(Y-\mu)^2}{\epsilon^2} \\
&= \frac{\sigma^2}{\epsilon^2},
\end{aligned}$$

and noticing that $\Pr\{(Y-\mu)^2 > \epsilon^2\} = \Pr\{|Y-\mu| > \epsilon\}$, we get,

$$\Pr\{|Y-\mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}.$$

(c) Letting Y in Chebyshev's inequality from part (b) equal $\bar{Z}_n$, and noticing that $E\bar{Z}_n = \mu$ and $\mathrm{Var}(\bar{Z}_n) = \frac{\sigma^2}{n}$ (ie. $\bar{Z}_n$ is the sum of n iid r.v.'s, $\frac{Z_i}{n}$, each with variance $\frac{\sigma^2}{n^2}$), we have,

$$\Pr\{|\bar{Z}_n - \mu| > \epsilon\} \leq \frac{\sigma^2}{n\epsilon^2}.$$

2. **AEP and mutual information**. Let (X_i, Y_i) be i.i.d. $\sim p(x,y)$. We form the log likelihood ratio of the hypothesis that X and Y are independent vs. the hypothesis that X and Y are dependent. What is the limit of

$$\frac{1}{n}\log\frac{p(X^n)p(Y^n)}{p(X^n,Y^n)}?$$

Solution:

$$\begin{aligned}
\frac{1}{n}\log\frac{p(X^n)p(Y^n)}{p(X^n,Y^n)} &= \frac{1}{n}\log\prod_{i=1}^{n}\frac{p(X_i)p(Y_i)}{p(X_i,Y_i)} \\
&= \frac{1}{n}\sum_{i=i}^{n}\log\frac{p(X_i)p(Y_i)}{p(X_i,Y_i)} \\
&\to E(\log\frac{p(X_i)p(Y_i)}{p(X_i,Y_i)}) \\
&= -I(X;Y)
\end{aligned}$$

Thus, $\frac{p(X^n)p(Y^n)}{p(X^n,Y^n)} \to 2^{-nI(X;Y)}$, which will converge to 1 if X and Y are indeed independent.

3. **Piece of cake**
A cake is sliced roughly in half, the largest piece being chosen each time, the other pieces discarded. We will assume that a random cut creates pieces of proportions:

$$P = \begin{cases} (\frac{2}{3},\frac{1}{3}) & \text{w.p. } \frac{3}{4} \\ (\frac{2}{5},\frac{3}{5}) & \text{w.p. } \frac{1}{4} \end{cases}$$

Thus, for example, the first cut (and choice of largest piece) may result in a piece of size $\frac{3}{5}$. Cutting and choosing from this piece might reduce it to size $(\frac{3}{5})(\frac{2}{3})$ at time 2, and so on.

How large, to first order in the exponent, is the piece of cake after n cuts?

Solution: Let C_i be the fraction of the piece of cake that is cut at the ith cut, and let T_n be the fraction of cake left after n cuts. Then we have $T_n = C_1C_2\ldots C_n = \prod_{i=1}^n C_i$. Hence, as in Question 2 of Homework Set #3,

$$\begin{aligned}
\lim\frac{1}{n}\log T_n &= \lim\frac{1}{n}\sum_{i=1}^{n}\log C_i \\
&= E[\log C_1] \\
&= \frac{3}{4}\log\frac{2}{3}+\frac{1}{4}\log\frac{3}{5}.
\end{aligned}$$

4. **AEP**
Let X_i be iid $\sim p(x)$, $x \in \{1,2,\ldots,m\}$. Let $\mu = EX$, and $H = -\sum p(x)\log p(x)$. Let $A^n = \{x^n \in \mathcal{X}^n : |-\frac{1}{n}\log p(x^n) - H| \le \epsilon\}$. Let $B^n = \{x^n \in \mathcal{X}^n : |\frac{1}{n}\sum_{i=1}^n X_i - \mu| \le \epsilon\}$.

   (a) Does $\Pr\{X^n \in A^n\} \longrightarrow 1$?

   (b) Does $\Pr\{X^n \in A^n \cap B^n\} \longrightarrow 1$?

   (c) Show $|A^n \cap B^n| \le 2^{n(H+\epsilon)}$, for all n.

   (d) Show $|A^n \cap B^n| \ge (\frac{1}{2})2^{n(H-\epsilon)}$, for n sufficiently large.

   Solution:

   (a) Yes, by the AEP for discrete random variables the probability X^n is typical goes to 1.

   (b) Yes, by the Strong Law of Large Numbers $Pr(X^n \in B^n) \to 1$. So there exists $\epsilon > 0$ and N_1 such that $Pr(X^n \in A^n) > 1 - \frac{\epsilon}{2}$ for all $n > N_1$, and there exists N_2 such that $Pr(X^n \in B^n) > 1 - \frac{\epsilon}{2}$ for all $n > N_2$. So for all $n > \max(N_1, N_2)$:

$$\begin{aligned} Pr(X^n \in A^n \cap B^n) &= Pr(X^n \in A^n) + Pr(X^n \in B^n) - Pr(X^n \in A^n \cup B^n) \\ &> 1 - \frac{\epsilon}{2} + 1 - \frac{\epsilon}{2} - 1 \\ &= 1 - \epsilon \end{aligned}$$

   So for any $\epsilon > 0$ there exists $N = \max(N_1, N_2)$ such that $Pr(X^n \in A^n \cap B^n) > 1 - \epsilon$ for all $n > N$, therefore $Pr(X^n \in A^n \cap B^n) \to 1$.

   (c) By the law of total probability $\sum_{x^n \in A^n \cap B^n} p(x^n) \le 1$. Also, for $x^n \in A^n$, from Theorem 3.1.2 in the text, $p(x^n) \ge 2^{-n(H+\epsilon)}$. Combining these two equations gives $1 \ge \sum_{x^n \in A^n \cap B^n} p(x^n) \ge \sum_{x^n \in A^n \cap B^n} 2^{-n(H+\epsilon)} = |A^n \cap B^n| 2^{-n(H+\epsilon)}$. Multiplying through by $2^{n(H+\epsilon)}$ gives the result $|A^n \cap B^n| \le 2^{n(H+\epsilon)}$.

   (d) Since from (b) $Pr\{X^n \in A^n \cap B^n\} \to 1$, there exists N such that $Pr\{X^n \in A^n \cap B^n\} \ge \frac{1}{2}$ for all $n > N$. From Theorem 3.1.2 in the text, for $x^n \in A^n$, $p(x^n) \le 2^{-n(H-\epsilon)}$. So combining these two gives $\frac{1}{2} \le \sum_{x^n \in A^n \cap B^n} p(x^n) \le \sum_{x^n \in A^n \cap B^n} 2^{-n(H-\epsilon)} = |A^n \cap B^n| 2^{-n(H-\epsilon)}$. Multiplying through by $2^{n(H-\epsilon)}$ gives the result $|A^n \cap B^n| \ge (\frac{1}{2})2^{n(H-\epsilon)}$ for n sufficiently large.

5. **Sets defined by probabilities.**
Let $X_1, X_2, \ldots$ be an i.i.d. sequence of discrete random variables with entropy $H(X)$. Let

$$C_n(t) = \{x^n \in \mathcal{X}^n : p(x^n) \ge 2^{-nt}\}$$

   denote the subset of n-sequences with probabilities $\ge 2^{-nt}$.

   (a) Show $|C_n(t)| \le 2^{nt}$.

   (b) For what values of t does $P(\{X^n \in C_n(t)\}) \to 1$?

   Solution:

(a) Since the total probability of all sequences is less than 1, $|C_n(t)| \min_{x^n \in C_n(t)} p(x^n) \leq 1$, and hence $|C_n(t)| \leq 2^{nt}$.

(b) Since $-\frac{1}{n} \log p(x^n) \to H$, if $t < H$, the probability that $p(x^n) > 2^{-nt}$ goes to 0, and if $t > H$, the probability goes to 1.

6. **An AEP-like limit.** Let $X_1, X_2, \ldots$ be i.i.d. drawn according to probability mass function $p(x)$. Find

$$\lim_{n \to \infty} [p(X_1, X_2, \ldots, X_n)]^{\frac{1}{n}} .$$

Solution: *An AEP-like limit.* $X_1, X_2, \ldots$, i.i.d. $\sim p(x)$. Hence $\log(X_i)$ are also i.i.d. and

$$\begin{aligned} \lim (p(X_1, X_2, \ldots, X_n))^{\frac{1}{n}} &= \lim 2^{\log(p(X_1, X_2, \ldots, X_n))^{\frac{1}{n}}} \\ &= 2^{\lim \frac{1}{n} \sum \log p(X_i)} \text{ a.e.} \\ &= 2^{E(\log(p(X)))} \text{ a.e.} \\ &= 2^{-H(X)} \text{ a.e.} \end{aligned}$$

by the strong law of large numbers (assuming of course that H(X) exists).

7. **The AEP and source coding.** A discrete memoryless source emits a sequence of statistically independent binary digits with probabilities $p(1) = 0.005$ and $p(0) = 0.995$. The digits are taken 100 at a time and a binary codeword is provided for every sequence of 100 digits containing three or fewer ones.

(a) Assuming that all codewords are the same length, find the minimum length required to provide codewords for all sequences with three or fewer ones.

(b) Calculate the probability of observing a source sequence for which no codeword has been assigned.

(c) Use Chebyshev's inequality to bound the probability of observing a source sequence for which no codeword has been assigned. Compare this bound with the actual probability computed in part (b).

Solution: *The AEP and source coding.*

(a) The number of 100-bit binary sequences with three or fewer ones is

$$\binom{100}{0} + \binom{100}{1} + \binom{100}{2} + \binom{100}{3} = 1 + 100 + 4950 + 161700 = 166751 .$$

The required codeword length is $\lceil \log_2 166751 \rceil = 18$. (Note that $H(0.005) = 0.0454$, so 18 is quite a bit larger than the 4.5 bits of entropy.)

(b) The probability that a 100-bit sequence has three or fewer ones is

$$\sum_{i=0}^{3} \binom{100}{i} (0.005)^i (0.995)^{100-i} = 0.60577 + 0.30441 + 0.7572 + 0.01243 = 0.99833$$

Thus the probability that the sequence that is generated cannot be encoded is $1 - 0.99833 = 0.00167$.

(c) In the case of a random variable S_n that is the sum of n i.i.d. random variables $X_1, X_2, \ldots, X_n$, Chebyshev's inequality states that

$$\Pr(|S_n - n\mu| \geq \epsilon) \;\leq\; \frac{n\sigma^2}{\epsilon^2}\,,$$

where μ and σ^2 are the mean and variance of X_i. (Therefore $n\mu$ and $n\sigma^2$ are the mean and variance of S_n.) In this problem, $n = 100$, $\mu = 0.005$, and $\sigma^2 = (0.005)(0.995)$. Note that $S_{100} \geq 4$ if and only if $|S_{100} - 100(0.005)| \geq 3.5$, so we should choose $\epsilon = 3.5$. Then

$$\Pr(S_{100} \geq 4) \leq \frac{100(0.005)(0.995)}{(3.5)^2} \approx 0.04061\,.$$

This bound is much larger than the actual probability 0.00167.

8. **Products.** Let

$$X = \begin{cases} 1, & \frac{1}{2} \\ 2, & \frac{1}{4} \\ 3, & \frac{1}{4} \end{cases}$$

Let $X_1, X_2, \ldots$ be drawn i.i.d. according to this distribution. Find the limiting behavior of the product

$$(X_1 X_2 \cdots X_n)^{\frac{1}{n}}\quad.$$

Solution: *Products.* Let

$$P_n = (X_1 X_2 \ldots X_n)^{\frac{1}{n}}\,. \tag{3.5}$$

Then

$$\log P_n = \frac{1}{n}\sum_{i=1}^{n} \log X_i \rightarrow E \log X, \tag{3.6}$$

with probability 1, by the strong law of large numbers. Thus $P_n \rightarrow 2^{E\log X}$ with prob. 1. We can easily calculate $E\log X = \frac{1}{2}\log 1 + \frac{1}{4}\log 2 + \frac{1}{4}\log 3 = \frac{1}{4}\log 6$, and therefore $P_n \rightarrow 2^{\frac{1}{4}\log 6} = 1.565$.

9. **AEP**. Let $X_1, X_2, \ldots$ be independent identically distributed random variables drawn according to the probability mass function $p(x), x \in \{1, 2, \ldots, m\}$. Thus $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n p(x_i)$. We know that $-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \rightarrow H(X)$ in probability. Let $q(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n q(x_i)$, where q is another probability mass function on $\{1, 2, \ldots, m\}$.

(a) Evaluate $\lim -\frac{1}{n}\log q(X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$.

(b) Now evaluate the limit of the log likelihood ratio $\frac{1}{n}\log\frac{q(X_1,\ldots,X_n)}{p(X_1,\ldots,X_n)}$ when $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. Thus the odds favoring q are exponentially small when p is true.

Solution: (*AEP*).

(a) Since the $X_1, X_2, \ldots, X_n$ are i.i.d., so are $q(X_1), q(X_2), \ldots, q(X_n)$, and hence we can apply the strong law of large numbers to obtain

$$\begin{align} \lim -\frac{1}{n} \log q(X_1, X_2, \ldots, X_n) &= \lim -\frac{1}{n} \sum \log q(X_i) \tag{3.7} \\ &= -E(\log q(X)) \text{ w.p. } 1 \tag{3.8} \\ &= -\sum p(x) \log q(x) \tag{3.9} \\ &= \sum p(x) \log \frac{p(x)}{q(x)} - \sum p(x) \log p(x) \tag{3.10} \\ &= D(\mathbf{p}||\mathbf{q}) + H(\mathbf{p}). \tag{3.11} \end{align}$$

(b) Again, by the strong law of large numbers,

$$\begin{align} \lim -\frac{1}{n} \log \frac{q(X_1, X_2, \ldots, X_n)}{p(X_1, X_2, \ldots, X_n)} &= \lim -\frac{1}{n} \sum \log \frac{q(X_i)}{p(X_i)} \tag{3.12} \\ &= -E(\log \frac{q(X)}{p(X)}) \text{ w.p. } 1 \tag{3.13} \\ &= -\sum p(x) \log \frac{q(x)}{p(x)} \tag{3.14} \\ &= \sum p(x) \log \frac{p(x)}{q(x)} \tag{3.15} \\ &= D(\mathbf{p}||\mathbf{q}). \tag{3.16} \end{align}$$

10. **Random box size.** An n-dimensional rectangular box with sides $X_1, X_2, X_3, \ldots, X_n$ is to be constructed. The volume is $V_n = \prod_{i=1}^n X_i$. The edge length l of a n-cube with the same volume as the random box is $l = V_n^{1/n}$. Let $X_1, X_2, \ldots$ be i.i.d. uniform random variables over the unit interval $[0,1]$. Find $\lim_{n\to\infty} V_n^{1/n}$, and compare to $(EV_n)^{\frac{1}{n}}$. Clearly the expected edge length does not capture the idea of the volume of the box. The geometric mean, rather than the arithmetic mean, characterizes the behavior of products.

Solution: *Random box size.* The volume $V_n = \prod_{i=1}^n X_i$ is a random variable, since the X_i are random variables uniformly distributed on $[0,1]$. V_n tends to 0 as $n \to \infty$. However

$$\log_e V_n^{\frac{1}{n}} = \frac{1}{n} \log_e V_n = \frac{1}{n} \sum \log_e X_i \to E(\log_e(X)) \text{ a.e.}$$

by the Strong Law of Large Numbers, since X_i and $\log_e(X_i)$ are i.i.d. and $E(\log_e(X)) < \infty$. Now

$$E(\log_e(X_i)) = \int_0^1 \log_e(x)\, dx = -1$$

Hence, since e^x is a continuous function,

$$\lim_{n\to\infty} V_n^{\frac{1}{n}} = e^{\lim_{n\to\infty} \frac{1}{n} \log_e V_n} = \frac{1}{e} < \frac{1}{2}.$$

Thus the "effective" edge length of this solid is e^{-1}. Note that since the X_i's are independent, $E(V_n) = \prod E(X_i) = (\frac{1}{2})^n$. Also $\frac{1}{2}$ is the arithmetic mean of the random variable, and $\frac{1}{e}$ is the geometric mean.

11. **Proof of Theorem 3.3.1**. This problem shows that the size of the smallest "probable" set is about 2^{nH}. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. Let $B_\delta^{(n)} \subset \mathcal{X}^n$ such that $\Pr(B_\delta^{(n)}) > 1-\delta$. Fix $\epsilon < \frac{1}{2}$.

(a) Given any two sets A, B such that $\Pr(A) > 1-\epsilon_1$ and $\Pr(B) > 1-\epsilon_2$, show that $\Pr(A \cap B) > 1-\epsilon_1-\epsilon_2$. Hence $\Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) \geq 1-\epsilon-\delta.$

(b) Justify the steps in the chain of inequalities

$$\begin{aligned} 1-\epsilon-\delta &\leq \Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) && (3.17)\\ &= \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) && (3.18)\\ &\leq \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\epsilon)} && (3.19)\\ &= |A_\epsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H-\epsilon)} && (3.20)\\ &\leq |B_\delta^{(n)}| 2^{-n(H-\epsilon)}. && (3.21) \end{aligned}$$

(c) Complete the proof of the theorem.

Solution: *Proof of Theorem 3.3.1.*

(a) Let A^c denote the complement of A. Then

$$P(A^c \cup B^c) \leq P(A^c) + P(B^c). \tag{3.22}$$

Since $P(A) \geq 1-\epsilon_1$, $P(A^c) \leq \epsilon_1$. Similarly, $P(B^c) \leq \epsilon_2$. Hence

$$\begin{aligned} P(A \cap B) &= 1 - P(A^c \cup B^c) && (3.23)\\ &\geq 1 - P(A^c) - P(B^c) && (3.24)\\ &\geq 1 - \epsilon_1 - \epsilon_2. && (3.25) \end{aligned}$$

(b) To complete the proof, we have the following chain of inequalities

$$\begin{aligned} 1-\epsilon-\delta &\overset{(a)}{\leq} \Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) && (3.26)\\ &\overset{(b)}{=} \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) && (3.27)\\ &\overset{(c)}{\leq} \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\epsilon)} && (3.28)\\ &\overset{(d)}{=} |A_\epsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H-\epsilon)} && (3.29)\\ &\overset{(e)}{\leq} |B_\delta^{(n)}| 2^{-n(H-\epsilon)}. && (3.30) \end{aligned}$$

where (a) follows from the previous part, (b) follows by definition of probability of a set, (c) follows from the fact that the probability of elements of the typical set are bounded by $2^{-n(H-\epsilon)}$, (d) from the definition of $|A_\epsilon^{(n)} \cap B_\delta^{(n)}|$ as the cardinality of the set $A_\epsilon^{(n)} \cap B_\delta^{(n)}$, and (e) from the fact that $A_\epsilon^{(n)} \cap B_\delta^{(n)} \subseteq B_\delta^{(n)}$.

12. **Monotonic convergence of the empirical distribution.** Let $\hat{p}_n$ denote the empirical probability mass function corresponding to $X_1, X_2, \ldots, X_n$ i.i.d. $\sim p(x), x \in \mathcal{X}$. Specifically,

$$\hat{p}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = x)$$

is the proportion of times that $X_i = x$ in the first n samples, where I is the indicator function.

(a) Show for $\mathcal{X}$ binary that

$$ED(\hat{p}_{2n} \parallel p) \le ED(\hat{p}_n \parallel p).$$

Thus the expected relative entropy "distance" from the empirical distribution to the true distribution decreases with sample size.
Hint: Write $\hat{p}_{2n} = \frac{1}{2}\hat{p}_n + \frac{1}{2}\hat{p}_n'$ and use the convexity of D.

(b) Show for an arbitrary discrete $\mathcal{X}$ that

$$ED(\hat{p}_n \parallel p) \le ED(\hat{p}_{n-1} \parallel p).$$

Hint: Write $\hat{p}_n$ as the average of n empirical mass functions with each of the n samples deleted in turn.

Solution: *Monotonic convergence of the empirical distribution.*

(a) Note that,

$$\begin{aligned}
\hat{p}_{2n}(x) &= \frac{1}{2n}\sum_{i=1}^{2n} I(X_i = x) \\
&= \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} I(X_i = x) + \frac{1}{2}\frac{1}{n}\sum_{i=n+1}^{2n} I(X_i = x) \\
&= \frac{1}{2}\hat{p}_n(x) + \frac{1}{2}\hat{p}_n'(x).
\end{aligned}$$

Using convexity of $D(p||q)$ we have that,

$$\begin{aligned}
D(\hat{p}_{2n}||p) &= D(\frac{1}{2}\hat{p}_n + \frac{1}{2}\hat{p}_n'||\frac{1}{2}p + \frac{1}{2}p) \\
&\le \frac{1}{2}D(\hat{p}_n||p) + \frac{1}{2}D(\hat{p}_n'||p).
\end{aligned}$$

Taking expectations and using the fact the X_i's are identically distributed we get,

$$ED(\hat{p}_{2n}||p) \le ED(\hat{p}_n||p).$$

(b) The trick to this part is similar to part a) and involves rewriting $\hat{p}_n$ in terms of $\hat{p}_{n-1}$. We see that,

$$\hat{p}_n = \frac{1}{n}\sum_{i=0}^{n-1} I(X_i = x) + \frac{I(X_n = x)}{n}$$

or in general,

$$\hat{p}_n = \frac{1}{n}\sum_{i \neq j} I(X_i = x) + \frac{I(X_j = x)}{n},$$

where j ranges from 1 to n.

Summing over j we get,

$$n\hat{p}_n = \frac{n-1}{n}\sum_{j=1}^{n} \hat{p}_{n-1}^{j} + \hat{p}_n,$$

or,

$$\hat{p}_n = \frac{1}{n}\sum_{j=1}^{n} \hat{p}_{n-1}^{j}$$

where,

$$\sum_{j=1}^{n} \hat{p}_{n-1}^{j} = \frac{1}{n-1}\sum_{i \neq j} I(X_i = x).$$

Again using the convexity of $D(p||q)$ and the fact that the $D(\hat{p}_{n-1}^{j}||p)$ are identically distributed for all j and hence have the same expected value, we obtain the final result.

13. **Calculation of typical set** To clarify the notion of a typical set $A_\epsilon^{(n)}$ and the smallest set of high probability $B_\delta^{(n)}$, we will calculate the set for a simple example. Consider a sequence of i.i.d. binary random variables, $X_1, X_2, \ldots, X_n$, where the probability that $X_i = 1$ is 0.6 (and therefore the probability that $X_i = 0$ is 0.4).

(a) Calculate $H(X)$.

(b) With $n = 25$ and $\epsilon = 0.1$, which sequences fall in the typical set $A_\epsilon^{(n)}$? What is the probability of the typical set? How many elements are there in the typical set? (This involves computation of a table of probabilities for sequences with k 1's, $0 \leq k \leq 25$, and finding those sequences that are in the typical set.)

k	$\binom{n}{k}$	$\binom{n}{k}p^k(1-p)^{n-k}$	$-\frac{1}{n}\log p(x^n)$
0	1	0.000000	1.321928
1	25	0.000000	1.298530
2	300	0.000000	1.275131
3	2300	0.000001	1.251733
4	12650	0.000007	1.228334
5	53130	0.000054	1.204936
6	177100	0.000227	1.181537
7	480700	0.001205	1.158139
8	1081575	0.003121	1.134740
9	2042975	0.013169	1.111342
10	3268760	0.021222	1.087943
11	4457400	0.077801	1.064545
12	5200300	0.075967	1.041146
13	5200300	0.267718	1.017748
14	4457400	0.146507	0.994349
15	3268760	0.575383	0.970951
16	2042975	0.151086	0.947552
17	1081575	0.846448	0.924154
18	480700	0.079986	0.900755
19	177100	0.970638	0.877357
20	53130	0.019891	0.853958
21	12650	0.997633	0.830560
22	2300	0.001937	0.807161
23	300	0.999950	0.783763
24	25	0.000047	0.760364
25	1	0.000003	0.736966

(c) How many elements are there in the smallest set that has probability 0.9?

(d) How many elements are there in the intersection of the sets in part (b) and (c)? What is the probability of this intersection?

Solution:

(a) $H(X) = -0.6\log 0.6 - 0.4\log 0.4 = 0.97095$ bits.

(b) By definition, $A_\epsilon^{(n)}$ for $\epsilon = 0.1$ is the set of sequences such that $-\frac{1}{n}\log p(x^n)$ lies in the range $(H(X)-\epsilon, H(X)+\epsilon)$, i.e., in the range (0.87095, 1.07095). Examining the last column of the table, it is easy to see that the typical set is the set of all sequences with the number of ones lying between 11 and 19.

The probability of the typical set can be calculated from cumulative probability column. The probability that the number of 1's lies between 11 and 19 is equal to $F(19) - F(10) = 0.970638 - 0.034392 = 0.936246$. Note that this is greater than $1-\epsilon$, i.e., the n is large enough for the probability of the typical set to be greater than $1-\epsilon$.

The number of elements in the typical set can be found using the third column.

$$|A_\epsilon^{(n)}| = \sum_{k=11}^{19} \binom{n}{k} = \sum_{k=0}^{19} \binom{n}{k} - \sum_{k=0}^{10} \binom{n}{k} = 33486026 - 7119516 = 26366510. \tag{3.31}$$

Note that the upper and lower bounds for the size of the $A_\epsilon^{(n)}$ can be calculated as $2^{n(H+\epsilon)} = 2^{25(0.97095+0.1)} = 2^{26.77} = 1.147365 \times 10^8$, and $(1-\epsilon)2^{n(H-\epsilon)} = 0.9 \times 2^{25(0.97095-0.1)} = 0.9 \times 2^{21.9875} = 3742308$. Both bounds are very loose!

(c) To find the smallest set $B_\delta^{(n)}$ of probability 0.9, we can imagine that we are filling a bag with pieces such that we want to reach a certain weight with the minimum number of pieces. To minimize the number of pieces that we use, we should use the largest possible pieces. In this case, it corresponds to using the sequences with the highest probability.

Thus we keep putting the high probability sequences into this set until we reach a total probability of 0.9. Looking at the fourth column of the table, it is clear that the probability of a sequence increases monotonically with k. Thus the set consists of sequences of $k = 25, 24, \ldots$, until we have a total probability 0.9.

Using the cumulative probability column, it follows that the set $B_\delta^{(n)}$ consist of sequences with $k \geq 13$ and some sequences with $k = 12$. The sequences with $k \geq 13$ provide a total probability of $1 - 0.153768 = 0.846232$ to the set $B_\delta^{(n)}$. The remaining probability of $0.9 - 0.846232 = 0.053768$ should come from sequences with $k = 12$. The number of such sequences needed to fill this probability is at least $0.053768/p(x^n) = 0.053768/1.460813 \times 10^{-8} = 3680690.1$, which we round up to 3680691. Thus the smallest set with probability 0.9 has $33554432 - 16777216 + 3680691 = 20457907$ sequences. Note that the set $B_\delta^{(n)}$ is not uniquely defined - it could include any 3680691 sequences with $k = 12$. However, the size of the smallest set is well defined.

(d) The intersection of the sets $A_\epsilon^{(n)}$ and $B_\delta^{(n)}$ in parts (b) and (c) consists of all sequences with k between 13 and 19, and 3680691 sequences with $k = 12$. The probability of this intersection $= 0.970638 - 0.153768 + 0.053768 = 0.870638$, and the size of this intersection $= 33486026 - 16777216 + 3680691 = 20389501$.

Chapter 4

Entropy Rates of a Stochastic Process

1. **Doubly stochastic matrices.** An $n \times n$ matrix $P = [P_{ij}]$ is said to be *doubly stochastic* if $P_{ij} \geq 0$ and $\sum_j P_{ij} = 1$ for all i and $\sum_i P_{ij} = 1$ for all j. An $n \times n$ matrix P is said to be a *permutation* matrix if it is doubly stochastic and there is precisely one $P_{ij} = 1$ in each row and each column.

   It can be shown that every doubly stochastic matrix can be written as the convex combination of permutation matrices.

   (a) Let $\mathbf{a}^t = (a_1, a_2, \ldots, a_n)$, $a_i \geq 0$, $\sum a_i = 1$, be a probability vector. Let $\mathbf{b} = \mathbf{a}P$, where P is doubly stochastic. Show that $\mathbf{b}$ is a probability vector and that $H(b_1, b_2, \ldots, b_n) \geq H(a_1, a_2, \ldots, a_n)$. Thus stochastic mixing increases entropy.

   (b) Show that a stationary distribution μ for a doubly stochastic matrix P is the uniform distribution.

   (c) Conversely, prove that if the uniform distribution is a stationary distribution for a Markov transition matrix P, then P is doubly stochastic.

   Solution: *Doubly Stochastic Matrices.*

   (a)

$$\begin{aligned} H(\mathbf{b}) - H(\mathbf{a}) &= -\sum_j b_j \log b_j + \sum_i a_i \log a_i & (4.1)\\ &= \sum_j \sum_i a_i P_{ij} \log(\sum_k a_k P_{kj}) + \sum_i a_i \log a_i & (4.2)\\ &= \sum_i \sum_j a_i P_{ij} \log \frac{a_i}{\sum_k a_k P_{kj}} & (4.3)\\ &\geq \left(\sum_{i,j} a_i P_{ij}\right) \log \frac{\sum_{i,j} a_i}{\sum_{i,j} b_j} & (4.4) \end{aligned}$$

$$= 1 \log \frac{m}{m} \tag{4.5}$$

$$= 0, \tag{4.6}$$

where the inequality follows from the log sum inequality.

(b) If the matrix is doubly stochastic, the substituting $\mu_i = \frac{1}{m}$, we can easily check that it satisfies $\mu = \mu P$.

(c) If the uniform is a stationary distribution, then

$$\frac{1}{m} = \mu_i = \sum_j \mu_j P_{ji} = \frac{1}{m} \sum_j P_{ji}, \tag{4.7}$$

or $\sum_j P_{ji} = 1$ or that the matrix is doubly stochastic.

2. **Time's arrow.** Let $\{X_i\}_{i=-\infty}^{\infty}$ be a stationary stochastic process. Prove that

$$H(X_0|X_{-1}, X_{-2}, \ldots, X_{-n}) = H(X_0|X_1, X_2, \ldots, X_n).$$

In other words, the present has a conditional entropy given the past equal to the conditional entropy given the future.

This is true even though it is quite easy to concoct stationary random processes for which the flow into the future looks quite different from the flow into the past. That is to say, one can determine the direction of time by looking at a sample function of the process. Nonetheless, given the present state, the conditional uncertainty of the next symbol in the future is equal to the conditional uncertainty of the previous symbol in the past.

Solution: *Time's arrow.* By the chain rule for entropy,

$$\begin{aligned} H(X_0|X_{-1}, \ldots, X_{-n}) &= H(X_0, X_{-1}, \ldots, X_{-n}) - H(X_{-1}, \ldots, X_{-n}) && (4.8)\\ &= H(X_0, X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n) && (4.9)\\ &= H(X_0|X_1, X_2, \ldots, X_n), && (4.10) \end{aligned}$$

where (4.9) follows from stationarity.

3. **Shuffles increase entropy.** Argue that for any distribution on shuffles T and any distribution on card positions X that

$$\begin{aligned} H(TX) &\geq H(TX|T) && (4.11)\\ &= H(T^{-1}TX|T) && (4.12)\\ &= H(X|T) && (4.13)\\ &= H(X), && (4.14) \end{aligned}$$

if X and T are independent.

Solution: *Shuffles increase entropy.*

$$\begin{aligned} H(TX) &\geq H(TX|T) && (4.15)\\ &= H(T^{-1}TX|T) && (4.16)\\ &= H(X|T) && (4.17)\\ &= H(X). && (4.18)\end{aligned}$$

The inequality follows from the fact that conditioning reduces entropy and the first equality follows from the fact that given T, we can reverse the shuffle.

4. **Second law of thermodynamics**. Let $X_1, X_2, X_3 \ldots$ be a stationary first-order Markov chain. In Section 4.4, it was shown that $H(X_n \mid X_1) \geq H(X_{n-1} \mid X_1)$ for $n = 2, 3 \ldots$. Thus conditional uncertainty about the future grows with time. This is true although the unconditional uncertainty $H(X_n)$ remains constant. However, show by example that $H(X_n|X_1 = x_1)$ does not necessarily grow with n for every x_1.

   Solution: *Second law of thermodynamics.*

$$\begin{aligned} H(X_n|X_1) &\leq H(X_n|X_1, X_2) \quad \text{(Conditioning reduces entropy)} && (4.19)\\ &= H(X_n|X_2) \quad \text{(by Markovity)} && (4.20)\\ &= H(X_{n-1}|X_1) \quad \text{(by stationarity)} && (4.21)\end{aligned}$$

   Alternatively, by an application of the data processing inequality to the Markov chain $X_1 \to X_{n-1} \to X_n$, we have

$$I(X_1; X_{n-1}) \geq I(X_1; X_n). \qquad (4.22)$$

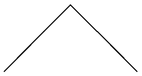
   Expanding the mutual informations in terms of entropies, we have

$$H(X_{n-1}) - H(X_{n-1}|X_1) \geq H(X_n) - H(X_n|X_1). \qquad (4.23)$$

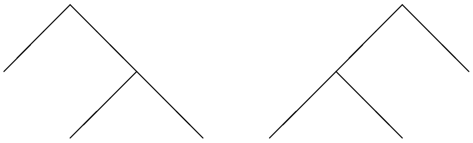
   By stationarity, $H(X_{n-1}) = H(X_n)$ and hence we have

$$H(X_{n-1}|X_1) \leq H(X_n|X_1). \qquad (4.24)$$

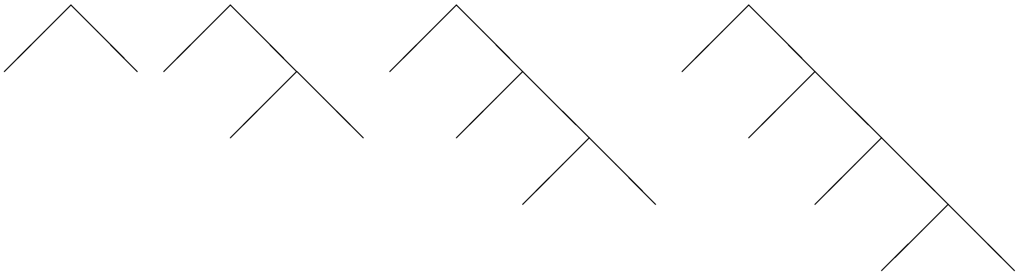
5. **Entropy of a random tree.** Consider the following method of generating a random tree with n nodes. First expand the root node:



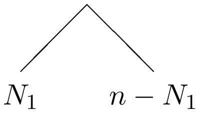
   Then expand one of the two terminal nodes at random:



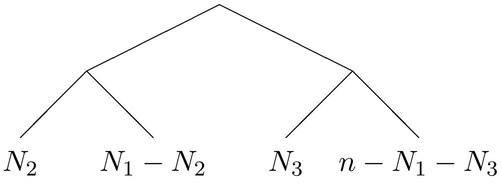
At time k, choose one of the $k-1$ terminal nodes according to a uniform distribution and expand it. Continue until n terminal nodes have been generated. Thus a sequence leading to a five node tree might look like this:



Surprisingly, the following method of generating random trees yields the same probability distribution on trees with n terminal nodes. First choose an integer N_1 uniformly distributed on $\{1, 2, \ldots, n-1\}$. We then have the picture.



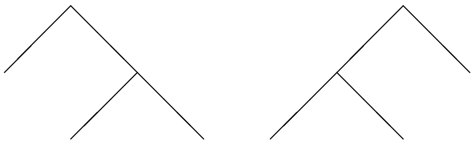
Then choose an integer N_2 uniformly distributed over $\{1, 2, \ldots, N_1 - 1\}$, and independently choose another integer N_3 uniformly over $\{1, 2, \ldots, (n - N_1) - 1\}$. The picture is now:



Continue the process until no further subdivision can be made. (The equivalence of these two tree generation schemes follows, for example, from Polya's urn model.)

Now let T_n denote a random n-node tree generated as described. The probability distribution on such trees seems difficult to describe, but we can find the entropy of this distribution in recursive form.

First some examples. For $n = 2$, we have only one tree. Thus $H(T_2) = 0$. For $n = 3$, we have two equally probable trees:



Thus $H(T_3) = \log 2$. For $n = 4$, we have five possible trees, with probabilities 1/3, 1/6, 1/6, 1/6, 1/6.

Now for the recurrence relation. Let $N_1(T_n)$ denote the number of terminal nodes of T_n in the right half of the tree. Justify each of the steps in the following:

$$\begin{aligned} H(T_n) &\overset{(a)}{=} H(N_1, T_n) && (4.25)\\ &\overset{(b)}{=} H(N_1) + H(T_n|N_1) && (4.26)\\ &\overset{(c)}{=} \log(n-1) + H(T_n|N_1) && (4.27)\\ &\overset{(d)}{=} \log(n-1) + \frac{1}{n-1}\sum_{k=1}^{n-1}[H(T_k) + H(T_{n-k})] && (4.28)\\ &\overset{(e)}{=} \log(n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H(T_k). && (4.29)\\ &= \log(n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H_k. && (4.30) \end{aligned}$$

(f) Use this to show that

$$(n-1)H_n = nH_{n-1} + (n-1)\log(n-1) - (n-2)\log(n-2), \tag{4.31}$$

or

$$\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + c_n, \tag{4.32}$$

for appropriately defined c_n. Since $\sum c_n = c < \infty$, you have proved that $\frac{1}{n}H(T_n)$ converges to a constant. Thus the expected number of bits necessary to describe the random tree T_n grows linearly with n.

Solution: *Entropy of a random tree.*

(a) $H(T_n, N_1) = H(T_n) + H(N_1|T_n) = H(T_n) + 0$ by the chain rule for entropies and since N_1 is a function of T_n.

(b) $H(T_n, N_1) = H(N_1) + H(T_n|N_1)$ by the chain rule for entropies.

(c) $H(N_1) = \log(n-1)$ since N_1 is uniform on $\{1, 2, \ldots, n-1\}$.

(d)

$$\begin{aligned} H(T_n|N_1) &= \sum_{k=1}^{n-1} P(N_1 = k)H(T_n|N_1 = k) && (4.33)\\ &= \frac{1}{n-1}\sum_{k=1}^{n-1} H(T_n|N_1 = k) && (4.34) \end{aligned}$$

by the definition of conditional entropy. Since conditional on N_1, the left subtree and the right subtree are chosen independently, $H(T_n|N_1 = k) = H(T_k, T_{n-k}|N_1 =$

$k) = H(T_k) + H(T_{n-k})$, so

$$H(T_n|N_1) = \frac{1}{n-1}\sum_{k=1}^{n-1}\left(H(T_k) + H(T_{n-k})\right). \tag{4.35}$$

(e) By a simple change of variables,

$$\sum_{k=1}^{n-1} H(T_{n-k}) = \sum_{k=1}^{n-1} H(T_k). \tag{4.36}$$

(f) Hence if we let $H_n = H(T_n)$,

$$\begin{aligned}(n-1)H_n &= (n-1)\log(n-1) + 2\sum_{k=1}^{n-1} H_k & (4.37)\\ (n-2)H_{n-1} &= (n-2)\log(n-2) + 2\sum_{k=1}^{n-2} H_k & (4.38)\\ & & (4.39)\end{aligned}$$

Subtracting the second equation from the first, we get

$$(n-1)H_n - (n-2)H_{n-1} = (n-1)\log(n-1) - (n-2)\log(n-2) + 2H_{n-1} \tag{4.40}$$

or

$$\begin{aligned}\frac{H_n}{n} &= \frac{H_{n-1}}{n-1} + \frac{\log(n-1)}{n} - \frac{(n-2)\log(n-2)}{n(n-1)} & (4.41)\\ &= \frac{H_{n-1}}{n-1} + C_n & (4.42)\end{aligned}$$

where

$$\begin{aligned}C_n &= \frac{\log(n-1)}{n} - \frac{(n-2)\log(n-2)}{n(n-1)} & (4.43)\\ &= \frac{\log(n-1)}{n} - \frac{\log(n-2)}{(n-1)} + \frac{2\log(n-2)}{n(n-1)} & (4.44)\end{aligned}$$

Substituting the equation for H_{n-1} in the equation for H_n and proceeding recursively, we obtain a telescoping sum

$$\begin{aligned}\frac{H_n}{n} &= \sum_{j=3}^{n} C_j + \frac{H_2}{2} & (4.45)\\ &= \sum_{j=3}^{n} \frac{2\log(j-2)}{j(j-1)} + \frac{1}{n}\log(n-1). & (4.46)\end{aligned}$$

Since $\lim_{n\to\infty} \frac{1}{n}\log(n-1) = 0$

$$\begin{aligned}\lim_{n\to\infty} \frac{H_n}{n} &= \sum_{j=3}^{\infty} \frac{2}{j(j-1)}\log(j-2) &(4.47)\\ &\le \sum_{j=3}^{\infty} \frac{2}{(j-1)^2}\log(j-1) &(4.48)\\ &= \sum_{j=2}^{\infty} \frac{2}{j^2}\log j &(4.49)\end{aligned}$$

For sufficiently large j, $\log j \le \sqrt{j}$ and hence the sum in (4.49) is dominated by the sum $\sum_j j^{-\frac{3}{2}}$ which converges. Hence the above sum converges. In fact, computer evaluation shows that

$$\lim \frac{H_n}{n} = \sum_{j=3}^{\infty} \frac{2}{j(j-1)}\log(j-2) = 1.736 \text{ bits.} \tag{4.50}$$

Thus the number of bits required to describe a random n-node tree grows linearly with n.

6. **Monotonicity of entropy per element.** For a stationary stochastic process $X_1, X_2, \ldots, X_n$, show that

   (a)
   $$\frac{H(X_1, X_2, \ldots, X_n)}{n} \le \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. \tag{4.51}$$

   (b)
   $$\frac{H(X_1, X_2, \ldots, X_n)}{n} \ge H(X_n|X_{n-1}, \ldots, X_1). \tag{4.52}$$

   Solution: *Monotonicity of entropy per element.*

   (a) By the chain rule for entropy,

   $$\begin{aligned}\frac{H(X_1, X_2, \ldots, X_n)}{n} &= \frac{\sum_{i=1}^{n} H(X_i|X^{i-1})}{n} &(4.53)\\ &= \frac{H(X_n|X^{n-1}) + \sum_{i=1}^{n-1} H(X_i|X^{i-1})}{n} &(4.54)\\ &= \frac{H(X_n|X^{n-1}) + H(X_1, X_2, \ldots, X_{n-1})}{n}. &(4.55)\end{aligned}$$

   From stationarity it follows that for all $1 \le i \le n$,

   $$H(X_n|X^{n-1}) \le H(X_i|X^{i-1}),$$

which further implies, by averaging both sides, that,

$$\begin{aligned} H(X_n|X^{n-1}) &\leq \frac{\sum_{i=1}^{n-1} H(X_i|X^{i-1})}{n-1} && (4.56)\\ &= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. && (4.57)\end{aligned}$$

Combining (4.55) and (4.57) yields,

$$\begin{aligned} \frac{H(X_1, X_2, \ldots, X_n)}{n} &\leq \frac{1}{n}\left[\frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1} + H(X_1, X_2, \ldots, X_{n-1})\right] \\ &= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. && (4.58)\end{aligned}$$

(b) By stationarity we have for all $1 \leq i \leq n$,

$$H(X_n|X^{n-1}) \leq H(X_i|X^{i-1}),$$

which implies that

$$\begin{aligned} H(X_n|X^{n-1}) &= \frac{\sum_{i=1}^{n} H(X_n|X^{n-1})}{n} && (4.59)\\ &\leq \frac{\sum_{i=1}^{n} H(X_i|X^{i-1})}{n} && (4.60)\\ &= \frac{H(X_1, X_2, \ldots, X_n)}{n}. && (4.61)\end{aligned}$$

7. **Entropy rates of Markov chains.**

(a) Find the entropy rate of the two-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1-p_{01} & p_{01} \\ p_{10} & 1-p_{10} \end{bmatrix}.$$

(b) What values of p_{01}, p_{10} maximize the rate of part (a)?

(c) Find the entropy rate of the two-state Markov chain with transition matrix

$$P = \begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix}.$$

(d) Find the maximum value of the entropy rate of the Markov chain of part (c). We expect that the maximizing value of p should be less than $1/2$, since the 0 state permits more information to be generated than the 1 state.

(e) Let $N(t)$ be the number of allowable state sequences of length t for the Markov chain of part (c). Find $N(t)$ and calculate

$$H_0 = \lim_{t\to\infty} \frac{1}{t} \log N(t).$$

Hint: Find a linear recurrence that expresses $N(t)$ in terms of $N(t-1)$ and $N(t-2)$. Why is H_0 an upper bound on the entropy rate of the Markov chain? Compare H_0 with the maximum entropy found in part (d).

Solution: *Entropy rates of Markov chains.*

(a) The stationary distribution is easily calculated. (See EIT pp. 62–63.)

$$\mu_0 = \frac{p_{10}}{p_{01}+p_{10}}, \quad \mu_0 = \frac{p_{01}}{p_{01}+p_{10}}\,.$$

Therefore the entropy rate is

$$H(X_2|X_1) = \mu_0 H(p_{01}) + \mu_1 H(p_{10}) = \frac{p_{10}H(p_{01}) + p_{01}H(p_{10})}{p_{01}+p_{10}}\,.$$

(b) The entropy rate is at most 1 bit because the process has only two states. This rate can be achieved if (and only if) $p_{01} = p_{10} = 1/2$, in which case the process is actually i.i.d. with $\Pr(X_i = 0) = \Pr(X_i = 1) = 1/2$.

(c) As a special case of the general two-state Markov chain, the entropy rate is

$$H(X_2|X_1) = \mu_0 H(p) + \mu_1 H(1) = \frac{H(p)}{p+1}\,.$$

(d) By straightforward calculus, we find that the maximum value of $H(X)$ of part (c) occurs for $p = (3-\sqrt{5})/2 = 0.382$. The maximum value is

$$H(p) = H(1-p) = H\left(\frac{\sqrt{5}-1}{2}\right) = 0.694 \text{ bits}\,.$$

Note that $(\sqrt{5}-1)/2 = 0.618$ is (the reciprocal of) the Golden Ratio.

(e) The Markov chain of part (c) forbids consecutive ones. Consider any allowable sequence of symbols of length t. If the first symbol is 1, then the next symbol must be 0; the remaining $N(t-2)$ symbols can form any allowable sequence. If the first symbol is 0, then the remaining $N(t-1)$ symbols can be any allowable sequence. So the number of allowable sequences of length t satisfies the recurrence

$$N(t) = N(t-1) + N(t-2) \qquad N(1) = 2,\ N(2) = 3$$

(The initial conditions are obtained by observing that for $t = 2$ only the sequence 11 is not allowed. We could also choose $N(0) = 1$ as an initial condition, since there is exactly one allowable sequence of length 0, namely, the empty sequence.) The sequence $N(t)$ grows exponentially, that is, $N(t) \approx c\lambda^t$, where λ is the maximum magnitude solution of the characteristic equation

$$1 = z^{-1} + z^{-2}\,.$$

Solving the characteristic equation yields $\lambda = (1+\sqrt{5})/2$, the Golden Ratio. (The sequence $\{N(t)\}$ is the sequence of Fibonacci numbers.) Therefore

$$H_0 = \lim_{n\to\infty} \frac{1}{t} \log N(t) = \log(1+\sqrt{5})/2 = 0.694 \text{ bits}\,.$$

Since there are only $N(t)$ possible outcomes for $X_1, \ldots, X_t$, an upper bound on $H(X_1, \ldots, X_t)$ is $\log N(t)$, and so the entropy rate of the Markov chain of part (c) is at most H_0. In fact, we saw in part (d) that this upper bound can be achieved.

8. **Maximum entropy process.** A discrete memoryless source has alphabet $\{1, 2\}$ where the symbol 1 has duration 1 and the symbol 2 has duration 2. The probabilities of 1 and 2 are p_1 and p_2, respectively. Find the value of p_1 that maximizes the source entropy per unit time $H(X)/El_X$. What is the maximum value H?

   Solution: *Maximum entropy process.* The entropy per symbol of the source is

$$H(p_1) = -p_1 \log p_1 - (1 - p_1) \log(1 - p_1)$$

   and the average symbol duration (or time per symbol) is

$$T(p_1) = 1 \cdot p_1 + 2 \cdot p_2 = p_1 + 2(1 - p_1) = 2 - p_1 = 1 + p_2 .$$

   Therefore the source entropy per unit time is

$$f(p_1) = \frac{H(p_1)}{T(p_1)} = \frac{-p_1 \log p_1 - (1 - p_1) \log(1 - p_1)}{2 - p_1} .$$

   Since $f(0) = f(1) = 0$, the maximum value of $f(p_1)$ must occur for some point p_1 such that $0 < p_1 < 1$ and $\partial f / \partial p_1 = 0$ and

$$\frac{\partial}{\partial p_1} \frac{H(p_1)}{T(p_1)} = \frac{T(\partial H/\partial p_1) - H(\partial T/\partial p_1)}{T^2}$$

   After some calculus, we find that the numerator of the above expression (assuming natural logarithms) is

$$T(\partial H/\partial p_1) - H(\partial T/\partial p_1) = \ln(1 - p_1) - 2 \ln p_1 ,$$

   which is zero when $1 - p_1 = p_1^2 = p_2$, that is, $p_1 = \frac{1}{2}(\sqrt{5} - 1) = 0.61803$, the reciprocal of the golden ratio, $\frac{1}{2}(\sqrt{5} + 1) = 1.61803$. The corresponding entropy per unit time is

$$\frac{H(p_1)}{T(p_1)} = \frac{-p_1 \log p_1 - p_1^2 \log p_1^2}{2 - p_1} = \frac{-(1 + p_1^2) \log p_1}{1 + p_1^2} = -\log p_1 = 0.69424 \text{ bits.}$$

   Note that this result is the same as the maximum entropy rate for the Markov chain in problem 4.7(d). This is because a source in which every 1 must be followed by a 0 is equivalent to a source in which the symbol 1 has duration 2 and the symbol 0 has duration 1.

9. **Initial conditions.** Show, for a Markov chain, that

$$H(X_0|X_n) \geq H(X_0|X_{n-1}).$$

   Thus initial conditions X_0 become more difficult to recover as the future X_n unfolds.

   Solution: *Initial conditions.* For a Markov chain, by the data processing theorem, we have

$$I(X_0; X_{n-1}) \geq I(X_0; X_n). \tag{4.62}$$

   Therefore

$$H(X_0) - H(X_0|X_{n-1}) \geq H(X_0) - H(X_0|X_n) \tag{4.63}$$

   or $H(X_0|X_n)$ increases with n.

10. **Pairwise independence.** Let $X_1, X_2, \ldots, X_{n-1}$ be i.i.d. random variables taking values in $\{0,1\}$, with $\Pr\{X_i = 1\} = \frac{1}{2}$. Let $X_n = 1$ if $\sum_{i=1}^{n-1} X_i$ is odd and $X_n = 0$ otherwise. Let $n \geq 3$.

    (a) Show that X_i and X_j are independent, for $i \neq j$, $i, j \in \{1, 2, \ldots, n\}$.

    (b) Find $H(X_i, X_j)$, for $i \neq j$.

    (c) Find $H(X_1, X_2, \ldots, X_n)$. Is this equal to $nH(X_1)$?

    Solution: (*Pairwise Independence*) $X_1, X_2, \ldots, X_{n-1}$ are i.i.d. Bernoulli(1/2) random variables. We will first prove that for any $k \leq n-1$, the probability that $\sum_{i=1}^{k} X_i$ is odd is $1/2$. We will prove this by induction. Clearly this is true for $k = 1$. Assume that it is true for $k-1$. Let $S_k = \sum_{i=1}^{k} X_i$. Then

$$\begin{aligned} P(S_k \text{ odd}) &= P(S_{k-1} \text{ odd})P(X_k = 0) + P(S_{k-1} \text{ even})P(X_k = 1) && (4.64) \\ &= \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} && (4.65) \\ &= \frac{1}{2}. && (4.66) \end{aligned}$$

    Hence for all $k \leq n-1$, the probability that S_k is odd is equal to the probability that it is even. Hence,

$$P(X_n = 1) = P(X_n = 0) = \frac{1}{2}. \quad (4.67)$$

    (a) It is clear that when i and j are both less than n, X_i and X_j are independent. The only possible problem is when $j = n$. Taking $i = 1$ without loss of generality,

$$\begin{aligned} P(X_1 = 1, X_n = 1) &= P(X_1 = 1, \sum_{i=2}^{n-1} X_i \text{ even}) && (4.68) \\ &= P(X_1 = 1)P(\sum_{i=2}^{n-1} X_i \text{ even}) && (4.69) \\ &= \frac{1}{2}\frac{1}{2} && (4.70) \\ &= P(X_1 = 1)P(X_n = 1) && (4.71) \end{aligned}$$

    and similarly for other possible values of the pair (X_1, X_n). Hence X_1 and X_n are independent.

    (b) Since X_i and X_j are independent and uniformly distributed on $\{0,1\}$,

$$H(X_i, X_j) = H(X_i) + H(X_j) = 1 + 1 = 2 \text{ bits}. \quad (4.72)$$

    (c) By the chain rule and the independence of $X_1, X_2, \ldots, X_{n_1}$, we have

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_1, X_2, \ldots, X_{n-1}) + H(X_n|X_{n-1}, \ldots, X_1) && (4.73) \\ &= \sum_{i=1}^{n-1} H(X_i) + 0 && (4.74) \\ &= n-1, && (4.75) \end{aligned}$$

since X_n is a function of the previous X_i 's. The total entropy is not n, which is what would be obtained if the X_i 's were all independent. This example illustrates that pairwise independence does not imply complete independence.

11. **Stationary processes.** Let $\ldots, X_{-1}, X_0, X_1, \ldots$ be a stationary (not necessarily Markov) stochastic process. Which of the following statements are true? Prove or provide a counterexample.

(a) $H(X_n|X_0) = H(X_{-n}|X_0)$.

(b) $H(X_n|X_0) \geq H(X_{n-1}|X_0)$.

(c) $H(X_n|X_1, X_2, \ldots, X_{n-1}, X_{n+1})$ is nonincreasing in n.

(d) $H(X_n|X_1, \ldots, X_{n-1}, X_{n+1}, \ldots, X_{2n})$ is non-increasing in n.

Solution: *Stationary processes.*

(a) $H(X_n|X_0) = H(X_{-n}|X_0)$.
This statement is true, since

$$H(X_n|X_0) = H(X_n, X_0) - H(X_0) \tag{4.76}$$
$$H(X_{-n}|X_0) = H(X_{-n}, X_0) - H(X_0) \tag{4.77}$$

and $H(X_n, X_0) = H(X_{-n}, X_0)$ by stationarity.

(b) $H(X_n|X_0) \geq H(X_{n-1}|X_0)$.
This statement is not true in general, though it is true for first order Markov chains. A simple counterexample is a periodic process with period n. Let $X_0, X_1, X_2, \ldots, X_{n-1}$ be i.i.d. uniformly distributed binary random variables and let $X_k = X_{k-n}$ for $k \geq n$. In this case, $H(X_n|X_0) = 0$ and $H(X_{n-1}|X_0) = 1$, contradicting the statement $H(X_n|X_0) \geq H(X_{n-1}|X_0)$.

(c) $H(X_n|X_1^{n-1}, X_{n+1})$ is non-increasing in n.
This statement is true, since by stationarity $H(X_n|X_1^{n-1}, X_{n+1}) = H(X_{n+1}|X_2^n, X_{n+2}) \geq H(X_{n+1}|X_1^n, X_{n+2})$ where the inequality follows from the fact that conditioning reduces entropy.

12. **The entropy rate of a dog looking for a bone.** A dog walks on the integers, possibly reversing direction at each step with probability $p = .1$. Let $X_0 = 0$. The first step is equally likely to be positive or negative. A typical walk might look like this:

$$(X_0, X_1, \ldots) = (0, -1, -2, -3, -4, -3, -2, -1, 0, 1, \ldots).$$

(a) Find $H(X_1, X_2, \ldots, X_n)$.

(b) Find the entropy rate of this browsing dog.

(c) What is the expected number of steps the dog takes before reversing direction?

Solution: *The entropy rate of a dog looking for a bone.*

(a) By the chain rule,

$$\begin{aligned} H(X_0, X_1, \ldots, X_n) &= \sum_{i=0}^{n} H(X_i|X^{i-1}) \\ &= H(X_0) + H(X_1|X_0) + \sum_{i=2}^{n} H(X_i|X_{i-1}, X_{i-2}), \end{aligned}$$

since, for $i > 1$, the next position depends only on the previous two (i.e., the dog's walk is 2nd order Markov, if the dog's position is the state). Since $X_0 = 0$ deterministically, $H(X_0) = 0$ and since the first step is equally likely to be positive or negative, $H(X_1|X_0) = 1$. Furthermore for $i > 1$,

$$H(X_i|X_{i-1}, X_{i-2}) \;=\; H(.1, .9).$$

Therefore,

$$H(X_0, X_1, \ldots, X_n) = 1 + (n-1)H(.1, .9).$$

(b) From a),

$$\begin{aligned} \frac{H(X_0, X_1, \ldots X_n)}{n+1} &= \frac{1 + (n-1)H(.1, .9)}{n+1} \\ &\to H(.1, .9). \end{aligned}$$

(c) The dog must take at least one step to establish the direction of travel from which it ultimately reverses. Letting S be the number of steps taken between reversals, we have

$$\begin{aligned} E(S) &= \sum_{s=1}^{\infty} s(.9)^{s-1}(.1) \\ &= 10. \end{aligned}$$

Starting at time 0, the expected number of steps to the first reversal is 11.

13. **The past has little to say about the future.** For a stationary stochastic process $X_1, X_2, \ldots, X_n, \ldots$, show that

$$\lim_{n\to\infty} \frac{1}{2n} I(X_1, X_2, \ldots, X_n; X_{n+1}, X_{n+2}, \ldots, X_{2n}) = 0. \qquad (4.78)$$

Thus the dependence between adjacent n-blocks of a stationary process does not grow linearly with n.

Solution:

$$\begin{aligned} &I(X_1, X_2, \ldots, X_n; X_{n+1}, X_{n+2}, \ldots, X_{2n}) \\ &\quad= H(X_1, X_2, \ldots, X_n) + H(X_{n+1}, X_{n+2}, \ldots, X_{2n}) - H(X_1, X_2, \ldots, X_n, X_{n+1}, X_{n+2}, \ldots, X_{2n}) \\ &\quad= 2H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n, X_{n+1}, X_{n+2}, \ldots, X_{2n}) \qquad (4.79) \end{aligned}$$

since $H(X_1, X_2, \ldots, X_n) = H(X_{n+1}, X_{n+2}, \ldots, X_{2n})$ by stationarity.

Thus

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{2n} I(X_1, X_2, \ldots, X_n; X_{n+1}, X_{n+2}, \ldots, X_{2n}) \\
&= \lim_{n\to\infty} \frac{1}{2n} 2H(X_1, X_2, \ldots, X_n) - \lim_{n\to\infty} \frac{1}{2n} H(X_1, X_2, \ldots, X_n, X_{n+1}, X_{n+2}, \ldots, X_{2n}) \qquad (4.80) \\
&= \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) - \lim_{n\to\infty} \frac{1}{2n} H(X_1, X_2, \ldots, X_n, X_{n+1}, X_{n+2}, \ldots, X_{2n}) \qquad (4.81)
\end{aligned}$$

Now $\lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) = \lim_{n\to\infty} \frac{1}{2n} H(X_1, X_2, \ldots, X_n, X_{n+1}, X_{n+2}, \ldots, X_{2n})$ since both converge to the entropy rate of the process, and therefore

$$\lim_{n\to\infty} \frac{1}{2n} I(X_1, X_2, \ldots, X_n; X_{n+1}, X_{n+2}, \ldots, X_{2n}) = 0. \tag{4.82}$$

14. **Functions of a stochastic process.**

(a) Consider a stationary stochastic process $X_1, X_2, \ldots, X_n$, and let $Y_1, Y_2, \ldots, Y_n$ be defined by

$$Y_i = \phi(X_i), \quad i = 1, 2, \ldots \tag{4.83}$$

for some function ϕ. Prove that

$$H(\mathcal{Y}) \le H(\mathcal{X}) \tag{4.84}$$

(b) What is the relationship between the entropy rates $H(\mathcal{Z})$ and $H(\mathcal{X})$ if

$$Z_i = \psi(X_i, X_{i+1}), \quad i = 1, 2, \ldots \tag{4.85}$$

for some function ψ.

Solution: The key point is that functions of a random variable have lower entropy. Since $(Y_1, Y_2, \ldots, Y_n)$ is a function of $(X_1, X_2, \ldots, X_n)$ (each Y_i is a function of the corresponding X_i), we have (from Problem 2.4)

$$H(Y_1, Y_2, \ldots, Y_n) \le H(X_1, X_2, \ldots, X_n) \tag{4.86}$$

Dividing by n, and taking the limit as $n \to \infty$, we have

$$\lim_{n\to\infty} \frac{H(Y_1, Y_2, \ldots, Y_n)}{n} \le \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} \tag{4.87}$$

or

$$\mathcal{H}(\mathcal{Y}) \le \mathcal{H}(\mathcal{X}) \tag{4.88}$$

15. **Entropy rate.** Let $\{X_i\}$ be a discrete stationary stochastic process with entropy rate $H(\mathcal{X})$. Show

$$\frac{1}{n}H(X_n,\ldots,X_1 \mid X_0,X_{-1},\ldots,X_{-k}) \to H(\mathcal{X}), \tag{4.89}$$

for $k=1,2,\ldots$.

Solution: *Entropy rate of a stationary process.* By the Cesáro mean theorem, the running average of the terms tends to the same limit as the limit of the terms. Hence

$$\begin{aligned}
\frac{1}{n}H(X_1,X_2,\ldots,X_n|X_0,X_{-1},\ldots,X_{-k}) &= \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1},X_{i-2},\ldots,X_{-k}) && (4.90)\\
&\to \lim H(X_n|X_{n-1},X_{n-2},\ldots,X_{-k}) && (4.91)\\
&= \mathcal{H}, && (4.92)
\end{aligned}$$

the entropy rate of the process.

16. **Entropy rate of constrained sequences.** In magnetic recording, the mechanism of recording and reading the bits imposes constraints on the sequences of bits that can be recorded. For example, to ensure proper sychronization, it is often necessary to limit the length of runs of 0's between two 1's. Also to reduce intersymbol interference, it may be necessary to require at least one 0 between any two 1's. We will consider a simple example of such a constraint.

Suppose that we are required to have at least one 0 and at most two 0's between any pair of 1's in a sequences. Thus, sequences like 101001 and 0101001 are valid sequences, but 0110010 and 0000101 are not. We wish to calculate the number of valid sequences of length n.

(a) Show that the set of constrained sequences is the same as the set of allowed paths on the following state diagram:

(b) Let $X_i(n)$ be the number of valid paths of length n ending at state i. Argue that $\mathbf{X}(n) = [X_1(n)\ X_2(n)\ X_3(n)]^t$ satisfies the following recursion:

$$\begin{bmatrix} X_1(n) \\ X_2(n) \\ X_3(n) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1(n-1) \\ X_2(n-1) \\ X_3(n-1) \end{bmatrix}, \tag{4.93}$$

with initial conditions $\mathbf{X}(1) = [1\ 1\ 0]^t$.

(c) Let

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{4.94}$$

Then we have by induction

$$\mathbf{X}(n) = A\mathbf{X}(n-1) = A^2\mathbf{X}(n-2) = \cdots = A^{n-1}\mathbf{X}(1). \tag{4.95}$$

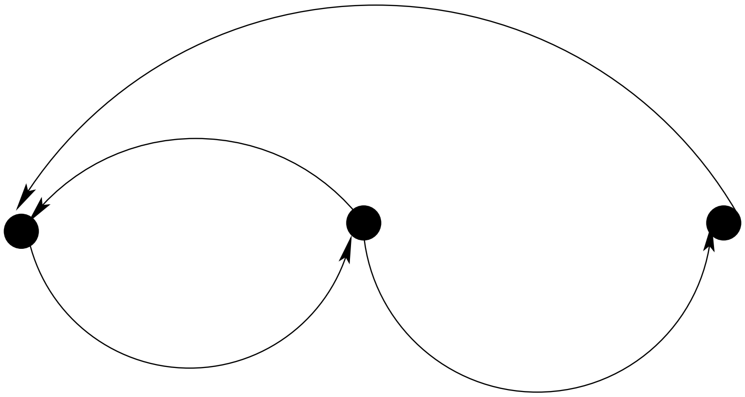


Figure 4.1: Entropy rate of constrained sequence

Using the eigenvalue decomposition of A for the case of distinct eigenvalues, we can write $A = U^{-1}\Lambda U$, where Λ is the diagonal matrix of eigenvalues. Then $A^{n-1} = U^{-1}\Lambda^{n-1}U$. Show that we can write

$$\mathbf{X}(n) = \lambda_1^{n-1}\mathbf{Y}_1 + \lambda_2^{n-1}\mathbf{Y}_2 + \lambda_3^{n-1}\mathbf{Y}_3, \tag{4.96}$$

where $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ do not depend on n. For large n, this sum is dominated by the largest term. Therefore argue that for $i = 1, 2, 3$, we have

$$\frac{1}{n}\log X_i(n) \to \log\lambda, \tag{4.97}$$

where λ is the largest (positive) eigenvalue. Thus the number of sequences of length n grows as λ^n for large n. Calculate λ for the matrix A above. (The case when the eigenvalues are not distinct can be handled in a similar manner.)

(d) We will now take a different approach. Consider a Markov chain whose state diagram is the one given in part (a), but with arbitrary transition probabilities. Therefore the probability transition matrix of this Markov chain is

$$P = \begin{bmatrix} 0 & 1 & 0 \\ \alpha & 0 & 1-\alpha \\ 1 & 0 & 0 \end{bmatrix}. \tag{4.98}$$

Show that the stationary distribution of this Markov chain is

$$\mu = \left[\frac{1}{3-\alpha}, \frac{1}{3-\alpha}, \frac{1-\alpha}{3-\alpha}\right]. \tag{4.99}$$

(e) Maximize the entropy rate of the Markov chain over choices of α. What is the maximum entropy rate of the chain?

(f) Compare the maximum entropy rate in part (e) with $\log \lambda$ in part (c). Why are the two answers the same?

Solution:

Entropy rate of constrained sequences.

(a) The sequences are constrained to have at least one 0 and at most two 0's between two 1's. Let the state of the system be the number of 0's that has been seen since the last 1. Then a sequence that ends in a 1 is in state 1, a sequence that ends in 10 in is state 2, and a sequence that ends in 100 is in state 3. From state 1, it is only possible to go to state 2, since there has to be at least one 0 before the next 1. From state 2, we can go to either state 1 or state 3. From state 3, we have to go to state 1, since there cannot be more than two 0's in a row. Thus we can the state diagram in the problem.

(b) Any valid sequence of length n that ends in a 1 must be formed by taking a valid sequence of length $n-1$ that ends in a 0 and adding a 1 at the end. The number of valid sequences of length $n-1$ that end in a 0 is equal to $X_2(n-1)+X_3(n-1)$ and therefore,

$$X_1(n) = X_2(n-1) + X_3(n-1). \tag{4.100}$$

By similar arguments, we get the other two equations, and we have

$$\begin{bmatrix} X_1(n) \\ X_2(n) \\ X_3(n) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1(n-1) \\ X_2(n-1) \\ X_3(n-1) \end{bmatrix}. \tag{4.101}$$

The initial conditions are obvious, since both sequences of length 1 are valid and therefore $\mathbf{X}(1) = [1\ 1\ 0]^T$.

(c) The induction step is obvious. Now using the eigenvalue decomposition of $A = U^{-1}\Lambda U$, it follows that $A^2 = U^{-1}\Lambda U U^{-1}\Lambda U = U^{-1}\Lambda^2 U$, etc. and therefore

$$\begin{aligned}
\mathbf{X}(n) &= A^{n-1}\mathbf{X}(1) = U^{-1}\Lambda^{n-1}U\mathbf{X}(1) && (4.102)\\
&= U^{-1}\begin{bmatrix} \lambda_1^{n-1} & 0 & 0 \\ 0 & \lambda_2^{n-1} & 0 \\ 0 & 0 & \lambda_3^{n-1} \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} && (4.103)\\
&= \lambda_1^{n-1}U^{-1}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \lambda_2^{n-1}U^{-1}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \\
&\quad + \lambda_3^{n-1}U^{-1}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} && (4.104)\\
&= \lambda_1^{n-1}\mathbf{Y}_1 + \lambda_2^{n-1}\mathbf{Y}_2 + \lambda_3^{n-1}\mathbf{Y}_3, && (4.105)
\end{aligned}$$

where $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ do not depend on n. Without loss of generality, we can assume that $\lambda_1 > \lambda_2 > \lambda_3$. Thus

$$X_1(n) = \lambda_1^{n-1}\mathbf{Y}_{11} + \lambda_2^{n-1}\mathbf{Y}_{21} + \lambda_3^{n-1}\mathbf{Y}_{31} \tag{4.106}$$

$$X_2(n) = \lambda_1^{n-1}\mathbf{Y}_{12} + \lambda_2^{n-1}\mathbf{Y}_{22} + \lambda_3^{n-1}\mathbf{Y}_{32} \tag{4.107}$$

$$X_3(n) = \lambda_1^{n-1}\mathbf{Y}_{13} + \lambda_2^{n-1}\mathbf{Y}_{23} + \lambda_3^{n-1}\mathbf{Y}_{33} \tag{4.108}$$

For large n, this sum is dominated by the largest term. Thus if $\mathbf{Y}_{1i} > 0$, we have

$$\frac{1}{n}\log X_i(n) \to \log \lambda_1. \tag{4.109}$$

To be rigorous, we must also show that $\mathbf{Y}_{1i} > 0$ for $i = 1, 2, 3$. It is not difficult to prove that if one of the $\mathbf{Y}_{1i}$ is positive, then the other two terms must also be positive, and therefore either

$$\frac{1}{n}\log X_i(n) \to \log \lambda_1. \tag{4.110}$$

for all $i = 1, 2, 3$ or they all tend to some other value.

The general argument is difficult since it is possible that the initial conditions of the recursion do not have a component along the eigenvector that corresponds to the maximum eigenvalue and thus $Y_{1i} = 0$ and the above argument will fail. In our example, we can simply compute the various quantities, and thus

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = U^{-1}\Lambda U, \tag{4.111}$$

where

$$\Lambda = \begin{bmatrix} 1.3247 & 0 & 0 \\ 0 & -0.6624 + 0.5623i & 0 \\ 0 & 0 & -0.6624 - 0.5623i \end{bmatrix}, \tag{4.112}$$

and

$$U = \begin{bmatrix} -0.5664 & -0.7503 & -0.4276 \\ 0.6508 - 0.0867i & -0.3823 + 0.4234i & -0.6536 - 0.4087i \\ 0.6508 + 0.0867i & -0.3823i0.4234i & -0.6536 + 0.4087i \end{bmatrix}, \tag{4.113}$$

and therefore

$$\mathbf{Y}_1 = \begin{bmatrix} 0.9566 \\ 0.7221 \\ 0.5451 \end{bmatrix}, \tag{4.114}$$

which has all positive components. Therefore,

$$\frac{1}{n}\log X_i(n) \to \log \lambda_i = \log 1.3247 = 0.4057 \text{ bits}. \tag{4.115}$$

(d) To verify the that

$$\mu = \left[\frac{1}{3-\alpha}, \frac{1}{3-\alpha}, \frac{1-\alpha}{3-\alpha} \right]^T . \tag{4.116}$$

is the stationary distribution, we have to verify that $P\mu = \mu$. But this is straightforward.

(e) The entropy rate of the Markov chain (in nats) is

$$\mathcal{H} = -\sum_i \mu_i \sum_j P_{ij} \ln P_{ij} = \frac{1}{3-\alpha} \left(-\alpha \ln \alpha - (1-\alpha) \ln(1-\alpha) \right), \tag{4.117}$$

and differentiating with respect to α to find the maximum, we find that

$$\frac{d\mathcal{H}}{d\alpha} = \frac{1}{(3-\alpha)^2} \left(-\alpha \ln \alpha - (1-\alpha) \ln(1-\alpha) \right) + \frac{1}{3-\alpha} \left(-1 - \ln \alpha + 1 + \ln(1-\alpha) \right) = 0, \tag{4.118}$$

or

$$(3-\alpha) \left(\ln a - \ln(1-\alpha) \right) = \left(-\alpha \ln \alpha - (1-\alpha) \ln(1-\alpha) \right) \tag{4.119}$$

which reduces to

$$3 \ln \alpha = 2 \ln(1-\alpha), \tag{4.120}$$

i.e.,

$$\alpha^3 = \alpha^2 - 2\alpha + 1, \tag{4.121}$$

which can be solved (numerically) to give $\alpha = 0.5698$ and the maximum entropy rate as 0.2812 nats $= 0.4057$ bits.

(f) The answers in parts (c) and (f) are the same. Why? A rigorous argument is quite involved, but the essential idea is that both answers give the asymptotics of the number of sequences of length n for the state diagram in part (a). In part (c) we used a direct argument to calculate the number of sequences of length n and found that asymptotically, $X(n) \approx \lambda_1^n$.

If we extend the ideas of Chapter 3 (typical sequences) to the case of Markov chains, we can see that there are approximately $2^{n\mathcal{H}}$ typical sequences of length n for a Markov chain of entropy rate $\mathcal{H}$. If we consider all Markov chains with state diagram given in part (a), the number of typical sequences should be less than the total number of sequences of length n that satisfy the state constraints. Thus, we see that $2^{n\mathcal{H}} \leq \lambda_1^n$ or $\mathcal{H} \leq \log \lambda_1$.

To complete the argument, we need to show that there exists a Markov transition matrix that achieves the upper bound. This can be done by two different methods. One is to derive the Markov transition matrix from the eigenvalues, etc. of parts (a)–(c). Instead, we will use an argument from the method of types. In Chapter 12, we show that there are at most a polynomial number of types, and that therefore, the largest type class has the same number of sequences (to the first order in the exponent) as the entire set. The same arguments can be applied to Markov types. There are only a polynomial number of Markov types and therefore of all the Markov type classes that satisfy the state constraints of part (a), at least one of them has the same exponent as the total number of sequences that satisfy the state constraint. For this Markov type, the number of sequences in the typeclass is $2^{n\mathcal{H}}$, and therefore for this type class, $\mathcal{H} = \log \lambda_1$.

This result is a very curious one that connects two apparently unrelated objects - the maximum eigenvalue of a state transition matrix, and the maximum entropy rate for a probability transition matrix with the same state diagram. We don't know a reference for a formal proof of this result.

17. **Waiting times are insensitive to distributions.** Let $X_0, X_1, X_2, \ldots$ be drawn i.i.d. $\sim p(x), x \in \mathcal{X} = \{1, 2, \ldots, m\}$ and let N be the waiting time to the next occurrence of X_0, where $N = \min_n\{X_n = X_0\}$.

    (a) Show that $EN = m$.

    (b) Show that $E \log N \le H(X)$.

    (c) (Optional) Prove part (a) for $\{X_i\}$ stationary and ergodic.

    Solution: *Waiting times are insensitive to distributions.* Since $X_0, X_1, X_2, \ldots, X_n$ are drawn i.i.d. $\sim p(x)$, the waiting time for the next occurrence of X_0 has a geometric distribution with probability of success $p(x_0)$.

    (a) Given $X_0 = i$, the expected time until we see it again is $1/p(i)$. Therefore,

    $$EN = E[E(N|X_0)] = \sum p(X_0 = i)\left(\frac{1}{p(i)}\right) = m. \tag{4.122}$$

    (b) By the same argument, since given $X_0 = i$, N has a geometric distribution with mean $p(i)$ and

    $$E(N|X_0 = i) = \frac{1}{p(i)}. \tag{4.123}$$

    Then using Jensen's inequality, we have

    $$\begin{aligned} E \log N &= \sum_i p(X_0 = i) E(\log N | X_0 = i) && (4.124)\\ &\le \sum_i p(X_0 = i) \log E(N|X_0 = i) && (4.125)\\ &= \sum_i p(i) \log \frac{1}{p(i)} && (4.126)\\ &= H(X). && (4.127)\end{aligned}$$

    (c) The property that $EN = m$ is essentially a combinatorial property rather than a statement about expectations. We prove this for stationary ergodic sources. In essence, we will calculate the empirical average of the waiting time, and show that this converges to m. Since the process is ergodic, the empirical average converges to the expected value, and thus the expected value must be m.

Let $X_1 = a$, and define a sequence of random variables $N_1, N_2, \ldots$, where $N_1 =$recurrence time for X_1, etc. It is clear that the N process is also stationary and ergodic.

Let $I_a(X_i)$ be the indicator that $X_i = a$ and $J_a(X_i)$ be the indicator that $X_i \neq a$. Then for all i, all $a \in \mathcal{X}$, $I_a(X_i) + J_a(X_i) = 1$.

Let $N_1(a), N_2(a), \ldots$ be the recurrence times of the symbol a in the sequence. Thus $X_1 = a$, $X_i \neq a, 1 < i < N_1(a)$, and $X_{N_1(a)} = a$, etc. Thus the sum of $J_a(X_i)$ over all i is equal to the $\sum_j (N_j(a) - 1)$. Or equivalently

$$\sum_j N_j(a) = \sum_i J_a(X_i) + \sum_i I_a(X_i) = n \tag{4.128}$$

Summing this over all $a \in \mathcal{X}$, we obtain

$$\sum_a \sum_j N_j(a) = nm \tag{4.129}$$

There are n terms in this sum, and therefore the empirical mean of $N_j(X_i)$ is m.

Thus the empirical average of N over any sample sequence is m and thus the expected value of N must also be m.

18. **Stationary but not ergodic process.** A bin has two biased coins, one with probability of heads p and the other with probability of heads $1-p$. One of these coins is chosen at random (i.e., with probability 1/2), and is then tossed n times. Let X denote the identity of the coin that is picked, and let Y_1 and Y_2 denote the results of the first two tosses.

   (a) Calculate $I(Y_1; Y_2|X)$.

   (b) Calculate $I(X; Y_1, Y_2)$.

   (c) Let $\mathcal{H}(\mathcal{Y})$ be the entropy rate of the Y process (the sequence of coin tosses). Calculate $\mathcal{H}(\mathcal{Y})$. (Hint: Relate this to $\lim_{n\to\infty} \frac{1}{n} H(X, Y_1, Y_2, \ldots, Y_n)$).

   You can check the answer by considering the behavior as $p \to 1/2$.

   Solution:

   (a) SInce the coin tosses are indpendent conditional on the coin chosen, $I(Y_1; Y_2|X) = 0$.

   (b) The key point is that if we did not know the coin being used, then Y_1 and Y_2 are not independent. The joint distribution of Y_1 and Y_2 can be easily calculated from the following table

X	Y_1	Y_2	Probability
1	H	H	p^2
1	H	T	$p(1-p)$
1	T	H	$p(1-p)$
1	T	T	$(1-p)^2$
2	H	H	$(1-p)^2$
2	H	T	$p(1-p)$
2	T	H	$p(1-p)$
2	T	T	p^2

Thus the joint distribution of (Y_1, Y_2) is $(\frac{1}{2}(p^2+(1-p)^2), p(1-p), p(1-p), \frac{1}{2}(p^2+(1-p)^2))$, and we can now calculate

$$I(X;Y_1,Y_2) = H(Y_1,Y_2) - H(Y_1,Y_2|X) \tag{4.130}$$

$$= H(Y_1,Y_2) - H(Y_1|X) - H(Y_2|X) \tag{4.131}$$

$$= H(Y_1,Y_2) - 2H(p) \tag{4.132}$$

$$= H\left(\frac{1}{2}(p^2+(1-p)^2), p(1-p), p(1-p), \frac{1}{2}(p^2+(1-p)^2)\right) - 2H(p)$$

$$= H(p(1-p)) + 1 - 2H(p) \tag{4.133}$$

where the last step follows from using the grouping rule for entropy.

(c)

$$\mathcal{H}(\mathcal{Y}) = \lim \frac{H(Y_1,Y_2,\ldots,Y_n)}{n} \tag{4.134}$$

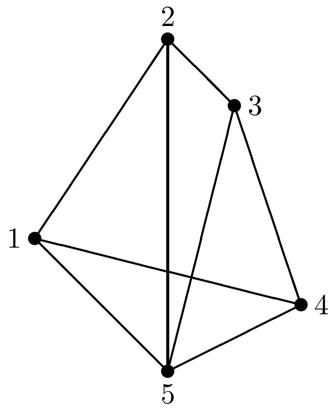
$$= \lim \frac{H(X,Y_1,Y_2,\ldots,Y_n) - H(X|Y_1,Y_2,\ldots,Y_n)}{n} \tag{4.135}$$

$$= \lim \frac{H(X) + H(Y_1,Y_2,\ldots,Y_n|X) - H(X|Y_1,Y_2,\ldots,Y_n)}{n} \tag{4.136}$$

Since $0 \leq H(X|Y_1,Y_2,\ldots,Y_n) \leq H(X) \leq 1$, we have $\lim \frac{1}{n}H(X) = 0$ and similarly $\frac{1}{n}H(X|Y_1,Y_2,\ldots,Y_n) = 0$. Also, $H(Y_1,Y_2,\ldots,Y_n|X) = nH(p)$, since the Y_i's are i.i.d. given X. Combining these terms, we get

$$\mathcal{H}(\mathcal{Y}) = \lim \frac{nH(p)}{n} = H(p) \tag{4.137}$$

19. **Random walk on graph**. Consider a random walk on the graph



(a) Calculate the stationary distribution.

(b) What is the entropy rate?

(c) Find the mutual information $I(X_{n+1}; X_n)$ assuming the process is stationary.

Solution:

(a) The stationary distribution for a connected graph of undirected edges with equal weight is given as $\mu_i = \frac{E_i}{2E}$ where E_i denotes the number of edges emanating from node i and E is the total number of edges in the graph. Hence, the stationary distribution is $[\frac{3}{16}, \frac{3}{16}, \frac{3}{16}, \frac{3}{16}, \frac{4}{16}]$; i.e., the first four nodes exterior nodes have steady state probability of $\frac{3}{16}$ while node 5 has steady state probability of $\frac{1}{4}$.

(b) Thus, the entropy rate of the random walk on this graph is $4\frac{3}{16}\log_2(3)+\frac{4}{16}\log_2(4) = \frac{3}{4}\log_2(3) + \frac{1}{2} = \log 16 - H(3/16, 3/16, 3/16, 3/16, 1/4)$

(c) The mutual information

$$\begin{aligned}
I(X_{n+1}; X_n) &= H(X_{n+1}) - H(X_{n+1}|X_n) && (4.138)\\
&= H(3/16, 3/16, 3/16, 3/16, 1/4) - (log 16 - H(3/16, 3/16, 3/16, 3/16, 1/4)) && (4.139)\\
&= 2H(3/16, 3/16, 3/16, 3/16, 1/4) - log 16 && (4.140)\\
&= 2(\frac{3}{4}\log\frac{16}{3} + \frac{1}{4}\log 4) - log 16 && (4.141)\\
&= 3 - \frac{3}{2}\log 3 && (4.142)
\end{aligned}$$

20. **Random walk on chessboard.** Find the entropy rate of the Markov chain associated with a random walk of a king on the 3 × 3 chessboard

1	2	3
4	5	6
7	8	9

What about the entropy rate of rooks, bishops and queens? There are two types of bishops.

Solution:

Random walk on the chessboard.

Notice that the king cannot remain where it is. It has to move from one state to the next. The stationary distribution is given by $\mu_i = E_i/E$, where $E_i =$ number of edges emanating from node i and $E = \sum_{i=1}^{9} E_i$. By inspection, $E_1 = E_3 = E_7 = E_9 = 3$, $E_2 = E_4 = E_6 = E_8 = 5$, $E_5 = 8$ and $E = 40$, so $\mu_1 = \mu_3 = \mu_7 = \mu_9 = 3/40$, $\mu_2 = \mu_4 = \mu_6 = \mu_8 = 5/40$ and $\mu_5 = 8/40$. In a random walk the next state is chosen with equal probability among possible choices, so $H(X_2|X_1 = i) = \log 3$ bits for $i = 1, 3, 7, 9$, $H(X_2|X_1 = i) = \log 5$ for $i = 2, 4, 6, 8$ and $H(X_2|X_1 = i) = \log 8$ bits for $i = 5$. Therefore, we can calculate the entropy rate of the king as

$$\begin{aligned} \mathcal{H} &= \sum_{i=1}^{9} \mu_i H(X_2|X_1 = i) && (4.143) \\ &= 0.3 \log 3 + 0.5 \log 5 + 0.2 \log 8 && (4.144) \\ &= 2.24 \text{ bits.} && (4.145) \end{aligned}$$

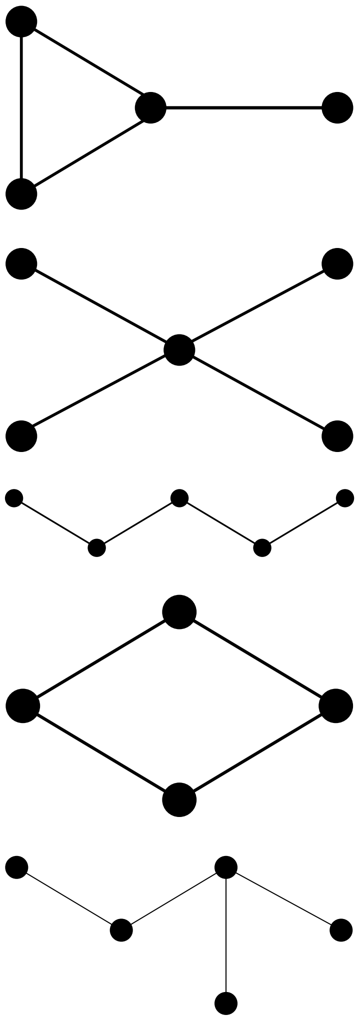
21. **Maximal entropy graphs.** Consider a random walk on a connected graph with 4 edges.

   (a) Which graph has the highest entropy rate?

   (b) Which graph has the lowest?

Solution: *Graph entropy.*

There are five graphs with four edges.



Where the entropy rates are $1/2 + 3/8 \log(3) \approx 1.094$, 1, $.75$, 1 and $1/4 + 3/8 \log(3) \approx .844$.

(a) From the above we see that the first graph maximizes entropy rate with and entropy rate of 1.094.

(b) From the above we see that the third graph minimizes entropy rate with and entropy rate of $.75$.

22. **3-D Maze.**
A bird is lost in a $3 \times 3 \times 3$ cubical maze. The bird flies from room to room going to adjoining rooms with equal probability through each of the walls. To be specific, the corner rooms have 3 exits.

(a) What is the stationary distribution?

(b) What is the entropy rate of this random walk?

Solution: *3D Maze.*
The entropy rate of a random walk on a graph with equal weights is given by equation 4.41 in the text:

$$H(\mathcal{X}) = \log(2E) - H\left(\frac{E_1}{2E}, \ldots, \frac{E_m}{2E}\right)$$

There are 8 corners, 12 edges, 6 faces and 1 center. Corners have 3 edges, edges have 4 edges, faces have 5 edges and centers have 6 edges. Therefore, the total number of edges $E = 54$. So,

$$\begin{aligned} H(\mathcal{X}) &= \log(108) + 8\left(\frac{3}{108}\log\frac{3}{108}\right) + 12\left(\frac{4}{108}\log\frac{4}{108}\right) + 6\left(\frac{5}{108}\log\frac{5}{108}\right) + 1\left(\frac{6}{108}\log\frac{6}{108}\right) \\ &= 2.03 \text{ bits} \end{aligned}$$

23. **Entropy rate**
Let $\{X_i\}$ be a stationary stochastic process with entropy rate $H(\mathcal{X})$.

(a) Argue that $H(\mathcal{X}) \le H(X_1)$.

(b) What are the conditions for equality?

Solution: Entropy Rate

(a) From Theorem 4.2.1

$$H(\mathcal{X}) = H(X_1|X_0, X_{-1}, \ldots) \le H(X_1) \tag{4.146}$$

since conditioning reduces entropy

(b) We have equality only if X_1 is independent of the past $X_0, X_{-1}, \ldots$, i.e., if and only if X_i is an i.i.d. process.

24. **Entropy rates**

Let $\{X_i\}$ be a stationary process. Let $Y_i = (X_i, X_{i+1})$. Let $Z_i = (X_{2i}, X_{2i+1})$. Let $V_i = X_{2i}$. Consider the entropy rates $H(\mathcal{X})$, $H(\mathcal{Y})$, $H(\mathcal{Z})$, and $H(\mathcal{V})$ of the processes $\{X_i\}, \{Y_i\}$, $\{Z_i\}$, and $\{V_i\}$. What is the inequality relationship $\le$, $=$, or $\ge$ between each of the pairs listed below:

(a) $H(\mathcal{X}) \overset{\geq}{\leq} H(\mathcal{Y})$.

(b) $H(\mathcal{X}) \overset{\geq}{\leq} H(\mathcal{Z})$.

(c) $H(\mathcal{X}) \overset{\geq}{\leq} H(\mathcal{V})$.

(d) $H(\mathcal{Z}) \overset{\geq}{\leq} H(\mathcal{X})$.

Solution: Entropy rates

$\{X_i\}$ is a stationary process, $Y_i = (X_i, X_{i+1})$. Let $Z_i = (X_{2i}, X_{2i+1})$. Let $V_i = X_{2i}$. Consider the entropy rates $H(\mathcal{X})$, $H(\mathcal{Y})$, $H(\mathcal{Z})$, and $H(\mathcal{V})$ of the processes $\{X_i\}$, $\{Z_i\}$, and $\{V_i\}$.

(a) $H(\mathcal{X}) = H(\mathcal{Y})$, since $H(X_1, X_2, \ldots, X_n, X_{n+1}) = H(Y_1, Y_2, \ldots, Y_n)$, and dividing by n and taking the limit, we get equality.

(b) $H(\mathcal{X}) < H(\mathcal{Z})$, since $H(X_1, \ldots, X_{2n}) = H(Z_1, \ldots, Z_n)$, and dividing by n and taking the limit, we get $2H(\mathcal{X}) = \mathcal{H}(\mathcal{Z})$.

(c) $H(\mathcal{X}) > H(\mathcal{V})$, since $H(V_1|V_0, \ldots) = H(X_2|X_0, X_{-2}, \ldots) \le H(X_2|X_1, X_0, X_{-1}, \ldots)$.

(d) $H(\mathcal{Z}) = 2H(\mathcal{X})$ since $H(X_1, \ldots, X_{2n}) = H(Z_1, \ldots, Z_n)$, and dividing by n and taking the limit, we get $2H(\mathcal{X}) = \mathcal{H}(\mathcal{Z})$.

25. **Monotonicity**.

(a) Show that $I(X; Y_1, Y_2, \ldots, Y_n)$ is non-decreasing in n.

(b) Under what conditions is the mutual information constant for all n?

Solution: *Monotonicity*

(a) Since conditioning reduces entropy,

$$H(X|Y_1, Y_2, \ldots, Y_n) \ge H(X|Y_1, Y_2, \ldots, Y_n, Y_{n+1}) \tag{4.147}$$

and hence

$$\begin{aligned} I(X; Y_1, Y_2, \ldots, Y_n) &= H(X) - H(X|Y_1, Y_2, \ldots, Y_n) && (4.148) \\ &\le H(X) - H(X|Y_1, Y_2, \ldots, Y_{n,n+1}) && (4.149) \\ &= I(X; Y_1, Y_2, \ldots, Y_n, Y_{n+1}) && (4.150) \end{aligned}$$

(b) We have equality if and only if $H(X|Y_1, Y_2, \ldots, Y_n) = H(X|Y_1)$ for all n, i.e., if X is conditionally independent of $Y_2, \ldots$ given Y_1.

26. **Transitions in Markov chains.** Suppose $\{X_i\}$ forms an irreducible Markov chain with transition matrix P and stationary distribution μ. Form the associated "edge-process" $\{Y_i\}$ by keeping track only of the transitions. Thus the new process $\{Y_i\}$ takes values in $\mathcal{X} \times \mathcal{X}$, and $Y_i = (X_{i-1}, X_i)$.

For example

$$X = 3, 2, 8, 5, 7, \ldots$$

becomes

$$Y = (\emptyset, 3), (3, 2), (2, 8), (8, 5), (5, 7), \ldots$$

Find the entropy rate of the edge process $\{Y_i\}$.

Solution: *Edge Process* $H(\mathcal{X}) = H(\mathcal{Y})$, since $H(X_1, X_2, \ldots, X_n, X_{n+1}) = H(Y_1, Y_2, \ldots, Y_n)$, and dividing by n and taking the limit, we get equality.

27. **Entropy rate**

    Let $\{X_i\}$ be a stationary $\{0,1\}$ valued stochastic process obeying

    $$X_{k+1} = X_k \oplus X_{k-1} \oplus Z_{k+1},$$

    where $\{Z_i\}$ is Bernoulli(p) and $\oplus$ denotes mod 2 addition. What is the entropy rate $H(\mathcal{X})$?

    Solution: *Entropy Rate*

    $$H(\mathcal{X}) = H(X_{k+1}|X_k, X_{k-1}, \ldots) = H(X_{k+1}|X_k, X_{k-1}) = H(Z_{k+1}) = H(p) \qquad (4.151)$$

28. **Mixture of processes**

    Suppose we observe one of two stochastic processes but don't know which. What is the entropy rate? Specifically, let $X_{11}, X_{12}, X_{13}, \ldots$ be a Bernoulli process with parameter p_1 and let $X_{21}, X_{22}, X_{23}, \ldots$ be Bernoulli(p_2). Let

    $$\theta = \begin{cases} 1, & \text{with probability } \frac{1}{2} \\ 2, & \text{with probability } \frac{1}{2} \end{cases}$$

    and let $Y_i = X_{\theta i}, \quad i = 1, 2, \ldots,$ be the observed stochastic process. Thus Y observes the process $\{X_{1i}\}$ or $\{X_{2i}\}$. Eventually Y will know which.

    (a) Is $\{Y_i\}$ stationary?

    (b) Is $\{Y_i\}$ an i.i.d. process?

    (c) What is the entropy rate H of $\{Y_i\}$?

    (d) Does

    $$-\frac{1}{n}\log p(Y_1, Y_2, \ldots Y_n) \longrightarrow H?$$

    (e) Is there a code that achieves an expected per-symbol description length $\frac{1}{n}EL_n \longrightarrow H$?

    Now let θ_i be Bern$(\frac{1}{2})$. Observe

    $$Z_i = X_{\theta_i i}, \quad i = 1, 2, \ldots,$$

    Thus θ is not fixed for all time, as it was in the first part, but is chosen i.i.d. each time. Answer (a), (b), (c), (d), (e) for the process $\{Z_i\}$, labeling the answers (a$'$), (b$'$), (c$'$), (d$'$), (e$'$).

    Solution: *Mixture of processes.*

    (a) Yes, $\{Y_i\}$ is stationary, since the scheme that we use to generate the Y_is doesn't change with time.

(b) No, it is not IID, since there's dependence now – all Y_is have been generated according to the same parameter θ.

Alternatively, we can arrive at the result by examining $I(Y_{n+1}; Y^n)$. If the process were to be IID, then the expression $I(Y_{n+1}; Y^n)$ would have to be 0. However, if we are given Y^n, then we can estimate what θ is, which in turn allows us to predict Y_{n+1}. Thus, $I(Y_{n+1}; Y^n)$ is nonzero.

(c) The process $\{Y_i\}$ is the mixture of two Bernoulli processes with different parameters, and its entropy rate is the mixture of the two entropy rates of the two processes so it's given by

$$\frac{H(p_1) + H(p_2)}{2}.$$

More rigorously,

$$\begin{aligned} H &= \lim_{n\to\infty} \frac{1}{n} H(Y^n) \\ &= \lim_{n\to\infty} \frac{1}{n}(H(\theta) + H(Y^n|\theta) - H(\theta|Y^n)) \\ &= \frac{H(p_1) + H(p_2)}{2} \end{aligned}$$

Note that only $H(Y^n|\theta)$ grows with n. The rest of the term is finite and will go to 0 as n goes to ∞.

(d) The process $\{Y_i\}$ is not ergodic, so the AEP does not apply and the quantity $-(1/n)\log P(Y_1, Y_2, \ldots, Y_n)$ does NOT converge to the entropy rate. (But it does converge to a random variable that equals $H(p_1)$ w.p. 1/2 and $H(p_2)$ w.p. 1/2.)

(e) Since the process is stationary, we can do Huffman coding on longer and longer blocks of the process. These codes will have an expected per-symbol length bounded above by $\frac{H(X_1, X_2, \ldots, X_n)+1}{n}$ and this converges to $H(\mathcal{X})$.

(a') Yes, $\{Y_i\}$ is stationary, since the scheme that we use to generate the Y_i's doesn't change with time.

(b') Yes, it is IID, since there's no dependence now – each Y_i is generated according to an independent parameter θ_i, and $Y_i \sim$ Bernoulli$((p_1 + p_2)/2)$.

(c') Since the process is now IID, its entropy rate is

$$H\left(\frac{p_1 + p_2}{2}\right).$$

(d') Yes, the limit exists by the AEP.

(e') Yes, as in (e) above.

29. **Waiting times.**
Let X be the waiting time for the first heads to appear in successive flips of a fair coin. Thus, for example, $Pr\{X = 3\} = (\frac{1}{2})^3$.

Let S_n be the waiting time for the n^{th} head to appear.
Thus,

$$\begin{aligned} S_0 &= 0 \\ S_{n+1} &= S_n + X_{n+1} \end{aligned}$$

where $X_1, X_2, X_3, \ldots$ are i.i.d according to the distribution above.

(a) Is the process $\{S_n\}$ stationary?

(b) Calculate $H(S_1, S_2, \ldots, S_n)$.

(c) Does the process $\{S_n\}$ have an entropy rate? If so, what is it? If not, why not?

(d) What is the expected number of fair coin flips required to generate a random variable having the same distribution as S_n?

Solution: Waiting time process.

(a) For the process to be stationary, the distribution must be time invariant. It turns out that process $\{S_n\}$ is not stationary. There are several ways to show this.

- S_0 is always 0 while S_i, $i \neq 0$ can take on several values. Since the marginals for S_0 and S_1, for example, are not the same, the process can't be stationary.
- It's clear that the variance of S_n grows with n, which again implies that the marginals are not time-invariant.
- Process $\{S_n\}$ is an independent increment process. An independent increment process is not stationary (not even wide sense stationary), since $\mathrm{var}(S_n) = \mathrm{var}(X_n) + \mathrm{var}(S_{n-1}) > \mathrm{var}(S_{n-1})$.

(b) We can use chain rule and Markov properties to obtain the following results.

$$\begin{aligned} H(S_1, S_2, \ldots, S_n) &= H(S_1) + \sum_{i=2}^{n} H(S_i|S^{i-1}) \\ &= H(S_1) + \sum_{i=2}^{n} H(S_i|S_{i-1}) \\ &= H(X_1) + \sum_{i=2}^{n} H(X_i) \\ &= \sum_{i=1}^{n} H(X_i) \\ &= 2n \end{aligned}$$

(c) It follows trivially from the previous part that

$$\begin{aligned} \mathcal{H}(S) &= \lim_{n\to\infty} \frac{H(S^n)}{n} \\ &= \lim_{n\to\infty} \frac{2n}{n} \\ &= 2 \end{aligned}$$

Note that the entropy rate can still exist even when the process is not stationary. Furthermore, the entropy rate (for this problem) is the same as the entropy of X.

(d) The expected number of flips required can be lower-bounded by $H(S_n)$ and upper-bounded by $H(S_n)+2$ (Theorem 5.12.3, page 115). S_n has a negative binomial distribution; i.e., $Pr(S_n = k) = \left(\begin{array}{c} k-1 \\ n-1 \end{array}\right)(\frac{1}{2})^k$ for $k \geq n$. (We have the nth success at the kth trial if and only if we have exactly $n-1$ successes in $k-1$ trials and a suceess at the kth trial.)

Since computing the exact value of $H(S_n)$ is difficult (and fruitless in the exam setting), it would be sufficient to show that the expected number of flips required is between $H(S_n)$ and $H(S_n)+2$, and set up the expression of $H(S_n)$ in terms of the pmf of S_n.

Note, however, that for large n, however, the distribution of S_n will tend to Gaussian with mean $\frac{n}{p} = 2n$ and variance $n(1-p)/p^2 = 2n$.
Let $p_k = Pr(S_n = k + ES_n) = Pr(S_n = k + 2n)$. Let $\phi(x)$ be the normal density function with mean zero and variance $2n$, i.e. $\phi(x) = \exp(-x^2/2\sigma^2)/\sqrt{2\pi\sigma^2}$, where $\sigma^2 = 2n$.
Then for large n, since the entropy is invariant under any constant shift of a random variable and $\phi(x)\log\phi(x)$ is Riemann integrable,

$$\begin{aligned}
H(S_n) &= H(S_n - E(S_n)) \\
&= -\sum p_k \log p_k \\
&\approx -\sum \phi(k)\log\phi(k) \\
&\approx -\int \phi(x)\log\phi(x)dx \\
&= (-\log e)\int \phi(x)\ln\phi(x)dx \\
&= (-\log e)\int \phi(x)(-\frac{x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}) \\
&= (\log e)(\frac{1}{2} + \frac{1}{2}\ln 2\pi\sigma^2) \\
&= \frac{1}{2}\log 2\pi e\sigma^2 \\
&= \frac{1}{2}\log n\pi e + 1.
\end{aligned}$$

(Refer to Chapter 9 for a more general discussion of the entropy of a continuous random variable and its relation to discrete entropy.)

Here is a specific example for $n = 100$. Based on earlier discussion, $Pr(S_{100} = k) = \left(\begin{array}{c} k-1 \\ 100-1 \end{array}\right)(\frac{1}{2})^k$. The Gaussian approximation of $H(S_n)$ is 5.8690 while the exact value of $H(S_n)$ is 5.8636. The expected number of flips required is somewhere between 5.8636 and 7.8636.

30. **Markov chain transitions.**

$$P = [P_{ij}] = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

Let X_1 be uniformly distributed over the states $\{0,1,2\}$. Let $\{X_i\}_1^\infty$ be a Markov chain with transition matrix P, thus $P(X_{n+1}=j|X_n=i) = P_{ij}, i,j \in \{0,1,2\}$.

(a) Is $\{X_n\}$ stationary?

(b) Find $\lim_{n\to\infty} \frac{1}{n} H(X_1,\ldots,X_n)$.

Now consider the derived process $Z_1, Z_2, \ldots, Z_n$, where

$$\begin{aligned} Z_1 &= X_1 \\ Z_i &= X_i - X_{i-1} \pmod 3, \qquad i = 2,\ldots,n. \end{aligned}$$

Thus Z^n encodes the transitions, not the states.

(c) Find $H(Z_1, Z_2, ..., Z_n)$.

(d) Find $H(Z_n)$ and $H(X_n)$, for $n \geq 2$.

(e) Find $H(Z_n|Z_{n-1})$ for $n \geq 2$.

(f) Are Z_{n-1} and Z_n independent for $n \geq 2$?

Solution:

(a) Let μ_n denote the probability mass function at time n. Since $\mu_1 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $\mu_2 = \mu_1 P = \mu_1$, $\mu_n = \mu_1 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ for all n and $\{X_n\}$ is stationary.
Alternatively, the observation P is doubly stochastic will lead the same conclusion.

(b) Since $\{X_n\}$ is stationary Markov,

$$\begin{aligned} \lim_{n\to\infty} H(X_1,\ldots,X_n) &= H(X_2|X_1) \\ &= \sum_{k=0}^{2} P(X_1 = k) H(X_2|X_1 = k) \\ &= 3 \times \frac{1}{3} \times H(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) \\ &= \frac{3}{2}. \end{aligned}$$

(c) Since $(X_1, \ldots, X_n)$ and $(Z_1, \ldots, Z_n)$ are one-to-one, by the chain rule of entropy and the Markovity,

$$\begin{aligned} H(Z_1, \ldots, Z_n) &= H(X_1, \ldots, X_n) \\ &= \sum_{k=1}^{n} H(X_k|X_1, \ldots, X_{k-1}) \\ &= H(X_1) + \sum_{k=2}^{n} H(X_k|X_{k-1}) \\ &= H(X_1) + (n-1)H(X_2|X_1) \\ &= \log 3 + \frac{3}{2}(n-1). \end{aligned}$$

Alternatively, we can use the results of parts (d), (e), and (f). Since $Z_1, \ldots, Z_n$ are independent and $Z_2, \ldots, Z_n$ are identically distributed with the probability distribution $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$,

$$\begin{aligned} H(Z_1, \ldots, Z_n) &= H(Z_1) + H(Z_2) + \ldots + H(Z_n) \\ &= H(Z_1) + (n-1)H(Z_2) \\ &= \log 3 + \frac{3}{2}(n-1). \end{aligned}$$

(d) Since $\{X_n\}$ is stationary with $\mu_n = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$,

$$H(X_n) = H(X_1) = H(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) = \log 3.$$

For $n \geq 2$, $Z_n = \begin{cases} 0, & \frac{1}{2}, \\ 1, & \frac{1}{4}, \\ 2, & \frac{1}{4}. \end{cases}$

Hence, $H(Z_n) = H(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) = \frac{3}{2}$.

(e) Due to the symmetry of P, $P(Z_n|Z_{n-1}) = P(Z_n)$ for $n \geq 2$. Hence, $H(Z_n|Z_{n-1}) = H(Z_n) = \frac{3}{2}$.

Alternatively, using the result of part (f), we can trivially reach the same conclusion.

(f) Let $k \geq 2$. First observe that by the symmetry of P, $Z_{k+1} = X_{k+1} - X_k$ is independent of X_k. Now that

$$\begin{aligned} H(Z_{k+1}|X_k, X_{k-1}) &= H(X_{k+1} - X_k|X_k, X_{k-1}) \\ &= H(X_{k+1} - X_k|X_k) \\ &= H(X_{k+1} - X_k) \\ &= H(Z_{k+1}), \end{aligned}$$

Z_{k+1} is independent of (X_k, X_{k-1}) and hence independent of $Z_k = X_k - X_{k-1}$. For $k = 1$, again by the symmetry of P, Z_2 is independent of $Z_1 = X_1$ trivially.

31. **Markov.**

    Let $\{X_i\} \sim$ Bernoulli(p). Consider the associated Markov chain $\{Y_i\}_{i=1}^n$ where $Y_i =$ (the number of 1's in the current run of 1's). For example, if $X^n = 101110\ldots$, we have $Y^n = 101230\ldots$.

    (a) Find the entropy rate of X^n.

    (b) Find the entropy rate of Y^n.

    Solution: Markov solution.

    (a) For an i.i.d. source, $H(\mathcal{X}) = H(X) = H(p)$.

    (b) Observe that X^n and Y^n have a one-to-one mapping. Thus, $H(\mathcal{Y}) = H(\mathcal{X}) = H(p)$.

32. **Time symmetry.**
    Let $\{X_n\}$ be a stationary Markov process. We condition on (X_0, X_1) and look into the past and future. For what index k is

$$H(X_{-n}|X_0, X_1) = H(X_k|X_0, X_1)?$$

Give the argument.

Solution: *Time symmetry.*
The trivial solution is $k = -n$. To find other possible values of k we expand

$$\begin{aligned}
H(X_{-n}|X_0, X_1) &= H(X_{-n}, X_0, X_1) - H(X_0, X_1) \\
&= H(X_{-n}) + H(X_0, X_1|X_{-n}) - H(X_0, X_1) \\
&= H(X_{-n}) + H(X_0|X_{-n}) + H(X_1|X_0, X_{-n}) - H(X_0, X_1) \\
&\overset{(a)}{=} H(X_{-n}) + H(X_0|X_{-n}) + H(X_1|X_0) - H(X_0, X_1) \\
&= H(X_{-n}) + H(X_0|X_{-n}) - H(X_0) \\
&\overset{(b)}{=} H(X_0) + H(X_0|X_{-n}) - H(X_0) \\
&\overset{(c)}{=} H(X_n|X_0) \\
&\overset{(d)}{=} H(X_n|X_0, X_{-1}) \\
&\overset{(e)}{=} H(X_{n+1}|X_1, X_0)
\end{aligned}$$

where (a) and (d) come from Markovity and (b), (c) and (e) come from stationarity. Hence $k = n + 1$ is also a solution. There are no other solution since for any other k, we can construct a periodic Markov process as a counterexample. Therefore $k \in \{-n, n + 1\}$.

33. **Chain inequality:** Let $X_1 \to X_2 \to X_3 \to X_4$ form a Markov chain. Show that

$$I(X_1;X_3) + I(X_2;X_4) \le I(X_1;X_4) + I(X_2;X_3) \tag{4.152}$$

Solution: *Chain inequality* $X_1 \to X_2 \to X_3 \to X_4$

$$\begin{aligned}
I(X_1;X_4) \quad & +I(X_2;X_3) - I(X_1;X_3) - I(X_2;X_4) && (4.153)\\
= \quad & H(X_1) - H(X_1|X_4) + H(X_2) - H(X_2|X_3) - (H(X_1) - H(X_1|X_3)) \\
& \quad -(H(X_2) - H(X_2|X_4)) && (4.154)\\
= \quad & H(X_1|X_3) - H(X_1|X_4) + H(X_2|X_4) - H(X_2|X_3) && (4.155)\\
= \quad & H(X_1,X_2|X_3) - H(X_2|X_1,X_3) - H(X_1,X_2|X_4) + H(X_2|X_1,X_4) && (4.156)\\
& \quad +H(X_1,X_2|X_4) - H(X_1|X_2,X_4) - H(X_1,X_2|X_3) + H(X_1|X_2,X_3) && (4.157)\\
= \quad & -H(X_2|X_1,X_3) + H(X_2|X_1,X_4) && (4.158)\\
- \quad & H(X_2|X_1,X_4) - H(X_2|X_1,X_3,X_4) && (4.159)\\
= \quad & I(X_2;X_3|X_1,X_4) && (4.160)\\
\ge \quad & 0 && (4.161)
\end{aligned}$$

where $H(X_1|X_2,X_3) = H(X_1|X_2,X_4)$ by the Markovity of the random variables.

34. **Broadcast channel**. Let $X \to Y \to (Z,W)$ form a Markov chain, i.e., $p(x,y,z,w) = p(x)p(y|x)p(z,w|y)$ for all x,y,z,w. Show that

$$I(X;Z) + I(X;W) \le I(X;Y) + I(Z;W) \tag{4.162}$$

Solution: *Broadcast Channel*

$X \to Y \to (Z,W)$, hence by the data processing inequality, $I(X;Y) \ge I(X;(Z,W))$, and hence

$$\begin{aligned}
I(X:Y) \quad & +I(Z;W) - I(X;Z) - I(X;W) && (4.163)\\
\ge \quad & I(X:Z,W) + I(Z;W) - I(X;Z) - I(X;W) && (4.164)\\
= \quad & H(Z,W) + H(X) - H(X,W,Z) + H(W) + H(Z) - H(W,Z) \\
& \quad -H(Z) - H(X) + H(X,Z)) - H(W) - H(X) + H(W,X) && (4.165)\\
= \quad & -H(X,W,Z) + H(X,Z) + H(X,W) - H(X) && (4.166)\\
= \quad & H(W|X) - H(W|X,Z) && (4.167)\\
= \quad & I(W;Z|X) && (4.168)\\
\ge \quad & 0 && (4.169)
\end{aligned}$$

35. **Concavity of second law.** Let $\{X_n\}_{-\infty}^{\infty}$ be a stationary Markov process. Show that $H(X_n|X_0)$ is concave in n. Specifically show that

$$\begin{aligned}
H(X_n|X_0) - H(X_{n-1}|X_0) - (H(X_{n-1}|X_0) - H(X_{n-2}|X_0)) \quad &= \quad -I(X_1;X_{n-1}|X_0,X_n) && (4.170)\\
&\le \quad 0 && (4.171)
\end{aligned}$$

Thus the second difference is negative, establishing that $H(X_n|X_0)$ is a concave function of n.

Solution: *Concavity of second law of thermodynamics*

Since $X_0 \to X_{n-2} \to X_{n-1} \to X_n$ is a Markov chain

$$\begin{align}
H(X_n|X_0) & \quad -H(X_{n-1}|X_0) - (H(X_{n-1}|X_0) - H(X_{n-2}|X_0) \tag{4.172}\\
&= H(X_n|X_0) - H(X_{n-1}|X_0, X_{-1}) - (H(X_{n-1}|X_0, X_{-1}) - H(X_{n-2}|X_0, X_{-1}) \tag{4.173}\\
&= H(X_n|X_0) - H(X_n|X_1, X_0) - (H(X_{n-1}|X_0) - H(X_{n-1}|X_1, X_0) \tag{4.174}\\
&= I(X_1; X_n|X_0) - I(X_1; X_{n-1}|X_0) \tag{4.175}\\
&= H(X_1|X_0) - H(X_1|X_n, X_0) - H(X_1|X_0) + H(X_1|X_{n-1}, X_0) \tag{4.176}\\
&= H(X_1|X_{n-1}, X_0) - H(X_1|X_n, X_0) \tag{4.177}\\
&= H(X_1, X_{n-1}, X_n, X_0) - H(X_1|X_n, X_0) \tag{4.178}\\
&= -I(X_1; X_{n-1}|X_n, X_0) \tag{4.179}\\
&\leq 0 \tag{4.180}
\end{align}$$

where (4.173) and (4.178) follows from Markovity and (4.174) follows from stationarity of the Markov chain.

If we define

$$\Delta_n = H(X_n|X_0) - H(X_{n-1}|X_0) \tag{4.181}$$

then the above chain of inequalities implies that $\Delta_n - \Delta_{n-1} \leq 0$, which implies that $H(X_n|X_0)$ is a concave function of n.

Chapter 5

Data Compression

1. **Uniquely decodable and instantaneous codes.** Let $L = \sum_{i=1}^{m} p_i l_i^{100}$ be the expected value of the 100th power of the word lengths associated with an encoding of the random variable X. Let $L_1 = \min L$ over all instantaneous codes; and let $L_2 = \min L$ over all uniquely decodable codes. What inequality relationship exists between L_1 and L_2?

   Solution: *Uniquely decodable and instantaneous codes.*

$$L = \sum_{i=1}^{m} p_i n_i^{100} \tag{5.1}$$

$$L_1 = \min_{\text{Instantaneous codes}} L \tag{5.2}$$

$$L_2 = \min_{\text{Uniquely decodable codes}} L \tag{5.3}$$

   Since all instantaneous codes are uniquely decodable, we must have $L_2 \leq L_1$. Any set of codeword lengths which achieve the minimum of L_2 will satisfy the Kraft inequality and hence we can construct an instantaneous code with the same codeword lengths, and hence the same L. Hence we have $L_1 \leq L_2$. From both these conditions, we must have $L_1 = L_2$.

2. **How many fingers has a Martian?** Let

$$S = \begin{pmatrix} S_1, \ldots, S_m \\ p_1, \ldots, p_m \end{pmatrix}.$$

   The S_i's are encoded into strings from a D-symbol output alphabet in a uniquely decodable manner. If $m = 6$ and the codeword lengths are $(l_1, l_2, \ldots, l_6) = (1, 1, 2, 3, 2, 3)$, find a good lower bound on D. You may wish to explain the title of the problem.

   Solution: *How many fingers has a Martian?*

Uniquely decodable codes satisfy Kraft's inequality. Therefore

$$f(D) = D^{-1} + D^{-1} + D^{-2} + D^{-3} + D^{-2} + D^{-3} \leq 1. \tag{5.4}$$

We have $f(2) = 7/4 > 1$, hence $D > 2$. We have $f(3) = 26/27 < 1$. So a possible value of D is 3. Our counting system is base 10, probably because we have 10 fingers. Perhaps the Martians were using a base 3 representation because they have 3 fingers. (Maybe they are like Maine lobsters ?)

3. **Slackness in the Kraft inequality.** An instantaneous code has word lengths $l_1, l_2, \ldots, l_m$ which satisfy the strict inequality

$$\sum_{i=1}^{m} D^{-l_i} < 1.$$

The code alphabet is $\mathcal{D} = \{0, 1, 2, \ldots, D-1\}$. Show that there exist arbitrarily long sequences of code symbols in $\mathcal{D}^*$ which cannot be decoded into sequences of codewords.

Solution:

Slackness in the Kraft inequality. Instantaneous codes are prefix free codes, i.e., no codeword is a prefix of any other codeword. Let $n_{max} = \max\{n_1, n_2, ..., n_q\}$. There are $D^{n_{max}}$ sequences of length n_{max}. Of these sequences, $D^{n_{max}-n_i}$ start with the i-th codeword. Because of the prefix condition no two sequences can start with the same codeword. Hence the total number of sequences which start with some codeword is $\sum_{i=1}^{q} D^{n_{max}-n_i} = D^{n_{max}} \sum_{i=1}^{q} D^{-n_i} < D^{n_{max}}$. Hence there are sequences which do not start with any codeword. These and all longer sequences with these length n_{max} sequences as prefixes cannot be decoded. (This situation can be visualized with the aid of a tree.)

Alternatively, we can map codewords onto dyadic intervals on the real line corresponding to real numbers whose decimal expansions start with that codeword. Since the length of the interval for a codeword of length n_i is D^{-n_i}, and $\sum D^{-n_i} < 1$, there exists some interval(s) not used by any codeword. The binary sequences in these intervals do not begin with any codeword and hence cannot be decoded.

4. **Huffman coding.** Consider the random variable

$$X = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ 0.49 & 0.26 & 0.12 & 0.04 & 0.04 & 0.03 & 0.02 \end{pmatrix}$$

(a) Find a binary Huffman code for X.

(b) Find the expected codelength for this encoding.

(c) Find a ternary Huffman code for X.

Solution: *Examples of Huffman codes.*

(a) The Huffman tree for this distribution is

Codeword								
1	x_1	0.49	0.49	0.49	0.49	0.49	0.51	1
00	x_2	0.26	0.26	0.26	0.26	0.26	0.49	
011	x_3	0.12	0.12	0.12	0.13	0.25		
01000	x_4	0.04	0.05	0.08	0.12			
01001	x_5	0.04	0.04	0.05				
01010	x_6	0.03	0.04					
01011	x_7	0.02						

(b) The expected length of the codewords for the binary Huffman code is 2.02 bits. ($H(X) = 2.01$ bits)

(c) The ternary Huffman tree is

Codeword					
0	x_1	0.49	0.49	0.49	1.0
1	x_2	0.26	0.26	0.26	
20	x_3	0.12	0.12	0.25	
22	x_4	0.04	0.09		
210	x_5	0.04	0.04		
211	x_6	0.03			
212	x_7	0.02			

This code has an expected length 1.34 ternary symbols. ($H_3(X) = 1.27$ ternary symbols).

5. **More Huffman codes.** Find the binary Huffman code for the source with probabilities $(1/3, 1/5, 1/5, 2/15, 2/15)$. Argue that this code is also optimal for the source with probabilities $(1/5, 1/5, 1/5, 1/5, 1/5)$.

Solution: *More Huffman codes.* The Huffman code for the source with probabilities $(\frac{1}{3}, \frac{1}{5}, \frac{1}{5}, \frac{2}{15}, \frac{2}{15})$ has codewords $\{00,10,11,010,011\}$.

To show that this code (*) is also optimal for (1/5, 1/5, 1/5, 1/5, 1/5) we have to show that it has minimum expected length, that is, no shorter code can be constructed without violating $H(X) \leq EL$.

$$H(X) = \log 5 = 2.32 \text{ bits.} \tag{5.5}$$

$$E(L(*)) = 2 \times \frac{3}{5} + 3 \times \frac{2}{5} = \frac{12}{5} \text{ bits.} \tag{5.6}$$

Since

$$E(L(\text{any code})) = \sum_{i=1}^{5} \frac{l_i}{5} = \frac{k}{5} \text{ bits} \tag{5.7}$$

for some integer k, the next lowest possible value of $E(L)$ is 11/5 = 2.2 bits ¡ 2.32 bits. Hence (*) is optimal.

Note that one could also prove the optimality of (*) by showing that the Huffman code for the (1/5, 1/5, 1/5, 1/5, 1/5) source has average length 12/5 bits. (Since each

Huffman code produced by the Huffman encoding algorithm is optimal, they all have the same average length.)

6. **Bad codes.** Which of these codes cannot be Huffman codes for any probability assignment?

   (a) $\{0, 10, 11\}$.
   (b) $\{00, 01, 10, 110\}$.
   (c) $\{01, 10\}$.

   Solution: *Bad codes*

   (a) {0,10,11} is a Huffman code for the distribution (1/2,1/4,1/4).
   (b) The code {00,01,10, 110} can be shortened to {00,01,10, 11} without losing its instantaneous property, and therefore is not optimal, so it cannot be a Huffman code. Alternatively, it is not a Huffman code because there is a unique longest codeword.
   (c) The code {01,10} can be shortened to {0,1} without losing its instantaneous property, and therefore is not optimal and not a Huffman code.

7. **Huffman 20 questions.** Consider a set of n objects. Let $X_i = 1$ or 0 accordingly as the i-th object is good or defective. Let $X_1, X_2, \ldots, X_n$ be independent with $\Pr\{X_i = 1\} = p_i$; and $p_1 > p_2 > \ldots > p_n > 1/2$. We are asked to determine the set of all defective objects. Any yes-no question you can think of is admissible.

   (a) Give a good lower bound on the minimum average number of questions required.
   (b) If the longest sequence of questions is required by nature's answers to our questions, what (in words) is the last question we should ask? And what two sets are we distinguishing with this question? Assume a compact (minimum average length) sequence of questions.
   (c) Give an upper bound (within 1 question) on the minimum average number of questions required.

   Solution: *Huffman 20 Questions.*

   (a) We will be using the questions to determine the sequence $X_1, X_2, \ldots, X_n$, where X_i is 1 or 0 according to whether the i-th object is good or defective. Thus the most likely sequence is all 1's, with a probability of $\prod_{i=1}^n p_i$, and the least likely sequence is the all 0's sequence with probability $\prod_{i=1}^n (1-p_i)$. Since the optimal set of questions corresponds to a Huffman code for the source, a good lower bound on the average number of questions is the entropy of the sequence $X_1, X_2, \ldots, X_n$. But since the X_i 's are independent Bernoulli random variables, we have

$$EQ \geq H(X_1, X_2, \ldots, X_n) = \sum H(X_i) = \sum H(p_i). \tag{5.8}$$

(b) The last bit in the Huffman code distinguishes between the least likely source symbols. (By the conditions of the problem, all the probabilities are different, and thus the two least likely sequences are uniquely defined.) In this case, the two least likely sequences are $000\ldots00$ and $000\ldots01$, which have probabilities $(1-p_1)(1-p_2)\ldots(1-p_n)$ and $(1-p_1)(1-p_2)\ldots(1-p_{n-1})p_n$ respectively. Thus the last question will ask "Is $X_n = 1$", i.e., "Is the last item defective?".

(c) By the same arguments as in Part (a), an upper bound on the minimum average number of questions is an upper bound on the average length of a Huffman code, namely $H(X_1, X_2, \ldots, X_n) + 1 = \sum H(p_i) + 1$.

8. **Simple optimum compression of a Markov source.** Consider the 3-state Markov process $U_1, U_2, \ldots,$ having transition matrix

$U_{n-1} \backslash U_n$	S_1	S_2	S_3
S_1	1/2	1/4	1/4
S_2	1/4	1/2	1/4
S_3	0	1/2	1/2

Thus the probability that S_1 follows S_3 is equal to zero. Design 3 codes C_1, C_2, C_3 (one for each state $1, 2$ and 3), each code mapping elements of the set of S_i's into sequences of 0's and 1's, such that this Markov process can be sent with maximal compression by the following scheme:

(a) Note the present symbol $X_n = i$.

(b) Select code C_i.

(c) Note the next symbol $X_{n+1} = j$ and send the codeword in C_i corresponding to j.

(d) Repeat for the next symbol.

What is the average message length of the next symbol conditioned on the previous state $X_n = i$ using this coding scheme? What is the unconditional average number of bits per source symbol? Relate this to the entropy rate $H(\mathcal{U})$ of the Markov chain.

Solution: *Simple optimum compression of a Markov source.*

It is easy to design an optimal code for each state. A possible solution is

Next state	S_1	S_2	S_3	
Code C_1	0	10	11	$E(L\|C_1) = 1.5$ bits/symbol
code C_2	10	0	11	$E(L\|C_2) = 1.5$ bits/symbol
code C_3	-	0	1	$E(L\|C_3) = 1$ bit/symbol

The average message lengths of the next symbol conditioned on the previous state being S_i are just the expected lengths of the codes C_i. Note that this code assignment achieves the conditional entropy lower bound.

To find the unconditional average, we have to find the stationary distribution on the states. Let μ be the stationary distribution. Then

$$\mu = \mu \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1/2 & 1/2 \end{bmatrix} \tag{5.9}$$

We can solve this to find that $\mu = (2/9, 4/9, 1/3)$. Thus the unconditional average number of bits per source symbol

$$\begin{aligned} EL &= \sum_{i=1}^{3} \mu_i E(L|C_i) && (5.10) \\ &= \frac{2}{9} \times 1.5 + \frac{4}{9} \times 1.5 + \frac{1}{3} \times 1 && (5.11) \\ &= \frac{4}{3} \text{ bits/symbol.} && (5.12) \end{aligned}$$

The entropy rate $\mathcal{H}$ of the Markov chain is

$$\begin{aligned} \mathcal{H} &= H(X_2|X_1) && (5.13) \\ &= \sum_{i=1}^{3} \mu_i H(X_2|X_1 = S_i) && (5.14) \\ &= 4/3 \text{ bits/symbol.} && (5.15) \end{aligned}$$

Thus the unconditional average number of bits per source symbol and the entropy rate $\mathcal{H}$ of the Markov chain are equal, because the expected length of each code C_i equals the entropy of the state after state i, $H(X_2|X_1 = S_i)$, and thus maximal compression is obtained.

9. **Optimal code lengths that require one bit above entropy.** The source coding theorem shows that the optimal code for a random variable X has an expected length less than $H(X)+1$. Give an example of a random variable for which the expected length of the optimal code is close to $H(X)+1$, i.e., for any $\epsilon > 0$, construct a distribution for which the optimal code has $L > H(X) + 1 - \epsilon$.

   Solution: *Optimal code lengths that require one bit above entropy.* There is a trivial example that requires almost 1 bit above its entropy. Let X be a binary random variable with probability of $X = 1$ close to 1. Then entropy of X is close to 0, but the length of its optimal code is 1 bit, which is almost 1 bit above its entropy.

10. **Ternary codes that achieve the entropy bound.** A random variable X takes on m values and has entropy $H(X)$. An instantaneous ternary code is found for this source, with average length

$$L = \frac{H(X)}{\log_2 3} = H_3(X). \tag{5.16}$$

(a) Show that each symbol of X has a probability of the form 3^{-i} for some i.

(b) Show that m is odd.

Solution: *Ternary codes that achieve the entropy bound.*

(a) We will argue that an optimal ternary code that meets the entropy bound corresponds to complete ternary tree, with the probability of each leaf of the form 3^{-i}. To do this, we essentially repeat the arguments of Theorem 5.3.1. We achieve the ternary entropy bound only if $D(\mathbf{p}||\mathbf{r}) = 0$ and $c = 1$, in (5.25). Thus we achieve the entropy bound if and only if $p_i = 3^{-j}$ for all i.

(b) We will show that any distribution that has $p_i = 3^{-l_i}$ for all i must have an odd number of symbols. We know from Theorem 5.2.1, that given the set of lengths, l_i, we can construct a ternary tree with nodes at the depths l_i. Now, since $\sum 3^{-l_i} = 1$, the tree must be complete. A complete ternary tree has an odd number of leaves (this can be proved by induction on the number of internal nodes). Thus the number of source symbols is odd.

Another simple argument is to use basic number theory. We know that for this distribution, $\sum 3^{-l_i} = 1$. We can write this as $3^{-l_{max}} \sum 3^{l_{max}-l_i} = 1$ or $\sum 3^{l_{max}-l_i} = 3^{l_{max}}$. Each of the terms in the sum is odd, and since their sum is odd, the number of terms in the sum has to be odd (the sum of an even number of odd terms is even). Thus there are an odd number of source symbols for any code that meets the ternary entropy bound.

11. **Suffix condition.** Consider codes that satisfy the suffix condition, which says that no codeword is a suffix of any other codeword. Show that a suffix condition code is uniquely decodable, and show that the minimum average length over all codes satisfying the suffix condition is the same as the average length of the Huffman code for that random variable.

Solution: *Suffix condition.* The fact that the codes are uniquely decodable can be seen easily be reversing the order of the code. For any received sequence, we work backwards from the end, and look for the reversed codewords. Since the codewords satisfy the suffix condition, the reversed codewords satisfy the prefix condition, and the we can uniquely decode the reversed code.

The fact that we achieve the same minimum expected length then follows directly from the results of Section 5.5. But we can use the same reversal argument to argue that corresponding to every suffix code, there is a prefix code of the same length and vice versa, and therefore we cannot achieve any lower codeword lengths with a suffix code than we can with a prefix code.

12. **Shannon codes and Huffman codes.** Consider a random variable X which takes on four values with probabilities $(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12})$.

(a) Construct a Huffman code for this random variable.

(b) Show that there exist two different sets of optimal lengths for the codewords, namely, show that codeword length assignments $(1,2,3,3)$ and $(2,2,2,2)$ are both optimal.

(c) Conclude that there are optimal codes with codeword lengths for some symbols that exceed the Shannon code length $\lceil \log \frac{1}{p(x)} \rceil$.

Solution: *Shannon codes and Huffman codes.*

(a) Applying the Huffman algorithm gives us the following table

Code	Symbol	Probability			
0	1	1/3	1/3	2/3	1
11	2	1/3	1/3	1/3	
101	3	1/4	1/3		
100	4	1/12			

which gives codeword lengths of 1,2,3,3 for the different codewords.

(b) Both set of lengths 1,2,3,3 and 2,2,2,2 satisfy the Kraft inequality, and they both achieve the same expected length (2 bits) for the above distribution. Therefore they are both optimal.

(c) The symbol with probability $1/4$ has an Huffman code of length 3, which is greater than $\lceil \log \frac{1}{p} \rceil$. Thus the Huffman code for a particular symbol may be longer than the Shannon code for that symbol. But on the average, the Huffman code cannot be longer than the Shannon code.

13. **Twenty questions.** Player A chooses some object in the universe, and player B attempts to identify the object with a series of yes-no questions. Suppose that player B is clever enough to use the code achieving the minimal expected length with respect to player A's distribution. We observe that player B requires an average of 38.5 questions to determine the object. Find a rough lower bound to the number of objects in the universe.

Solution: *Twenty questions.*

$$37.5 = L^* - 1 < H(X) \leq \log |\mathcal{X}| \tag{5.17}$$

and hence number of objects in the universe $> 2^{37.5} = 1.94 \times 10^{11}$.

14. **Huffman code.** Find the (a) *binary* and (b) *ternary* Huffman codes for the random variable X with probabilities

$$p = (\frac{1}{21}, \frac{2}{21}, \frac{3}{21}, \frac{4}{21}, \frac{5}{21}, \frac{6}{21}) \quad .$$

(c) Calculate $L = \sum p_i l_i$ in each case.

Solution: *Huffman code.*

(a) The Huffman tree for this distribution is

Codeword							
00	x_1	6/21	6/21	6/21	9/21	12/21	1
10	x_2	5/21	5/21	6/21	6/21	9/21	
11	x_3	4/21	4/21	5/21	6/21		
010	x_4	3/21	3/21	4/21			
0110	x_5	2/21	3/21				
0111	x_6	1/21					

(b) The ternary Huffman tree is

Codeword					
1	x_1	6/21	6/21	10/21	1
2	x_2	5/21	5/21	6/21	
00	x_3	4/21	4/21	5/21	
01	x_4	3/21	3/21		
020	x_5	2/21	3/21		
021	x_6	1/21			
022	x_7	0/21			

(c) The expected length of the codewords for the binary Huffman code is $51/21 = 2.43$ bits.

The ternary code has an expected length of $34/21 = 1.62$ ternary symbols.

15. **Huffman codes.**

(a) Construct a binary Huffman code for the following distribution on 5 symbols $\mathbf{p} = (0.3, 0.3, 0.2, 0.1, 0.1)$. What is the average length of this code?

(b) Construct a probability distribution $\mathbf{p}'$ on 5 symbols for which the code that you constructed in part (a) has an average length (under $\mathbf{p}'$) equal to its entropy $H(\mathbf{p}')$.

Solution: *Huffman codes*

(a) The code constructed by the standard Huffman procedure

Codeword	X	Probability				
10	1	0.3	0.3	0.4	0.6	1
11	2	0.3	0.3	0.3	0.4	
00	3	0.2	0.2	0.3		
010	4	0.1	0.2			
011	5	0.1				

The average length $= 2 * 0.8 + 3 * 0.2 = 2.2$ bits/symbol.

(b) The code would have a rate equal to the entropy if each of the codewords was of length $1/p(X)$. In this case, the code constructed above would be efficient for the distrubution (0.25.0.25,0.25,0.125,0.125).

16. **Huffman codes:** Consider a random variable X which takes 6 values $\{A, B, C, D, E, F\}$ with probabilities $(0.5, 0.25, 0.1, 0.05, 0.05, 0.05)$ respectively.

(a) Construct a binary Huffman code for this random variable. What is its average length?

(b) Construct a quaternary Huffman code for this random variable, i.e., a code over an alphabet of four symbols (call them a, b, c and d). What is the average length of this code?

(c) One way to construct a binary code for the random variable is to start with a quaternary code, and convert the symbols into binary using the mapping $a \rightarrow 00$, $b \rightarrow 01$, $c \rightarrow 10$ and $d \rightarrow 11$. What is the average length of the binary code for the above random variable constructed by this process?

(d) For any random variable X, let L_H be the average length of the binary Huffman code for the random variable, and let L_{QB} be the average length code constructed by first building a quaternary Huffman code and converting it to binary. Show that

$$L_H \leq L_{QB} < L_H + 2 \tag{5.18}$$

(e) The lower bound in the previous example is tight. Give an example where the code constructed by converting an optimal quaternary code is also the optimal binary code.

(f) The upper bound, i.e., $L_{QB} < L_H + 2$ is not tight. In fact, a better bound is $L_{QB} \leq L_H + 1$. Prove this bound, and provide an example where this bound is tight.

Solution: *Huffman codes:* Consider a random variable X which takes 6 values $\{A, B, C, D, E, F\}$ with probabilities $(0.5, 0.25, 0.1, 0.05, 0.05, 0.05)$ respectively.

(a) Construct a binary Huffman code for this random variable. What is its average length?

Solution:

Code	Source symbol	Prob.					
0	A	0.5	0.5	0.5	0.5	0.5	1.0
10	B	0.25	0.25	0.25	0.25	0.5	
1100	C	0.1	0.1	0.15	0.25		
1101	D	0.05	0.1	0.1			
1110	E	0.05	0.05				
1111	F	0.05					

The average length of this code is $1 \times 0.5 + 2 \times 0.25 + 4 \times (0.1 + 0.05 + 0.05 + 0.05) = 2$ bits. The entropy $H(X)$ in this case is 1.98 bits.

(b) Construct a quaternary Huffman code for this random variable, i.e., a code over an alphabet of four symbols (call them a, b, c and d). What is the average length of this code?

Solution: Since the number of symbols, i.e., 6 is not of the form $1 + k(D-1)$, we need to add a dummy symbol of probability 0 to bring it to this form. In this case, drawing up the Huffman tree is straightforward.

Code	Symbol	Prob.		
a	A	0.5	0.5	1.0
b	B	0.25	0.25	
d	C	0.1	0.15	
ca	D	0.05	0.1	
cb	E	0.05		
cc	F	0.05		
cd	G	0.0		

The average length of this code is $1 \times 0.85 + 2 \times 0.15 = 1.15$ quaternary symbols.

(c) One way to construct a binary code for the random variable is to start with a quaternary code, and convert the symbols into binary using the mapping $a \to 00$, $b \to 01$, $c \to 10$ and $d \to 11$. What is the average length of the binary code for the above random variable constructed by this process?

Solution: The code constructed by the above process is $A \to 00$, $B \to 01$, $C \to 11$, $D \to 1000$, $E \to 1001$, and $F \to 1010$, and the average length is $2 \times 0.85 + 4 \times 0.15 = 2.3$ bits.

(d) For any random variable X, let L_H be the average length of the binary Huffman code for the random variable, and let L_{QB} be the average length code constructed by firsting building a quaternary Huffman code and converting it to binary. Show that

$$L_H \leq L_{QB} < L_H + 2 \tag{5.19}$$

Solution: Since the binary code constructed from the quaternary code is also instantaneous, its average length cannot be better than the average length of the best instantaneous code, i.e., the Huffman code. That gives the lower bound of the inequality above.

To prove the upper bound, the L_Q be the length of the optimal quaternary code. Then from the results proved in the book, we have

$$H_4(X) \leq L_Q < H_4(X) + 1 \tag{5.20}$$

Also, it is easy to see that $L_{QB} = 2L_Q$, since each symbol in the quaternary code is converted into two bits. Also, from the properties of entropy, it follows that $H_4(X) = H_2(X)/2$. Substituting these in the previous equation, we get

$$H_2(X) \leq L_{QB} < H_2(X) + 2. \tag{5.21}$$

Combining this with the bound that $H_2(X) \leq L_H$, we obtain $L_{QB} < L_H + 2$.

(e) The lower bound in the previous example is tight. Give an example where the code constructed by converting an optimal quaternary code is also the optimal binary code?

Solution: Consider a random variable that takes on four equiprobable values. Then the quaternary Huffman code for this is 1 quaternary symbol for each source symbol, with average length 1 quaternary symbol. The average length L_{QB} for this code is then 2 bits. The Huffman code for this case is also easily seen to assign 2 bit codewords to each symbol, and therefore for this case, $L_H = L_{QB}$.

(f) (*Optional, no credit*) The upper bound, i.e., $L_{QB} < L_H + 2$ is not tight. In fact, a better bound is $L_{QB} \leq L_H + 1$. Prove this bound, and provide an example where this bound is tight.

Solution: Consider a binary Huffman code for the random variable X and consider all codewords of odd length. Append a 0 to each of these codewords, and we will obtain an instantaneous code where all the codewords have even length. Then we can use the inverse of the mapping mentioned in part (c) to construct a quaternary code for the random variable - it is easy to see that the quatnerary code is also instantaneous. Let L_{BQ} be the average length of this quaternary code. Since the length of the quaternary codewords of BQ are half the length of the corresponding binary codewords, we have

$$L_{BQ} = \frac{1}{2}\left(L_H + \sum_{i:l_i \text{ is odd}} p_i \right) < \frac{L_H + 1}{2} \tag{5.22}$$

and since the BQ code is at best as good as the quaternary Huffman code, we have

$$L_{BQ} \geq L_Q \tag{5.23}$$

Therefore $L_{QB} = 2L_Q \leq 2L_{BQ} < L_H + 1$.

An example where this upper bound is tight is the case when we have only two possible symbols. Then $L_H = 1$, and $L_{QB} = 2$.

17. **Data compression.** Find an optimal set of binary codeword lengths $l_1, l_2, \ldots$ (minimizing $\sum p_i l_i$) for an instantaneous code for each of the following probability mass functions:

(a) $\mathbf{p} = (\frac{10}{41}, \frac{9}{41}, \frac{8}{41}, \frac{7}{41}, \frac{7}{41})$

(b) $\mathbf{p} = (\frac{9}{10}, (\frac{9}{10})(\frac{1}{10}), (\frac{9}{10})(\frac{1}{10})^2, (\frac{9}{10})(\frac{1}{10})^3, \ldots)$

Solution: *Data compression*

(a)

Code	Source symbol	Prob.				
10	A	10/41	14/41	17/41	24/41	41/41
00	B	9/41	10/41	14/41	17/41	
01	C	8/41	9/41	10/41		
110	D	7/41	8/41			
111	E	7/41				

(b) This is case of an Huffman code on an infinite alphabet. If we consider an initial subset of the symbols, we can see that the cumulative probability of all symbols $\{x : x > i\}$ is $\sum_{j>i} 0.9 * (0.1)^{j-1} = 0.9(0.1)^{i-1}(1/(1-0.1)) = (0.1)^{i-1}$. Since this is less than $0.9 * (0.1)^{i-1}$, the cumulative sum of all the remaining terms is less than the last term used. Thus Huffman coding will always merge the last two terms. This in terms implies that the Huffman code in this case is of the form 1,01,001,0001, etc.

18. **Classes of codes.** Consider the code $\{0, 01\}$

    (a) Is it instantaneous?

    (b) Is it uniquely decodable?

    (c) Is it nonsingular?

    Solution: *Codes.*

    (a) No, the code is not instantaneous, since the first codeword, 0, is a prefix of the second codeword, 01.

    (b) Yes, the code is uniquely decodable. Given a sequence of codewords, first isolate occurrences of 01 (i.e., find all the ones) and then parse the rest into 0's.

    (c) Yes, all uniquely decodable codes are non-singular.

19. **The game of Hi-Lo.**

    (a) A computer generates a number X according to a known probability mass function $p(x), x \in \{1, 2, \ldots, 100\}$. The player asks a question, "Is $X = i$?" and is told "Yes", "You're too high," or "You're too low." He continues for a total of six questions. If he is right (i.e., he receives the answer "Yes") during this sequence, he receives a prize of value $v(X)$. How should the player proceed to maximize his expected winnings?

    (b) The above doesn't have much to do with information theory. Consider the following variation: $X \sim p(x)$, prize $= v(x)$, $p(x)$ known, as before. But *arbitrary* Yes-No questions are asked sequentially until X is determined. ("Determined" doesn't mean that a "Yes" answer is received.) Questions cost one unit each. How should the player proceed? What is the expected payoff?

    (c) Continuing (b), what if $v(x)$ is fixed, but $p(x)$ can be chosen by the computer (and then announced to the player)? The computer wishes to minimize the player's expected return. What should $p(x)$ be? What is the expected return to the player?

    Solution: *The game of Hi-Lo.*

    (a) The first thing to recognize in this problem is that the player cannot cover more than 63 values of X with 6 questions. This can be easily seen by induction. With one question, there is only one value of X that can be covered. With two questions, there is one value of X that can be covered with the first question, and depending on the answer to the first question, there are two possible values of X that can be asked in the next question. By extending this argument, we see that we can ask at more 63 different questions of the form "Is $X = i$?" with 6 questions. (The fact that we have narrowed the range at the end is irrelevant, if we have not isolated the value of X.)

    Thus if the player seeks to maximize his return, he should choose the 63 most valuable outcomes for X, and play to isolate these values. The probabilities are irrelevant to this procedure. He will choose the 63 most valuable outcomes, and his first question will be "Is $X = i$?" where i is the median of these 63 numbers. After isolating to either half, his next question will be "Is $X = j$?", where j is the median of that half. Proceeding this way, he will win if X is one of the 63 most valuable outcomes, and lose otherwise. This strategy maximizes his expected winnings.

(b) Now if arbitrary questions are allowed, the game reduces to a game of 20 questions to determine the object. The return in this case to the player is $\sum_x p(x)(v(x) - l(x))$, where $l(x)$ is the number of questions required to determine the object. Maximizing the return is equivalent to minimizing the expected number of questions, and thus, as argued in the text, the optimal strategy is to construct a Huffman code for the source and use that to construct a question strategy. His expected return is therefore between $\sum p(x)v(x) - H$ and $\sum p(x)v(x) - H - 1$.

(c) A computer wishing to minimize the return to player will want to minimize $\sum p(x)v(x) - H(X)$ over choices of $p(x)$. We can write this as a standard minimization problem with constraints. Let

$$J(p) = \sum p_i v_i + \sum p_i \log p_i + \lambda \sum p_i \tag{5.24}$$

and differentiating and setting to 0, we obtain

$$v_i + \log p_i + 1 + \lambda = 0 \tag{5.25}$$

or after normalizing to ensure that the p_i's form a probability distribution,

$$p_i = \frac{2^{-v_i}}{\sum_j 2^{-v_j}}. \tag{5.26}$$

To complete the proof, we let $r_i = \frac{2^{-v_i}}{\sum_j 2^{-v_j}}$, and rewrite the return as

$$\begin{aligned}
\sum p_i v_i + \sum p_i \log p_i &= \sum p_i \log p_i - \sum p_i \log 2^{-v_i} && (5.27)\\
&= \sum p_i \log p_i - \sum p_i \log r_i - \log(\sum 2^{-v_j}) && (5.28)\\
&= D(p||r) - \log(\sum 2^{-v_j}), && (5.29)
\end{aligned}$$

and thus the return is minimized by choosing $p_i = r_i$. This is the distribution that the computer must choose to minimize the return to the player.

20. **Huffman codes with costs.** Words like Run! Help! and Fire! are short, not because they are frequently used, but perhaps because time is precious in the situations in which these words are required. Suppose that $X = i$ with probability $p_i, i = 1, 2, \ldots, m$. Let l_i be the number of binary symbols in the codeword associated with $X = i$, and let c_i denote the cost per letter of the codeword when $X = i$. Thus the average cost C of the description of X is $C = \sum_{i=1}^m p_i c_i l_i$.

(a) Minimize C over all $l_1, l_2, \ldots, l_m$ such that $\sum 2^{-l_i} \le 1$. Ignore any implied integer constraints on l_i. Exhibit the minimizing $l_1^*, l_2^*, \ldots, l_m^*$ and the associated minimum value C^*.

(b) How would you use the Huffman code procedure to minimize C over all uniquely decodable codes? Let $C_{Huffman}$ denote this minimum.

(c) Can you show that

$$C^* \le C_{Huffman} \le C^* + \sum_{i=1}^{m} p_i c_i ?$$

Solution: *Huffman codes with costs.*

(a) We wish to minimize $C = \sum p_i c_i n_i$ subject to $\sum 2^{-n_i} \le 1$. We will assume equality in the constraint and let $r_i = 2^{-n_i}$ and let $Q = \sum_i p_i c_i$. Let $q_i = (p_i c_i)/Q$. Then $\mathbf{q}$ also forms a probability distribution and we can write C as

$$\begin{aligned} C &= \sum p_i c_i n_i && (5.30)\\ &= Q \sum q_i \log \frac{1}{r_i} && (5.31)\\ &= Q\left(\sum q_i \log \frac{q_i}{r_i} - \sum q_i \log q_i\right) && (5.32)\\ &= Q(D(\mathbf{q}||\mathbf{r}) + H(\mathbf{q})). && (5.33)\end{aligned}$$

Since the only freedom is in the choice of r_i, we can minimize C by choosing $\mathbf{r} = \mathbf{q}$ or

$$n_i^* = -\log \frac{p_i c_i}{\sum p_j c_j}, \tag{5.34}$$

where we have ignored any integer constraints on n_i. The minimum cost C^* for this assignment of codewords is

$$C^* = QH(\mathbf{q}) \tag{5.35}$$

(b) If we use $\mathbf{q}$ instead of $\mathbf{p}$ for the Huffman procedure, we obtain a code minimizing expected cost.

(c) Now we can account for the integer constraints.
Let

$$n_i = \lceil -\log q_i \rceil \tag{5.36}$$

Then

$$-\log q_i \le n_i < -\log q_i + 1 \tag{5.37}$$

Multiplying by $p_i c_i$ and summing over i, we get the relationship

$$C^* \le C_{Huffman} < C^* + Q. \tag{5.38}$$

21. **Conditions for unique decodability.** Prove that a code C is uniquely decodable if (and only if) the extension

$$C^k(x_1, x_2, \ldots, x_k) = C(x_1)C(x_2)\cdots C(x_k)$$

is a one-to-one mapping from $\mathcal{X}^k$ to D^* for every $k \geq 1$. (The only if part is obvious.)

Solution: *Conditions for unique decodability.* If C^k is not one-to-one for some k, then C is not UD, since there exist two distinct sequences, $(x_1, \ldots, x_k)$ and $(x'_1, \ldots, x'_k)$ such that

$$C^k(x_1, \ldots, x_k) = C(x_1)\cdots C(x_k) = C(x'_1)\cdots C(x'_k) = C(x'_1, \ldots, x'_k).$$

Conversely, if C is not UD then by definition there exist distinct sequences of source symbols, $(x_1, \ldots, x_i)$ and $(y_1, \ldots, y_j)$, such that

$$C(x_1)C(x_2)\cdots C(x_i) = C(y_1)C(y_2)\cdots C(y_j).$$

Concatenating the input sequences $(x_1, \ldots, x_i)$ and $(y_1, \ldots, y_j)$, we obtain

$$C(x_1)\cdots C(x_i)C(y_1)\cdots C(y_j) = C(y_1)\cdots C(y_j)C(x_1)\cdots C(x_i),$$

which shows that C^k is not one-to-one for $k = i + j$.

22. **Average length of an optimal code.** Prove that $L(p_1, \ldots, p_m)$, the average codeword length for an optimal D-ary prefix code for probabilities $\{p_1, \ldots, p_m\}$, is a continuous function of $p_1, \ldots, p_m$. This is true even though the optimal code changes discontinuously as the probabilities vary.

Solution: *Average length of an optimal code.* The longest possible codeword in an optimal code has $n-1$ binary digits. This corresponds to a completely unbalanced tree in which each codeword has a different length. Using a D-ary alphabet for codewords can only decrease its length. Since we know the maximum possible codeword length, there are only a finite number of possible codes to consider. For each candidate code $\mathcal{C}$, the average codeword length is determined by the probability distribution $p_1, p_2, \ldots, p_n$:

$$L(\mathcal{C}) = \sum_{i=1}^{n} p_i \ell_i.$$

This is a linear, and therefore continuous, function of $p_1, p_2, \ldots, p_n$. The optimal code is the candidate code with the minimum L, and its length is the minimum of a finite number of continuous functions and is therefore itself a continuous function of $p_1, p_2, \ldots, p_n$.

23. **Unused code sequences.** Let C be a variable length code that satisfies the Kraft inequality with equality but does *not* satisfy the prefix condition.

    (a) Prove that some finite sequence of code alphabet symbols is not the prefix of any sequence of codewords.

    (b) (Optional) Prove or disprove: C has infinite decoding delay.

Solution: *Unused code sequences.* Let C be a variable length code that satisfies the Kraft inequality with equality but does *not* satisfy the prefix condition.

(a) When a prefix code satisfies the Kraft inequality with equality, every (infinite) sequence of code alphabet symbols corresponds to a sequence of codewords, since the probability that a random generated sequence begins with a codeword is

$$\sum_{i=1}^{m} D^{-\ell_i} = 1\,.$$

If the code does not satisfy the prefix condition, then at least one codeword, say $C(x_1)$, is a prefix of another, say $C(x_m)$. Then the probability that a random generated sequence begins with a codeword is at most

$$\sum_{i=1}^{m-1} D^{-\ell_i} \le 1 - D^{-\ell_m} < 1\,,$$

which shows that not every sequence of code alphabet symbols is the beginning of a sequence of codewords.

(b) (Optional) A reference to a paper proving that C has infinite decoding delay will be supplied later. It is easy to see by example that the decoding delay cannot be finite. An simple example of a code that satisfies the Kraft inequality, but not the prefix condition is a suffix code (see problem 11). The simplest non-trivial suffix code is one for three symbols $\{0, 01, 11\}$. For such a code, consider decoding a string 011111. . . 1110. If the number of one's is even, then the string must be parsed 0,11,11,. . . ,11,0, whereas if the number of 1's is odd, the string must be parsed 01,11,. . . ,11. Thus the string cannot be decoded until the string of 1's has ended, and therefore the decoding delay could be infinite.

24. **Optimal codes for uniform distributions.** Consider a random variable with m equiprobable outcomes. The entropy of this information source is obviously $\log_2 m$ bits.

(a) Describe the optimal instantaneous binary code for this source and compute the average codeword length L_m.

(b) For what values of m does the average codeword length L_m equal the entropy $H = \log_2 m$?

(c) We know that $L < H + 1$ for any probability distribution. The *redundancy* of a variable length code is defined to be $\rho = L - H$. For what value(s) of m, where $2^k \le m \le 2^{k+1}$, is the redundancy of the code maximized? What is the limiting value of this worst case redundancy as $m \to \infty$?

Solution: *Optimal codes for uniform distributions.*

(a) For uniformly probable codewords, there exists an optimal binary variable length prefix code such that the longest and shortest codewords differ by at most one bit.

If two codes differ by 2 bits or more, call m_s the message with the shorter codeword C_s and m_ℓ the message with the longer codeword C_ℓ. Change the codewords for these two messages so that the new codeword C'_s is the old C_s with a zero appended ($C'_s = C_s0$) and C'_ℓ is the old C_s with a one appended ($C'_\ell = C_s1$). C'_s and C'_ℓ are legitimate codewords since no other codeword contained C_s as a prefix (by definition of a prefix code), so obviously no other codeword could contain C'_s or C'_ℓ as a prefix. The length of the codeword for m_s increases by 1 and the length of the codeword for m_ℓ decreases by at least 1. Since these messages are equally likely, $L' \le L$. By this method we can transform any optimal code into a code in which the length of the shortest and longest codewords differ by at most one bit. (In fact, it is easy to see that every optimal code has this property.)

For a source with n messages, $\ell(m_s) = \lfloor \log_2 n \rfloor$ and $\ell(m_\ell) = \lceil \log_2 n \rceil$. Let d be the difference between n and the next smaller power of 2:

$$d = n - 2^{\lfloor \log_2 n \rfloor}.$$

Then the optimal code has $2d$ codewords of length $\lceil \log_2 n \rceil$ and $n{-}2d$ codewords of length $\lfloor \log_2 n \rfloor$. This gives

$$\begin{aligned} L &= \frac{1}{n}\left(2d\lceil \log_2 n \rceil + (n-2d)\lfloor \log_2 n \rfloor\right) \\ &= \frac{1}{n}\left(n\lfloor \log_2 n \rfloor + 2d\right) \\ &= \lfloor \log_2 n \rfloor + \frac{2d}{n}. \end{aligned}$$

Note that $d = 0$ is a special case in the above equation.

(b) The average codeword length equals the entropy if and only if n is a power of 2. To see this, consider the following calculation of L:

$$L = \sum_i p_i \ell_i = -\sum_i p_i \log_2 2^{-\ell_i} = H + D(p\|q),$$

where $q_i = 2^{-\ell_i}$. Therefore $L = H$ only if $p_i = q_i$, that is, when all codewords have equal length, or when $d = 0$.

(c) For $n = 2^m + d$, the redundancy $r = L - H$ is given by

$$\begin{aligned} r &= L - \log_2 n \\ &= \lfloor \log_2 n \rfloor + \frac{2d}{n} - \log_2 n \\ &= m + \frac{2d}{n} - \log_2(2^m + d) \\ &= m + \frac{2d}{2^m + d} - \frac{\ln(2^m + d)}{\ln 2}. \end{aligned}$$

Therefore

$$\frac{\partial r}{\partial d} = \frac{(2^m + d)(2) - 2d}{(2^m + d)^2} - \frac{1}{\ln 2} \cdot \frac{1}{2^m + d}$$

Setting this equal to zero implies $d^* = 2^m(2\ln 2 - 1)$. Since there is only one maximum, and since the function is convex $\cap$, the maximizing d is one of the two integers nearest $(.3862)(2^m)$. The corresponding maximum redundancy is

$$\begin{aligned} r^* &\approx m + \frac{2d^*}{2^m + d^*} - \frac{\ln(2^m + d^*)}{\ln 2} \\ &= m + \frac{2(.3862)(2^m)}{2^m + (.3862)(2^m)} - \frac{\ln(2^m + (.3862)2^m)}{\ln 2} \\ &= .0861\,. \end{aligned}$$

This is achieved with arbitrary accuracy as $n \to \infty$. (The quantity $\sigma = 0.0861$ is one of the lesser fundamental constants of the universe. See Robert Gallager[7]).

25. **Optimal codeword lengths.** Although the codeword lengths of an optimal variable length code are complicated functions of the message probabilities $\{p_1, p_2, \ldots, p_m\}$, it can be said that less probable symbols are encoded into longer codewords. Suppose that the message probabilities are given in decreasing order $p_1 > p_2 \geq \cdots \geq p_m$.

(a) Prove that for any binary Huffman code, if the most probable message symbol has probability $p_1 > 2/5$, then that symbol must be assigned a codeword of length 1.

(b) Prove that for any binary Huffman code, if the most probable message symbol has probability $p_1 < 1/3$, then that symbol must be assigned a codeword of length ≥ 2.

Solution: *Optimal codeword lengths.* Let $\{c_1, c_2, \ldots, c_m\}$ be codewords of respective lengths $\{\ell_1, \ell_2, \ldots, \ell_m\}$ corresponding to probabilities $\{p_1, p_2, \ldots, p_m\}$.

(a) We prove that if $p_1 > p_2$ and $p_1 > 2/5$ then $\ell_1 = 1$. Suppose, for the sake of contradiction, that $\ell_1 \geq 2$. Then there are no codewords of length 1; otherwise c_1 would not be the shortest codeword. Without loss of generality, we can assume that c_1 begins with 00. For $x, y \in \{0, 1\}$ let C_{xy} denote the set of codewords beginning with xy. Then the sets C_{01}, C_{10}, and C_{11} have total probability $1 - p_1 < 3/5$, so some two of these sets (without loss of generality, C_{10} and C_{11}) have total probability less 2/5. We can now obtain a better code by interchanging the subtree of the decoding tree beginning with 1 with the subtree beginning with 00; that is, we replace codewords of the form $1x\ldots$ by $00x\ldots$ and codewords of the form $00y\ldots$ by $1y\ldots$. This improvement contradicts the assumption that $\ell_1 \geq 2$, and so $\ell_1 = 1$. (Note that $p_1 > p_2$ was a hidden assumption for this problem; otherwise, for example, the probabilities $\{.49, .49, .02\}$ have the optimal code $\{00, 1, 01\}$.)

(b) The argument is similar to that of part (a). Suppose, for the sake of contradiction, that $\ell_1 = 1$. Without loss of generality, assume that $c_1 = 0$. The total probability of C_{10} and C_{11} is $1 - p_1 > 2/3$, so at least one of these two sets (without loss of generality, C_{10}) has probability greater than 2/3. We can now obtain a better code by interchanging the subtree of the decoding tree beginning with 0 with the subtree beginning with 10; that is, we replace codewords of the form $10x\ldots$ by $0x\ldots$ and we let $c_1 = 10$. This improvement contradicts the assumption that $\ell_1 = 1$, and so $\ell_1 \geq 2$.

26. **Merges**. Companies with values $W_1, W_2, \ldots, W_m$ are merged as follows. The two least valuable companies are merged, thus forming a list of $m-1$ companies. The *value of the merge* is the sum of the values of the two merged companies. This continues until one supercompany remains. Let V equal the sum of the values of the merges. Thus V represents the total reported dollar volume of the merges. For example, if $\mathbf{W} = (3,3,2,2)$, the merges yield $(3,3,2,2) \rightarrow (4,3,3) \rightarrow (6,4) \rightarrow (10)$, and $V = 4+6+10 = 20$.

    (a) Argue that V is the minimum volume achievable by sequences of pair-wise merges terminating in one supercompany. (*Hint:* Compare to Huffman coding.)

    (b) Let $W = \sum W_i$, $\quad \tilde{W}_i = W_i/W$, and show that the minimum merge volume V satisfies

$$W H(\tilde{\mathbf{W}}) \leq V \leq W H(\tilde{\mathbf{W}}) + W \tag{5.39}$$

Solution: *Problem: Merges*

(a) We first normalize the values of the companies to add to one. The total volume of the merges is equal to the sum of value of each company times the number of times it takes part in a merge. This is identical to the average length of a Huffman code, with a tree which corresponds to the merges. Since Huffman coding minimizes average length, this scheme of merges minimizes total merge volume.

(b) Just as in the case of Huffman coding, we have

$$H \leq EL < H+1, \tag{5.40}$$

we have in this case for the corresponding merge scheme

$$W H(\tilde{\mathbf{W}}) \leq V \leq W H(\tilde{\mathbf{W}}) + W \tag{5.41}$$

27. **The Sardinas-Patterson test for unique decodability.** A code is not uniquely decodable if and only if there exists a finite sequence of code symbols which can be resolved in two different ways into sequences of codewords. That is, a situation such as

$$\begin{array}{l} | \qquad A_1 \qquad | \qquad\quad A_2 \qquad\quad | \qquad\quad A_3 \qquad \ldots \qquad\quad A_m \qquad | \\ \hline | \;\; B_1 \;\; | \quad B_2 \quad | \quad B_3 \qquad\qquad\quad \ldots \qquad\qquad B_n \qquad\qquad | \end{array}$$

must occur where each A_i and each B_i is a codeword. Note that B_1 must be a prefix of A_1 with some resulting "dangling suffix." Each dangling suffix must in turn be either a prefix of a codeword or have another codeword as its prefix, resulting in another dangling suffix. Finally, the last dangling suffix in the sequence must also be a codeword. Thus one can set up a test for unique decodability (which is essentially the Sardinas-Patterson test[11]) in the following way: Construct a set S of all possible dangling suffixes. The code is uniquely decodable if and only if S contains no codeword.

(a) State the precise rules for building the set S.

(b) Suppose the codeword lengths are l_i, $i = 1, 2, \ldots, m$. Find a good upper bound on the number of elements in the set S.

(c) Determine which of the following codes is uniquely decodable:

i. $\{0, 10, 11\}$.

ii. $\{0, 01, 11\}$.

iii. $\{0, 01, 10\}$.

iv. $\{0, 01\}$.

v. $\{00, 01, 10, 11\}$.

vi. $\{110, 11, 10\}$.

vii. $\{110, 11, 100, 00, 10\}$.

(d) For each uniquely decodable code in part (c), construct, if possible, an infinite encoded sequence with a known starting point, such that it can be resolved into codewords in two different ways. (This illustrates that unique decodability does not imply finite decodability.) Prove that such a sequence cannot arise in a prefix code.

Solution: *Test for unique decodability.*

The proof of the Sardinas-Patterson test has two parts. In the first part, we will show that if there is a code string that has two different interpretations, then the code will fail the test. The simplest case is when the concatenation of two codewords yields another codeword. In this case, S_2 will contain a codeword, and hence the test will fail.

In general, the code is not uniquely decodeable, iff there exists a string that admits two different parsings into codewords, e.g.

$$x_1x_2x_3x_4x_5x_6x_7x_8 = x_1x_2, x_3x_4x_5, x_6x_7x_8 = x_1x_2x_3x_4, x_5x_6x_7x_8. \tag{5.42}$$

In this case, S_2 will contain the string x_3x_4, S_3 will contain x_5, S_4 will contain $x_6x_7x_8$, which is a codeword. It is easy to see that this procedure will work for any string that has two different parsings into codewords; a formal proof is slightly more difficult and using induction.

In the second part, we will show that if there is a codeword in one of the sets $S_i, i \geq 2$, then there exists a string with two different possible interpretations, thus showing that the code is not uniquely decodeable. To do this, we essentially reverse the construction of the sets. We will not go into the details - the reader is referred to the original paper.

(a) Let S_1 be the original set of codewords. We construct S_{i+1} from S_i as follows: A string y is in S_{i+1} iff there is a codeword x in S_1, such that xy is in S_i or if there exists a $z \in S_i$ such that zy is in S_1 (i.e., is a codeword). Then the code is uniquely decodable iff none of the S_i, $i \geq 2$ contains a codeword. Thus the set $S = \cup_{i \geq 2} S_i$.

(b) A simple upper bound can be obtained from the fact that all strings in the sets S_i have length less than l_{max}, and therefore the maximum number of elements in S is less than $2^{l_{max}}$.

(c) i. $\{0, 10, 11\}$. This code is instantaneous and hence uniquely decodable.

ii. $\{0, 01, 11\}$. This code is a suffix code (see problem 11). It is therefore uniquely decodable. The sets in the Sardinas-Patterson test are $S_1 = \{0, 01, 11\}$, $S_2 = \{1\} = S_3 = S_4 = \ldots$.

iii. $\{0, 01, 10\}$. This code is not uniquely decodable. The sets in the test are $S_1 = \{0, 01, 10\}$, $S_2 = \{1\}$, $S_3 = \{0\}$, ... Since 0 is codeword, this code fails the test. It is easy to see otherwise that the code is not UD - the string 010 has two valid parsings.

iv. $\{0, 01\}$. This code is a suffix code and is therefore UD. THe test produces sets $S_1 = \{0, 01\}$, $S_2 = \{1\}$, $S_3 = \phi$.

v. $\{00, 01, 10, 11\}$. This code is instantaneous and therefore UD.

vi. $\{110, 11, 10\}$. This code is uniquely decodable, by the Sardinas-Patterson test, since $S_1 = \{110, 11, 10\}$, $S_2 = \{0\}$, $S_3 = \phi$.

vii. $\{110, 11, 100, 00, 10\}$. This code is UD, because by the Sardinas Patterson test, $S_1 = \{110, 11, 100, 00, 10\}$, $S_2 = \{0\}$, $S_3 = \{0\}$, etc.

(d) We can produce infinite strings which can be decoded in two ways only for examples where the Sardinas Patterson test produces a repeating set. For example, in part (ii), the string 011111... could be parsed either as 0,11,11,... or as 01,11,11,.... Similarly for (viii), the string 10000... could be parsed as 100,00,00,... or as 10,00,00,.... For the instantaneous codes, it is not possible to construct such a string, since we can decode as soon as we see a codeword string, and there is no way that we would need to wait to decode.

28. **Shannon code.** Consider the following method for generating a code for a random variable X which takes on m values $\{1, 2, \ldots, m\}$ with probabilities $p_1, p_2, \ldots, p_m$. Assume that the probabilities are ordered so that $p_1 \geq p_2 \geq \cdots \geq p_m$. Define

$$F_i = \sum_{k=1}^{i-1} p_k, \tag{5.43}$$

the sum of the probabilities of all symbols less than i. Then the codeword for i is the number $F_i \in [0, 1]$ rounded off to l_i bits, where $l_i = \lceil \log \frac{1}{p_i} \rceil$.

(a) Show that the code constructed by this process is prefix-free and the average length satisfies

$$H(X) \leq L < H(X) + 1. \tag{5.44}$$

(b) Construct the code for the probability distribution $(0.5, 0.25, 0.125, 0.125)$.

Solution: *Shannon code.*

(a) Since $l_i = \lceil \log \frac{1}{p_i} \rceil$, we have

$$\log \frac{1}{p_i} \le l_i < \log \frac{1}{p_i} + 1 \tag{5.45}$$

which implies that

$$H(X) \le L = \sum p_i l_i < H(X) + 1. \tag{5.46}$$

The difficult part is to prove that the code is a prefix code. By the choice of l_i, we have

$$2^{-l_i} \le p_i < 2^{-(l_i - 1)}. \tag{5.47}$$

Thus F_j, $j > i$ differs from F_i by at least 2^{-l_i}, and will therefore differ from F_i is at least one place in the first l_i bits of the binary expansion of F_i. Thus the codeword for F_j, $j > i$, which has length $l_j \ge l_i$, differs from the codeword for F_i at least once in the first l_i places. Thus no codeword is a prefix of any other codeword.

(b) We build the following table

Symbol	Probability	F_i in decimal	F_i in binary	l_i	Codeword
1	0.5	0.0	0.0	1	0
2	0.25	0.5	0.10	2	10
3	0.125	0.75	0.110	3	110
4	0.125	0.875	0.111	3	111

The Shannon code in this case achieves the entropy bound (1.75 bits) and is optimal.

29. **Optimal codes for dyadic distributions.** For a Huffman code tree, define the probability of a node as the sum of the probabilities of all the leaves under that node. Let the random variable X be drawn from a dyadic distribution, i.e., $p(x) = 2^{-i}$, for some i, for all $x \in \mathcal{X}$. Now consider a binary Huffman code for this distribution.

(a) Argue that for any node in the tree, the probability of the left child is equal to the probability of the right child.

(b) Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. $\sim p(x)$. Using the Huffman code for $p(x)$, we map $X_1, X_2, \ldots, X_n$ to a sequence of bits $Y_1, Y_2, \ldots, Y_{k(X_1, X_2, \ldots, X_n)}$. (The length of this sequence will depend on the outcome $X_1, X_2, \ldots, X_n$.) Use part (a) to argue that the sequence $Y_1, Y_2, \ldots,$ forms a sequence of fair coin flips, i.e., that $\Pr\{Y_i = 0\} = \Pr\{Y_i = 1\} = \frac{1}{2}$, independent of $Y_1, Y_2, \ldots, Y_{i-1}$.

Thus the entropy rate of the coded sequence is 1 bit per symbol.

(c) Give a heuristic argument why the encoded sequence of bits for any code that achieves the entropy bound cannot be compressible and therefore should have an entropy rate of 1 bit per symbol.

Solution: *Optimal codes for dyadic distributions.*

(a) For a dyadic distribution, the Huffman code acheives the entropy bound. The code tree constructed be the Huffman algorithm is a complete tree with leaves at depth l_i with probability $p_i = 2^{-l_i}$.

For such a complete binary tree, we can prove the following properties

- The probability of any internal node at depth k is 2^{-k}.

  We can prove this by induction. Clearly, it is true for a tree with 2 leaves. Assume that it is true for all trees with n leaves. For any tree with $n+1$ leaves, at least two of the leaves have to be siblings on the tree (else the tree would not be complete). Let the level of these siblings be j. The probability of the parent of these two siblings (at level $j-1$) has probability $2^j + 2^j = 2^{j-1}$. We can now replace the two siblings with their parent, without changing the probability of any other internal node. But now we have a tree with n leaves which satisfies the required property. Thus, by induction, the property is true for all complete binary trees.
- From the above property, it follows immediately the the probability of the left child is equal to the probability of the right child.

(b) For a sequence X_1, X_2, we can construct a code tree by first constructing the optimal tree for X_1, and then attaching the optimal tree for X_2 to each leaf of the optimal tree for X_1. Proceeding this way, we can construct the code tree for $X_1, X_2, \ldots, X_n$. When X_i are drawn i.i.d. according to a dyadic distribution, it is easy to see that the code tree constructed will be also be a complete binary tree with the properties in part (a). Thus the probability of the first bit being 1 is 1/2, and at any internal node, the probability of the next bit produced by the code being 1 is equal to the probability of the next bit being 0. Thus the bits produced by the code are i.i.d. Bernoulli(1/2), and the entropy rate of the coded sequence is 1 bit per symbol.

(c) Assume that we have a coded sequence of bits from a code that met the entropy bound with equality. If the coded sequence were compressible, then we could used the compressed version of the coded sequence as our code, and achieve an average length less than the entropy bound, which will contradict the bound. Thus the coded sequence cannot be compressible, and thus must have an entropy rate of 1 bit/symbol.

30. **Relative entropy is cost of miscoding:** Let the random variable X have five possible outcomes $\{1, 2, 3, 4, 5\}$. Consider two distributions $p(x)$ and $q(x)$ on this random variable

Symbol	$p(x)$	$q(x)$	$C_1(x)$	$C_2(x)$
1	1/2	1/2	0	0
2	1/4	1/8	10	100
3	1/8	1/8	110	101
4	1/16	1/8	1110	110
5	1/16	1/8	1111	111

(a) Calculate $H(p)$, $H(q)$, $D(p||q)$ and $D(q||p)$.

(b) The last two columns above represent codes for the random variable. Verify that the average length of C_1 under p is equal to the entropy $H(p)$. Thus C_1 is optimal for p. Verify that C_2 is optimal for q.

(c) Now assume that we use code C_2 when the distribution is p. What is the average length of the codewords. By how much does it exceed the entropy p?

(d) What is the loss if we use code C_1 when the distribution is q?

Solution: *Cost of miscoding*

(a) $H(p) = \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{8}\log 8 + \frac{1}{16}\log 16 + \frac{1}{16}\log 16 = 1.875$ bits.
$H(q) = \frac{1}{2}\log 2 + \frac{1}{8}\log 8 + \frac{1}{8}\log 8 + \frac{1}{8}\log 8 + \frac{1}{8}\log 8 = 2$ bits.
$D(p||q) = \frac{1}{2}\log\frac{1/2}{1/2} + \frac{1}{4}\log\frac{1/4}{1/8} + \frac{1}{8}\log\frac{1/8}{1/8} + \frac{1}{16}\log\frac{1/16}{1/8} + \frac{1}{16}\log\frac{1/16}{1/8} = 0.125$ bits.
$D(p||q) = \frac{1}{2}\log\frac{1/2}{1/2} + \frac{1}{8}\log\frac{1/8}{1/4} + \frac{1}{8}\log\frac{1/8}{1/8} + \frac{1}{8}\log\frac{1/8}{1/16} + \frac{1}{8}\log\frac{1/8}{1/16} = 0.125$ bits.

(b) The average length of C_1 for $p(x)$ is 1.875 bits, which is the entropy of p. Thus C_1 is an efficient code for $p(x)$. Similarly, the average length of code C_2 under $q(x)$ is 2 bits, which is the entropy of q. Thus C_2 is an efficient code for q.

(c) If we use code C_2 for $p(x)$, then the average length is $\frac{1}{2} * 1 + \frac{1}{4} * 3 + \frac{1}{8} * 3 + \frac{1}{16} * 3 + \frac{1}{16} * 3 = 2$ bits. It exceeds the entropy by 0.125 bits, which is the same as $D(p||q)$.

(d) Similary, using code C_1 for q has an average length of 2.125 bits, which exceeds the entropy of q by 0.125 bits, which is $D(q||p)$.

31. **Non-singular codes:** The discussion in the text focused on instantaneous codes, with extensions to uniquely decodable codes. Both these are required in cases when the code is to be used repeatedly to encode a sequence of outcomes of a random variable. But if we need to encode only one outcome and we know when we have reached the end of a codeword, we do not need unique decodability - only the fact that the code is non-singular would suffice. For example, if a random variable X takes on 3 values a, b and c, we could encode them by 0, 1, and 00. Such a code is non-singular but not uniquely decodable.

In the following, assume that we have a random variable X which takes on m values with probabilities $p_1, p_2, \ldots, p_m$ and that the probabilities are ordered so that $p_1 \geq p_2 \geq \ldots \geq p_m$.

(a) By viewing the non-singular binary code as a ternary code with three symbols, 0,1 and "STOP", show that the expected length of a non-singular code $L_{1:1}$ for a random variable X satisfies the following inequality:

$$L_{1:1} \geq \frac{H_2(X)}{\log_2 3} - 1 \tag{5.48}$$

where $H_2(X)$ is the entropy of X in bits. Thus the average length of a non-singular code is at least a constant fraction of the average length of an instantaneous code.

(b) Let L_{INST} be the expected length of the best instantaneous code and $L^*_{1:1}$ be the expected length of the best non-singular code for X. Argue that $L^*_{1:1} \leq L^*_{INST} \leq H(X) + 1$.

(c) Give a simple example where the average length of the non-singular code is less than the entropy.

(d) The set of codewords available for an non-singular code is $\{0, 1, 00, 01, 10, 11, 000, \ldots\}$. Since $L_{1:1} = \sum_{i=1}^m p_i l_i$, show that this is minimized if we allot the shortest codewords to the most probable symbols.

Thus $l_1 = l_2 = 1$, $l_3 = l_4 = l_5 = l_6 = 2$, etc. Show that in general $l_i = \lceil \log \left(\frac{i}{2} + 1\right) \rceil$, and therefore $L^*_{1:1} = \sum_{i=1}^m p_i \lceil \log \left(\frac{i}{2} + 1\right) \rceil$.

(e) The previous part shows that it is easy to find the optimal non-singular code for a distribution. However, it is a little more tricky to deal with the average length of this code. We now bound this average length. It follows from the previous part that $L^*_{1:1} \geq \tilde{L} \overset{\triangle}{=} \sum_{i=1}^m p_i \log \left(\frac{i}{2} + 1\right)$. Consider the difference

$$F(\mathbf{p}) = H(X) - \tilde{L} = -\sum_{i=1}^m p_i \log p_i - \sum_{i=1}^m p_i \log \left(\frac{i}{2} + 1\right). \tag{5.49}$$

Prove by the method of Lagrange multipliers that the maximum of $F(\mathbf{p})$ occurs when $p_i = c/(i+2)$, where $c = 1/(H_{m+2} - H_2)$ and H_k is the sum of the harmonic series, i.e.,

$$H_k \overset{\triangle}{=} \sum_{i=1}^k \frac{1}{i} \tag{5.50}$$

(This can also be done using the non-negativity of relative entropy.)

(f) Complete the arguments for

$$\begin{aligned} H(X) - L^*_{1:1} &\leq H(X) - \tilde{L} & (5.51)\\ &\leq \log(2(H_{m+2} - H_2)) & (5.52) \end{aligned}$$

Now it is well known (see, e.g. Knuth, "Art of Computer Programming", Vol. 1) that $H_k \approx \ln k$ (more precisely, $H_k = \ln k + \gamma + \frac{1}{2k} - \frac{1}{12k^2} + \frac{1}{120k^4} - \epsilon$ where $0 < \epsilon < 1/252n^6$, and $\gamma =$ Euler's constant $= 0.577\ldots$). Either using this or a simple approximation that $H_k \leq \ln k + 1$, which can be proved by integration of $\frac{1}{x}$, it can be shown that $H(X) - L^*_{1:1} < \log \log m + 2$. Thus we have

$$H(X) - \log \log |\mathcal{X}| - 2 \leq L^*_{1:1} \leq H(X) + 1. \tag{5.53}$$

A non-singular code cannot do much better than an instantaneous code!

Solution:

(a) In the text, it is proved that the average length of any prefix-free code in a D-ary alphabet was greater than $H_D(X)$, the D-ary entropy. Now if we start with any binary non-singular code and add the additional symbol "STOP" at the end, the new code is prefix-free in the alphabet of 0,1, and "STOP" (since "STOP" occurs only at the end of codewords, and every codeword has a "STOP" symbol, so the only way a code word can be a prefix of another is if they were equal). Thus each code word in the new alphabet is one symbol longer than the binary codewords, and the average length is 1 symbol longer.

Thus we have $L_{1:1}+1 \geq H_3(X)$, or $L_{1:1} \geq \frac{H_2(X)}{\log 3} - 1 = 0.63H(X)-1$.

(b) Since an instantaneous code is also a non-singular code, the best non-singular code is at least as good as the best instantaneous code. Since the best instantaneous code has average length $\leq H(X)+1$, we have $L^*_{1:1} \leq L^*_{INST} \leq H(X)+1$.

(c) For a 2 symbol alphabet, the best non-singular code and the best instantaneous code are the same. So the simplest example where they differ is when $|\mathcal{X}|=3$. In this case, the simplest (and it turns out, optimal) non-singular code has three codewords $0, 1, 00$. Assume that each of the symbols is equally likely. Then $H(X) = \log 3 = 1.58$ bits, whereas the average length of the non-singular code is $\frac{1}{3}.1 + \frac{1}{3}.1 + \frac{1}{3}.2 = 4/3 = 1.3333 < H(X)$. Thus a non-singular code could do better than entropy.

(d) For a given set of codeword lengths, the fact that allotting the shortest codewords to the most probable symbols is proved in Lemma 5.8.1, part 1 of EIT.

This result is a general version of what is called the Hardy-Littlewood-Polya inequality, which says that if $a<b$, $c<d$, then $ad+bc<ac+bd$. The general version of the Hardy-Littlewood-Polya inequality states that if we were given two sets of numbers $A=\{a_j\}$ and $B=\{b_j\}$ each of size m, and let $a_{[i]}$ be the i-th largest element of A and $b_{[i]}$ be the i-th largest element of set B. Then

$$\sum_{i=1}^{m} a_{[i]}b_{[m+1-i]} \leq \sum_{i=1}^{m} a_i b_i \leq \sum_{i=1}^{m} a_{[i]}b_{[i]} \tag{5.54}$$

An intuitive explanation of this inequality is that you can consider the a_i's to the position of hooks along a rod, and b_i's to be weights to be attached to the hooks. To maximize the moment about one end, you should attach the largest weights to the furthest hooks.

The set of available codewords is the set of all possible sequences. Since the only restriction is that the code be non-singular, each source symbol could be alloted to any codeword in the set $\{0, 1, 00, \ldots\}$.

Thus we should allot the codewords 0 and 1 to the two most probable source symbols, i.e., to probablities p_1 and p_2. Thus $l_1 = l_2 = 1$. Similarly, $l_3 = l_4 = l_5 = l_6 = 2$ (corresponding to the codewords 00, 01, 10 and 11). The next 8 symbols will use codewords of length 3, etc.

We will now find the general form for l_i. We can prove it by induction, but we will derive the result from first principles. Let $c_k = \sum_{j=1}^{k-1} 2^j$. Then by the arguments of the previous paragraph, all source symbols of index $c_k+1, c_k+2, \ldots, c_k+2^k = c_{k+1}$

use codewords of length k. Now by using the formula for the sum of the geometric series, it is easy to see that

$$c_k = \sum j = 1^{k-1} 2^j = 2 \sum j = 0^{k-2} 2^j = 2 \frac{2^{k-1}-1}{2-1} = 2^k - 2 \tag{5.55}$$

Thus all sources with index i, where $2^k - 1 \le i \le 2^k - 2 + 2^k = 2^{k+1} - 2$ use codewords of length k. This corresponds to $2^k < i+2 \le 2^{k+1}$ or $k < \log(i+2) \le k+1$ or $k - 1 < \log \frac{i+2}{2} \le k$. Thus the length of the codeword for the i-th symbol is $k = \lceil \log \frac{i+2}{2} \rceil$. Thus the best non-singular code assigns codeword length $l_i^* = \lceil \log(i/2+1) \rceil$ to symbol i, and therefore $L_{1:1}^* = \sum_{i=1}^m p_i \lceil \log(i/2+1) \rceil$.

(e) Since $\lceil \log(i/2+1) \rceil \ge \log(i/2+1)$, it follows that $L_{1:1}^* \ge \tilde{L} \stackrel{\triangle}{=} \sum_{i=1}^m p_i \log\left(\frac{i}{2}+1\right)$. Consider the difference

$$F(\mathbf{p}) = H(X) - \tilde{L} = -\sum_{i=1}^m p_i \log p_i - \sum_{i=1}^m p_i \log\left(\frac{i}{2}+1\right). \tag{5.56}$$

We want to maximize this function over all probability distributions, and therefore we use the method of Lagrange multipliers with the constraint $\sum p_i = 1$. Therefore let

$$J(\mathbf{p}) = -\sum_{i=1}^m p_i \log p_i - \sum_{i=1}^m p_i \log\left(\frac{i}{2}+1\right) + \lambda(\sum_{i=1}^m p_i - 1) \tag{5.57}$$

Then differentiating with respect to p_i and setting to 0, we get

$$\frac{\partial J}{\partial p_i} = -1 - \log p_i - \log\left(\frac{i}{2}+1\right) + \lambda = 0 \tag{5.58}$$

$$\log p_i = \lambda - 1 - \log \frac{i+2}{2} \tag{5.59}$$

$$p_i = 2^{\lambda-1} \frac{2}{i+2} \tag{5.60}$$

Now substituting this in the constraint that $\sum p_i = 1$, we get

$$2^\lambda \sum_{i=1}^m \frac{1}{i+2} = 1 \tag{5.61}$$

or $2^\lambda = 1/(\sum_i \frac{1}{i+2})$. Now using the definition $H_k = \sum_{j=1}^k \frac{1}{j}$, it is obvious that

$$\sum_{i=1}^m \frac{1}{i+2} = \sum_{i=1}^{m+2} \frac{1}{i} - 1 - \frac{1}{2} = H_{m+2} - H_2. \tag{5.62}$$

Thus $2^\lambda = \frac{1}{H_{m+2} - H_2}$, and

$$p_i = \frac{1}{H_{m+2} - H_2} \frac{1}{i+2} \tag{5.63}$$

Substituting this value of p_i in the expression for $F(\mathbf{p})$, we obtain

$$\begin{aligned} F(\mathbf{p}) &= -\sum_{i=1}^{m} p_i \log p_i - \sum_{i=1}^{m} p_i \log\left(\frac{i}{2}+1\right) &(5.64)\\ &= -\sum_{i=1}^{m} p_i \log p_i \frac{i+2}{2} &(5.65)\\ &= -\sum_{i=1}^{m} p_i \log \frac{1}{2(H_{m+2}-H_2)} &(5.66)\\ &= \log 2(H_{m+2}-H_2) &(5.67)\end{aligned}$$

Thus the extremal value of $F(\mathbf{p})$ is $\log 2(H_{m+2}-H_2)$. We have not showed that it is a maximum - that can be shown be taking the second derivative. But as usual, it is easier to see it using relative entropy. Looking at the expressions above, we can see that if we define $q_i = \frac{1}{H_{m+2}-H_2}\frac{1}{i+2}$, then q_i is a probability distribution (i.e., $q_i \geq 0$, $\sum q_i = 1$). Also, $\frac{i+2}{2 = \frac{1}{2(H_{m+2}-H_2)}\frac{1}{q_i}}$, and substuting this in the expression for $F(\mathbf{p})$, we obtain

$$\begin{aligned} F(\mathbf{p}) &= -\sum_{i=1}^{m} p_i \log p_i - \sum_{i=1}^{m} p_i \log\left(\frac{i}{2}+1\right) &(5.68)\\ &= -\sum_{i=1}^{m} p_i \log p_i \frac{i+2}{2} &(5.69)\\ &= -\sum_{i=1}^{m} p_i \log p_i \frac{1}{2(H_{m+2}-H_2)}\frac{1}{q_i} &(5.70)\\ &= -\sum_{i=1}^{m} p_i \log \frac{p_i}{q_i} - \sum_{i=1}^{m} p_i \log \frac{1}{2(H_{m+2}-H_2)} &(5.71)\\ &= \log 2(H_{m+2}-H_2) - D(p\|q) &(5.72)\\ &\leq \log 2(H_{m+2}-H_2) &(5.73)\end{aligned}$$

with equality iff $p = q$. Thus the maximum value of $F(\mathbf{p})$ is $\log 2(H_{m+2}-H_2)$

(f)

$$\begin{aligned} H(X) - L^*_{1:1} &\leq H(X) - \tilde{L} &(5.74)\\ &\leq \log 2(H_{m+2}-H_2) &(5.75)\end{aligned}$$

The first inequality follows from the definition of $\tilde{L}$ and the second from the result of the previous part.

To complete the proof, we will use the simple inequality $H_k \leq \ln k + 1$, which can be shown by integrating $\frac{1}{x}$ between 1 and k. Thus $H_{m+2} \leq \ln(m+2)+1$, and $2(H_{m+2}-H_2) = 2(H_{m+2}-1-\frac{1}{2}) \leq 2(\ln(m+2)+1-1-\frac{1}{2}) \leq 2(\ln(m+2)) = 2\log(m+2)/\log e \leq 2\log(m+2) \leq 2\log m^2 = 4\log m$ where the last inequality is true for $m \geq 2$. Therefore

$$H(X) - L_{1:1} \leq \log 2(H_{m+2}-H_2) \leq \log(4\log m) = \log\log m + 2 \qquad (5.76)$$

We therefore have the following bounds on the average length of a non-singular code

$$H(X) - \log\log|\mathcal{X}| - 2 \le L^*_{1:1} \le H(X) + 1 \tag{5.77}$$

A non-singular code cannot do much better than an instantaneous code!

32. **Bad wine.** One is given 6 bottles of wine. It is known that precisely one bottle has gone bad (tastes terrible). From inspection of the bottles it is determined that the probability p_i that the i^{th} bottle is bad is given by $(p_1, p_2, \ldots, p_6) = (\frac{8}{23}, \frac{6}{23}, \frac{4}{23}, \frac{2}{23}, \frac{2}{23}, \frac{1}{23})$. Tasting will determine the bad wine.

Suppose you taste the wines one at a time. Choose the order of tasting to minimize the expected number of tastings required to determine the bad bottle. Remember, if the first 5 wines pass the test you don't have to taste the last.

(a) What is the expected number of tastings required?

(b) Which bottle should be tasted first?

Now you get smart. For the first sample, you mix some of the wines in a fresh glass and sample the mixture. You proceed, mixing and tasting, stopping when the bad bottle has been determined.

(c) What is the minimum expected number of tastings required to determine the bad wine?

(d) What mixture should be tasted first?

Solution: *Bad Wine*

(a) If we taste one bottle at a time, to minimize the expected number of tastings the order of tasting should be from the most likely wine to be bad to the least. The expected number of tastings required is

$$\begin{aligned}\sum_{i=1}^{6} p_i l_i &= 1\times\frac{8}{23} + 2\times\frac{6}{23} + 3\times\frac{4}{23} + 4\times\frac{2}{23} + 5\times\frac{2}{23} + 5\times\frac{1}{23} \\ &= \frac{55}{23} \\ &= 2.39\end{aligned}$$

(b) The first bottle to be tasted should be the one with probability $\frac{8}{23}$.

(c) The idea is to use Huffman coding. With Huffman coding, we get codeword lengths as $(2,2,2,3,4,4)$. The expected number of tastings required is

$$\begin{aligned}\sum_{i=1}^{6} p_i l_i &= 2\times\frac{8}{23} + 2\times\frac{6}{23} + 2\times\frac{4}{23} + 3\times\frac{2}{23} + 4\times\frac{2}{23} + 4\times\frac{1}{23} \\ &= \frac{54}{23} \\ &= 2.35\end{aligned}$$

(d) The mixture of the first and second bottles should be tasted first.

33. **Huffman vs. Shannon.** A random variable X takes on three values with probabilities 0.6, 0.3, and 0.1.

(a) What are the lengths of the binary Huffman codewords for X? What are the lengths of the binary Shannon codewords $(l(x) = \lceil \log(\frac{1}{p(x)}) \rceil)$ for X?

(b) What is the smallest integer D such that the expected Shannon codeword length with a D-ary alphabet equals the expected Huffman codeword length with a D-ary alphabet?

Solution: *Huffman vs. Shannon*

(a) It is obvious that an Huffman code for the distribution (0.6,0.3,0.1) is (1,01,00), with codeword lengths (1,2,2). The Shannon code would use lengths $\lceil \log \frac{1}{p} \rceil$, which gives lengths (1,2,4) for the three symbols.

(b) For any $D > 2$, the Huffman code for the three symbols are all one character. The Shannon code length $\lceil \log_D \frac{1}{p} \rceil$ would be equal to 1 for all symbols if $\log_D \frac{1}{0.1} = 1$, i.e., if $D = 10$. Hence for $D \geq 10$, the Shannon code is also optimal.

34. **Huffman algorithm for tree construction.** Consider the following problem: m binary signals $S_1, S_2, \ldots, S_m$ are available at times $T_1 \leq T_2 \leq \ldots \leq T_m$, and we would like to find their sum $S_1 \oplus S_2 \oplus \cdots \oplus S_m$ using 2-input gates, each gate with 1 time unit delay, so that the final result is available as quickly as possible. A simple greedy algorithm is to combine the earliest two results, forming the partial result at time $\max(T_1, T_2) + 1$. We now have a new problem with $S_1 \oplus S_2, S_3, \ldots, S_m$, available at times $\max(T_1, T_2) + 1, T_3, \ldots, T_m$. We can now sort this list of T's, and apply the same merging step again, repeating this until we have the final result.

(a) Argue that the above procedure is optimal, in that it constructs a circuit for which the final result is available as quickly as possible.

(b) Show that this procedure finds the tree that minimizes

$$C(T) = \max_i (T_i + l_i) \tag{5.78}$$

where T_i is the time at which the result alloted to the i-th leaf is available, and l_i is the length of the path from the i-th leaf to the root.

(c) Show that

$$C(T) \geq \log_2 \left(\sum_i 2^{T_i} \right) \tag{5.79}$$

for any tree T.

(d) Show that there exists a tree such that

$$C(T) \leq \log_2 \left(\sum_i 2^{T_i} \right) + 1 \tag{5.80}$$

Thus $\log_2\left(\sum_i 2^{T_i}\right)$ is the analog of entropy for this problem.

Solution:

Tree construction:

(a) The proof is identical to the proof of optimality of Huffman coding. We first show that for the optimal tree if $T_i < T_j$, then $l_i \geq l_j$. The proof of this is, as in the case of Huffman coding, by contradiction. Assume otherwise, i.e., that if $T_i < T_j$ and $l_i < l_j$, then by exchanging the inputs, we obtain a tree with a lower total cost, since

$$\max\{T_i + l_i, T_j + l_j\} \geq \max\{T_i + l_j, T_j + l_i\} \tag{5.81}$$

Thus the longest branches are associated with the earliest times.

The rest of the proof is identical to the Huffman proof. We show that the longest branches correspond to the two earliest times, and that they could be taken as siblings (inputs to the same gate). Then we can reduce the problem to constructing the optimal tree for a smaller problem. By induction, we extend the optimality to the larger problem, proving the optimality of the above algorithm.

Given any tree of gates, the earliest that the output corresponding to a particular signal would be available is $T_i + l_i$, since the signal undergoes l_i gate delays. Thus $\max_i(T_i + l_i)$ is a lower bound on the time at which the final answer is available.

The fact that the tree achieves this bound can be shown by induction. For any internal node of the tree, the output is available at time equal to the maximum of the input times plus 1. Thus for the gates connected to the inputs T_i and T_j, the output is available at time $\max(T_i, T_j) + 1$. For any node, the output is available at time equal to maximum of the times at the leaves plus the gate delays to get from the leaf to the node. This result extneds to the complete tree, and for the root, the time at which the final result is available is $\max_i(T_i + l_i)$. The above algorithm minimizes this cost.

(b) Let $c_1 = \sum_i 2^{T_i}$ and $c_2 = \sum_i 2^{-l_i}$. By the Kraft inequality, $c_2 \leq 1$. Now let $p_i = \frac{2^{T_i}}{\sum_j 2^{T_j}}$, and let $r_i = \frac{2^{-l_i}}{\sum_j 2^{-l_j}}$. Clearly, p_i and r_i are probability mass functions. Also, we have $T_i = \log(p_i c_1)$ and $l_i = -\log(r_i c_2)$. Then

$$\begin{aligned} C(T) &= \max_i (T_i + l_i) &(5.82)\\ &= \max_i \left(\log(p_i c_1) - \log(r_i c_2)\right) &(5.83)\\ &= \log c_1 - \log c_2 + \max_i \log\frac{p_i}{r_i} &(5.84)\end{aligned}$$

Now the maximum of any random variable is greater than its average under any distribution, and therefore

$$\begin{aligned} C(T) &\geq \log c_1 - \log c_2 + \sum_i p_i \log\frac{p_i}{r_i} &(5.85)\\ &\geq \log c_1 - \log c_2 + D(p||r) &(5.86)\end{aligned}$$

Since $-logc_2 \geq 0$ and $D(p||r) \geq 0$, we have

$$C(T) \geq \log c_1 \tag{5.87}$$

which is the desired result.

(c) From the previous part, we achieve the lower bound if $p_i = r_i$ and $c_2 = 1$. However, since the l_i's are constrained to be integers, we cannot achieve equality in all cases.

Instead, if we let

$$l_i = \left\lceil \log \frac{1}{p_i} \right\rceil = \left\lceil \log \frac{\sum_j 2^{T_j}}{2^{T_i}} \right\rceil, \tag{5.88}$$

it is easy to verify that $\sum 2^{-l_i} \leq \sum p_i = 1$, and that thus we can construct a tree that achieves

$$T_i + l_i \leq \log(\sum_j 2^{T_j}) + 1 \tag{5.89}$$

for all i. Thus this tree achieves within 1 unit of the lower bound.

Clearly, $\log(\sum_j 2^{T_j})$ is the equivalent of entropy for this problem!

35. **Generating random variables.** One wishes to generate a random variable X

$$X = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases} \tag{5.90}$$

You are given fair coin flips $Z_1, Z_2, \ldots$. Let N be the (random) number of flips needed to generate X. Find a good way to use $Z_1, Z_2, \ldots$ to generate X. Show that $EN \leq 2$.

Solution: We expand $p = 0.p_1p_2\ldots$ as a binary number. Let $U = 0.Z_1Z_2\ldots$, the sequence Z treated as a binary number. It is well known that U is uniformly distributed on $[0,1)$. Thus, we generate $X = 1$ if $U < p$ and 0 otherwise.

The procedure for generated X would therefore examine $Z_1, Z_2, \ldots$ and compare with $p_1, p_2, \ldots$, and generate a 1 at the first time one of the Z_i's is less than the corresponding p_i and generate a 0 the first time one of the Z_i's is greater than the corresponding p_i's. Thus the probability that X is generated after seeing the first bit of Z is the probability that $Z_1 \neq p_1$, i.e., with probability 1/2. Similarly, X is generated after 2 bits of Z if $Z_1 = p_1$ and $Z_2 \neq p_2$, which occurs with probability 1/4. Thus

$$\begin{aligned} EN &= 1.\frac{1}{2} + 2\frac{1}{4} + 3\frac{1}{8} + \ldots + && (5.91) \\ &= 2 && (5.92) \end{aligned}$$

36. **Optimal word lengths.**

(a) Can $l = (1,2,2)$ be the word lengths of a binary Huffman code. What about (2,2,3,3)?

(b) What word lengths $l = (l_1, l_2, \ldots)$ can arise from binary Huffman codes?

Solution: *Optimal Word Lengths*

We first answer (b) and apply the result to (a).

(b) Word lengths of a binary Huffman code *must* satisfy the Kraft inequality with equality, i.e., $\sum_i 2^{-l_i} = 1$. An easy way to see this is the following: every node in the tree has a sibling (property of optimal binary code), and if we assign each node a 'weight', namely 2^{-l_i}, then 2×2^{-l_i} is the weight of the father (mother) node. Thus, 'collapsing' the tree back, we have that $\sum_i 2^{-l_i} = 1$.

(a) Clearly, $(1, 2, 2)$ satisfies Kraft with equality, while $(2, 2, 3, 3)$ does not. Thus, $(1, 2, 2)$ can arise from Huffman code, while $(2, 2, 3, 3)$ cannot.

37. **Codes.** Which of the following codes are

    (a) uniquely decodable?

    (b) instantaneous?

$$\begin{aligned}
C_1 &= \{00, 01, 0\} \\
C_2 &= \{00, 01, 100, 101, 11\} \\
C_3 &= \{0, 10, 110, 1110, \ldots\} \\
C_4 &= \{0, 00, 000, 0000\}
\end{aligned}$$

Solution: *Codes.*

(a) $C_1 = \{00, 01, 0\}$ is uniquely decodable (suffix free) but not instantaneous.

(b) $C_2 = \{00, 01, 100, 101, 11\}$ is prefix free (instantaneous).

(c) $C_3 = \{0, 10, 110, 1110, \ldots\}$ is instantaneous

(d) $C_4 = \{0, 00, 000, 0000\}$ is neither uniquely decodable or instantaneous.

38. **Huffman.** Find the Huffman D-ary code for $(p_1, p_2, p_3, p_4, p_5, p_6) = (\frac{6}{25}, \frac{6}{25}, \frac{4}{25}, \frac{4}{25}, \frac{3}{25}, \frac{2}{25})$ and the expected word length

    (a) for $D = 2$.

    (b) for $D = 4$.

Solution: Huffman Codes.

(a) D=2

6	6	6	8	11	14	25
6	6	6	6	8	11	
4	4	5	6	6		
4	4	4	5			
2	3	4				
2	2					
1						

p_i	$\frac{6}{25}$	$\frac{6}{25}$	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$
l_i	2	2	3	3	3	4	4

$$\begin{aligned}
\mathbf{E}(l) &= \sum_{i=1}^{7} p_i l_i \\
&= \frac{1}{25}(6\times 2 + 6\times 2 + 4\times 3 + 4\times 3 + 2\times 3 + 2\times 4 + 1\times 4) \\
&= \frac{66}{25} = 2.66
\end{aligned}$$

(b) D=4

6	9	25
6	6	
4	6	
4	4	
2		
2		
1		

p_i	$\frac{6}{25}$	$\frac{6}{25}$	$\frac{4}{25}$	$\frac{4}{25}$	$\frac{2}{25}$	$\frac{2}{25}$	$\frac{1}{25}$
l_i	1	1	1	2	2	2	2

$$\begin{aligned}
\mathbf{E}(l) &= \sum_{i=1}^{7} p_i l_i \\
&= \frac{1}{25}(6\times 1 + 6\times 1 + 4\times 1 + 4\times 2 + 2\times 2 + 2\times 2 + 1\times 2) \\
&= \frac{34}{25} = 1.36
\end{aligned}$$

39. **Entropy of encoded bits.** Let $C : X \longrightarrow \{0,1\}^*$ be a nonsingular but nonuniquely decodable code. Let X have entropy $H(X)$.

(a) Compare $H(C(X))$ to $H(X)$.

(b) Compare $H(C(X^n))$ to $H(X^n)$.

Solution: *Entropy of encoded bits*

(a) Since the code is non-singular, the function $X \to C(X)$ is one to one, and hence $H(X) = H(C(X))$. (Problem 2.4)

(b) Since the code is not uniquely decodable, the function $X^n \to C(X^n)$ is many to one, and hence $H(X^n) \geq H(C(X^n))$.

40. **Code rate.**
Let X be a random variable with alphabet $\{1, 2, 3\}$ and distribution

$$X = \begin{cases} 1, & \text{with probability } 1/2 \\ 2, & \text{with probability } 1/4 \\ 3, & \text{with probability } 1/4. \end{cases}$$

The data compression code for X assigns codewords

$$C(x) = \begin{cases} 0, & \text{if } x = 1 \\ 10, & \text{if } x = 2 \\ 11, & \text{if } x = 3. \end{cases}$$

Let $X_1, X_2, \ldots$ be independent identically distributed according to this distribution and let $Z_1 Z_2 Z_3 \ldots = C(X_1)C(X_2)\ldots$ be the string of binary symbols resulting from concatenating the corresponding codewords. For example, 122 becomes 01010.

(a) Find the entropy rate $H(\mathcal{X})$ and the entropy rate $H(\mathcal{Z})$ in bits per symbol. Note that Z is not compressible further.

(b) Now let the code be

$$C(x) = \begin{cases} 00, & \text{if } x = 1 \\ 10, & \text{if } x = 2 \\ 01, & \text{if } x = 3. \end{cases}$$

and find the entropy rate $H(\mathcal{Z})$.

(c) Finally, let the code be

$$C(x) = \begin{cases} 00, & \text{if } x = 1 \\ 1, & \text{if } x = 2 \\ 01, & \text{if } x = 3. \end{cases}$$

and find the entropy rate $H(\mathcal{Z})$.

Solution: Code rate.

This is a slightly tricky question. There's no straightforward rigorous way to calculate the entropy rates, so you need to do some guessing.

(a) First, since the X_i's are independent, $H(\mathcal{X}) = H(X_1) = 1/2 \log 2 + 2(1/4)\log(4) = 3/2$.

Now we observe that this is an optimal code for the given distribution on X, and since the probabilities are dyadic there is no gain in coding in blocks. So the resulting process *has to be* i.i.d. Bern(1/2), (for otherwise we could get further compression from it).

Therefore $H(\mathcal{Z}) = H(\text{Bern}(1/2)) = 1$.

(b) Here it's easy.

$$\begin{aligned} H(\mathcal{Z}) &= \lim_{n\to\infty} \frac{H(Z_1, Z_2, \ldots, Z_n)}{n} \\ &= \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_{n/2})}{n} \\ &= \lim_{n\to\infty} \frac{H(\mathcal{X})\frac{n}{2}}{n} \\ &= 3/4. \end{aligned}$$

(We're being a little sloppy and ignoring the fact that n above may not be a even, but in the limit as $n \to \infty$ this doesn't make a difference).

(c) This is the tricky part.

Suppose we encode the first n symbols $X_1 X_2 \cdots X_n$ into

$$Z_1 Z_2 \cdots Z_m = C(X_1)C(X_2)\cdots C(X_n).$$

Here $m = L(C(X_1)) + L(C(X_2)) + \cdots + L(C(X_n))$ is the total length of the encoded sequence (in bits), and L is the (binary) length function. Since the concatenated codeword sequence is an invertible function of $(X_1, \ldots, X_n)$, it follows that

$$nH(\mathcal{X}) \;=\; H(X_1 X_2 \cdots X_n) \;=\; H(Z_1 Z_2 \cdots Z_{\sum_1^n L(C(X_i))}) \tag{5.93}$$

The first equality above is trivial since the X_i's are independent. Similarly, may guess that the right-hand-side above can be written as

$$\begin{aligned} H(Z_1 Z_2 \cdots Z_{\sum_1^n L(C(X_i))}) &= E[\sum_{i=1}^{n} L(C(X_i))]H(\mathcal{Z}) \\ &= nE[L(C(X_1))]H(\mathcal{Z}) \end{aligned} \tag{5.94}$$

(This is not trivial to prove, but it *is* true.)

Combining the left-hand-side of (5.93) with the right-hand-side of (5.94) yields

$$\begin{aligned} H(\mathcal{Z}) &= \frac{H(\mathcal{X})}{E[L(C(X_1))]} \\ &= \frac{3/2}{7/4} \\ &= \frac{6}{7}, \end{aligned}$$

where $E[L(C(X_1))] = \sum_{x=1}^{3} p(x)L(C(x)) = 7/4$.

41. **Optimal codes.** Let $l_1, l_2, \ldots, l_{10}$ be the binary Huffman codeword lengths for the probabilities $p_1 \geq p_2 \geq \ldots \geq p_{10}$. Suppose we get a new distribution by splitting the last probability mass. What can you say about the optimal binary codeword lengths $\tilde{l_1}, \tilde{l_2}, \ldots, \tilde{l_{11}}$ for the probabilities $p_1, p_2, \ldots, p_9, \alpha p_{10}, (1-\alpha)p_{10}$, where $0 \leq \alpha \leq 1$.

    Solution: Optimal codes.

    To construct a Huffman code, we first combine the two smallest probabilities. In this case, we would combine αp_{10} and $(1-\alpha)p_{10}$. The result of the sum of these two probabilities is p_{10}. Note that the resulting probability distribution is now exactly the same as the original probability distribution. The key point is that an optimal code for $p_1, p_2, \ldots, p_{10}$ yields an optimal code (when expanded) for $p_1, p_2, \ldots, p_9, \alpha p_{10}, (1-\alpha)p_{10}$. In effect, the first 9 codewords will be left unchanged, while the 2 new codewords will be $XXX0$ and $XXX1$ where XXX represents the last codeword of the original distribution.

    In short, the lengths of the first 9 codewords remain unchanged, while the lengths of the last 2 codewords (new codewords) are equal to $l_{10} + 1$.

42. **Ternary codes.** Which of the following codeword lengths can be the word lengths of a 3-ary Huffman code and which cannot?

    (a) $(1, 2, 2, 2, 2)$

    (b) $(2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3)$

    Solution: Ternary codes.

    (a) The word lengths $(1, 2, 2, 2, 2)$ CANNOT be the word lengths for a 3-ary Huffman code. This can be seen by drawing the tree implied by these lengths, and seeing that one of the codewords of length 2 can be reduced to a codeword of length 1 which is shorter. Since the Huffman tree produces the minimum expected length tree, these codeword lengths cannot be the word lengths for a Huffman tree.

    (b) The word lengths $(2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3)$ ARE the word lengths for a 3-ary Huffman code. Again drawing the tree will verify this. Also, $\sum_i 3^{-l_i} = 8 \times 3^{-2} + 3 \times 3^{-3} = 1$, so these word lengths satisfy the Kraft inequality with equality. Therefore the word lengths are optimal for some distribution, and are the word lengths for a 3-ary Huffman code.

43. **Piecewise Huffman.** Suppose the codeword that we use to describe a random variable $X \sim p(x)$ always starts with a symbol chosen from the set $\{A, B, C\}$, followed by binary digits $\{0, 1\}$. Thus we have a ternary code for the first symbol and binary thereafter. Give the optimal uniquely decodeable code (minimum expected number of symbols) for the probability distribution

$$p = \left(\frac{16}{69}, \frac{15}{69}, \frac{12}{69}, \frac{10}{69}, \frac{8}{69}, \frac{8}{69}\right). \tag{5.95}$$

Solution: Piecewise Huffman.

Codeword						
a	x_1	16	16	22	31	69
b1	x_2	15	16	16	22	
c1	x_3	12	15	16	16	
c0	x_4	10	12	15		
b01	x_5	8	10			
b00	x_6	8				

Note that the above code is not only uniquely decodable, but it is also instantaneously decodable. Generally given a uniquely decodable code, we can construct an instantaneous code with the same codeword lengths. This is not the case with the piecewise Huffman construction. There exists a code with smaller expected lengths that is uniquely decodable, but not instantaneous.

Codeword
a
b
c
a0
b0
c0

44. **Huffman.** Find the word lengths of the optimal binary encoding of $p = \left(\frac{1}{100}, \frac{1}{100}, \ldots, \frac{1}{100}\right)$.

Solution: Huffman.

Since the distribution is uniform the Huffman tree will consist of word lengths of $\lceil \log(100) \rceil = 7$ and $\lfloor \log(100) \rfloor = 6$. There are 64 nodes of depth 6, of which (64-k) will be leaf nodes; and there are k nodes of depth 6 which will form $2k$ leaf nodes of depth 7. Since the total number of leaf nodes is 100, we have

$$(64 - k) + 2k = 100 \Rightarrow k = 36.$$

So there are 64 - 36 $= 28$ codewords of word length 6, and $2 \times 36 = 72$ codewords of word length 7.

45. **Random "20" questions.** Let X be uniformly distributed over $\{1, 2, \ldots, m\}$. Assume $m = 2^n$. We ask random questions: Is $X \in S_1$? Is $X \in S_2$?...until only one integer remains. All 2^m subsets of $\{1, 2, \ldots, m\}$ are equally likely.

(a) How many deterministic questions are needed to determine X?

(b) Without loss of generality, suppose that $X = 1$ is the random object. What is the probability that object 2 yields the same answers for k questions as object 1?

(c) What is the expected number of objects in $\{2, 3, \ldots, m\}$ that have the same answers to the questions as does the correct object 1?

(d) Suppose we ask $n + \sqrt{n}$ *random* questions. What is the expected number of wrong objects agreeing with the answers?

(e) Use Markov's inequality $\Pr\{X \geq t\mu\} \leq \frac{1}{t}$, to show that the probability of error (one or more wrong object remaining) goes to zero as $n \longrightarrow \infty$.

Solution: *Random "20" questions.*

(a) Obviously, Huffman codewords for X are all of length n. Hence, with n deterministic questions, we can identify an object out of 2^n candidates.

(b) Observe that the total number of subsets which include both object 1 and object 2 or neither of them is 2^{m-1}. Hence, the probability that object 2 yields the same answers for k questions as object 1 is $(2^{m-1}/2^m)^k = 2^{-k}$.

More information theoretically, we can view this problem as a channel coding problem through a noiseless channel. Since all subsets are equally likely, the probability the object 1 is in a specific random subset is $1/2$. Hence, the question whether object 1 belongs to the kth subset or not corresponds to the kth bit of the random codeword for object 1, where codewords X^k are Bern($1/2$) random k-sequences.

Object	Codeword
1	$0110\ldots1$
2	$0010\ldots0$
$\vdots$	

Now we observe a noiseless output Y^k of X^k and figure out which object was sent. From the same line of reasoning as in the achievability proof of the channel coding theorem, i.e. joint typicality, it is obvious the probability that object 2 has the same codeword as object 1 is 2^{-k}.

(c) Let

$$1_j = \begin{cases} 1, & \text{object } j \text{ yields the same answers for } k \text{ questions as object 1} \\ 0, & \text{otherwise} \end{cases},$$

for $j = 2, \ldots, m.$

Then,

$$\begin{aligned} E(\#\text{ of objects in } \{2,3,\ldots,m\} \text{ with the same answers}) &= E(\sum_{j=2}^{m} 1_j) \\ &= \sum_{j=2}^{m} E(1_j) \\ &= \sum_{j=2}^{m} 2^{-k} \\ &= (m-1)2^{-k} \\ &= (2^n-1)2^{-k}. \end{aligned}$$

(d) Plugging $k = n + \sqrt{n}$ into (c) we have the expected number of $(2^n - 1)2^{-n-\sqrt{n}}$.

(e) Let N by the number of wrong objects remaining. Then, by Markov's inequality

$$\begin{aligned} P(N \geq 1) &\leq EN \\ &= (2^n - 1)2^{-n-\sqrt{n}} \\ &\leq 2^{-\sqrt{n}} \\ &\to 0, \end{aligned}$$

where the first equality follows from part (d).

Chapter 6

Gambling and Data Compression

1. **Horse race.** Three horses run a race. A gambler offers 3-for-1 odds on each of the horses. These are fair odds under the assumption that all horses are equally likely to win the race. The true win probabilities are known to be

$$\mathbf{p} = (p_1, p_2, p_3) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right). \tag{6.1}$$

Let $\mathbf{b} = (b_1, b_2, b_3)$, $b_i \geq 0$, $\sum b_i = 1$, be the amount invested on each of the horses. The expected log wealth is thus

$$W(\mathbf{b}) = \sum_{i=1}^{3} p_i \log 3b_i. \tag{6.2}$$

(a) Maximize this over $\mathbf{b}$ to find $\mathbf{b}^*$ and W^*. Thus the wealth achieved in repeated horse races should grow to infinity like 2^{nW^*} with probability one.

(b) Show that if instead we put all of our money on horse 1, the most likely winner, we will eventually go broke with probability one.

Solution: *Horse race.*

(a) The doubling rate

$$\begin{aligned} W(\mathbf{b}) &= \sum_i p_i \log b_i o_i & (6.3)\\ &= \sum_i p_i \log 3b_i & (6.4)\\ &= \sum p_i \log 3 + \sum p_i \log p_i - \sum p_i \log \frac{p_i}{b_i} & (6.5)\\ &= \log 3 - H(\mathbf{p}) - D(\mathbf{p}||\mathbf{b}) & (6.6)\\ &\leq \log 3 - H(\mathbf{p}), & (6.7) \end{aligned}$$

with equality iff $\mathbf{p} = \mathbf{b}$. Hence $\mathbf{b}^* = \mathbf{p} = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ and $W^* = \log 3 - H(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) = \frac{1}{2}\log\frac{9}{8} = 0.085$.

By the strong law of large numbers,

$$\begin{aligned} S_n &= \prod_j 3b(X_j) && (6.8)\\ &= 2^{n(\frac{1}{n}\sum_j \log 3b(X_j))} && (6.9)\\ &\to 2^{nE\log 3b(X)} && (6.10)\\ &= 2^{nW(\mathbf{b})} && (6.11)\\ & && (6.12)\end{aligned}$$

When $\mathbf{b} = \mathbf{b}^*$, $W(\mathbf{b}) = W^*$ and $S_n \doteq 2^{nW^*} = 2^{0.085n} = (1.06)^n$.

(b) If we put all the money on the first horse, then the probability that we do not go broke in n races is $(\frac{1}{2})^n$. Since this probability goes to zero with n, the probability of the set of outcomes where we do not ever go broke is zero, and we will go broke with probability 1.

Alternatively, if $\mathbf{b} = (1, 0, 0)$, then $W(\mathbf{b}) = -\infty$ and

$$S_n \to 2^{nW} = 0 \quad \text{w.p.1} \tag{6.13}$$

by the strong law of large numbers.

2. **Horse race with subfair odds.** If the odds are bad (due to a track take) the gambler may wish to keep money in his pocket. Let $b(0)$ be the amount in his pocket and let $b(1), b(2), \ldots, b(m)$ be the amount bet on horses $1, 2, \ldots, m$, with odds $o(1), o(2), \ldots, o(m)$, and win probabilities $p(1), p(2), \ldots, p(m)$. Thus the resulting wealth is $S(x) = b(0) + b(x)o(x)$, with probability $p(x), x = 1, 2, \ldots, m$.

   (a) Find $\mathbf{b}^*$ maximizing $E\log S$ if $\sum 1/o(i) < 1$.

   (b) Discuss $\mathbf{b}^*$ if $\sum 1/o(i) > 1$. (There isn't an easy closed form solution in this case, but a "water-filling" solution results from the application of the Kuhn-Tucker conditions.)

Solution: (*Horse race with a cash option*).

Since in this case, the gambler is allowed to keep some of the money as cash, the mathematics becomes more complicated. In class, we used two different approaches to prove the optimality of proportional betting when the gambler is not allowed keep any of the money as cash. We will use both approaches for this problem. But in the case of subfair odds, the relative entropy approach breaks down, and we have to use the calculus approach.

The setup of the problem is straight-forward. We want to maximize the expected log return, i.e.,

$$W(\mathbf{b}, \mathbf{p}) = E\log S(X) = \sum_{i=1}^m p_i \log(b_0 + b_i o_i) \tag{6.14}$$

over all choices $\mathbf{b}$ with $b_i \geq 0$ and $\sum_{i=0}^m b_i = 1$.

Approach 1: Relative Entropy

We try to express $W(\mathbf{b}, \mathbf{p})$ as a sum of relative entropies.

$$\begin{aligned} W(\mathbf{b}, \mathbf{p}) &= \sum p_i \log(b_0 + b_i o_i) && (6.15) \\ &= \sum p_i \log \left(\frac{\frac{b_0}{o_i} + b_i}{\frac{1}{o_i}} \right) && (6.16) \\ &= \sum p_i \log \left(\frac{\frac{b_0}{o_i} + b_i}{p_i} \frac{p_i}{\frac{1}{o_i}} \right) && (6.17) \\ &= \sum p_i \log p_i o_i + \log K - D(\mathbf{p}||\mathbf{r}), && (6.18) \end{aligned}$$

where

$$K = \sum (\frac{b_0}{o_i} + b_i) = b_0 \sum \frac{1}{o_i} + \sum b_i = b_0 (\sum \frac{1}{o_i} - 1) + 1, \qquad (6.19)$$

and

$$r_i = \frac{\frac{b_0}{o_i} + b_i}{K} \qquad (6.20)$$

is a kind of normalized portfolio. Now both K and $\mathbf{r}$ depend on the choice of $\mathbf{b}$. To maximize $W(\mathbf{b}, \mathbf{p})$, we must maximize $\log K$ and at the same time minimize $D(\mathbf{p}||\mathbf{r})$. Let us consider the two cases:

(a) $\sum \frac{1}{o_i} \leq 1$. This is the case of superfair or fair odds. In these cases, it seems intuitively clear that we should put all of our money in the race. For example, in the case of a superfair gamble, one could invest any cash using a "Dutch book" (investing inversely proportional to the odds) and do strictly better with probability 1.

Examining the expression for K, we see that K is maximized for $b_0 = 0$. In this case, setting $b_i = p_i$ would imply that $r_i = p_i$ and hence $D(\mathbf{p}||\mathbf{r}) = 0$. We have succeeded in simultaneously maximizing the two variable terms in the expression for $W(\mathbf{b}, \mathbf{p})$ and this must be the optimal solution.

Hence, for fair or superfair games, the gambler should invest all his money in the race using proportional gambling, and not leave anything aside as cash.

(b) $\frac{1}{o_i} > 1$. In this case, sub-fair odds, the argument breaks down. Looking at the expression for K, we see that it is maximized for $b_0 = 1$. However, we cannot simultaneously minimize $D(\mathbf{p}||\mathbf{r})$.

If $p_i o_i \leq 1$ for all horses, then the first term in the expansion of $W(\mathbf{b}, \mathbf{p})$, that is, $\sum p_i \log p_i o_i$ is negative. With $b_0 = 1$, the best we can achieve is proportional betting, which sets the last term to be 0. Hence, with $b_0 = 1$, we can only achieve a negative expected log return, which is strictly worse than the 0 log return achieved be setting $b_0 = 1$. This would indicate, but not prove, that in this case, one should leave all one's money as cash. A more rigorous approach using calculus will prove this.

We can however give a simple argument to show that in the case of sub-fair odds, the gambler should leave at least some of his money as cash and that there is at least one horse on which he does not bet any money. We will prove this by contradiction—starting with a portfolio that does not satisfy these criteria, we will generate one which does better with probability one.

Let the amount bet on each of the horses be $(b_1, b_2, \ldots, b_m)$ with $\sum_{i=1}^m b_i = 1$, so that there is no money left aside. Arrange the horses in order of decreasing $b_i o_i$, so that the m-th horse is the one with the minimum product.

Consider a new portfolio with

$$b_i' = b_i - \frac{b_m o_m}{o_i} \tag{6.21}$$

for all i. Since $b_i o_i \geq b_m o_m$ for all i, $b_i' \geq 0$. We keep the remaining money, i.e.,

$$\begin{aligned} 1 - \sum_{i=1}^m b_i' &= 1 - \sum_{i=1}^m \left(b_i - \frac{b_m o_m}{o_i} \right) && (6.22) \\ &= \sum_{i=1}^m \frac{b_m o_m}{o_i} && (6.23) \end{aligned}$$

as cash.

The return on the new portfolio if horse i wins is

$$\begin{aligned} b_i' o_i &= \left(b_i - \frac{b_m o_m}{o_i} \right) o_i + \sum_{i=1}^m \frac{b_m o_m}{o_i} && (6.24) \\ &= b_i o_i + b_m o_m \left(\sum_{i=1}^m \frac{1}{o_i} - 1 \right) && (6.25) \\ &> b_i o_i, && (6.26) \end{aligned}$$

since $\sum 1/o_i > 1$. Hence irrespective of which horse wins, the new portfolio does better than the old one and hence the old portfolio could not be optimal.

Approach 2: Calculus

We set up the functional using Lagrange multipliers as before:

$$J(\mathbf{b}) = \sum_{i=1}^m p_i \log(b_0 + b_i o_i) + \lambda \left(\sum_{i=0}^m b_i \right) \tag{6.27}$$

Differentiating with respect to b_i, we obtain

$$\frac{\partial J}{\partial b_i} = \frac{p_i o_i}{b_0 + b_i o_i} + \lambda = 0. \tag{6.28}$$

Differentiating with respect to b_0, we obtain

$$\frac{\partial J}{\partial b_0} = \sum_{i=1}^m \frac{p_i}{b_0 + b_i o_i} + \lambda = 0. \tag{6.29}$$

Differentiating w.r.t. λ, we get the constraint

$$\sum b_i = 1. \tag{6.30}$$

The solution to these three equations, if they exist, would give the optimal portfolio $\mathbf{b}$. But substituting the first equation in the second, we obtain the following equation

$$\lambda \sum \frac{1}{o_i} = \lambda. \tag{6.31}$$

Clearly in the case when $\sum \frac{1}{o_i} \neq 1$, the only solution to this equation is $\lambda = 0$, which indicates that the solution is on the boundary of the region over which the maximization is being carried out. Actually, we have been quite cavalier with the setup of the problem—in addition to the constraint $\sum b_i = 1$, we have the inequality constraints $b_i \geq 0$. We should have allotted a Lagrange multiplier to each of these. Rewriting the functional with Lagrange multipliers

$$J(\mathbf{b}) = \sum_{i=1}^{m} p_i \log(b_0 + b_i o_i) + \lambda \left(\sum_{i=0}^{m} b_i \right) + \sum \gamma_i b_i \tag{6.32}$$

Differentiating with respect to b_i, we obtain

$$\frac{\partial J}{\partial b_i} = \frac{p_i o_i}{b_0 + b_i o_i} + \lambda + \gamma_i = 0. \tag{6.33}$$

Differentiating with respect to b_0, we obtain

$$\frac{\partial J}{\partial b_0} = \sum_{i=1}^{m} \frac{p_i}{b_0 + b_i o_i} + \lambda + \gamma_0 = 0. \tag{6.34}$$

Differentiating w.r.t. λ, we get the constraint

$$\sum b_i = 1. \tag{6.35}$$

Now, carrying out the same substitution, we get

$$\lambda + \gamma_0 = \lambda \sum \frac{1}{o_i} + \sum \frac{\gamma_i}{o_i}, \tag{6.36}$$

which indicates that if $\sum \frac{1}{o_i} \neq 1$, at least one of the γ's is non-zero, which indicates that the corresponding constraint has become active, which shows that the solution is on the boundary of the region.

In the case of solutions on the boundary, we have to use the Kuhn-Tucker conditions to find the maximum. These conditions are described in Gallager[6], pg. 87. The conditions describe the behavior of the derivative at the maximum of a concave function over a convex region. For any coordinate which is in the interior of the region, the derivative should be 0. For any coordinate on the boundary, the derivative should be negative in the direction towards the interior of the region. More formally, for a concave function $F(x_1, x_2, \ldots, x_n)$ over the region $x_i \geq 0$,

$$\frac{\partial F}{\partial x_i} \begin{array}{ll} \leq 0 & \text{if } x_i = 0 \\ = 0 & \text{if } x_i > 0 \end{array} \tag{6.37}$$

Applying the Kuhn-Tucker conditions to the present maximization, we obtain

$$\frac{p_i o_i}{b_0 + b_i o_i} + \lambda \begin{array}{ll} \leq 0 & \text{if } b_i = 0 \\ = 0 & \text{if } b_i > 0 \end{array} \tag{6.38}$$

and

$$\sum \frac{p_i}{b_0 + b_i o_i} + \lambda \begin{array}{ll} \leq 0 & \text{if } b_0 = 0 \\ = 0 & \text{if } b_0 > 0 \end{array} \tag{6.39}$$

Theorem 4.4.1 in Gallager[6] proves that if we can find a solution to the Kuhn-Tucker conditions, then the solution is the maximum of the function in the region. Let us consider the two cases:

(a) $\sum \frac{1}{o_i} \leq 1$. In this case, we try the solution we expect, $b_0 = 0$, and $b_i = p_i$. Setting $\lambda = -1$, we find that all the Kuhn-Tucker conditions are satisfied. Hence, this is the optimal portfolio for superfair or fair odds.

(b) $\sum \frac{1}{o_i} > 1$. In this case, we try the expected solution, $b_0 = 1$, and $b_i = 0$. We find that all the Kuhn-Tucker conditions are satisfied if all $p_i o_i \leq 1$. Hence under this condition, the optimum solution is to not invest anything in the race but to keep everything as cash.

In the case when some $p_i o_i > 1$, the Kuhn-Tucker conditions are no longer satisfied by $b_0 = 1$. We should then invest some money in the race; however, since the denominator of the expressions in the Kuhn-Tucker conditions also changes, more than one horse may now violate the Kuhn-Tucker conditions. Hence, the optimum solution may involve investing in some horses with $p_i o_i \leq 1$. There is no explicit form for the solution in this case.

The Kuhn Tucker conditions for this case do not give rise to an explicit solution. Instead, we can formulate a procedure for finding the optimum distribution of capital:

Order the horses according to $p_i o_i$, so that

$$p_1 o_1 \geq p_2 o_2 \geq \cdots \geq p_m o_m. \tag{6.40}$$

Define

$$C_k = \begin{cases} \frac{1 - \sum_{i=1}^k p_i}{1 - \sum_{i=1}^k \frac{1}{o_i}} & \text{if } k \geq 1 \\ 1 & \text{if } k = 0 \end{cases} \tag{6.41}$$

Define

$$t = \min\{n | p_{n+1} o_{n+1} \leq C_n\}. \tag{6.42}$$

Clearly $t \geq 1$ since $p_1 o_1 > 1 = C_0$.

Claim: The optimal strategy for the horse race when the odds are subfair and some of the $p_i o_i$ are greater than 1 is: set

$$b_0 = C_t, \tag{6.43}$$

and for $i = 1, 2, \ldots, t$, set

$$b_i = p_i - \frac{C_t}{o_i}, \tag{6.44}$$

and for $i = t+1, \ldots, m$, set

$$b_i = 0. \tag{6.45}$$

The above choice of $\mathbf{b}$ satisfies the Kuhn-Tucker conditions with $\lambda = 1$. For b_0, the Kuhn-Tucker condition is

$$\sum \frac{p_i}{b_o + b_i o_i} = \sum_{i=1}^{t} \frac{1}{o_i} + \sum_{i=t+1}^{m} \frac{p_i}{C_t} = \sum_{i=1}^{t} \frac{1}{o_i} + \frac{1 - \sum_{i=1}^{t} p_i}{C_t} = 1. \tag{6.46}$$

For $1 \le i \le t$, the Kuhn Tucker conditions reduce to

$$\frac{p_i o_i}{b_0 + b_i o_i} = \frac{p_i o_i}{p_i o_i} = 1. \tag{6.47}$$

For $t+1 \le i \le m$, the Kuhn Tucker conditions reduce to

$$\frac{p_i o_i}{b_0 + b_i o_i} = \frac{p_i o_i}{C_t} \le 1, \tag{6.48}$$

by the definition of t. Hence the Kuhn Tucker conditions are satisfied, and this is the optimal solution.

3. **Cards.** An ordinary deck of cards containing 26 red cards and 26 black cards is shuffled and dealt out one card at at time without replacement. Let X_i be the color of the ith card.

   (a) Determine $H(X_1)$.

   (b) Determine $H(X_2)$.

   (c) Does $H(X_k \mid X_1, X_2, \ldots, X_{k-1})$ increase or decrease?

   (d) Determine $H(X_1, X_2, \ldots, X_{52})$.

   Solution:

   (a) P(first card red) $=$ P(first card black)$=$ $1/2$. Hence $H(X_1) = (1/2)\log 2 + (1/2)\log 2 = \log 2 = 1$ bit.

   (b) P(second card red) $=$ P(second card black) $= 1/2$ by symmetry. Hence $H(X_2) = (1/2)\log 2 + (1/2)\log 2 = \log 2 = 1$ bit. There is no change in the probability from X_1 to X_2 (or to X_i, $1 \le i \le 52$) since all the permutations of red and black cards are equally likely.

(c) Since all permutations are equally likely, the joint distribution of X_k and $X_1, \ldots, X_{k-1}$ is the same as the joint distribution of X_{k+1} and $X_1, \ldots, X_{k-1}$. Therefore

$$H(X_k|X_1, \ldots, X_{k-1}) = H(X_{k+1}|X_1, \ldots, X_{k-1}) \geq H(X_{k+1}|X_1, \ldots, X_k) \quad (6.49)$$

and so the conditional entropy decreases as we proceed along the sequence.

Knowledge of the past reduces uncertainty and thus means that the conditional entropy of the k-th card's color given all the previous cards will decrease as k increases.

(d) All $\binom{52}{26}$ possible sequences of 26 red cards and 26 black cards are equally likely. Thus

$$H(X_1, X_2, \ldots, X_{52}) = \log \binom{52}{26} = 48.8 \text{ bits (3.2 bits less than 52)} \quad (6.50)$$

4. **Gambling.** Suppose one gambles sequentially on the card outcomes in Problem 3. Even odds of 2-for-1 are paid. Thus the wealth S_n at time n is $S_n = 2^n b(x_1, x_2, \ldots, x_n)$, where $b(x_1, x_2, \ldots, x_n)$ is the proportion of wealth bet on $x_1, x_2, \ldots, x_n$. Find $\max_{b(\cdot)} E \log S_{52}$.

Solution: *Gambling on red and black cards.*

$$\begin{aligned} E[\log S_n] &= E[\log[2^n b(X_1, X_2, ..., X_n)]] && (6.51) \\ &= n \log 2 + E[\log b(\mathbf{X})] && (6.52) \\ &= n + \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \log b(\mathbf{x}) && (6.53) \\ &= n + \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x})[\log \frac{b(\mathbf{x})}{p(\mathbf{x})} - \log p(\mathbf{x})] && (6.54) \\ &= n + D(p(\mathbf{x})||b(\mathbf{x})) - H(X). && (6.55) \end{aligned}$$

Taking $p(\mathbf{x}) = b(\mathbf{x})$ makes $D(p(\mathbf{x})||b(\mathbf{x})) = 0$ and maximizes $E \log S_{52}$.

$$\begin{aligned} \max_{b(\mathbf{x})} E \log S_{52} &= 52 - H(X) && (6.56) \\ &= 52 - \log \frac{52!}{26!26!} && (6.57) \\ &= 3.2 && (6.58) \end{aligned}$$

Alternatively, as in the horse race, proportional betting is log-optimal. Thus $b(\mathbf{x}) = p(\mathbf{x})$ and, regardless of the outcome,

$$S_{52} = \frac{2^{52}}{\binom{52}{26}} = 9.08. \quad (6.59)$$

and hence

$$\log S_{52} = \max_{b(\mathbf{x})} E \log S_{52} = \log 9.08 = 3.2. \quad (6.60)$$

5. **Beating the public odds.** Consider a 3-horse race with win probabilities

$$(p_1, p_2, p_3) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$$

and fair odds with respect to the (false) distribution

$$(r_1, r_2, r_3) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2}) \quad .$$

Thus the odds are

$$(o_1, o_2, o_3) = (4, 4, 2) \quad .$$

(a) What is the entropy of the race?

(b) Find the set of bets (b_1, b_2, b_3) such that the compounded wealth in repeated plays will grow to infinity.

Solution: *Beating the public odds.*

(a) The entropy of the race is given by

$$\begin{aligned} H(\mathbf{p}) &= \frac{1}{2}\log 2 + \frac{1}{4}\log 4 + \frac{1}{4}\log 4 \\ &= \frac{3}{2}. \end{aligned}$$

(b) Compounded wealth will grow to infinity for the set of bets (b_1, b_2, b_3) such that $W(\mathbf{b}, \mathbf{p}) > 0$ where

$$\begin{aligned} W(\mathbf{b}, \mathbf{p}) &= D(\mathbf{p}\|\mathbf{r}) - D(\mathbf{p}\|\mathbf{b}) \\ &= \sum_{i=1}^{3} p_i \log \frac{b_i}{r_i}. \end{aligned}$$

Calculating $D(\mathbf{p}\|\mathbf{r})$, this criterion becomes

$$D(\mathbf{p}\|\mathbf{b}) \quad < \quad \frac{1}{4}.$$

6. **Horse race**: A 3 horse race has win probabilities $\mathbf{p} = (p_1, p_2, p_3)$, and odds $\mathbf{o} = (1, 1, 1)$. The gambler places bets $\mathbf{b} = (b_1, b_2, b_3)$, $b_i \geq 0, \sum b_i = 1$, where b_i denotes the proportion on wealth bet on horse i. These odds are very bad. The gambler gets his money back on the winning horse and loses the other bets. Thus the wealth S_n at time n resulting from independent gambles goes expnentially to zero.

(a) Find the exponent.

(b) Find the optimal gambling scheme $\mathbf{b}$, i.e., the bet $\mathbf{b}^*$ that maximizes the exponent.

(c) Assuming $\mathbf{b}$ is chosen as in (b), what distribution $\mathbf{p}$ causes S_n to go to zero at the fastest rate?

Solution: *Minimizing losses.*

(a) Despite the bad odds, the optimal strategy is still proportional gambling. Thus the optimal bets are $\mathbf{b} = \mathbf{p}$, and the exponent in this case is

$$W^* = \sum_i p_i \log p_i = -H(\mathbf{p}). \tag{6.61}$$

(b) The optimal gambling strategy is still proportional betting.

(c) The worst distribution (the one that causes the doubling rate to be as negative as possible) is that distribution that maximizes the entropy. Thus the worst W^* is $-\log 3$, and the gambler's money goes to zero as 3^{-n}.

7. **Horse race.** Consider a horse race with 4 horses. Assume that each of the horses pays 4-for-1 if it wins. Let the probabilities of winning of the horses be $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\}$. If you started with \$100 and bet optimally to maximize your long term growth rate, what are your optimal bets on each horse? Approximately how much money would you have after 20 races with this strategy ?

   Solution: *Horse race.* The optimal betting strategy is proportional betting, i.e., dividing the investment in proportion to the probabilities of each horse winning. Thus the bets on each horse should be (50%, 25%,12.5%,12.5%), and the growth rate achieved by this strategy is equal to $\log 4 - H(p) = \log 4 - H(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}) = 2 - 1.75 = 0.25$. After 20 races with this strategy, the wealth is approximately $2^{nW} = 2^5 = 32$, and hence the wealth would grow approximately 32 fold over 20 races.

8. **Lotto.** The following analysis is a crude approximation to the games of Lotto conducted by various states. Assume that the player of the game is required pay \$1 to play and is asked to choose 1 number from a range 1 to 8. At the end of every day, the state lottery commission picks a number uniformly over the same range. The jackpot, i.e., all the money collected that day, is split among all the people who chose the same number as the one chosen by the state. E.g., if 100 people played today, and 10 of them chose the number 2, and the drawing at the end of the day picked 2, then the \$100 collected is split among the 10 people, i.e., each of persons who picked 2 will receive \$10, and the others will receive nothing.

   The general population does not choose numbers uniformly - numbers like 3 and 7 are supposedly lucky and are more popular than 4 or 8. Assume that the fraction of people choosing the various numbers $1, 2, \ldots, 8$ is $(f_1, f_2, \ldots, f_8)$, and assume that n people play every day. Also assume that n is very large, so that any single person's choice choice does not change the proportion of people betting on any number.

   (a) What is the optimal strategy to divide your money among the various possible tickets so as to maximize your long term growth rate? (Ignore the fact that you cannot buy fractional tickets.)

(b) What is the optimal growth rate that you can achieve in this game?

(c) If $(f_1, f_2, \ldots, f_8) = (1/8, 1/8, 1/4, 1/16, 1/16, 1/16, 1/4, 1/16)$, and you start with \$1, how long will it be before you become a millionaire?

Solution:

(a) The probability of winning does not depend on the number you choose, and therefore, irrespective of the proportions of the other players, the log optimal strategy is to divide your money uniformly over all the tickets.

(b) If there are n people playing, and f_i of them choose number i, then the number of people sharing the jackpot of n dollars is nf_i, and therefore each person gets $n/nf_i = 1/f_i$ dollars if i is picked at the end of the day. Thus the odds for number i is $1/f_i$, and does not depend on the number of people playing.

Using the results of Section 6.1, the optimal growth rate is given by

$$W^*(\mathbf{p}) = \sum p_i \log o_i - H(\mathbf{p}) = \sum \frac{1}{8} \log \frac{1}{f_i} - \log 8 \tag{6.62}$$

(c) Substituing these fraction in the previous equation we get

$$\begin{aligned} W^*(\mathbf{p}) &= \frac{1}{8} \sum \log \frac{1}{f_i} - \log 8 && (6.63) \\ &= \frac{1}{8}(3+3+2+4+4+4+2+4) - 3 && (6.64) \\ &= 0.25 && (6.65) \end{aligned}$$

and therefore after N days, the amount of money you would have would be approximately $2^{0.25N}$. The number of days before this crosses a million $= \log_2(1,000,000)/0.25 = 79.7$, i.e., in 80 days, you should have a million dollars.

There are many problems with the analysis, not the least of which is that the state governments take out about half the money collected, so that the jackpot is only half of the total collections. Also there are about 14 million different possible tickets, and it is therefore possible to use a uniform distribution using \$1 tickets only if we use capital of the order of 14 million dollars. And with such large investments, the proportions of money bet on the different possibilities will change, which would further complicate the analysis.

However, the fact that people's choices are not uniform does leave a loophole that can be exploited. Under certain conditions, i.e., if the accumulated jackpot has reached a certain size, the expected return can be greater than 1, and it is worthwhile to play, despite the 50% cut taken by the state. But under normal circumstances, the 50% cut of the state makes the odds in the lottery very unfair, and it is not a worthwhile investment.

9. **Horse race.** Suppose one is interested in maximizing the doubling rate for a horse race. Let $p_1, p_2, \ldots, p_m$ denote the win probabilities of the m horses. When do the odds $(o_1, o_2, \ldots, o_m)$ yield a higher doubling rate than the odds $(o'_1, o'_2, \ldots, o'_m)$?

Solution: *Horse Race*

Let W and W' denote the optimal doubling rates for the odds $(o_1, o_2, \ldots, o_m)$ and $(o'_1, o'_2, \ldots, o'_m)$ respectively. By Theorem 6.1.2 in the book,

$$\begin{aligned} W &= \sum p_i \log o_i - H(p), \text{ and} \\ W' &= \sum p_i \log o'_i - H(p) \end{aligned}$$

where p is the probability vector $(p_1, p_2, \ldots, p_m)$. Then $W > W'$ exactly when $\sum p_i \log o_i > \sum p_i \log o'_i$; that is, when

$$E \log o_i > E \log o'_i.$$

10. **Horse race with probability estimates**

    (a) Three horses race. Their probabilities of winning are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. The odds are (4-for-1, 3-for-1 and 3-for-1). Let W^* be the optimal doubling rate.

    Suppose you believe the probabilities are $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. If you try to maximize the doubling rate, what doubling rate W will you achieve? By how much has your doubling rate decreased due to your poor estimate of the probabilities, i.e., what is $\Delta W = W^* - W$?

    (b) Now let the horse race be among m horses, with probabilities $p = (p_1, p_2, \ldots, p_m)$ and odds $o = (o_1, o_2, \ldots, o_m)$. If you believe the true probabilities to be $q = (q_1, q_2, \ldots, q_m)$, and try to maximize the doubling rate W, what is $W^* - W$?

Solution: *Horse race with probability estimates*

(a) If you believe that the probabilities of winning are $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, you would bet proportional to this, and would achieve a growth rate $\sum p_i \log b_i o_i = \frac{1}{2} \log 4\frac{1}{4} + \frac{1}{4} \log 3\frac{1}{2} + \frac{1}{4} \log 3\frac{1}{4} = \frac{1}{4} \log \frac{9}{8}$. If you bet according to the true probabilities, you would bet $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ on the three horses, achieving a growth rate $\sum p_i \log b_i o_i = \frac{1}{2} \log 4\frac{1}{2} + \frac{1}{4} \log 3\frac{1}{4} + \frac{1}{4} \log 3\frac{1}{4} = \frac{1}{2} \log \frac{3}{2}$. The loss in growth rate due to incorrect estimation of the probabilities is the difference between the two growth rates, which is $\frac{1}{4} \log 2 = 0.25$.

(b) For m horses, the growth rate with the true distribution is $\sum p_i \log p_i o_i$, and with the incorrect estimate is $\sum p_i \log q_i o_i$. The difference between the two is $\sum p_i \log \frac{p_1}{q_i} = D(p||q)$.

11. **The two envelope problem:** One envelope contains b dollars, the other $2b$ dollars. The amount b is unknown. An envelope is selected at random. Let X be the amount observed in this envelope, and let Y be the amount in the other envelope.

    Adopt the strategy of switching to the other envelope with probability $p(x)$, where $p(x) = \frac{e^{-x}}{e^{-x}+e^{x}}$. Let Z be the amount that the player receives. Thus

$$(X, Y) = \begin{cases} (b, 2b), & \text{with probability } 1/2 \\ (2b, b), & \text{with probability } 1/2 \end{cases} \tag{6.66}$$

$$Z = \begin{cases} X, & \text{with probability } 1 - p(x) \\ Y, & \text{with probability } p(x) \end{cases} \tag{6.67}$$

(a) Show that $E(X) = E(Y) = \frac{3b}{2}$.

(b) Show that $E(Y/X) = 5/4$. Since the expected ratio of the amount in the other envelope to the one in hand is 5/4, it seems that one should always switch. (This is the origin of the switching paradox.) However, observe that $E(Y) \neq E(X)E(Y/X)$. Thus, although $E(Y/X) > 1$, it does not follow that $E(Y) > E(X)$.

(c) Let J be the index of the envelope containing the maximum amount of money, and let J' be the index of the envelope chosen by the algorithm. Show that for any b, $I(J; J') > 0$. Thus the amount in the first envelope always contains some information about which envelope to choose.

(d) Show that $E(Z) > E(X)$. Thus you can do better than always staying or always switching. In fact, this is true for any monotonic decreasing switching function $p(x)$. By randomly switching according to $p(x)$, you are more likely to trade up than trade down.

Solution: *Two envelope problem:*

(a) $X = b$ or $2b$ with prob. 1/2, and therefore $E(X) = 1.5b$. Y has the same unconditional distribution.

(b) Given $X = x$, the other envelope contains $2x$ with probability 1/2 and contains $x/2$ with probability 1/2. Thus $E(Y/X) = 5/4$.

(c) Without any conditioning, $J = 1$ or 2 with probability (1/2,1/2). By symmetry, it is not difficult to see that the unconditional probability distribution of J' is also the same. We will now show that the two random variables are not independent, and therefore $I(J; J') \neq 0$. To do this, we will calculate the conditional probability $P(J' = 1|J = 1)$.

Conditioned on $J = 1$, the probability that $X = b$ or $2b$ is still (1/2,1/2). However, conditioned on $(J = 1, X = 2b)$, the probability that $Z = X$, and therefore $J' = 1$ is $p(2b)$. Similary, conditioned on $(J = 1, X = b)$, the probability that $J' = 1$ is $1 - p(b)$. Thus,

$$\begin{aligned} P(J' = 1|J = 1) &= P(X = b|J = 1)P(J' = 1|X = b, J = 1) \\ &\quad + P(X = 2b|J = 1)P(J' = 1|X = 2b, J = 1) && (6.68) \\ &= \frac{1}{2}(1 - p(b)) + \frac{1}{2}p(2b) && (6.69) \\ &= \frac{1}{2} + \frac{1}{2}(p(2b) - p(b)) && (6.70) \\ &> \frac{1}{2} && (6.71) \end{aligned}$$

Thus the conditional distribution is not equal to the unconditional distribution and J and J' are not independent.

(d) We use the above calculation of the conditional distribution to calculate $E(Z)$. Without loss of generality, we assume that $J=1$, i.e., the first envelope contains $2b$. Then

$$\begin{aligned}
E(Z|J=1) &= P(X=b|J=1)E(Z|X=b,J=1) \\
&\qquad +P(X=2b|J=1)E(Z|X=2b,J=1) && (6.72)\\
&= \frac{1}{2}E(Z|X=b,J=1)+\frac{1}{2}E(Z|X=2b,J=1) && (6.73)\\
&= \frac{1}{2}\big(p(J'=1|X=b,J=1)E(Z|J'=1,X=b,J=1) \\
&\qquad +p(J'=2|X=b,J=1)E(Z|J'=2,X=b,J=1) \\
&\qquad +p(J'=1|X=2b,J=1)E(Z|J'=1,X=2b,J=1) \\
&\qquad +p(J'=2|X=2b,J=1)E(Z|J'=2,X=2b,J=1)\big) && (6.74)\\
&= \frac{1}{2}\left([1-p(b)]2b+p(b)b+p(2b)2b+[1-p(2b)]b\right) && (6.75)\\
&= \frac{3b}{2}+\frac{1}{2}b(p(2b)-p(b)) && (6.76)\\
&> \frac{3b}{2} && (6.77)
\end{aligned}$$

as long as $p(2b)-p(b)>0$. Thus $E(Z)>E(X)$.

12. **Gambling.** Find the horse win probabilities $p_1,p_2,\ldots,p_m$

(a) maximizing the doubling rate W^* for given *fixed* known odds $o_1,o_2,\ldots,o_m$.

(b) minimizing the doubling rate for given fixed odds $o_1,o_2,\ldots,o_m$.

Solution: *Gambling*

(a) From Theorem 6.1.2, $W^*=\sum p_i\log o_i-H(p)$. We can also write this as

$$\begin{aligned}
W^* &= \sum_i p_i\log p_io_i && (6.78)\\
&= \sum_i p_i\log\frac{p_i}{\frac{1}{o_i}} && (6.79)\\
&= \sum_i p_i\log\frac{p_i}{q_i}-\sum_i p_i\log\left(\sum_j\frac{1}{o_j}\right) && (6.80)\\
&= \sum_i p_i\log\frac{p_i}{q_i}-\log\left(\sum_j\frac{1}{o_j}\right) && (6.81)
\end{aligned}$$

where

$$q_i=\frac{\frac{1}{o_i}}{\sum_j\frac{1}{o_j}} \tag{6.82}$$

Therefore the minimum value of the growth rate occurs when $p_i = q_i$. This is the distribution that minimizes the growth rate, and the minimum value is $-\log\left(\sum_j \frac{1}{o_j}\right)$.

(b) The maximum growth rate occurs when the horse with the maximum odds wins in all the races, i.e., $p_i = 1$ for the horse that provides the maximum odds

13. **Dutch book.** Consider a horse race with $m = 2$ horses,

$$\begin{aligned} X &= 1,\ 2 \\ p &= 1/2,\ 1/2 \\ \text{Odds (for one)} &= 10,\ 30 \\ \text{Bets} &= b,\ 1-b. \end{aligned}$$

The odds are super fair.

(a) There is a bet b which guarantees the same payoff regardless of which horse wins. Such a bet is called a Dutch book. Find this b and the associated wealth factor $S(X)$.

(b) What is the maximum growth rate of the wealth for this gamble? Compare it to the growth rate for the Dutch book.

Solution: Solution: Dutch book.

(a)

$$\begin{aligned} 10b_D &= 30(1-b_D) \\ 40b_D &= 30 \\ b_D &= 3/4. \end{aligned}$$

Therefore,

$$\begin{aligned} W(b_D, P) &= \frac{1}{2}\log\left(10\frac{3}{4}\right) + \frac{1}{2}\log\left(30\frac{1}{4}\right) \\ &= 2.91 \end{aligned}$$

and

$$S_D(X) = 2^{W(b_D,P)} = 7.5.$$

(b) In general,

$$W(b,p) = \frac{1}{2}\log(10b) + \frac{1}{2}\log(30(1-b)).$$

Setting the $\frac{\partial W}{\partial b}$ to zero we get

$$\frac{1}{2}\left(\frac{10}{10b^*}\right) + \frac{1}{2}\left(\frac{-30}{30-30b^*}\right) = 0$$

$$\begin{aligned}
\frac{1}{2b^*} + \frac{1}{2(b^*-1)} &= 0 \\
\frac{(b^*-1)+b^*}{2b^*(b^*-1)} &= 0 \\
\frac{2b^*-1}{4b^*(1-b^*)} &= 0 \\
b^* &= \frac{1}{2}.
\end{aligned}$$

Hence

$$\begin{aligned}
W^*(p) &= \frac{1}{2}\log(5) + \frac{1}{2}\log(15) = 3.11 \\
W(b_D, p) &= 2.91
\end{aligned}$$

and

$$\begin{aligned}
S^* &= 2^{W^*} = 8.66 \\
S_D &= 2^{W_D} = 7.5
\end{aligned}$$

Thus gambling (a little) with b^* beats the sure win of 7.5 given by the Dutch book

14. **Horse race.** Suppose one is interested in maximizing the doubling rate for a horse race. Let $p_1, p_2, \ldots, p_m$ denote the win probabilities of the m horses. When do the odds $(o_1, o_2, \ldots, o_m)$ yield a higher doubling rate than the odds $(o'_1, o'_2, \ldots, o'_m)$?

    Solution: *Horse Race* (Repeat of problem 9)

    Let W and W' denote the optimal doubling rates for the odds $(o_1, o_2, \ldots, o_m)$ and $(o'_1, o'_2, \ldots, o'_m)$ respectively. By Theorem 6.1.2 in the book,

$$\begin{aligned}
W &= \sum p_i \log o_i - H(p), \quad \text{and} \\
W' &= \sum p_i \log o'_i - H(p)
\end{aligned}$$

    where p is the probability vector $(p_1, p_2, \ldots, p_m)$. Then $W > W'$ exactly when $\sum p_i \log o_i > \sum p_i \log o'_i$; that is, when

$$E \log o_i > E \log o'_i.$$

15. **Entropy of a fair horse race.** Let $X \sim p(x)$, $x = 1, 2, \ldots, m$, denote the winner of a horse race. Suppose the odds $o(x)$ are fair with respect to $p(x)$, i.e., $o(x) = \frac{1}{p(x)}$. Let $b(x)$ be the amount bet on horse x, $b(x) \geq 0$, $\sum_1^m b(x) = 1$. Then the resulting wealth factor is $S(x) = b(x)o(x)$, with probability $p(x)$.

(a) Find the expected wealth $ES(X)$.

(b) Find W^*, the optimal growth rate of wealth.

(c) Suppose

$$Y = \begin{cases} 1, & X = 1 \text{ or } 2 \\ 0, & \text{otherwise} \end{cases}$$

If this side information is available before the bet, how much does it increase the growth rate W^*?

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

Solution: Entropy of a fair horse race.

(a) The expected wealth $ES(X)$ is

$$\begin{aligned} ES(X) &= \sum_{x=1}^{m} S(x)p(x) && (6.83) \\ &= \sum_{x=1}^{m} b(x)o(x)p(x) && (6.84) \\ &= \sum_{x=1}^{m} b(x), \quad (\text{since } o(x) = 1/p(x)) && (6.85) \\ &= 1. && (6.86) \end{aligned}$$

(b) The optimal growth rate of wealth, W^*, is achieved when $b(x) = p(x)$ for all x, in which case,

$$\begin{aligned} W^* &= E(\log S(X)) && (6.87) \\ &= \sum_{x=1}^{m} p(x)\log(b(x)o(x)) && (6.88) \\ &= \sum_{x=1}^{m} p(x)\log(p(x)/p(x)) && (6.89) \\ &= \sum_{x=1}^{m} p(x)\log(1) && (6.90) \\ &= 0, && (6.91) \end{aligned}$$

so we maintain our current wealth.

(c) The increase in our growth rate due to the side information is given by $I(X;Y)$. Let $q = \Pr(Y = 1) = p(1) + p(2)$.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) && (6.92) \\ &= H(Y) \quad (\text{since } Y \text{ is a deterministic function of } X) && (6.93) \\ &= H(q). && (6.94) \end{aligned}$$

(d) Already computed above.

16. **Negative horse race** Consider a horse race with m horses with win probabilities $p_1, p_2, \ldots p_m$. Here the gambler hopes a given horse will lose. He places bets $(b_1, b_2, \ldots, b_m), \sum_{i=1}^{m} b_i = 1$, on the horses, loses his bet b_i if horse i wins, and retains the rest of his bets. (No odds.) Thus $S = \sum_{j \neq i} b_j$, with probability p_i, and one wishes to maximize $\sum p_i \ln(1-b_i)$ subject to the constraint $\sum b_i = 1$.

    (a) Find the growth rate optimal investment strategy b^*. Do *not* constrain the bets to be positive, but do constrain the bets to sum to 1. (This effectively allows short selling and margin.)

    (b) What is the optimal growth rate?

    Solution: Negative horse race

    (a) Let $b_i' = 1 - b_i \geq 0$, and note that $\sum_i b_i' = m - 1$. Let $q_i = b_i' / \sum_j b_j'$. Then, $\{q_i\}$ is a probability distribution on $\{1, 2, \ldots, m\}$. Now,

$$\begin{aligned} W &= \sum_i p_i \log(1 - b_i) \\ &= \sum_i p_i \log q_i (m-1) \\ &= \log(m-1) + \sum_i p_i \log p_i \frac{q_i}{p_i} \\ &= \log(m-1) - H(p) - D(p\|q) \quad . \end{aligned}$$

    Thus, W^* is obtained upon setting $D(p\|q) = 0$, which means making the bets such that $p_i = q_i = b_i'/(m-1)$, or $b_i = 1 - (m-1)p_i$. Alternatively, one can use Lagrange multipliers to solve the problem.

    (b) From (a) we directly see that setting $D(p\|q) = 0$ implies $W^* = \log(m-1) - H(p)$.

17. **The St. Petersburg paradox.** Many years ago in ancient St. Petersburg the following gambling proposition caused great consternation. For an entry fee of c units, a gambler receives a payoff of 2^k units with probability $2^{-k}, k = 1, 2, \ldots$.

    (a) Show that the expected payoff for this game is infinite. For this reason, it was argued that $c = \infty$ was a "fair" price to pay to play this game. Most people find this answer absurd.

    (b) Suppose that the gambler can buy a share of the game. For example, if he invests $c/2$ units in the game, he receives $1/2$ a share and a return $X/2$, where $\Pr(X = 2^k) = 2^{-k}, k = 1, 2, \ldots$. Suppose $X_1, X_2, \ldots$ are i.i.d. according to this distribution and the gambler reinvests all his wealth each time. Thus his wealth S_n at time n is given by

$$S_n = \prod_{i=1}^{n} \frac{X_i}{c}. \tag{6.95}$$

Show that this limit is ∞ or 0, with probability one, accordingly as $c < c^*$ or $c > c^*$. Identify the "fair" entry fee c^*.

More realistically, the gambler should be allowed to keep a proportion $\overline{b} = 1 - b$ of his money in his pocket and invest the rest in the St. Petersburg game. His wealth at time n is then

$$S_n = \prod_{i=1}^{n} \left(\overline{b} + \frac{bX_i}{c} \right). \tag{6.96}$$

Let

$$W(b,c) = \sum_{k=1}^{\infty} 2^{-k} \log \left(1 - b + \frac{b2^k}{c} \right). \tag{6.97}$$

We have

$$S_n \doteq 2^{nW(b,c)} \tag{6.98}$$

Let

$$W^*(c) = \max_{0 \le b \le 1} W(b,c). \tag{6.99}$$

Here are some questions about $W^*(c)$.

(c) For what value of the entry fee c does the optimizing value b^* drop below 1?

(d) How does b^* vary with c?

(e) How does $W^*(c)$ fall off with c?

Note that since $W^*(c) > 0$, for all c , we can conclude that any entry fee c is fair.

Solution: *The St. Petersburg paradox.*

(a) The expected return,

$$EX = \sum_{k=1}^{\infty} p(X = 2^k) 2^k = \sum_{k=1}^{\infty} 2^{-k} 2^k = \sum_{k=1}^{\infty} 1 = \infty. \tag{6.100}$$

Thus the expected return on the game is infinite.

(b) By the strong law of large numbers, we see that

$$\frac{1}{n} \log S_n = \frac{1}{n} \sum_{i=1}^{n} \log X_i - \log c \to E \log X - \log c, \text{w.p.1} \tag{6.101}$$

and therefore S_n goes to infinity or 0 according to whether $E \log X$ is greater or less than $\log c$. Therefore

$$\log c^* = E \log X = \sum_{k=1}^{\infty} k 2^{-k} = 2. \tag{6.102}$$

Therefore a fair entry fee is 2 units if the gambler is forced to invest all his money.

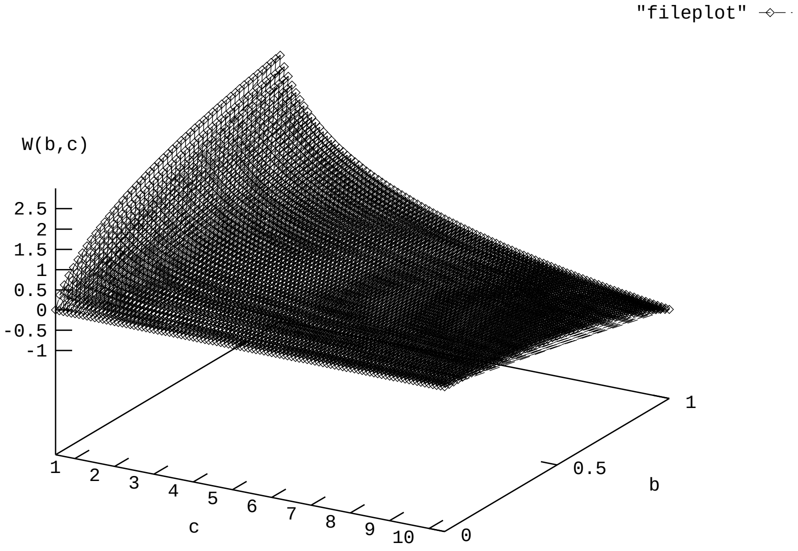


Figure 6.1: St. Petersburg: $W(b,c)$ as a function of b and c.

(c) If the gambler is not required to invest all his money, then the growth rate is

$$W(b,c)=\sum_{k=1}^{\infty}2^{-k}\log\left(1-b+\frac{b2^k}{c}\right). \tag{6.103}$$

For $b=0$, $W=1$, and for $b=1$, $W=E\log X-\log c=2-\log c$. Differentiating to find the optimum value of b, we obtain

$$\frac{\partial W(b,c)}{\partial b}=\sum_{k=1}^{\infty}2^{-k}\frac{1}{\left(1-b+\frac{b2^k}{c}\right)}\left(-1+\frac{2^k}{c}\right) \tag{6.104}$$

Unfortunately, there is no explicit solution for the b that maximizes W for a given value of c, and we have to solve this numerically on the computer.

We have illustrated the results with three plots. The first (Figure 6.1) shows $W(b,c)$ as a function of b and c. The second (Figure 6.2)shows b^* as a function of c and the third (Figure 6.3) shows W^* as a function of c.

From Figure 2, it is clear that b^* is less than 1 for $c>3$. We can also see this analytically by calculating the slope $\frac{\partial W(b,c)}{\partial b}$ at $b=1$.

$$\frac{\partial W(b,c)}{\partial b} \quad = \quad \sum_{k=1}^{\infty}2^{-k}\frac{1}{\left(1-b+\frac{b2^k}{c}\right)}\left(-1+\frac{2^k}{c}\right) \tag{6.105}$$

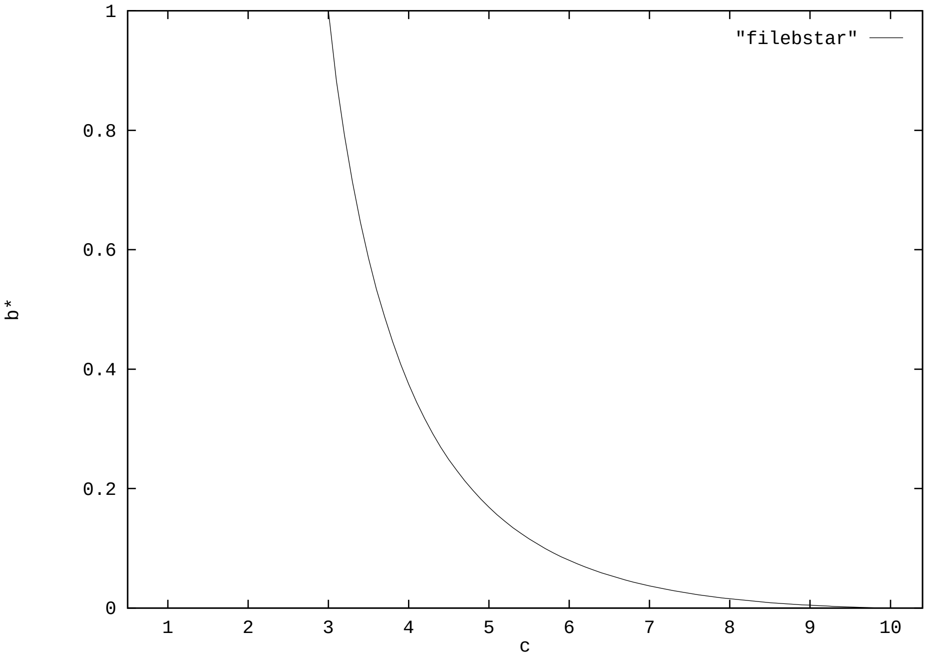


Figure 6.2: St. Petersburg: b^* as a function of c.

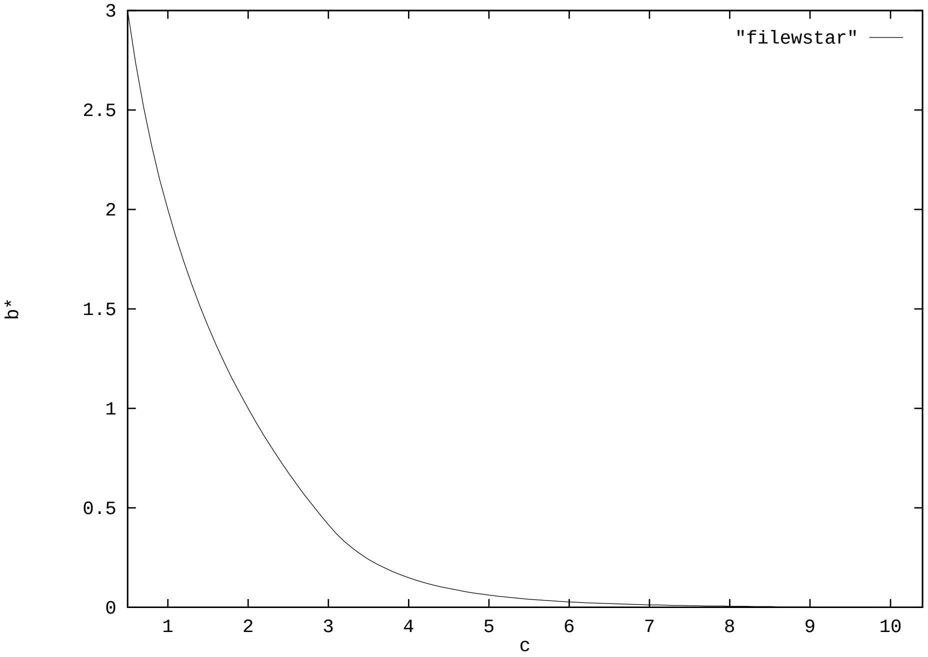


Figure 6.3: St. Petersburg: $W^*(b^*, c)$ as a function of c.

$$= \sum_k \frac{2^{-k}}{\frac{2^k}{c}} \left(\frac{2^k}{d} - 1 \right) \tag{6.106}$$

$$= \sum_{k=1}^{\infty} 2^{-k} - \sum_{k=1}^{\infty} c 2^{-2k} \tag{6.107}$$

$$= 1 - \frac{c}{3} \tag{6.108}$$

which is positive for $c < 3$. Thus for $c < 3$, the optimal value of b lies on the boundary of the region of b's, and for $c > 3$, the optimal value of b lies in the interior.

(d) The variation of b^* with c is shown in Figure 6.2. As $c \to \infty$, $b^* \to 0$. We have a conjecture (based on numerical results) that $b^* \to \frac{1}{\sqrt{2}} c 2^{-c}$ as $c \to \infty$, but we do not have a proof.

(e) The variation of W^* with c is shown in Figure 6.3.

18. **Super St. Petersburg.** Finally, we have the super St. Petersburg paradox, where $\Pr(X = 2^{2^k}) = 2^{-k}, k = 1, 2, \ldots$. Here the expected log wealth is infinite for all $b > 0$, for all c, and the gambler's wealth grows to infinity faster than exponentially for any $b > 0$. But that doesn't mean all investment ratios b are equally good. To see this, we wish to maximize the relative growth rate with respect to some other portfolio, say, $\mathbf{b} = (\frac{1}{2}, \frac{1}{2})$. Show that there exists a unique b maximizing

$$E \ln \frac{(\bar{b} + bX/c)}{(\frac{1}{2} + \frac{1}{2}X/c)}$$

and interpret the answer.

Solution: *Super St. Petersburg.* With $\Pr(X = 2^{2^k}) = 2^{-k}, k = 1, 2, \ldots$, we have

$$E \log X = \sum_k 2^{-k} \log 2^{2^k} = \infty, \tag{6.109}$$

and thus with any constant entry fee, the gambler's money grows to infinity faster than exponentially, since for any $b > 0$,

$$W(b, c) = \sum_{k=1}^{\infty} 2^{-k} \log \left(1 - b + \frac{b 2^{2^k}}{c} \right) > \sum 2^{-k} \log \frac{b 2^{2^k}}{c} = \infty. \tag{6.110}$$

But if we wish to maximize the wealth relative to the $(\frac{1}{2}, \frac{1}{2})$ portfolio, we need to maximize

$$J(b, c) = \sum_k 2^{-k} \log \frac{(1-b) + \frac{b 2^{2^k}}{c}}{\frac{1}{2} + \frac{1}{2} \frac{2^{2^k}}{c}} \tag{6.111}$$

As in the case of the St. Petersburg problem, we cannot solve this problem explicitly. In this case, a computer solution is fairly straightforward, although there are some

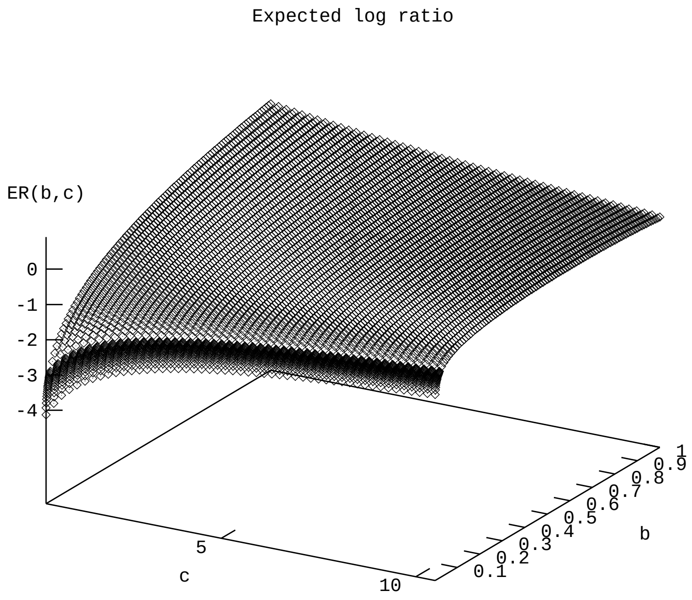


Figure 6.4: Super St. Petersburg: $J(b,c)$ as a function of b and c.

complications. For example, for $k = 6$, 2^{2^k} is outside the normal range of numbers representable on a standard computer. However, for $k \geq 6$, we can approximate the ratio within the log by $\frac{b}{0.5}$ without any loss of accuracy. Using this, we can do a simple numerical computation as in the previous problem.

As before, we have illustrated the results with three plots. The first (Figure 6.4) shows $J(b,c)$ as a function of b and c. The second (Figure 6.5)shows b^* as a function of c and the third (Figure 6.6) shows J^* as a function of c.

These plots indicate that for large values of c, the optimum strategy is not to put all the money into the game, even though the money grows at an infinite rate. There exists a unique b^* which maximizes the expected ratio, which therefore causes the wealth to grow to infinity at the fastest possible rate. Thus there exists an optimal b^* even when the log optimal portfolio is undefined.

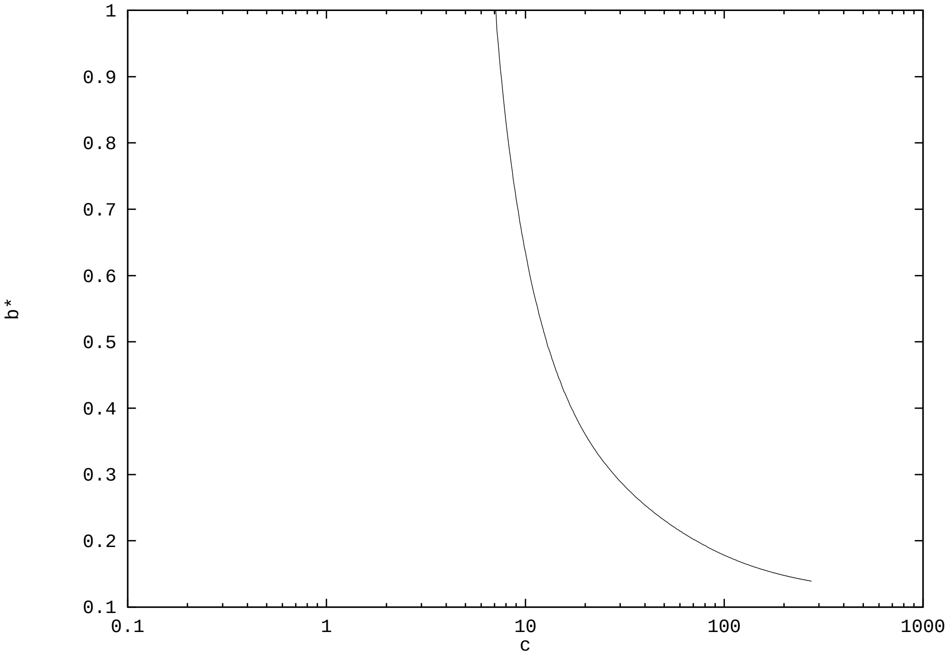


Figure 6.5: Super St. Petersburg: b^* as a function of c.

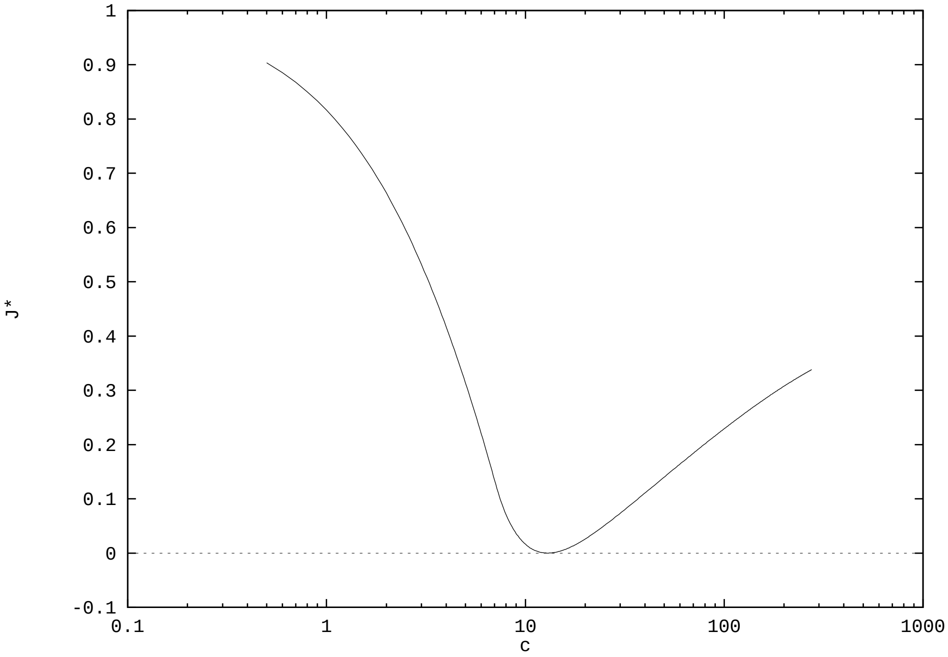


Figure 6.6: Super St. Petersburg: $J^*(b^*, c)$ as a function of c.

Chapter 7

Channel Capacity

1. **Preprocessing the output.** One is given a communication channel with transition probabilities $p(y|x)$ and channel capacity $C = \max_{p(x)} I(X;Y)$. A helpful statistician preprocesses the output by forming $\tilde{Y} = g(Y)$. He claims that this will strictly improve the capacity.

   (a) Show that he is wrong.

   (b) Under what conditions does he not strictly decrease the capacity?

   Solution: *Preprocessing the output.*

   (a) The statistician calculates $\tilde{Y} = g(Y)$. Since $X \to Y \to \tilde{Y}$ forms a Markov chain, we can apply the data processing inequality. Hence for every distribution on x,

   $$I(X;Y) \geq I(X;\tilde{Y}). \tag{7.1}$$

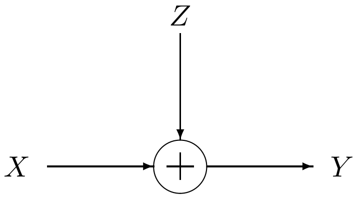
   Let $\tilde{p}(x)$ be the distribution on x that maximizes $I(X;\tilde{Y})$. Then

   $$C = \max_{p(x)} I(X;Y) \geq I(X;Y)_{p(x)=\tilde{p}(x)} \geq I(X;\tilde{Y})_{p(x)=\tilde{p}(x)} = \max_{p(x)} I(X;\tilde{Y}) = \tilde{C}. \tag{7.2}$$

   Thus, the statistician is wrong and processing the output does not increase capacity.

   (b) We have equality (no decrease in capacity) in the above sequence of inequalities only if we have equality in the data processing inequality, i.e., for the distribution that maximizes $I(X;\tilde{Y})$, we have $X \to \tilde{Y} \to Y$ forming a Markov chain.

2. **An additive noise channel.** Find the channel capacity of the following discrete memoryless channel:



where $\Pr\{Z=0\} = \Pr\{Z=a\} = \frac{1}{2}$. The alphabet for x is $\mathbf{X} = \{0,1\}$. Assume that Z is independent of X.

Observe that the channel capacity depends on the value of a.

Solution: *A sum channel.*

$$Y = X + Z \qquad X \in \{0,1\}, \quad Z \in \{0,a\} \tag{7.3}$$

We have to distinguish various cases depending on the values of a.

$a=0$ In this case, $Y=X$, and $\max I(X;Y) = \max H(X) = 1$. Hence the capacity is 1 bit per transmission.

$a \neq 0, \pm 1$ In this case, Y has four possible values $0, 1, a$ and $1+a$. Knowing Y, we know the X which was sent, and hence $H(X|Y)=0$. Hence $\max I(X;Y) = \max H(X) = 1$, achieved for an uniform distribution on the input X.

$a=1$ In this case Y has three possible output values, $0, 1$ and 2, and the channel is identical to the binary erasure channel discussed in class, with $a = 1/2$. As derived in class, the capacity of this channel is $1-a = 1/2$ bit per transmission.

$a=-1$ This is similar to the case when $a=1$ and the capacity here is also $1/2$ bit per transmission.

3. **Channels with memory have higher capacity.** Consider a binary symmetric channel with $Y_i = X_i \oplus Z_i$, where $\oplus$ is mod 2 addition, and $X_i, Y_i \in \{0,1\}$.

   Suppose that $\{Z_i\}$ has constant marginal probabilities $\Pr\{Z_i = 1\} = p = 1 - \Pr\{Z_i = 0\}$, but that $Z_1, Z_2, \ldots, Z_n$ are not necessarily independent. Assume that Z^n is independent of the input X^n. Let $C = 1 - H(p, 1-p)$. Show that $\max_{p(x_1,x_2,\ldots,x_n)} I(X_1, X_2, \ldots, X_n; Y_1, Y_2, ..$ nC.

   Solution: *Channels with memory have a higher capacity.*

$$Y_i = X_i \oplus Z_i, \tag{7.4}$$

where

$$Z_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases} \tag{7.5}$$

and Z_i are not independent.

$$\begin{aligned}
& I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) \\
&= H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n | Y_1, Y_2, \ldots, Y_n) \\
&= H(X_1, X_2, \ldots, X_n) - H(Z_1, Z_2, \ldots, Z_n | Y_1, Y_2, \ldots, Y_n) \\
&\geq H(X_1, X_2, \ldots, X_n) - H(Z_1, Z_2, \ldots, Z_n) && (7.6) \\
&\geq H(X_1, X_2, \ldots, X_n) - \sum H(Z_i) && (7.7) \\
&= H(X_1, X_2, \ldots, X_n) - nH(p) && (7.8) \\
&= n - nH(p), && (7.9)
\end{aligned}$$

if $X_1, X_2, \ldots, X_n$ are chosen i.i.d. $\sim$ Bern($\frac{1}{2}$). The capacity of the channel with memory over n uses of the channel is

$$\begin{aligned}
nC^{(n)} &= \max_{p(x_1, x_2, \ldots, x_n)} I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) && (7.10) \\
&\geq I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n)_{p(x_1, x_2, \ldots, x_n) = \mathrm{Bern}(\frac{1}{2})} && (7.11) \\
&\geq n(1 - H(p)) && (7.12) \\
&= nC. && (7.13)
\end{aligned}$$

Hence channels with memory have higher capacity. The intuitive explanation for this result is that the correlation between the noise decreases the effective noise; one could use the information from the past samples of the noise to combat the present noise.

4. **Channel capacity.** Consider the discrete memoryless channel $Y = X + Z$ (mod 11), where

$$Z = \begin{pmatrix} 1, & 2, & 3 \\ 1/3, & 1/3, & 1/3 \end{pmatrix}$$

and $X \in \{0, 1, \ldots, 10\}$. Assume that Z is independent of X.

(a) Find the capacity.

(b) What is the maximizing $p^*(x)$?

Solution: *Channel capacity.*

$$Y = X + Z (\text{mod } 11) \qquad (7.14)$$

where

$$Z = \begin{cases} 1 & \text{with probability} 1/3 \\ 2 & \text{with probability} 1/3 \\ 3 & \text{with probability} 1/3 \end{cases} \qquad (7.15)$$

In this case,

$$H(Y|X) = H(Z|X) = H(Z) = \log 3, \qquad (7.16)$$

independent of the distribution of X, and hence the capacity of the channel is

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) && (7.17)\\ &= \max_{p(x)} H(Y) - H(Y|X) && (7.18)\\ &= \max_{p(x)} H(Y) - \log 3 && (7.19)\\ &= \log 11 - \log 3, && (7.20) \end{aligned}$$

which is attained when Y has a uniform distribution, which occurs (by symmetry) when X has a uniform distribution.

(a) The capacity of the channel is $\log \frac{11}{3}$ bits/transmission.

(b) The capacity is achieved by an uniform distribution on the inputs. $p(X = i) = \frac{1}{11}$ for $i = 0, 1, \ldots, 10$.

5. **Using two channels at once.** Consider two discrete memoryless channels $(\mathcal{X}_1, p(y_1 \mid x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2 \mid x_2), \mathcal{Y}_2)$ with capacities C_1 and C_2 respectively. A new channel $(\mathcal{X}_1 \times \mathcal{X}_2, p(y_1 \mid x_1) \times p(y_2 \mid x_2), \mathcal{Y}_1 \times \mathcal{Y}_2)$ is formed in which $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$, are simultaneously sent, resulting in y_1, y_2. Find the capacity of this channel.

Solution: *Using two channels at once.* Suppose we are given two channels, $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2)$, which we can use at the same time. We can define the product channel as the channel, $(\mathcal{X}_1 \times \mathcal{X}_2, p(y_1, y_2|x_1, x_2) = p(y_1|x_1)p(y_2|x_2), \mathcal{Y}_1 \times \mathcal{Y}_2)$. To find the capacity of the product channel, we must find the distribution $p(x_1, x_2)$ on the input alphabet $\mathcal{X}_1 \times \mathcal{X}_2$ that maximizes $I(X_1, X_2; Y_1, Y_2)$. Since the joint distribution

$$p(x_1, x_2, y_1, y_2) = p(x_1, x_2)p(y_1|x_1)p(y_2|x_2), \qquad (7.21)$$

$Y_1 \to X_1 \to X_2 \to Y_2$ forms a Markov chain and therefore

$$\begin{aligned} I(X_1, X_2; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) && (7.22)\\ &= H(Y_1, Y_2) - H(Y_1|X_1, X_2) - H(Y_2|X_1, X_2) && (7.23)\\ &= H(Y_1, Y_2) - H(Y_1|X_1) - H(Y_2|X_2) && (7.24)\\ &\le H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) && (7.25)\\ &= I(X_1; Y_1) + I(X_2; Y_2), && (7.26) \end{aligned}$$

where (7.23) and (7.24) follow from Markovity, and we have equality in (7.25) if Y_1 and Y_2 are independent. Equality occurs when X_1 and X_2 are independent. Hence

$$\begin{aligned} C &= \max_{p(x_1,x_2)} I(X_1, X_2; Y_1, Y_2) && (7.27)\\ &\le \max_{p(x_1,x_2)} I(X_1; Y_1) + \max_{p(x_1,x_2)} I(X_2; Y_2) && (7.28)\\ &= \max_{p(x_1)} I(X_1; Y_1) + \max_{p(x_2)} I(X_2; Y_2) && (7.29)\\ &= C_1 + C_2. && (7.30) \end{aligned}$$

with equality iff $p(x_1, x_2) = p^*(x_1)p^*(x_2)$ and $p^*(x_1)$ and $p^*(x_2)$ are the distributions that maximize C_1 and C_2 respectively.

6. **Noisy typewriter.** Consider a 26-key typewriter.

    (a) If pushing a key results in printing the associated letter, what is the capacity C in bits?

    (b) Now suppose that pushing a key results in printing that letter or the next (with equal probability). Thus $A \to A$ or $B, \ldots, Z \to Z$ or A. What is the capacity?

    (c) What is the highest rate code with block length one that you can find that achieves *zero* probability of error for the channel in part (b) .

    Solution: *Noisy typewriter.*

    (a) If the typewriter prints out whatever key is struck, then the output, Y, is the same as the input, X, and

    $$C = \max I(X;Y) = \max H(X) = \log 26, \tag{7.31}$$

    attained by a uniform distribution over the letters.

    (b) In this case, the output is either equal to the input (with probability $\frac{1}{2}$) or equal to the next letter (with probability $\frac{1}{2}$). Hence $H(Y|X) = \log 2$ independent of the distribution of X, and hence

    $$C = \max I(X;Y) = \max H(Y) - \log 2 = \log 26 - \log 2 = \log 13, \tag{7.32}$$

    attained for a uniform distribution over the output, which in turn is attained by a uniform distribution on the input.

    (c) A simple zero error block length one code is the one that uses every alternate letter, say A,C,E,. . . ,W,Y. In this case, none of the codewords will be confused, since A will produce either A or B, C will produce C or D, etc. The rate of this code,

    $$R = \frac{\log(\#\text{ codewords})}{\text{Block length}} = \frac{\log 13}{1} = \log 13. \tag{7.33}$$

    In this case, we can achieve capacity with a simple code with zero error.

7. **Cascade of binary symmetric channels.** Show that a cascade of n identical independent binary symmetric channels,

    $$X_0 \to \boxed{\text{BSC}} \to_1 \to \cdots \to X_{n-1} \to \boxed{\text{BSC}} \to_n$$

    each with raw error probability p, is equivalent to a single BSC with error probability $\frac{1}{2}(1-(1-2p)^n)$ and hence that $\lim_{n\to\infty} I(X_0;X_n) = 0$ if $p \neq 0, 1$. No encoding or decoding takes place at the intermediate terminals $X_1, \ldots, X_{n-1}$. Thus the capacity of the cascade tends to zero.

    Solution: *Cascade of binary symmetric channels.* There are many ways to solve this problem. One way is to use the singular value decomposition of the transition probability matrix for a single BSC.

Let,

$$A = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

be the transition probability matrix for our BSC. Then the transition probability matrix for the cascade of n of these BSC's is given by,

$$A_n = A^n.$$

Now check that,

$$A = T^{-1} \begin{bmatrix} 1 & 0 \\ 0 & 1-2p \end{bmatrix} T$$

where,

$$T = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Using this we have,

$$\begin{aligned} A_n &= A^n \\ &= T^{-1} \begin{bmatrix} 1 & 0 \\ 0 & (1-2p)^n \end{bmatrix} T \\ &= \begin{bmatrix} \frac{1}{2}(1+(1-2p)^n) & \frac{1}{2}(1-(1-2p)^n) \\ \frac{1}{2}(1-(1-2p)^n) & \frac{1}{2}(1+(1-2p)^n) \end{bmatrix}. \end{aligned}$$

From this we see that the cascade of n BSC's is also a BSC with probablility of error,

$$p_n = \frac{1}{2}(1-(1-2p)^n).$$

The matrix, T, is simply the matrix of eigenvectors of A.

This problem can also be solved by induction on n.

Probably the simplest way to solve the problem is to note that the probability of error for the cascade channel is simply the sum of the odd terms of the binomial expansion of $(x+y)^n$ with $x = p$ and $y = 1-p$. But this can simply be written as $\frac{1}{2}(x+y)^n - \frac{1}{2}(y-x)^n = \frac{1}{2}(1-(1-2p)^n$.

8. **The Z channel.** The Z-channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by the following matrix:

$$Q = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix} \qquad x, y \in \{0,1\}$$

Find the capacity of the Z-channel and the maximizing input probability distribution.

Solution: *The Z channel.* First we express $I(X;Y)$, the mutual information between the input and output of the Z-channel, as a function of $x = \Pr(X = 1)$:

$$\begin{aligned} H(Y|X) &= \Pr(X = 0) \cdot 0 + \Pr(X = 1) \cdot 1 = x \\ H(Y) &= H(\Pr(Y = 1)) = H(x/2) \\ I(X;Y) &= H(Y) - H(Y|X) = H(x/2) - x \end{aligned}$$

Since $I(X;Y) = 0$ when $x = 0$ and $x = 1$, the maximum mutual information is obtained for some value of x such that $0 < x < 1$.

Using elementary calculus, we determine that

$$\frac{d}{dx} I(X;Y) = \frac{1}{2} \log_2 \frac{1 - x/2}{x/2} - 1 \,,$$

which is equal to zero for $x = 2/5$. (It is reasonable that $\Pr(X = 1) < 1/2$ because $X = 1$ is the noisy input to the channel.) So the capacity of the Z-channel in bits is $H(1/5) - 2/5 = 0.722 - 0.4 = 0.322$.

9. **Suboptimal codes**. For the Z channel of the previous problem, assume that we choose a $(2^{nR}, n)$ code at random, where each codeword is a sequence of *fair* coin tosses. This will not achieve capacity. Find the maximum rate R such that the probability of error $P_e^{(n)}$, averaged over the randomly generated codes, tends to zero as the block length n tends to infinity.

   Solution: *Suboptimal codes.* From the proof of the channel coding theorem, it follows that using a random code with codewords generated according to probability $p(x)$, we can send information at a rate $I(X;Y)$ corresponding to that $p(x)$ with an arbitrarily low probability of error. For the Z channel described in the previous problem, we can calculate $I(X;Y)$ for a uniform distribution on the input. The distribution on Y is (3/4, 1/4), and therefore

$$I(X;Y) = H(Y) - H(Y|X) = H(\frac{3}{4}, \frac{1}{4}) - \frac{1}{2} H(\frac{1}{2}, \frac{1}{2}) = \frac{3}{2} - \frac{3}{4} \log 3. \qquad (7.34)$$

10. **Zero-error capacity.** A channel with alphabet $\{0, 1, 2, 3, 4\}$ has transition probabilities of the form

$$p(y|x) = \begin{cases} 1/2 & \text{if } y = x \pm 1 \bmod 5 \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Compute the capacity of this channel in bits.

   (b) The zero-error capacity of a channel is the number of bits per channel use that can be transmitted with zero probability of error. Clearly, the zero-error capacity of this pentagonal channel is at least 1 bit (transmit 0 or 1 with probability 1/2). Find a block code that shows that the zero-error capacity is greater than 1 bit. Can you estimate the exact value of the zero-error capacity?

   (*Hint*: Consider codes of length 2 for this channel.)

   The zero-error capacity of this channel was finally found by Lovasz[8].

Solution: *Zero-error capacity.*

(a) Since the channel is symmetric, it is easy to compute its capacity:

$$\begin{aligned} H(Y|X) &= 1 \\ I(X;Y) &= H(Y) - H(Y|X) = H(Y) - 1\,. \end{aligned}$$

So mutual information is maximized when Y is uniformly distributed, which occurs when the input X is uniformly distributed. Therefore the capacity in bits is $C = \log_2 5 - 1 = \log_2 2.5 = 1.32$.

(b) Let us construct a block code consisting of 2-tuples. We need more than 4 codewords in order to achieve capacity greater than 2 bits, so we will pick 5 codewords with distinct first symbols: $\{0a, 1b, 2c, 3d, 4e\}$. We must choose a, b, c, d, e so that the receiver will be able to determine which codeword was transmitted. A simple repetition code will not work, since if, say, 22 is transmitted, then 11 might be received, and the receiver could not tell whether the codeword was 00 or 22. Instead, using codewords of the form (i+1 mod 5, 2i+1 mod 5) yields the code 11,23,30,42,04.

Here is the decoding table for the pentagon channel:

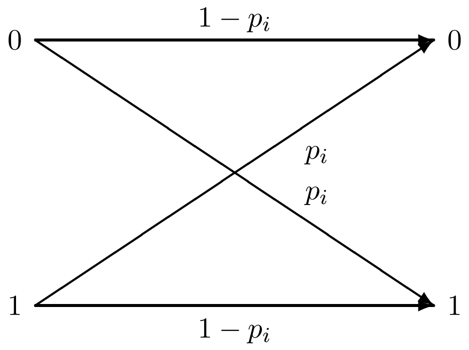
0 4 0 . 4 3 . 2 3 2 0 1 0 1 . 3 4 . 3 4 . 1 2 1 2

It is amusing to note that the five pairs that cannot be received are exactly the 5 codewords.

Then whenever xy is received, there is exactly one possible codeword. (Each codeword will be received as one of 4 possible 2-tuples; so there are 20 possible received 2-tuples, out of a total of 25.) Since there are 5 possible error-free messages with 2 channel uses, the zero-error capacity of this channel is at least $\frac{1}{2}\log_2 5 = 1.161$ bits.

In fact, the zero-error capacity of this channel is exactly $\frac{1}{2}\log_2 5$. This result was obtained by László Lovász, "On the Shannon capacity of a graph," *IEEE Transactions on Information Theory*, Vol IT-25, pp. 1–7, January 1979. The first results on zero-error capacity are due to Claude E. Shannon, "The zero-error capacity of a noisy channel," *IEEE Transactions on Information Theory*, Vol IT-2, pp. 8–19, September 1956, reprinted in *Key Papers in the Development of Information Theory*, David Slepian, editor, IEEE Press, 1974.

11. **Time-varying channels.** Consider a time-varying discrete *memoryless* channel. Let $Y_1, Y_2, \ldots, Y_n$ be conditionally independent given $X_1, X_2, \ldots, X_n$, with conditional distribution given by $p(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{n} p_i(y_i \mid x_i)$.



Let $\mathbf{X} = (X_1, X_2, \ldots, X_n), \mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. Find $\max_{p(\mathbf{x})} I(\mathbf{X};\mathbf{Y})$.

Solution: *Time-varying channels.*

We can use the same chain of inequalities as in the proof of the converse to the channel coding theorem. Hence

$$\begin{aligned} I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) && (7.35)\\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1,\ldots,Y_{i-1},X^n) && (7.36)\\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i), && (7.37)\end{aligned}$$

since by the definition of the channel, Y_i depends only on X_i and is conditionally independent of everything else. Continuing the series of inequalities, we have

$$\begin{aligned} I(X^n;Y^n) &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) && (7.38)\\ &\le \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) && (7.39)\\ &\le \sum_{i=1}^{n} (1-h(p_i)), && (7.40)\end{aligned}$$

with equality if $X_1, X_2, \ldots, X_n$ is chosen i.i.d. $\sim$ Bern(1/2). Hence

$$\max_{p(\mathbf{x})} I(X_1,X_2,\ldots,X_n;Y_1,Y_2,\ldots,Y_n) = \sum_{i=1}^{n}(1-h(p_i)). \qquad (7.41)$$

12. **Unused symbols.** Show that the capacity of the channel with probability transition matrix

$$P_{y|x} = \begin{bmatrix} 2/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 \\ 0 & 1/3 & 2/3 \end{bmatrix} \qquad (7.42)$$

is achieved by a distribution that places zero probability on one of input symbols. What is the capacity of this channel? Give an intuitive reason why that letter is not used.

Solution: *Unused symbols* Let the probabilities of the three input symbols be p_1, p_2 and p_3. Then the probabilities of the three output symbols can be easily calculated to be $(\frac{2}{3}p_1 + \frac{1}{3}p_2, \frac{1}{3}, \frac{1}{3}p_2 + \frac{2}{3}p_3)$, and therefore

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) && (7.43)\\ &= H(\frac{2}{3}p_1 + \frac{1}{3}p_2, \frac{1}{3}, \frac{1}{3}p_2 + \frac{2}{3}p_3) - (p_1+p_3)H(\frac{2}{3},\frac{1}{3}) - p_2 \log 3 && (7.44)\\ &= H(\frac{1}{3} + \frac{1}{3}(p_1-p_3), \frac{1}{3}, \frac{1}{3} - \frac{1}{3}(p_1-p_3)) - (p_1+p_3)H(\frac{2}{3},\frac{1}{3}) - (1-p_1-p_3)\log 3 && (7.45) \end{aligned}$$

where we have substituted $p_2 = 1 - p_1 - p_3$. Now if we fix $p_1 + p_3$, then the second and third terms are fixed, and the first term is maximized if $p_1 - p_3 = 0$, i.e., if $p_1 = p_3$. (The same conclusion can be drawn from the symmetry of the problem.)

Now setting $p_1 = p_3$, we have

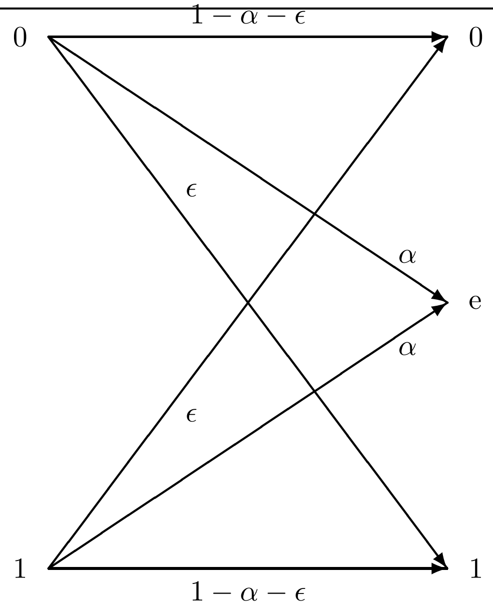
$$\begin{aligned} I(X;Y) &= H(\frac{1}{3},\frac{1}{3},\frac{1}{3}) - (p_1+p_3)H(\frac{2}{3},\frac{1}{3}) - (1-p_1-p_3)\log 3 && (7.46)\\ &= \log 3 - (p_1+p_3)H(\frac{2}{3},\frac{1}{3}) - (1-p_1-p_3)\log 3 && (7.47)\\ &= (p_1+p_3)(\log 3 - H(\frac{2}{3},\frac{1}{3})) && (7.48) \end{aligned}$$

which is maximized if $p_1 + p_3$ is as large as possible (since $\log 3 > H(\frac{2}{3},\frac{1}{3})$). Therefore the maximizing distribution corresponds to $p_1 + p_3 = 1$, $p_1 = p_3$, and therefore $(p_1, p_2, p_3) = (\frac{1}{2}, 0, \frac{1}{2})$. The capacity of this channel $= \log 3 - H(\frac{2}{3},\frac{1}{3}) = \log 3 - (\log 3 - \frac{2}{3}) = \frac{2}{3}$ bits.

The intuitive reason why $p_2 = 0$ for the maximizing distribution is that conditional on the input being 2, the output is uniformly distributed. The same uniform output distribution can be achieved without using the symbol 2 (by setting $p_1 = p_3$), and therefore the use of symbol 2 does not add any information (it does not change the entropy of the output and the conditional entropy $H(Y|X = 2)$ is the maximum possible, i.e., $\log 3$, so any positive probability for symbol 2 will only reduce the mutual information.

Note that not using a symbol is optimal only if the uniform output distribution can be achieved without use of that symbol. For example, in the Z channel example above, both symbols are used, even though one of them gives a conditionally uniform distribution on the output.

13. **Erasures and errors in a binary channel.** Consider a channel with binary inputs that has both erasures and errors. Let the probability of error be ϵ and the probability of erasure be α, so the the channel is as illustrated below:



(a) Find the capacity of this channel.

(b) Specialize to the case of the binary symmetric channel ($\alpha = 0$).

(c) Specialize to the case of the binary erasure channel ($\epsilon = 0$).

Solution:

(a) As with the examples in the text, we set the input distribution for the two inputs to be π and $1-\pi$. Then

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) && (7.49)\\ &= \max_{p(x)} (H(Y) - H(Y|X)) && (7.50)\\ &= \max_{p(x)} H(Y) - H(1-\epsilon-\alpha, \alpha, \epsilon). && (7.51) \end{aligned}$$

As in the case of the erasure channel, the maximum value for $H(Y)$ cannot be $\log 3$, since the probability of the erasure symbol is α independent of the input distribution. Thus,

$$\begin{aligned} H(Y) &= H(\pi(1-\epsilon-\alpha) + (1-\pi)\epsilon, \alpha, (1-\pi)(1-\epsilon-\alpha) + \pi\epsilon) && (7.52)\\ &= H(\alpha) + (1-\alpha) H\left(\frac{\pi+\epsilon-\pi\alpha-2\pi\epsilon}{1-\alpha}, \frac{1-\pi-\epsilon+2\epsilon\pi-\alpha+\alpha\pi}{1-\alpha}\right) && (7.53)\\ &\le H(\alpha) + (1-\alpha) && (7.54) \end{aligned}$$

with equality iff $\frac{\pi+\epsilon-\pi\alpha-2\pi\epsilon}{1-\alpha} = \frac{1}{2}$, which can be achieved by setting $\pi = \frac{1}{2}$. (The fact that $\pi = 1-\pi = \frac{1}{2}$ is the optimal distribution should be obvious from the symmetry of the problem, even though the channel is not weakly symmetric.)

Therefore the capacity of this channel is

$$\begin{aligned} C &= H(\alpha) + 1 - \alpha - H(1-\alpha-\epsilon, \alpha, \epsilon) && (7.55) \\ &= H(\alpha) + 1 - \alpha - H(\alpha) - (1-\alpha) H\left(\frac{1-\alpha-\epsilon}{1-\alpha}, \frac{\epsilon}{1-\alpha}\right) && (7.56) \\ &= (1-\alpha)\left(1 - H\left(\frac{1-\alpha-\epsilon}{1-\alpha}, \frac{\epsilon}{1-\alpha}\right)\right) && (7.57) \end{aligned}$$

(b) Setting $\alpha = 0$, we get

$$C = 1 - H(\epsilon), \qquad (7.58)$$

which is the capacity of the binary symmetric channel.

(c) Setting $\epsilon = 0$, we get

$$C = 1 - \alpha \qquad (7.59)$$

which is the capacity of the binary erasure channel.

14. **Channels with dependence between the letters.** Consider the following channel over a binary alphabet that takes in two bit symbols and produces a two bit output, as determined by the following mapping: $00 \to 01$, $01 \to 10$, $10 \to 11$, and $11 \to 00$. Thus if the two bit sequence 01 is the input to the channel, the output is 10 with probability 1. Let X_1, X_2 denote the two input symbols and Y_1, Y_2 denote the corresponding output symbols.

(a) Calculate the mutual information $I(X_1, X_2; Y_1, Y_2)$ as a function of the input distribution on the four possible pairs of inputs.

(b) Show that the capacity of a pair of transmissions on this channel is 2 bits.

(c) Show that under the maximizing input distribution, $I(X_1; Y_1) = 0$.

Thus the distribution on the input sequences that achieves capacity does not necessarily maximize the mutual information between individual symbols and their corresponding outputs.

Solution:

(a) If we look at pairs of inputs and pairs of outputs, this channel is a noiseless four input four output channel. Let the probabilities of the four input pairs be p_{00}, p_{01}, p_{10} and p_{11} respectively. Then the probability of the four pairs of output bits is p_{11}, p_{00}, p_{01} and p_{10} respectively, and

$$\begin{aligned} I(X_1, X_2; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2 | X_1, X_2) && (7.60) \\ &= H(Y_1, Y_2) - 0 && (7.61) \\ &= H(p_{11}, p_{00}, p_{01}, p_{10}) && (7.62) \end{aligned}$$

(b) The capacity of the channel is achieved by a uniform distribution over the inputs, which produces a uniform distribution on the output pairs

$$C = \max_{p(x_1,x_2)} I(X_1, X_2; Y_1, Y_2) = 2 \text{ bits} \tag{7.63}$$

and the maximizing $p(x_1, x_2)$ puts probability $\frac{1}{4}$ on each of the pairs 00, 01, 10 and 11.

(c) To calculate $I(X_1;Y_1)$, we need to calculate the joint distribution of X_1 and Y_1. The joint distribution of X_1X_2 and Y_1Y_2 under an uniform input distribution is given by the following matrix

$X_1X_2 \backslash Y_1Y_2$	00	01	10	11
00	0	$\frac{1}{4}$	0	0
01	0	0	$\frac{1}{4}$	0
10	0	0	0	$\frac{1}{4}$
11	$\frac{1}{4}$	0	0	0

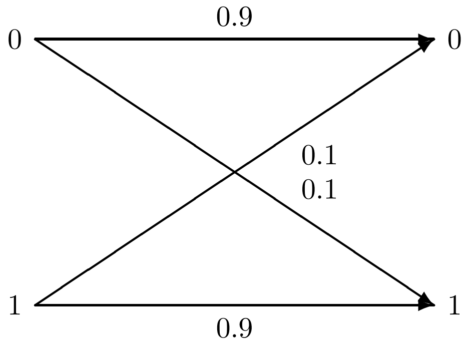
From this, it is easy to calculate the joint distribution of X_1 and Y_1 as

$X_1 \backslash Y_1$	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{4}$	$\frac{1}{4}$

and therefore we can see that the marginal distributions of X_1 and Y_1 are both $(1/2, 1/2)$ and that the joint distribution is the product of the marginals, i.e., X_1 is independent of Y_1, and therefore $I(X_1;Y_1) = 0$.

Thus the distribution on the input sequences that achieves capacity does not necessarily maximize the mutual information between individual symbols and their corresponding outputs.

15. **Jointly typical sequences.** As we did in problem 13 of Chapter 3 for the typical set for a single random variable, we will calculate the jointly typical set for a pair of random variables connected by a binary symmetric channel, and the probability of error for jointly typical decoding for such a channel.



We will consider a binary symmetric channel with crossover probability 0.1. The input distribution that achieves capacity is the uniform distribution, i.e., $p(x) = (1/2, 1/2)$, which yields the joint distribution $p(x,y)$ for this channel is given by

$X \backslash Y$	0	1
0	0.45	0.05
1	0.05	0.45

The marginal distribution of Y is also $(1/2, 1/2)$.

(a) Calculate $H(X)$, $H(Y)$, $H(X,Y)$ and $I(X;Y)$ for the joint distribution above.

(b) Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. according the Bernoulli(1/2) distribution. Of the 2^n possible input sequences of length n, which of them are typical, i.e., member of $A_\epsilon^{(n)}(X)$ for $\epsilon = 0.2$? Which are the typical sequences in $A_\epsilon^{(n)}(Y)$?

(c) The jointly typical set $A_\epsilon^{(n)}(X,Y)$ is defined as the set of sequences that satisfy equations (7.35-7.37). The first two equations correspond to the conditions that x^n and y^n are in $A_\epsilon^{(n)}(X)$ and $A_\epsilon^{(n)}(Y)$ respectively. Consider the last condition, which can be rewritten to state that $-\frac{1}{n}\log p(x^n, y^n) \in (H(X,Y)-\epsilon, H(X,Y)+\epsilon)$. Let k be the number of places in which the sequence x^n differs from y^n (k is a function of the two sequences). Then we can write

$$\begin{aligned} p(x^n, y^n) &= \prod_{i=1}^n p(x_i, y_i) && (7.64)\\ &= (0.45)^{n-k}(0.05)^k && (7.65)\\ &= \left(\frac{1}{2}\right)^n (1-p)^{n-k}p^k && (7.66) \end{aligned}$$

An alternative way at looking at this probability is to look at the binary symmetric channel as in additive channel $Y = X \oplus Z$, where Z is a binary random variable that is equal to 1 with probability p, and is independent of X. In this case,

$$\begin{aligned} p(x^n, y^n) &= p(x^n)p(y^n|x^n) && (7.67)\\ &= p(x^n)p(z^n|x^n) && (7.68)\\ &= p(x^n)p(z^n) && (7.69)\\ &= \left(\frac{1}{2}\right)^n (1-p)^{n-k}p^k && (7.70) \end{aligned}$$

Show that the condition that (x^n, y^n) being jointly typical is equivalent to the condition that x^n is typical and $z^n = y^n - x^n$ is typical.

(d) We now calculate the size of $A_\epsilon^{(n)}(Z)$ for $n = 25$ and $\epsilon = 0.2$. As in problem 13 of Chapter 3, here is a table of the probabilities and numbers of sequences of with k ones

k	$\binom{n}{k}$	$\binom{n}{k}p^k(1-p)^{n-k}$	$-\frac{1}{n}\log p(x^n)$
0	1	0.071790	0.152003
1	25	0.199416	0.278800
2	300	0.265888	0.405597
3	2300	0.226497	0.532394
4	12650	0.138415	0.659191
5	53130	0.064594	0.785988
6	177100	0.023924	0.912785
7	480700	0.007215	1.039582
8	1081575	0.001804	1.166379
9	2042975	0.000379	1.293176
10	3268760	0.000067	1.419973
11	4457400	0.000010	1.546770
12	5200300	0.000001	1.673567

(Sequences with more than 12 ones are omitted since their total probability is negligible (and they are not in the typical set).)

What is the size of the set $A_\epsilon^{(n)}(Z)$?

(e) Now consider random coding for the channel, as in the proof of the channel coding theorem. Assume that 2^{nR} codewords $X^n(1), X^n(2), \ldots, X^n(2^{nR})$ are chosen uniformly over the 2^n possible binary sequences of length n. One of these codewords is chosen and sent over the channel. The receiver looks at the received sequence and tries to find a codeword in the code that is jointly typical with the received sequence. As argued above, this corresponds to finding a codeword $X^n(i)$ such that $Y^n - X^n(i) \in A_\epsilon^{(n)}(Z)$. For a fixed codeword $x^n(i)$, what is the probability that the received sequence Y^n is such that $(x^n(i), Y^n)$ is jointly typical?

(f) Now consider a particular received sequence $y^n = 000000\ldots0$, say. Assume that we choose a sequence X^n at random, uniformly distributed among all the 2^n possible binary n-sequences. What is the probability that the chosen sequence is jointly typical with this y^n? (Hint: this is the probability of all sequences x^n such that $y^n - x^n \in A_\epsilon^{(n)}(Z)$.)

(g) Now consider a code with $2^9 = 512$ codewords of length 12 chosen at random, uniformly distributed among all the 2^n sequences of length $n = 25$. One of these codewords, say the one corresponding to $i = 1$, is chosen and sent over the channel. As calculated in part (e), the received sequence, with high probability, is jointly typical with the codeword that was sent. What is probability that one or more of the other codewords (which were chosen at random, independently of the sent codeword) is jointly typical with the received sequence? (Hint: You could use the union bound but you could also calculate this probability exactly, using the result of part (f) and the independence of the codewords)

(h) Given that a particular codeword was sent, the probability of error (averaged over the probability distribution of the channel and over the random choice of other codewords) can be written as

$$\Pr(\text{Error}|x^n(1)\text{ sent}) \quad = \sum_{y^n : y^n \text{causes error}} p(y^n|x^n(1)) \qquad (7.71)$$

There are two kinds of error: the first occurs if the received sequence y^n is not jointly typical with the transmitted codeword, and the second occurs if there is another codeword jointly typical with the received sequence. Using the result of the previous parts, calculate this probability of error.

By the symmetry of the random coding argument, this does not depend on which codeword was sent.

The calculations above show that average probability of error for a random code with 512 codewords of length 25 over the binary symmetric channel of crossover probability 0.1 is about 0.34. This seems quite high, but the reason for this is that the value of ϵ that we have chosen is too large. By choosing a smaller ϵ, and a larger n in the definitions of $A_\epsilon^{(n)}$, we can get the probability of error to be as small as we want, as long as the rate of the code is less than $I(X;Y) - 3\epsilon$.

Also note that the decoding procedure described in the problem is not optimal. The optimal decoding procedure is maximum likelihood, i.e., to choose the codeword that is closest to the received sequence. It is possible to calculate the average probability of error for a random code for which the decoding is based on an approximation to maximum likelihood decoding, where we decode a received sequence to the unique codeword that differs from the received sequence in ≤ 4 bits, and declare an error otherwise. The only difference with the jointly typical decoding described above is that in the case when the codeword is equal to the received sequence! The average probability of error for this decoding scheme can be shown to be about 0.285.

Solution: *Jointly typical set*

(a) Calculate $H(X)$, $H(Y)$, $H(X,Y)$ and $I(X;Y)$ for the joint distribution above.

Solution: $H(X) = H(Y) = 1$ bit, $H(X,Y) = H(X) + H(Y|X) = 1 + H(p) = 1 - 0.9 \log 0.9 - 0.1 \log 0.1 = 1 + 0.469 = 1.469$ bits, and $I(X;Y) = H(Y) - H(Y|X) = 0.531$ bits.

(b) Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. according the Bernoulli(1/2) distribution. Of the 2^n possible sequences of length n, which of them are typical, i.e., member of $A_\epsilon^{(n)}(X)$ for $\epsilon = 0.2$? Which are the typical sequences in $A_\epsilon^{(n)}(Y)$?

Solution: In the case for the uniform distribution, every sequence has probability $(1/2)^n$, and therefore for every sequence, $-\frac{1}{n}\log p(x^n) = 1 = H(X)$, and therefore every sequence is typical, i.e., $\in A_\epsilon^{(n)}(X)$.

Similarly, every sequence y^n is typical, i.e., $\in A_\epsilon^{(n)}(Y)$.

(c) The jointly typical set $A_\epsilon^{(n)}(X,Y)$ is defined as the set of sequences that satisfy equations (7.35-7.37) of EIT. The first two equations correspond to the conditions that x^n and y^n are in $A_\epsilon^{(n)}(X)$ and $A_\epsilon^{(n)}(Y)$ respectively. Consider the last condition, which can be rewritten to state that $-\frac{1}{n}\log p(x^n, y^n) \in (H(X,Y) - \epsilon, H(X,Y)+\epsilon)$. Let k be the number of places in which the sequence x^n differs from y^n (k is a function of the two sequences). Then we can write

$$\begin{aligned} p(x^n, y^n) &= \prod_{i=1}^{n} p(x_i, y_i) && (7.72) \\ &= (0.45)^{n-k}(0.05)^k && (7.73) \\ &= \left(\frac{1}{2}\right)^n (1-p)^{n-k}p^k && (7.74) \end{aligned}$$

An alternative way at looking at this probability is to look at the binary symmetric channel as in additive channel $Y = X \oplus Z$, where Z is a binary random variable that is equal to 1 with probability p, and is independent of X. In this case,

$$\begin{aligned} p(x^n, y^n) &= p(x^n)p(y^n|x^n) && (7.75) \\ &= p(x^n)p(z^n|x^n) && (7.76) \\ &= p(x^n)p(z^n) && (7.77) \\ &= \left(\frac{1}{2}\right)^n (1-p)^{n-k}p^k && (7.78) \end{aligned}$$

Show that the condition that (x^n, y^n) being jointly typical is equivalent to the condition that x^n is typical and $z^n = y^n - x^n$ is typical.

Solution: The conditions for $(x^n, y^n) \in A_\epsilon^{(n)}(X,Y)$ are

$$\begin{aligned} A_\epsilon^{(n)} &= \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : && (7.79) \\ & \left|-\frac{1}{n}\log p(x^n) - H(X)\right| < \epsilon, && (7.80) \\ & \left|-\frac{1}{n}\log p(y^n) - H(Y)\right| < \epsilon, && (7.81) \\ & \left|-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)\right| < \epsilon\}, && (7.82) \end{aligned}$$

But, as argued above, every sequence x^n and y^n satisfies the first two conditions. Thereofre, the only condition that matters is the last one. As argued above,

$$\begin{aligned} -\frac{1}{n}\log p(x^n, y^n) &= -\frac{1}{n}\log\left(\left(\frac{1}{2}\right)^n p^k(1-p)^{n-k}\right) && (7.83) \\ &= 1 - \frac{k}{n}\log p - \frac{n-k}{n}\log(1-p) && (7.84) \end{aligned}$$

Thus the pair (x^n, y^n) is jointly typical iff $|1 - \frac{k}{n}\log p - \frac{n-k}{n}\log(1-p) - H(X,Y)| < \epsilon$, i.e., iff $|-\frac{k}{n}\log p - \frac{n-k}{n}\log(1-p) - H(p)| < \epsilon$, which is exactly the condition for $z^n = y^n \oplus x^n$ to be typical. Thus the set of jointly typical pairs (x^n, y^n) is the set such that the number of places in which x^n differs from y^n is close to np.

(d) We now calculate the size of $A_\epsilon^{(n)}(Z)$ for $n = 25$ and $\epsilon = 0.2$. As in problem 7 of Homework 4, here is a table of the probabilities and numbers of sequences of with k ones

k	$\binom{n}{k}$	$\sum_{j\leq k}\binom{n}{j}$	$p(x^n) = p^k(1-p)^{n-k}$	$\binom{n}{k}p^k(1-p)^{n-k}$	Cumul. pr.	$-\frac{1}{n}\log p(x^n)$
0	1	1	7.178975e-02	0.071790	0.071790	0.152003
1	25	26	7.976639e-03	0.199416	0.271206	0.278800
2	300	326	8.862934e-04	0.265888	0.537094	0.405597
3	2300	2626	9.847704e-05	0.226497	0.763591	0.532394
4	12650	15276	1.094189e-05	0.138415	0.902006	0.659191
5	53130	68406	1.215766e-06	0.064594	0.966600	0.785988
6	177100	245506	1.350851e-07	0.023924	0.990523	0.912785
7	480700	726206	1.500946e-08	0.007215	0.997738	1.039582
8	1081575	1807781	1.667718e-09	0.001804	0.999542	1.166379
9	2042975	3850756	1.853020e-10	0.000379	0.999920	1.293176
10	3268760	7119516	2.058911e-11	0.000067	0.999988	1.419973
11	4457400	11576916	2.287679e-12	0.000010	0.999998	1.546770
12	5200300	16777216	2.541865e-13	0.000001	0.999999	1.673567

(Sequences with more than 12 ones are omitted since their total probability is negligible (and they are not in the typical set).)

What is the size of the set $A_\epsilon^{(n)}(Z)$?

Solution: $H(Z) = H(0.1) = 0.469$.

Setting $\epsilon = 0.2$, the typical set for Z is the set sequences for which $-\frac{1}{n}\log p(z^n) \in (H(Z)-\epsilon, H(Z)+\epsilon) = (0.269, 0.669)$. Looking at the table above for $n = 25$, it follows that the typical Z sequences are those with 1,2,3 or 4 ones.

The total probability of the set $A_\epsilon^{(n)}(Z) = 0.902006 - 0.071790 = 0.830216$ and the size of this set is $15276 - 1 = 15275$.

(e) Now consider random coding for the channel, as in the proof of the channel coding theorem. Assume that 2^{nR} codewords $X^n(1), X^n(2), \ldots, X^n(2^{nR})$ are chosen uniformly over the 2^n possible binary sequences of length n. One of these codewords is chosen and sent over the channel. The receiver looks at the received sequence and tries to find a codeword in the code that is jointly typical with the received sequence. As argued above, this corresponds to finding a codeword $X^n(i)$ such that $Y^n - X^n(i) \in A_\epsilon^{(n)}(Z)$. For a fixed codeword $x^n(i)$, what is the probability that the received sequence Y^n is such that $(x^n(i), Y^n)$ is jointly typical?

Solution: The easiest way to calculate this probability is to view the BSC as an additive channel $Y = X \oplus Z$, where Z is Bernoulli(p). Then the probability that for a given codeword, $x^n(i)$, that the output Y^n is jointly typical with it is equal to the probability that the noise sequence Z^n is typical, i.e., in $A_\epsilon^{(n)}(Z)$. The noise sequence is drawn i.i.d. according to the distribution $(1-p, p)$, and as calculated above, the probability that the sequence is typical, i.e., $\Pr(A_\epsilon^{(n)}(Z)) = 0.830216$. Therefore the probability that the received sequence is not jointly typical with the transmitted codeword is 0.169784.

(f) Now consider a particular received sequence $y^n = 000000\ldots0$, say. Assume that we choose a sequence X^n at random, uniformly distributed among all the 2^n possible binary n-sequences. What is the probability that the chosen sequence is jointly typical with this y^n? (Hint: this is the probability of all sequences x^n such that $y^n - x^n \in A_\epsilon^{(n)}(Z)$.)

Solution: Since all x^n sequences are chosen with the same probability ($(1/2)^n$), the probability that the x^n sequence chosen is jointly typical with the received y^n is equal to the number of possible jointly typical (x^n, y^n) pairs times $(1/2)^n$. The number of sequences x^n that are jointly typical with a given y^n is equal to number of typical z^n, where $z^n = x^n \oplus y^n$. Thus the probability that a randomly chosen x^n is typical with the given y^n is $|A_\epsilon^{(n)}(Z)| * (\frac{1}{2})^n = 15275 * 2^{-25} = 4.552 \times 10^{-4}$.

(g) Now consider a code with $2^9 = 512$ codewords of length 12 chosen at random, uniformly distributed among all the 2^n sequences of length $n = 25$. One of these codewords, say the one corresponding to $i = 1$, is chosen and sent over the channel. As calculated in part (e), the received sequence, with high probability, is jointly typical with the codeword that was sent. What is probability that one or more of the other codewords (which were chosen at random, independently of the sent codeword) is jointly typical with the received sequence? (Hint: You could use the union bound but you could also calculate this probability exactly, using the result of part (f) and the independence of the codewords)

Solution: Each of the other codewords is jointly typical with received sequence with probability 4.552×10^{-4}, and each of these codewords is independent. The probability that none of the 511 codewords are jointly typical with the received sequence is therefore $(1 - 4.552 \times 10^{-4})^{511} = 0.79241$, and the probability that at least one of them is jointly typical with the received sequence is therefore $1 - 0.79241 = 0.20749$.

Using the simple union of events bound gives the probability of another codeword being jointly typical with the received sequence to be $4.552 \times 10^{-4} \times 511 = 0.23262$. The previous calculation gives the more exact answer.

(h) Given that a particular codeword was sent, the probability of error (averaged over the probability distribution of the channel and over the random choice of other codewords) can be written as

$$\Pr(\text{Error}|x^n(1) \text{ sent}) = \sum_{y^n : y^n \text{causes error}} p(y^n|x^n(1)) \tag{7.85}$$

There are two kinds of error: the first occurs if the received sequence y^n is not jointly typical with the transmitted codeword, and the second occurs if there is another codeword jointly typical with the received sequence. Using the result of the previous parts, calculate this probability of error.

By the symmetry of the random coding argument, this does not depend on which codeword was sent.

Solution: There are two error events, which are conditionally independent, given the received sequence. In the previous part, we showed that the conditional probability of error of the second kind was 0.20749, irrespective of the received sequence y^n.

The probability of error of the first kind is 0.1698, conditioned on the input codeword. In part (e), we calculated the probability that $(x^n(i), Y^n) \notin A_\epsilon^{(n)}(X,Y)$, but this was conditioned on a particular input sequence. Now by the symmetry and uniformity of the random code construction, this probability does not depend on $x^n(i)$, and therefore the probability that $(X^n, Y^n) \notin A_\epsilon^{(n)}(X,Y)$ is also equal to this probability, i.e., to 0.1698.

We can therefore use a simple union of events bound to bound the total probability of error $\leq 0.1698 + 0.2075 = 0.3773$.

Thus we can send 512 codewords of length 25 over a BSC with crossover probability 0.1 with probability of error less than 0.3773.

A little more accurate calculation can be made of the probability of error using the fact that conditioned on the received sequence, both kinds of error are independent. Using the symmetry of the code construction process, the probability of error of the first kind conditioned on the received sequence does not depend on the received sequence, and is therefore $= 0.1698$. Therefore the probability that neither type of error occurs is (using their independence) $= (1 - 0.1698)(1 - 0.2075) = 0.6579$ and therefore, the probability of error is $1 - 0.6579 = 0.3421$

The calculations above show that average probability of error for a random code with 512 codewords of length 25 over the binary symmetric channel of crossover probability 0.1 is about 0.34. This seems quite high, but the reason for this is that the value of ϵ that we have chosen is too large. By choosing a smaller ϵ, and a larger n in the definitions of $A_\epsilon^{(n)}$, we can get the probability of error to be as small as we want, as long as the rate of the code is less than $I(X;Y) - 3\epsilon$.

Also note that the decoding procedure described in the problem is not optimal. The optimal decoding procedure is maximum likelihood, i.e., to choose the codeword that is closest to the received sequence. It is possible to calculate the average probability of error for a random code for which the decoding is based on an approximation to maximum likelihood decoding, where we decode a received sequence to the unique codeword that differs from the received sequence in ≤ 4 bits, and declare an error otherwise. The only difference with the jointly typical decoding described above is that in the case when the codeword is equal to the received sequence! The average probability of error for this decoding scheme can be shown to be about 0.285.

16. **Encoder and decoder as part of the channel:** Consider a binary symmetric channel with crossover probability 0.1. A possible coding scheme for this channel with two codewords of length 3 is to encode message a_1 as 000 and a_2 as 111. With this coding scheme, we can consider the combination of encoder, channel and decoder as forming a new BSC, with two inputs a_1 and a_2 and two outputs a_1 and a_2.

    (a) Calculate the crossover probability of this channel.

(b) What is the capacity of this channel in bits per transmission of the original channel?

(c) What is the capacity of the original BSC with crossover probability 0.1?

(d) Prove a general result that for any channel, considering the encoder, channel and decoder together as a new channel from messages to estimated messages will not increase the capacity in bits per transmission of the original channel.

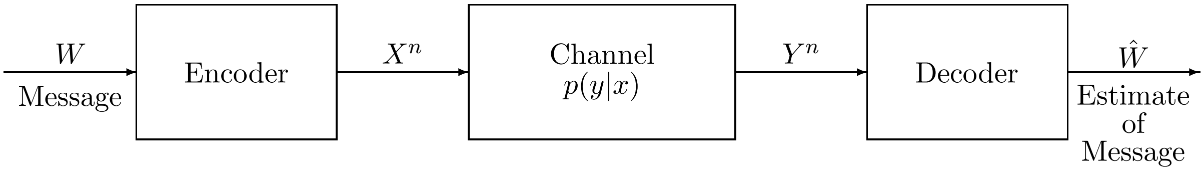
Solution: *Encoder and Decoder as part of the channel:*

(a) The probability of error with these 3 bits codewords was 2.8%, and thus the crossover probability of this channel is 0.028.

(b) The capacity of a BSC with crossover probability 0.028 is $1 - H(0.028)$, i.e., 1-0.18426 or 0.81574 bits for each 3 bit codeword. This corresponds to 0.27191 bits per transmission of the original channel.

(c) The original channel had capacity $1 - H(0.1)$, i.e., 0.531 bits/transmission.

(d) The general picture for the channel with encoder and decoder is shown below



By the data processing inequality, $I(W;\hat{W}) \leq I(X^n;Y^n)$, and therefore

$$C_W = \frac{1}{n} \max_{p(w)} I(W;\hat{W}) \leq \frac{1}{n} \max_{p(x^n)} I(X^n;Y^n) = C \tag{7.86}$$

Thus the capacity of the channel per transmission is not increased by the addition of the encoder and decoder.

17. **Codes of length 3 for a BSC and BEC:** In Problem 16, the probability of error was calculated for a code with two codewords of length 3 (000 and 111) sent over a binary symmetric channel with crossover probability ϵ. For this problem, take $\epsilon = 0.1$.

(a) Find the best code of length 3 with four codewords for this channel. What is the probability of error for this code? (Note that all possible received sequences should be mapped onto possible codewords)

(b) What is the probability of error if we used all the 8 possible sequences of length 3 as codewords?

(c) Now consider a binary erasure channel with erasure probability 0.1. Again, if we used the two codeword code 000 and 111, then received sequences 00E,0E0,E00,0EE,E0E,EE0 would all be decoded as 0, and similarly we would decode 11E,1E1,E11,1EE,E1E,EE1 as 1. If we received the sequence EEE we would not know if it was a 000 or a 111 that was sent - so we choose one of these two at random, and are wrong half the time.

What is the probability of error for this code over the erasure channel?

(d) What is the probability of error for the codes of parts (a) and (b) when used over the binary erasure channel?

Solution: *Codes of length 3 for a BSC and BEC:*

(a) To minimize the probability of confusion, the codewords should be as far apart as possible. With four codewords, the minimum distance is at most 2, and there are various sets of codewords that achieve this minimum distance. An example set is 000, 011, 110 and 101. Each of these codewords differs from the other codewords in at least two places.

To calculate the probability of error, we need to find the best decoding rule, i.e,. we need to map all possible recieved sequences onto codewords. As argued in the previous homework, the best decoding rule assigns to each received sequence the nearest codeword, with ties being broken arbitrarily. Of the 8 possible received sequences, 4 are codewords, and each of the other 4 sequences has three codewords within distance one of them. We can assign these received sequences to any of the nearest codewords, or alternatively, for symmetry, we might toss a three sided coin on receiving the sequence, and choose one of the nearest codewords with probability (1/3, 1/3, 1/3). All these decoding strategies will give the same average probability of error.

In the current example, there are 8 possible received sequences, and we will use the following decoder mapping 000, 001 → 000; 011, 010 → 011; 110, 100 → 110; and 101, 111 → 101.

Under this symmetric mapping, the codeword and one received sequence at distance 1 from the codeword are mapped on to the codeword. The probability therefore that the codeword is decoded correctly is $0.9{*}0.9{*}0.9{+}0.9{*}0.9{*}0.1 = 0.81$ and the probability of error (for each codeword) is 0.19. Thus the average probability of error is also 0.19.

(b) If we use all possible input sequences as codewords, then we have an error if any of the bits is changed. The probability that all the three bits are received correctly is $0.9 * 0.9 * 0.9 = 0.729$ and therefore the probability of error is 0.271.

(c) There will be an error only if all three bits of the codeword are erased, and on receiveing EEE, the decoder choses the wrong codeword. The probability of receiving EEE is 0.001 and conditioned on that, the probability of error is 0.5, so the probability of error for this code over the BEC is 0.0005.

(d) For the code of part (a), the four codewords are 000, 011,110, and 101. We use the following decoder mapping:

Received Sequences	codeword
000, 00E, 0E0, E00	000
011, 01E, 0E1, E11	011
110, 11E, 1E0, E10	110
101, 10E, 1E1, E01	101
0EE	000 or 011 with prob. 0.5
EE0	000 or 110 with prob. 0.5
⋮	
EE1	011 or 101 with prob. 0.5
EEE	000 or 011 or 110 or 101 with prob. 0.25

Essentially all received sequences with only one erasure can be decoded correctly. If there are two erasures, then there are two possible codewords that could have caused the received sequence, and the conditional probability of error is 0.5. If there are three erasures, any of the codewords could have caused it, and the conditional probability of error is 0.75. Thus the probability of error given that 000 was sent is the probability of two erasures times 0.5 plus the probability of 3 erasures times 0.75, i.e, $3 * 0.9 * 0.1 * 0.1 * 0.5 + 0.1 * 0.1 * 0.1 * 0.75 = 0.01425$. This is also the average probability of error.

If all input sequences are used as codewords, then we will be confused if there is any erasure in the received sequence. The conditional probability of error if there is one erasure is 0.5, two erasures is 0.75 and three erasures is 0.875 (these corrospond to the numbers of other codewords that could have caused the received sequence). Thus the probability of error given any codeword is $3 * 0.9 * 0.9 * 0.1 * 0.5 + 3 * 0.9 * 0.1 * 0.1 * 0.75 + 0.1 * 0.1 * 0.1 * 0.875 = 0.142625$. This is also the average probability of error.

18. **Channel capacity:** Calculate the capacity of the following channels with probability transition matrices:

    (a) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

    $$p(y|x) = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \qquad (7.87)$$

    (b) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

    $$p(y|x) = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{bmatrix} \qquad (7.88)$$

    (c) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2, 3\}$

    $$p(y|x) = \begin{bmatrix} p & 1-p & 0 & 0 \\ 1-p & p & 0 & 0 \\ 0 & 0 & q & 1-q \\ 0 & 0 & 1-q & q \end{bmatrix} \qquad (7.89)$$

Solution: *Channel Capacity:*

(a) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \tag{7.90}$$

This is a symmetric channel and by the results of Section 8.2,

$$C = \log|\mathcal{Y}| - H(\mathbf{r}) = \log 3 - \log 3 = 0. \tag{7.91}$$

In this case, the output is independent of the input.

(b) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{bmatrix} \tag{7.92}$$

Again the channel is symmetric, and by the results of Section 8.2,

$$C = \log|\mathcal{Y}| - H(\mathbf{r}) = \log 3 - \log = 0.58 \text{ bits} \tag{7.93}$$

(c) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2, 3\}$

$$p(y|x) = \begin{bmatrix} p & 1-p & 0 & 0 \\ 1-p & p & 0 & 0 \\ 0 & 0 & q & 1-q \\ 0 & 0 & 1-q & q \end{bmatrix} \tag{7.94}$$

This channel consists of a sum of two BSC's, and using the result of Problem 2 of Homework 9, the capacity of the channel is

$$C = \log\left(2^{1-H(p)} + 2^{1-H(q)}\right) \tag{7.95}$$

19. **Capacity of the carrier pigeon channel.** Consider a commander of an army besieged a fort for whom the only means of communication to his allies is a set of carrier pigeons. Assume that each carrier pigeon can carry one letter (8 bits), and assume that pigeons are released once every 5 minutes, and that each pigeon takes exactly 3 minutes to reach its destination.

(a) Assuming all the pigeons reach safely, what is the capacity of this link in bits/hour?

(b) Now assume that the enemies try to shoot down the pigeons, and that they manage to hit a fraction α of them. Since the pigeons are sent at a constant rate, the receiver knows when the pigeons are missing. What is the capacity of this link?

(c) Now assume that the enemy is more cunning, and every time they shoot down a pigeon, they send out a dummy pigeon carrying a random letter (chosen uniformly from all 8-bit letters). What is the capacity of this link in bits/hour?

Set up an appropriate model for the channel in each of the above cases, and indicate how to go about finding the capacity.

Solution: *Capacity of the carrier pigeon channel.*

(a) The channel sends 8 bits every 5 minutes, or 96 bits/hour.

(b) This is the equivalent of an erasure channel with an input alphabet of 8 bit symbols, i.e., 256 different symbols. For any symbols sent, a fraction α of them are received as an erasure. We would expect that the capacity of this channel is $(1-\alpha)8$ bits/pigeon. We will justify it more formally by mimicking the derivation for the binary erasure channel.

Consider a erasure channel with 256 symbol inputs and 257 symbol output - the extra symbol is the erasure symbol, which occurs with probability α. Then

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(\alpha) \tag{7.96}$$

since the probability of erasure is independent of the input.

However, we cannot get $H(Y)$ to attain its maximum value, i.e., $\log 257$, since the probability of the erasure channel is α independent of our input distribution. However, if we let E be the erasure event, then

$$H(Y) = H(Y,E) = H(E) + H(Y|E) = H(\alpha) + \alpha \times 0 + (1-\alpha)H(Y|E=0) \tag{7.97}$$

and we can maximize $H(Y)$ by maximizing $H(Y|E=0)$. However, $H(Y|E=0)$ is just the entropy of the input distribution, and this is maximized by the uniform. Thus the maximum value of $H(Y)$ is $H(\alpha)+(1-\alpha)\log 256$, and the capacity of this channel is $(1-\alpha)\log 256$ bits/pigeon, or $(1-\alpha)96$ bits/hour, as we might have expected from intuitive arguments.

(c) In this case, we have a symmetric channel with 256 inputs and 256 output. With probability $(1-\alpha)+\alpha/256$, the output symbol is equal to the input, and with probability $\alpha/256$, it is transformed to any of the other 255 symbols. This channel is symmetric in the sense of Section 8.2, and therefore the capacity of the channel is

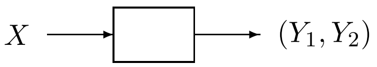
$$\begin{aligned}
C &= \log|\mathcal{Y}| - H(\mathbf{r}) & (7.98)\\
&= \log 256 - H(1-\alpha+\alpha/256, \alpha/256, \alpha/256, \ldots, \alpha/256) & (7.99)\\
&= 8 - H(1-\frac{255}{256}\alpha) - \frac{255}{256}\alpha H(1/255, 1/255, \ldots, 1/255) & (7.100)\\
&= 8 - H(1-\frac{255}{256}\alpha) - \frac{255}{256}\alpha \log 255 & (7.101)
\end{aligned}$$

We have to multiply this by 12 to get the capacity in bits/hour.

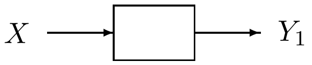
20. **A channel with two independent looks at Y.** Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X.

(a) Show $I(X;Y_1,Y_2) = 2I(X;Y_1) - I(Y_1,Y_2)$.

(b) Conclude that the capacity of the channel



is less than twice the capacity of the channel



Solution: *A channel with two independent looks at Y*

(a)

$$\begin{aligned}
I(X;Y_1,Y_2) &= H(Y_1,Y_2) - H(Y_1,Y_2|X) && (7.102)\\
&= H(Y_1) + H(Y_2) - I(Y_1;Y_2) - H(Y_1|X) - H(Y_2|X) && (7.103)\\
&\quad \text{(since } Y_1 \text{ and } Y_2 \text{ are conditionally independent given } X) && (7.104)\\
&= I(X;Y_1) + I(X;Y_2) - I(Y_1;Y_2) && (7.105)\\
&= 2I(X;Y_1) - I(Y_1;Y_2) \quad \text{(since } Y_1 \text{ and } Y_2 \text{ are conditionally identically distributed)} && (7.106)
\end{aligned}$$

(b) The capacity of the single look channel $X \to Y_1$ is

$$C_1 = \max_{p(x)} I(X;Y_1). \qquad (7.107)$$

The capacity of the channel $X \to (Y_1, Y_2)$ is

$$\begin{aligned}
C_2 &= \max_{p(x)} I(X;Y_1,Y_2) && (7.108)\\
&= \max_{p(x)} 2I(X;Y_1) - I(Y_1;Y_2) && (7.109)\\
&\le \max_{p(x)} 2I(X;Y_1) && (7.110)\\
&= 2C_1. && (7.111)
\end{aligned}$$

Hence, two independent looks cannot be more than twice as good as one look.

21. **Tall, fat people**

Suppose that average height of people in a room is 5 feet. Suppose the average weight is 100 lbs.

(a) Argue that no more than $\frac{1}{3}$ of the population is 15 feet tall.

(b) Find an upper bound on the fraction of 300 lb, 10 footers in the room.

Solution:

Tall, fat people.

(a) The average height of the individuals in the population is 5 feet. So $\frac{1}{n}\sum h_i = 5$ where n is the population size and h_i is the height of the i-th person. If more than $\frac{1}{3}$ of the population is at least 15 feet tall, then the average will be greater than $\frac{1}{3}15 = 5$ feet since each person is at least 0 feet tall. Thus no more than $\frac{1}{3}$ of the population is 15 feet tall.

(b) By the same reasoning as in part (a), at most $\frac{1}{2}$ of the poplulation is 10 feet tall and at most $\frac{1}{3}$ of the population weighs 300 lbs. Therefore at most $\frac{1}{3}$ are both 10 feet tall and weigh 300 lbs.

22. **Can signal alternatives lower capacity?** Show that adding a row to a channel transition matrix does not decrease capacity.

Solution: *Can signal alternatives lower capacity?*
Adding a row to the channel transition matrix is equivalent to adding a symbol to the input alphabet $\mathcal{X}$. Suppose there were m symbols and we add an $(m+1)$-st. We can always choose not to use this extra symbol.

Specifically, let C_m and C_{m+1} denote the capacity of the original channel and the new channel, respectively. Then

$$\begin{aligned} C_{m+1} &= \max_{p(x_1,\ldots,x_{m+1})} I(X;Y) \\ &\geq \max_{p(x_1,\ldots,x_m,0)} I(X;Y) \\ &= C_m. \end{aligned}$$

23. **Binary multiplier channel**

(a) Consider the channel $Y = XZ$ where X and Z are independent binary random variables that take on values 0 and 1. Z is Bernoulli(α), i.e. $P(Z=1)=\alpha$. Find the capacity of this channel and the maximizing distribution on X.

(b) Now suppose the receiver can observe Z as well as Y. What is the capacity?

Solution: *Binary Multiplier Channel*

(a) Let $P(X=1)=p$. Then $P(Y=1)=P(X=1)P(Z=1)=\alpha p$.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - P(X=1)H(Z) \\ &= H(\alpha p) - pH(\alpha) \end{aligned}$$

We find that $p^* = \frac{1}{\alpha(2^{\frac{H(\alpha)}{\alpha}}+1)}$ maximizes $I(X;Y)$. The capacity is calculated to be $\log(2^{\frac{H(\alpha)}{\alpha}}+1) - \frac{H(\alpha)}{\alpha}$.

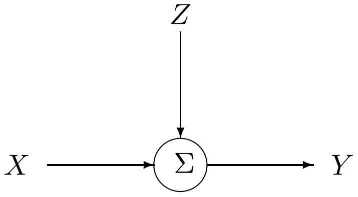
(b) Let $P(X=1)=p$. Then

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

The expression is maximized for $p=1/2$, resulting in $C=\alpha$. Intuitively, we can only get X through when Z is 1, which happens α of the time.

24. **Noise alphabets**

Consider the channel



$\mathcal{X}=\{0,1,2,3\}$, where $Y=X+Z$, and Z is uniformly distributed over three distinct integer values $\mathcal{Z}=\{z_1,z_2,z_3\}$.

(a) What is the maximum capacity over all choices of the $\mathcal{Z}$ alphabet? Give distinct integer values z_1,z_2,z_3 and a distribution on $\mathcal{X}$ achieving this.

(b) What is the minimum capacity over all choices for the $\mathcal{Z}$ alphabet? Give distinct integer values z_1,z_2,z_3 and a distribution on $\mathcal{X}$ achieving this.

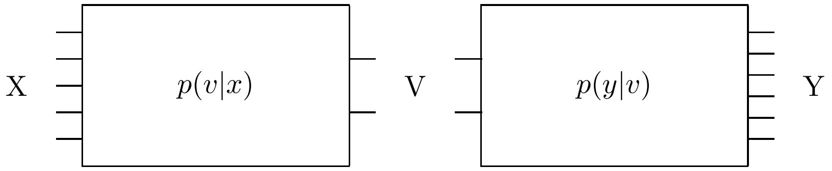
Solution: *Noise alphabets*

(a) Maximum capacity is $C=2$ bits. $\mathcal{Z}=\{10,20,30\}$ and $p(X)=(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4})$.

(b) Minimum capacity is $C=1$ bit. $\mathcal{Z}=\{0,1,2\}$ and $p(X)=(\frac{1}{2},0,0,\frac{1}{2})$.

25. **Bottleneck channel**

Suppose a signal $X\in\mathcal{X}=\{1,2,\ldots,m\}$ goes through an intervening transition $X\longrightarrow V\longrightarrow Y$:



where $x = \{1, 2, \ldots, m\}$, $y = \{1, 2, \ldots, m\}$, and $v = \{1, 2, \ldots, k\}$. Here $p(v|x)$ and $p(y|v)$ are arbitrary and the channel has transition probability $p(y|x) = \sum_v p(v|x)p(y|v)$.

Show $C \leq \log k$.

Solution: *Bottleneck channel*

The capacity of the cascade of channels is $C = \max_{p(x)} I(X;Y)$. By the data processing inequality, $I(X;Y) \leq I(V;Y) = H(V) - H(V|Y) \leq H(V) \leq \log k$.

26. **Noisy typewriter.** Consider the channel with $x, y \in \{0, 1, 2, 3\}$ and transition probabilities $p(y|x)$ given by the following matrix:

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{bmatrix}$$

(a) Find the capacity of this channel.

(b) Define the random variable $z = g(y)$ where

$$g(y) = \begin{cases} A & \text{if} \quad y \in \{0,1\} \\ B & \text{if} \quad y \in \{2,3\} \end{cases}.$$

For the following two PMFs for x, compute $I(X;Z)$

i.

$$p(x) = \begin{cases} \frac{1}{2} & \text{if} \quad x \in \{1,3\} \\ 0 & \text{if} \quad x \in \{0,2\} \end{cases}$$

ii.

$$p(x) = \begin{cases} 0 & \text{if} \quad x \in \{1,3\} \\ \frac{1}{2} & \text{if} \quad x \in \{0,2\} \end{cases}$$

(c) Find the capacity of the channel between x and z, specifically where $x \in \{0, 1, 2, 3\}$, $z \in \{A, B\}$, and the transition probabilities $P(z|x)$ are given by

$$p(Z = z|X = x) = \sum_{g(y_0)=z} P(Y = y_0|X = x)$$

(d) For the X distribution of part i. of b, does $X \to Z \to Y$ form a Markov chain?

Solution: *Noisy typewriter*

(a) This is a noisy typewriter channel with 4 inputs, and is also a symmetric channel. Capacity of the channel by Theorem 7.2.1 is $\log 4 - 1 = 1$ bit per transmission.

(b) i. The resulting conditional distribution of Z given X is

$$\begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

If

$$p(x) = \begin{cases} \frac{1}{2} & \text{if} \quad x \in \{1,3\} \\ 0 & \text{if} \quad x \in \{0,2\} \end{cases}$$

then it is easy to calculate $H(Z|X) = 0$, and $I(X;Z) = 1$. If

$$p(x) = \begin{cases} 0 & \text{if} \quad x \in \{1,3\} \\ \frac{1}{2} & \text{if} \quad x \in \{0,2\} \end{cases}$$

then $H(Z|X) = 1$ and $I(X;Y) = 0$.

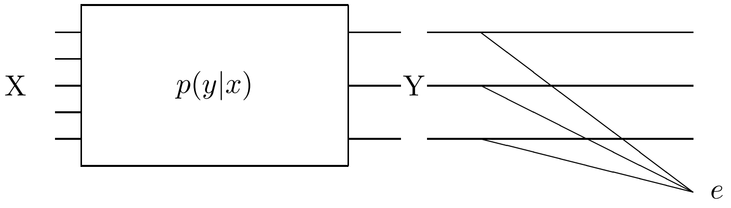
ii. Since $I(X;Z) \leq H(Z) \leq 1$, the capacity of the channel is 1, achieved by the input distribution

$$p(x) = \begin{cases} \frac{1}{2} & \text{if} \quad x \in \{1,3\} \\ 0 & \text{if} \quad x \in \{0,2\} \end{cases}$$

(c) For the input distribution that achieves capacity, $X \leftrightarrow Z$ is a one-to-one function, and hence $p(x,z) = 1$ or 0. We can therefore see the that $p(x,y,z) = p(z,x)p(y|x,z) = p(z,x)p(y|z)$, and hence $X \to Z \to Y$ forms a Markov chain.

27. **Erasure channel**

Let $\{\mathcal{X}, p(y|x), \mathcal{Y}\}$ be a discrete memoryless channel with capacity C. Suppose this channel is immediately cascaded with an erasure channel $\{\mathcal{Y}, p(s|y), \mathcal{S}\}$ that erases α of its symbols.



Specifically, $\mathcal{S} = \{y_1, y_2, \ldots, y_m, e\}$, and

$$\begin{aligned} \Pr\{S = y|X = x\} &= \bar{\alpha}p(y|x), \quad y \in \mathcal{Y}, \\ \Pr\{S = e|X = x\} &= \alpha. \end{aligned}$$

Determine the capacity of this channel.

Solution: *Erasure channel*

The capacity of the channel is

$$C = \max_{p(x)} I(X;S) \tag{7.112}$$

Define a new random variable Z, a function of S, where $Z = 1$ if $S = e$ and $Z = 0$ otherwise. Note that $p(Z = 1) = \alpha$ independent of X. Expanding the mutual information,

$$\begin{aligned} I(X;S) &= H(S) - H(S|X) && (7.113)\\ &= H(S,Z) - H(S,Z|X) && (7.114)\\ &+ H(Z) + H(S|Z) - H(Z|X) - H(S|X,Z) && (7.115)\\ &= I(X;Z) + I(S;X|Z) && (7.116)\\ &= 0 + \alpha I(X;S|Z=1) + (1-\alpha) I(X;S|Z=0) && (7.117) \end{aligned}$$

When $Z = 1$, $S = e$ and $H(S|Z = 1) = H(S|X, Z = 1) = 0$. When $Z = 0$, $S = Y$, and $I(X;S|Z = 0) = I(X;Y)$. Thus

$$I(X;S) = (1-\alpha)I(X;Y) \tag{7.118}$$

and therefore the capacity of the cascade of a channel with an erasure channel is $(1-\alpha)$ times the capacity of the original channel.

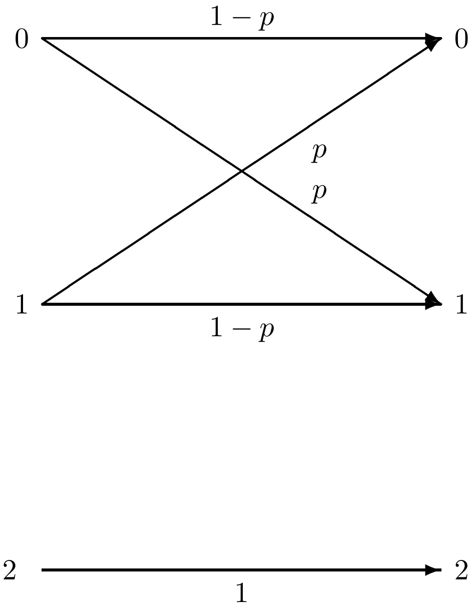
28. **Choice of channels.**

Find the capacity C of the union of 2 channels $(\mathcal{X}_1, p_1(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p_2(y_2|x_2), \mathcal{Y}_2)$ where, at each time, one can send a symbol over channel 1 or over channel 2 but not both. Assume the output alphabets are distinct and do not intersect.

(a) Show $2^C = 2^{C_1} + 2^{C_2}$. Thus 2^C is the effective alphabet size of a channel with capacity C.

(b) Compare with problem 10 of Chapter 2 where $2^H = 2^{H_1} + 2^{H_2}$, and interpret (a) in terms of the effective number of noise-free symbols.

(c) Use the above result to calculate the capacity of the following channel



Solution: *Choice of Channels*
(a) This is solved by using the very same trick that was used to solve problem 2.10.

Consider the following communication scheme:

$$X = \begin{cases} X_1 & \text{Probability } \alpha \\ X_2 & \text{Probability } (1-\alpha) \end{cases}$$

Let

$$\theta(X) = \begin{cases} 1 & X = X_1 \\ 2 & X = X_2 \end{cases}$$

Since the output alphabets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are disjoint, θ is a function of Y as well, i.e. $X \to Y \to \theta$.

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

Since $X \to Y \to \theta$, $I(X;\theta|Y) = 0$. Therefore,

$$\begin{aligned} I(X;Y) &= I(X;\theta) + I(X;Y|\theta) \\ &= H(\theta) - H(\theta|X) + \alpha I(X_1;Y_1) + (1-\alpha) I(X_2;Y_2) \\ &= H(\alpha) + \alpha I(X_1;Y_1) + (1-\alpha) I(X_2;Y_2) \end{aligned}$$

Thus, it follows that

$$C = \sup_{\alpha} \left\{ H(\alpha) + \alpha C_1 + (1-\alpha) C_2 \right\}.$$

Maximizing over α one gets the desired result. The maximum occurs for $H'(\alpha)+C_1-C_2=0$, or $\alpha = 2^{C_1}/(2^{C_1}+2^{C_2})$.

(b) If one interprets $M = 2^C$ as the effective number of noise free symbols, then the above result follows in a rather intuitive manner: we have $M_1 = 2^{C_1}$ noise free symbols from channel 1, and $M_2 = 2^{C_2}$ noise free symbols from channel 2. Since at each step we get to chose which channel to use, we essentially have $M_1 + M_2 = 2^{C_1} + 2^{C_2}$ noise free symbols for the new channel. Therefore, the capacity of this channel is $C = \log_2(2^{C_1} + 2^{C_2})$.

This argument is very similiar to the effective alphabet argument given in Problem 10, Chapter 2 of the text.

29. **Binary multiplier channel.**

(a) Consider the discrete memoryless channel $Y = XZ$ where X and Z are independent binary random variables that take on values 0 and 1. Let $P(Z=1)=\alpha$. Find the capacity of this channel and the maximizing distribution on X.

(b) Now suppose the receiver can observe Z as well as Y. What is the capacity?

Solution: *Binary Multiplier Channel* (Repeat of problem 7.23)

(a) Let $P(X=1)=p$. Then $P(Y=1)=P(X=1)P(Z=1)=\alpha p$.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - P(X=1)H(Z) \\ &= H(\alpha p) - pH(\alpha) \end{aligned}$$

We find that $p^* = \frac{1}{\alpha(2^{\frac{H(\alpha)}{\alpha}}+1)}$ maximizes $I(X;Y)$. The capacity is calculated to be $\log(2^{\frac{H(\alpha)}{\alpha}}+1) - \frac{H(\alpha)}{\alpha}$.

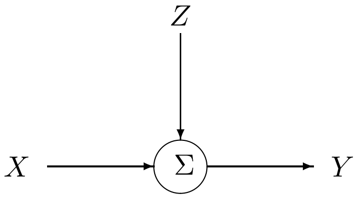
(b) Let $P(X=1)=p$. Then

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

The expression is maximized for $p = 1/2$, resulting in $C = \alpha$. Intuitively, we can only get X through when Z is 1, which happens α of the time.

30. **Noise alphabets.**

Consider the channel



$\mathcal{X} = \{0,1,2,3\}$, where $Y = X + Z$, and Z is uniformly distributed over three distinct integer values $\mathcal{Z} = \{z_1, z_2, z_3\}$.

(a) What is the maximum capacity over all choices of the $\mathcal{Z}$ alphabet? Give distinct integer values z_1, z_2, z_3 and a distribution on $\mathcal{X}$ achieving this.

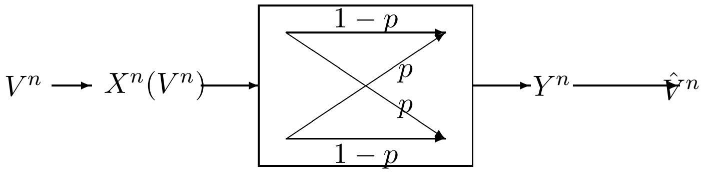
(b) What is the minimum capacity over all choices for the $\mathcal{Z}$ alphabet? Give distinct integer values z_1, z_2, z_3 and a distribution on $\mathcal{X}$ achieving this.

Solution: *Noise alphabets* (Repeat of problem 7.24)

(a) Maximum capacity is $C = 2$ bits. $\mathcal{Z} = \{10, 20, 30\}$ and $p(X) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.

(b) Minimum capacity is $C = 1$ bit. $\mathcal{Z} = \{0, 1, 2\}$ and $p(X) = (\frac{1}{2}, 0, 0, \frac{1}{2})$.

31. **Source and channel.**
We wish to encode a Bernoulli(α) process $V_1, V_2, \ldots$ for transmission over a binary symmetric channel with crossover probability p.



Find conditions on α and p so that the probability of error $P(\hat{V}^n \neq V^n)$ can be made to go to zero as $n \longrightarrow \infty$.

Solution: *Source And Channel*
Suppose we want to send a binary i.i.d. Bernoulli(α) source over a binary symmetric channel with error probability p.

By the source-channel separation theorem, in order to achieve an error rate that vanishes asymptotically, $P(\hat{V}^n \neq V^n) \to 0$, we need the entropy of the source to be smaller than the capacity of the channel. In this case this translates to

$$H(\alpha) + H(p) < 1,$$

or, equivalently,

$$\alpha^\alpha (1-\alpha)^{1-\alpha} p^p (1-p)^{1-p} < \frac{1}{2}.$$

32. **Random "20" questions**

Let X be uniformly distributed over $\{1, 2, \ldots, m\}$. Assume $m = 2^n$. We ask random questions: Is $X \in S_1$? Is $X \in S_2$?...until only one integer remains. All 2^m subsets S of $\{1, 2, \ldots, m\}$ are equally likely.

(a) How many deterministic questions are needed to determine X?

(b) Without loss of generality, suppose that $X = 1$ is the random object. What is the probability that object 2 yields the same answers for k questions as object 1?

(c) What is the expected number of objects in $\{2, 3, \ldots, m\}$ that have the same answers to the questions as does the correct object 1?

(d) Suppose we ask $n + \sqrt{n}$ *random* questions. What is the expected number of wrong objects agreeing with the answers?

(e) Use Markov's inequality $\Pr\{X \geq t\mu\} \leq \frac{1}{t}$, to show that the probability of error (one or more wrong object remaining) goes to zero as $n \longrightarrow \infty$.

Solution: *Random "20" questions.* (Repeat of Problem 5.45)

(a) Obviously, Huffman codewords for X are all of length n. Hence, with n deterministic questions, we can identify an object out of 2^n candidates.

(b) Observe that the total number of subsets which include both object 1 and object 2 or neither of them is 2^{m-1}. Hence, the probability that object 2 yields the same answers for k questions as object 1 is $(2^{m-1}/2^m)^k = 2^{-k}$.

More information theoretically, we can view this problem as a channel coding problem through a noiseless channel. Since all subsets are equally likely, the probability the object 1 is in a specific random subset is $1/2$. Hence, the question whether object 1 belongs to the kth subset or not corresponds to the kth bit of the random codeword for object 1, where codewords X^k are Bern$(1/2)$ random k-sequences.

Object	Codeword
1	$0110\ldots1$
2	$0010\ldots0$
$\vdots$	

Now we observe a noiseless output Y^k of X^k and figure out which object was sent. From the same line of reasoning as in the achievability proof of the channel coding theorem, i.e. joint typicality, it is obvious the probability that object 2 has the same codeword as object 1 is 2^{-k}.

(c) Let

$$1_j = \begin{cases} 1, & \text{object } j \text{ yields the same answers for } k \text{ questions as object 1} \\ 0, & \text{otherwise} \end{cases},$$

for $j = 2, \ldots, m.$

Then,

$$\begin{aligned} E(\#\text{ of objects in } \{2,3,\ldots,m\} \text{ with the same answers}) &= E(\sum_{j=2}^{m} 1_j) \\ &= \sum_{j=2}^{m} E(1_j) \\ &= \sum_{j=2}^{m} 2^{-k} \\ &= (m-1)2^{-k} \\ &= (2^n-1)2^{-k}. \end{aligned}$$

(d) Plugging $k = n + \sqrt{n}$ into (c) we have the expected number of $(2^n-1)2^{-n-\sqrt{n}}$.

(e) Let N by the number of wrong objects remaining. Then, by Markov's inequality

$$\begin{aligned} P(N \geq 1) &\leq EN \\ &= (2^n-1)2^{-n-\sqrt{n}} \\ &\leq 2^{-\sqrt{n}} \\ &\to 0, \end{aligned}$$

where the first equality follows from part (d).

33. **BSC with feedback.** Suppose that feedback is used on a binary symmetric channel with parameter p. Each time a Y is received, it becomes the next transmission. Thus X_1 is Bern(1/2), $X_2 = Y_1$, $X_3 = Y_2$, $\ldots$, $X_n = Y_{n-1}$.

(a) Find $\lim_{n\to\infty} \frac{1}{n} I(X^n;Y^n)$.

(b) Show that for some values of p, this can be higher than capacity.

(c) Using this feedback transmission scheme, $X^n(W,Y^n) = (X_1(W), Y_1, Y_2, \ldots, Y_{m-1})$, what is the asymptotic communication rate achieved; that is, what is $\lim_{n\to\infty} \frac{1}{n} I(W;Y^n)$?

Solution: BSC with feedback solution.

(a)

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

$$H(Y^n|X^n) = \sum_i H(Y_i|Y^{i-1},X^n) = H(Y_1|X_1) + \sum_i H(Y_i|Y^n) = H(p) + 0.$$

$$H(Y^n) = \sum_i H(Y_i|Y^{i-1}) = H(Y_1) + \sum_i H(Y_i|X_i) = 1 + (n-1)H(p)$$

So,

$$I(X^n;Y^n) = 1 + (n-1)H(p) - H(p) = 1 + (n-2)H(p)$$

and,

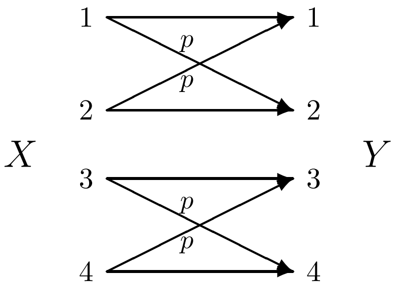
$$\lim_{n\to\infty} \frac{1}{n} I(X^n;Y^n) = \lim_{n\to\infty} \frac{1+(n-2)H(p)}{n} = H(p)$$

(b) For the BSC $C = 1 - H(p)$. For $p = 0.5$, $C = 0$, while $\lim_{n\to\infty} \frac{1}{n} I(X^n;Y^n) = H(0.5) = 1$.

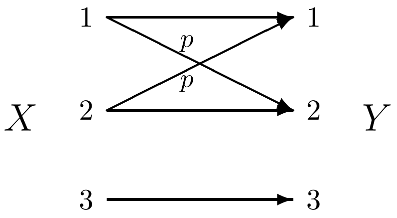
(c) Using this scheme $\frac{1}{n} I(W;Y^n) \to 0$.

34. **Capacity.** Find the capacity of

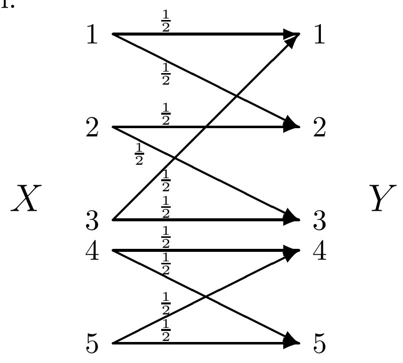
(a) Two parallel BSC's



(b) BSC and single symbol.



(c) BSC and ternary channel.



(d) Ternary channel.

$$p(y|x) = \left[\begin{array}{ccc} 2/3 & 1/3 & 0 \\ 0 & 1/3 & 2/3 \end{array} \right]. \tag{7.119}$$

Solution: *Capacity*

Recall the parallel channels problem (problem 7.28 showed that for two channels in parallel with capacities C_1 and C_2, the capacity C of the new channel satisfies

$$2^C = 2^{C_1} + 2^{C_2}$$

(a) Here $C_1 = C_2 = 1 - H(p)$, and hence $2^C = 2^{C_1+1}$, or,

$$C = 2 - H(p).$$

(b) Here $C_1 = 1 - H(p)$ but $C_2 = 0$ and so $2^C = 2^{C_1} + 1$, or,

$$C = \log\left(2^{1-H(p)} + 1\right).$$

(c) The p in the figure is a typo. All the transition probabilities are 1/2. The capacity of the ternary channel (which is symmetric) is $\log 3 - H(\frac{1}{2}) = \log 3 - 1$. The capacity of the BSC is 0, and together the parallel channels have a capacity $2^C = 3/2 + 1$, or $C = \log\frac{5}{2}$.

(d) The channel is weakly symmetric and hence the capacity is $log 3 - H(\frac{1}{3}, \frac{2}{3}) = \log 3 - (\log 3 - \frac{2}{3}) = \frac{2}{3}$.

35. **Capacity.**
Suppose channel $\mathcal{P}$ has capacity C, where $\mathcal{P}$ is an $m \times n$ channel matrix.

(a) What is the capacity of

$$\tilde{\mathcal{P}} = \left[\begin{array}{cc} \mathcal{P} & 0 \\ 0 & 1 \end{array} \right]$$

(b) What about the capacity of

$$\hat{\mathcal{P}} = \left[\begin{array}{cc} \mathcal{P} & 0 \\ 0 & I_k \end{array} \right]$$

where I_k if the $k \times k$ identity matrix.

Solution: Solution: Capacity.

(a) By adding the extra column and row to the transition matrix, we have two channels in parallel. You can transmit on either channel. From problem 7.28, it follows that

$$\begin{aligned} \tilde{C} &= \log(2^0 + 2^C) \\ \tilde{C} &= \log(1 + 2^C) \end{aligned}$$

(b) This part is also an application of the conclusion problem 7.28. Here the capacity of the added channel is $\log k$.

$$\begin{aligned} \hat{C} &= \log(2^{\log k} + 2^C) \\ \hat{C} &= \log(k + 2^C) \end{aligned}$$

36. **Channel with memory.**

Consider the discrete memoryless channel $Y_i = Z_i X_i$ with input alphabet $X_i \in \{-1, 1\}$.

(a) What is the capacity of this channel when $\{Z_i\}$ is i.i.d. with

$$Z_i = \begin{cases} 1, & p = 0.5 \\ -1, & p = 0.5 \end{cases} ? \tag{7.120}$$

Now consider the channel with memory. Before transmission begins, Z is randomly chosen and fixed for all time. Thus $Y_i = Z X_i$.

(b) What is the capacity if

$$Z = \begin{cases} 1, & p = 0.5 \\ -1, & p = 0.5 \end{cases} ? \tag{7.121}$$

Solution: Channel with memory solution.

(a) This is a BSC with cross over probability 0.5, so $C = 1 - H(p) = 0$.

(b) Consider the coding scheme of sending $X^n = (1, b_1, b_2, \ldots, b_{n-1})$ where the first symbol is always a zero and the rest of the $n-1$ symbols are ± 1 bits. For the first symbol $Y_1 = Z$, so the receiver knows Z exactly. After that the receiver can recover the remaining bits error free. So in n symbol transmissions n bits are sent, for a rate $R = \frac{n-1}{n} \rightarrow 1$. The capacity C is bounded by $\log |\mathcal{X}| = 1$, therefore the capacity is 1 bit per symbol.

37. **Joint typicality.**
Let (X_i, Y_i, Z_i) be *i.i.d.* according to $p(x, y, z)$. We will say that (x^n, y^n, z^n) is jointly typical (written $(x^n, y^n, z^n) \in A_\epsilon^{(n)}$) if

- $p(x^n) \in 2^{-n(H(X) \pm \epsilon)}$
- $p(y^n) \in 2^{-n(H(Y) \pm \epsilon)}$
- $p(z^n) \in 2^{-n(H(Z) \pm \epsilon)}$
- $p(x^n, y^n) \in 2^{-n(H(X,Y) \pm \epsilon)}$
- $p(x^n, z^n) \in 2^{-n(H(X,Z) \pm \epsilon)}$
- $p(y^n, z^n) \in 2^{-n(H(Y,Z) \pm \epsilon)}$
- $p(x^n, y^n, z^n) \in 2^{-n(H(X,Y,Z) \pm \epsilon)}$

Now suppose $(\tilde{X}^n, \tilde{Y}^n, \tilde{Z}^n)$ is drawn according to $p(x^n)p(y^n)p(z^n)$. Thus $\tilde{X}^n, \tilde{Y}^n, \tilde{Z}^n$ have the same marginals as $p(x^n, y^n, z^n)$ but are independent. Find (bounds on) $Pr\{(\tilde{X}^n, \tilde{Y}^n, \tilde{Z}^n) \in A_\epsilon^{(n)}\}$ in terms of the entropies $H(X), H(Y), H(Z), H(X,Y), H(X,Z), H(Y,Z)$ and $H(X,Y,Z)$.

Solution: *Joint typicality.*

$$\begin{aligned}
Pr\{(\tilde{X}^n, \tilde{Y}^n, \tilde{Z}^n) \in A_\epsilon^{(n)}\} &= \sum_{(x^n,y^n,z^n)\in A_\epsilon^{(n)}} p(x^n)p(y^n)p(z^n) \\
&\le \sum_{(x^n,y^n,z^n)\in A_\epsilon^{(n)}} 2^{-n(H(X)+H(Y)+H(Z)-3\epsilon)} \\
&\le |A_\epsilon^{(n)}| 2^{-n(H(X)+H(Y)+H(Z)-3\epsilon)} \\
&\le 2^{n(H(X,Y,Z)+\epsilon)} 2^{-n(H(X)+H(Y)+H(Z)-3\epsilon)} \\
&\le 2^{n(H(X,Y,Z)-H(X)-H(Y)-H(Z)+4\epsilon)}
\end{aligned}$$

$$\begin{aligned}
Pr\{(\tilde{X}^n, \tilde{Y}^n, \tilde{Z}^n) \in A_\epsilon^{(n)}\} &= \sum_{(x^n,y^n,z^n)\in A_\epsilon^{(n)}} p(x^n)p(y^n)p(z^n) \\
&\ge \sum_{(x^n,y^n,z^n)\in A_\epsilon^{(n)}} 2^{-n(H(X)+H(Y)+H(Z)+3\epsilon)} \\
&\ge |A_\epsilon^{(n)}| 2^{-n(H(X)+H(Y)+H(Z)-3\epsilon)} \\
&\ge (1-\epsilon) 2^{n(H(X,Y,Z)-\epsilon)} 2^{-n(H(X)+H(Y)+H(Z)-3\epsilon)} \\
&\ge (1-\epsilon) 2^{n(H(X,Y,Z)-H(X)-H(Y)-H(Z)-4\epsilon)}
\end{aligned}$$

Note that the upper bound is true for all n, but the lower bound only hold for n large.

Chapter 8

Differential Entropy

1. **Differential entropy.** Evaluate the differential entropy $h(X) = -\int f \ln f$ for the following:

   (a) The exponential density, $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$.

   (b) The Laplace density, $f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}$.

   (c) The sum of X_1 and X_2, where X_1 and X_2 are independent normal random variables with means μ_i and variances $\sigma_i^2, i = 1, 2$.

   Solution: *Differential Entropy.*

   (a) Exponential distribution.

$$\begin{aligned} h(f) &= -\int_0^\infty \lambda e^{-\lambda x}[\ln \lambda - \lambda x]dx && (8.1)\\ &= -\ln \lambda + 1 \text{ nats.} && (8.2)\\ &= \log \frac{e}{\lambda} \text{ bits.} && (8.3) \end{aligned}$$

   (b) Laplace density.

$$\begin{aligned} h(f) &= -\int_{-\infty}^\infty \frac{1}{2}\lambda e^{-\lambda|x|}[\ln \frac{1}{2} + \ln \lambda - \lambda|x|]\, dx && (8.4)\\ &= -\ln \frac{1}{2} - \ln \lambda + 1 && (8.5)\\ &= \ln \frac{2e}{\lambda} \text{ nats.} && (8.6)\\ &= \log \frac{2e}{\lambda} \text{ bits.} && (8.7) \end{aligned}$$

(c) Sum of two normal distributions.

The sum of two normal random variables is also normal, so applying the result derived the class for the normal distribution, since $X_1+X_2 \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$,

$$h(f) = \frac{1}{2}\log 2\pi e(\sigma_1^2+\sigma_2^2) \text{ bits.} \tag{8.8}$$

2. **Concavity of determinants.** Let K_1 and K_2 be two symmetric nonnegative definite $n \times n$ matrices. Prove the result of Ky Fan[4]:

$$\mid \lambda K_1 + \overline{\lambda} K_2 \mid \geq \mid K_1 \mid^{\lambda} \mid K_2 \mid^{\overline{\lambda}}, \quad \text{for } 0 \leq \lambda \leq 1, \ \overline{\lambda} = 1-\lambda,$$

where $\mid K \mid$ denotes the determinant of K.

Hint: Let $\mathbf{Z} = \mathbf{X}_\theta$, where $\mathbf{X}_1 \sim N(0, K_1)$, $\mathbf{X}_2 \sim N(0, K_2)$ and $\theta = \text{Bernoulli}(\lambda)$. Then use $h(\mathbf{Z} \mid \theta) \leq h(\mathbf{Z})$.

Solution: *Concavity of Determinants.* Let X_1 and X_2 be normally distributed n-vectors, $\mathbf{X}_i \sim \phi_{K_i}(\mathbf{x})$, $i = 1, 2$. Let the random variable θ have distribution $\Pr\{\theta = 1\} = \lambda$, $\Pr\{\theta = 2\} = 1-\lambda$, $0 \leq \lambda \leq 1$. Let θ, $\mathbf{X}_1$, and $\mathbf{X}_2$ be independent and let $\mathbf{Z} = \mathbf{X}_\theta$. Then $\mathbf{Z}$ has covariance $K_Z = \lambda K_1 + (1-\lambda)K_2$. However, $\mathbf{Z}$ will not be multivariate normal. However, since a normal distribution maximizes the entropy for a given variance, we have

$$\frac{1}{2}\ln(2\pi e)^n|\lambda K_1 + (1-\lambda)K_2| \geq h(\mathbf{Z}) \geq h(\mathbf{Z}|\theta) = \lambda\frac{1}{2}\ln(2\pi e)^n|K_1| + (1-\lambda)\frac{1}{2}\ln(2\pi e)^n|K_2| \,. \tag{8.9}$$

Thus

$$|\lambda K_1 + (1-\lambda)K_2| \geq |K_1|^{\lambda}|K_2|^{1-\lambda} \,, \tag{8.10}$$

as desired.

3. **Uniformly distributed noise.** Let the input random variable X to a channel be uniformly distributed over the interval $-1/2 \leq x \leq +1/2$. Let the output of the channel be $Y = X + Z$, where the noise random variable is uniformly distributed over the interval $-a/2 \leq z \leq +a/2$.

(a) Find $I(X;Y)$ as a function of a.

(b) For $a = 1$ find the capacity of the channel when the input X is peak-limited; that is, the range of X is limited to $-1/2 \leq x \leq +1/2$. What probability distribution on X maximizes the mutual information $I(X;Y)$?

(c) (Optional) Find the capacity of the channel for all values of a, again assuming that the range of X is limited to $-1/2 \leq x \leq +1/2$.

Solution: *Uniformly distributed noise.* The probability density function for $Y = X+Z$ is the convolution of the densities of X and Z. Since both X and Z have rectangular densities, the density of Y is a trapezoid. For $a < 1$ the density for Y is

$$p_Y(y) = \begin{cases} (1/2a)(y+(1+a)/2) & -(1+a)/2 \leq y \leq -(1-a)/2 \\ 1 & -(1-a)/2 \leq y \leq +(1-a)/2 \\ (1/2a)(-y-(1+a)/2) & +(1-a)/2 \leq y \leq +(1+a)/2 \end{cases}$$

and for $a > 1$ the density for Y is

$$p_Y(y) = \begin{cases} y + (a+1)/2 & -(a+1)/2 \le y \le -(a-1)/2 \\ 1/a & -(a-1)/2 \le y \le +(a-1)/2 \\ -y - (a+1)/2 & +(a-1)/2 \le y \le +(a+1)/2 \end{cases}$$

(When $a = 1$, the density of Y is triangular over the interval $[-1, +1]$.)

(a) We use the identity $I(X;Y) = h(Y) - h(Y|X)$. It is easy to compute $h(Y)$ directly, but it is even easier to use the grouping property of entropy. First suppose that $a < 1$. With probability $1 - a$, the output Y is conditionally uniformly distributed in the interval $[-(1-a)/2, +(1-a)/2]$; whereas with probability a, Y has a split triangular density where the base of the triangle has width a.

$$\begin{aligned} h(Y) &= H(a) + (1-a)\ln(1-a) + a(\frac{1}{2} + \ln a) \\ &= -a\ln a - (1-a)\ln(1-a) + (1-a)\ln(1-a) + \frac{a}{2} + a\ln a = \frac{a}{2} \text{ nats.} \end{aligned}$$

If $a > 1$ the trapezoidal density of Y can be scaled by a factor a, which yields $h(Y) = \ln a + 1/2a$. Given any value of x, the output Y is conditionally uniformly distributed over an interval of length a, so the conditional differential entropy in nats is $h(Y|X) = h(Z) = \ln a$ for all $a > 0$. Therefore the mutual information in nats is

$$I(X;Y) = \begin{cases} a/2 - \ln a & \text{if } a \le 1 \\ 1/2a & \text{if } a \ge 0\,. \end{cases}$$

As expected, $I(X;Y) \to \infty$ as $a \to 0$ and $I(X;Y) \to 0$ as $a \to \infty$.

(b) As usual with additive noise, we can express $I(X;Y)$ in terms of $h(Y)$ and $h(Z)$:

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

Since both X and Z are limited to the interval $[-1/2, +1/2]$, their sum Y is limited to the interval $[-1, +1]$. The differential entropy of Y is at most that of a random variable uniformly distributed on that interval; that is, $h(Y) \le 1$. This maximum entropy can be achieved if the input X takes on its extreme values $x = \pm 1$ each with probability 1/2. In this case, $I(X;Y) = h(Y) - h(Z) = 1 - 0 = 1$. Decoding for this channel is quite simple:

$$\hat{X} = \begin{cases} -1/2 & \text{if } y < 0 \\ +1/2 & \text{if } y \ge 0\,. \end{cases}$$

This coding scheme transmits one bit per channel use with zero error probability. (Only a received value $y = 0$ is ambiguous, and this occurs with probability 0.)

(c) When a is of the form $1/m$ for $m = 2, 3, \ldots$, we can achieve the maximum possible value $I(X;Y) = \log m$ when X is uniformly distributed over the discrete points $\{-1, -1+2/(m-1), \ldots, +1-2/(m-1), +1\}$. In this case Y has a uniform probability density on the interval $[-1-1/(m-1), +1+1/(m-1)]$. Other values of a are left as an exercise.

4. **Quantized random variables.** Roughly how many bits are required on the average to describe to 3 digit accuracy the decay time (in years) of a radium atom if the half-life of radium is 80 years? Note that half-life is the median of the distribution.

Solution: *Quantized random variables.* The differential entropy of an exponentially distributed random variable with mean $1/\lambda$ is $\log \frac{e}{\lambda}$ bits. If the median is 80 years, then

$$\int_0^{80} \lambda e^{-\lambda x}\, dx = \frac{1}{2} \tag{8.11}$$

or

$$\lambda = \frac{\ln 2}{80} = 0.00866 \tag{8.12}$$

and the differential entropy is $\log e/\lambda$. To represent the random variable to 3 digits $\approx$ 10 bits accuracy would need $\log e/\lambda + 10$ bits $=$ 18.3 bits.

5. **Scaling.** Let $h(\mathbf{X}) = -\int f(\mathbf{x}) \log f(\mathbf{x})\, d\mathbf{x}$. Show $h(A\mathbf{X}) = \log \mid \det(A) \mid + h(\mathbf{X})$.

Solution: *Scaling.* Let $\mathbf{Y} = A\mathbf{X}$. Then the density of $\mathbf{Y}$ is

$$g(\mathbf{y}) = \frac{1}{|A|} f(A^{-1}\mathbf{y}). \tag{8.13}$$

Hence

$$\begin{aligned} h(A\mathbf{X}) &= -\int g(\mathbf{y}) \ln g(\mathbf{y})\, d\mathbf{y} && (8.14)\\ &= -\int \frac{1}{|A|} f(A^{-1}\mathbf{y}) \left[\ln f(A^{-1}\mathbf{y}) - \log|A|\right] d\mathbf{y} && (8.15)\\ &= -\int \frac{1}{|A|} f(\mathbf{x}) \left[\ln f(\mathbf{x}) - \log|A|\right] |A|\, d\mathbf{x} && (8.16)\\ &= h(\mathbf{X}) + \log|A|. && (8.17) \end{aligned}$$

6. **Variational inequality:** Verify, for positive random variables X, that

$$\log E_P(X) = \sup_Q \left[E_Q(\log X) - D(Q||P)\right] \tag{8.18}$$

where $E_P(X) = \sum_x xP(x)$ and $D(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$, and the supremum is over all $Q(x) \geq 0$, $\sum Q(x) = 1$. It is enough to extremize $J(Q) = E_Q \ln X - D(Q||P) + \lambda(\sum Q(x) - 1)$.

Solution: *Variational inequality*

Using the calculus of variations to extremize

$$J(Q) = \sum_x q(x) \ln x - \sum_x q(x) \ln \frac{q(x)}{p(x)} + \lambda(\sum_x q(x) - 1) \tag{8.19}$$

we differentiate with respect to $q(x)$ to obtain

$$\frac{\partial J}{\partial q(x)} = \ln x - \ln \frac{q(x)}{p(x)} - 1 + \lambda = 0 \tag{8.20}$$

or

$$q(x) = c' x p(x) \tag{8.21}$$

where c' has to be chosen to satisfy the constraint, $\sum_x q(x) = 1$. Thus

$$c' = \frac{1}{\sum_x xp(x)} \tag{8.22}$$

Substituting this in the expression for J, we obtain

$$\begin{aligned} J^* &= \sum_x c'xp(x)\ln x - \sum_x c'xp(x)\ln\frac{c'xp(x)}{p(x)} && (8.23)\\ &= -\ln c' + \sum_x c'xp(x)\ln x - \sum_x c'xp(x)\ln x && (8.24)\\ &= \ln\sum_x xp(x) && (8.25)\end{aligned}$$

To verify this is indeed a maximum value, we use the standard technique of writing it as a relative entropy. Thus

$$\begin{aligned} \ln\sum_x xp(x) - \sum_x q(x)\ln x + \sum_x q(x)\ln\frac{q(x)}{p(x)} &= \sum_x q(x)\ln\frac{q(x)}{\frac{xp(x)}{\sum_y yp(y)}} && (8.26)\\ &= D(q||p') && (8.27)\\ &\geq 0 && (8.28)\end{aligned}$$

Thus

$$\ln\sum_x xp(x) = \sup_Q\left(E_Q\ln(X) - D(Q||P)\right) \tag{8.29}$$

This is a special case of a general relationship that is a key in the theory of large deviations.

7. **Differential entropy bound on discrete entropy:** Let X be a discrete random variable on the set $\mathcal{X} = \{a_1, a_2, \ldots\}$ with $\Pr(X = a_i) = p_i$. Show that

$$H(p_1, p_2, \ldots) \leq \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{\infty} p_i i^2 - \left(\sum_{i=1}^{\infty} ip_i\right)^2 + \frac{1}{12}\right). \tag{8.30}$$

Moreover, for every permutation σ,

$$H(p_1, p_2, \ldots) \leq \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{\infty} p_{\sigma(i)} i^2 - \left(\sum_{i=1}^{\infty} ip_{\sigma(i)}\right)^2 + \frac{1}{12}\right). \tag{8.31}$$

Hint: Construct a random variable X' such that $\Pr(X' = i) = p_i$. Let U be an uniform(0,1] random variable and let $Y = X' + U$, where X' and U are independent.

Use the maximum entropy bound on Y to obtain the bounds in the problem. This bound is due to Massey (unpublished) and Willems(unpublished).

Solution: *Differential entropy bound on discrete entropy*

Of all distributions with the same variance, the normal maximizes the entropy. So the entropy of the normal gives a good bound on the differential entropy in terms of the variance of the random variable.

Let X be a discrete random variable on the set $\mathcal{X} = \{a_1, a_2, \ldots\}$ with

$$\Pr(X = a_i) = p_i. \tag{8.32}$$

$$H(p_1, p_2, \ldots) \le \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{\infty} p_i i^2 - \left(\sum_{i=1}^{\infty} i p_i\right)^2 + \frac{1}{12}\right). \tag{8.33}$$

Moreover, for every permutation σ,

$$H(p_1, p_2, \ldots) \le \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{\infty} p_{\sigma(i)} i^2 - \left(\sum_{i=1}^{\infty} i p_{\sigma(i)}\right)^2 + \frac{1}{12}\right). \tag{8.34}$$

Define two new random variables. The first, X_0, is an integer-valued discrete random variable with the distribution

$$\Pr(X_0 = i) = p_i. \tag{8.35}$$

Let U be a random variable uniformly distributed on the range $[0,1]$, independent of X_0. Define the continuous random variable $\tilde{X}$ by

$$\tilde{X} = X_0 + U. \tag{8.36}$$

The distribution of the r.v. $\tilde{X}$ is shown in Figure 8.1.

It is clear that $H(X) = H(X_0)$, since discrete entropy depends only on the probabilities and not on the values of the outcomes. Now

$$\begin{aligned}
H(X_0) &= -\sum_{i=1}^{\infty} p_i \log p_i && (8.37)\\
&= -\sum_{i=1}^{\infty}\left(\int_i^{i+1} f_{\tilde{X}}(x)\,dx\right)\log\left(\int_i^{i+1} f_{\tilde{X}}(x)\,dx\right) && (8.38)\\
&= -\sum_{i=1}^{\infty}\int_i^{i+1} f_{\tilde{X}}(x)\log f_{\tilde{X}}(x)\,dx && (8.39)\\
&= -\int_1^{\infty} f_{\tilde{X}}(x)\log f_{\tilde{X}}(x)\,dx && (8.40)\\
&= h(\tilde{X}), && (8.41)
\end{aligned}$$

since $f_{\tilde{X}}(x) = p_i$ for $i \le x < i+1$.

Figure 8.1: Distribution of $\tilde{X}$.

Hence we have the following chain of inequalities:

$$\begin{aligned} H(X) &= H(X_0) && (8.42)\\ &= h(\tilde{X}) && (8.43)\\ &\le \frac{1}{2}\log(2\pi e)\mathrm{Var}(\tilde{X}) && (8.44)\\ &= \frac{1}{2}\log(2\pi e)\left(\mathrm{Var}(X_0)+\mathrm{Var}(U)\right) && (8.45)\\ &= \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{\infty} p_i i^2 - \left(\sum_{i=1}^{\infty} i p_i\right)^2 + \frac{1}{12}\right). && (8.46) \end{aligned}$$

Since entropy is invariant with respect to permutation of $p_1, p_2, \ldots$, we can also obtain a bound by a permutation of the p_i's. We conjecture that a good bound on the variance will be achieved when the high probabilities are close together, i.e, by the assignment $\ldots, p_5, p_3, p_1, p_2, p_4, \ldots$ for $p_1 \ge p_2 \ge \cdots$.

How good is this bound? Let X be a Bernoulli random variable with parameter $\frac{1}{2}$, which implies that $H(X) = 1$. The corresponding random variable X_0 has variance $\frac{1}{4}$, so the bound is

$$H(X) \le \frac{1}{2}\log(2\pi e)\left(\frac{1}{4}+\frac{1}{12}\right) = 1.255 \text{ bits.} \qquad (8.47)$$

8. **Channel with uniformly distributed noise:** Consider a additive channel whose input alphabet $\mathcal{X} = \{0, \pm 1, \pm 2\}$, and whose output $Y = X + Z$, where Z is uniformly distributed over the interval $[-1, 1]$. Thus the input of the channel is a discrete random

variable, while the output is continuous. Calculate the capacity $C = \max_{p(x)} I(X;Y)$ of this channel.

Solution: *Uniformly distributed noise*

We can expand the mutual information

$$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(Z) \tag{8.48}$$

and $h(Z) = \log 2$, since $Z \sim U(-1,1)$.

The output Y is a sum a of a discrete and a continuous random variable, and if the probabilities of X are $p_{-2}, p_{-1}, \ldots, p_2$, then the output distribution of Y has a uniform distribution with weight $p_{-2}/2$ for $-3 \le Y \le -2$, uniform with weight $(p_{-2}+p_{-1})/2$ for $-2 \le Y \le -1$, etc. Given that Y ranges from -3 to 3, the maximum entropy that it can have is an uniform over this range. This can be achieved if the distribution of X is (1/3, 0, 1/3,0,1/3). Then $h(Y) = \log 6$ and the capacity of this channel is $C = \log 6 - \log 2 = \log 3$ bits.

9. **Gaussian mutual information**. Suppose that (X,Y,Z) are jointly Gaussian and that $X \rightarrow Y \rightarrow Z$ forms a Markov chain. Let X and Y have correlation coefficient ρ_1 and let Y and Z have correlation coefficient ρ_2. Find $I(X;Z)$.

Solution: *Gaussian Mutual Information*

First note that we may without any loss of generality assume that the means of X, Y and Z are zero. If in fact the means are not zero one can subtract the vector of means without affecting the mutual information or the conditional independence of X, Z given Y. Let

$$\Lambda = \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_z\rho_{xz} \\ \sigma_x\sigma_z\rho_{xz} & \sigma_z^2 \end{pmatrix},$$

be the covariance matrix of X and Z. We can now use Eq. (8.34) to compute

$$\begin{aligned} I(X;Z) &= h(X) + h(Z) - h(X,Z) \\ &= \frac{1}{2}\log(2\pi e\sigma_x^2) + \frac{1}{2}\log(2\pi e\sigma_x^2) - \frac{1}{2}\log(2\pi e|\Lambda|) \\ &= -\frac{1}{2}\log(1-\rho_{xz}^2) \end{aligned}$$

Now,

$$\begin{aligned} \rho_{xz} &= \frac{\mathbf{E}\{XZ\}}{\sigma_x\sigma_z} \\ &= \frac{\mathbf{E}\{\mathbf{E}\{XZ|Y\}\}}{\sigma_x\sigma_z} \\ &= \frac{\mathbf{E}\{\mathbf{E}\{X|Y\}\mathbf{E}\{Z|Y\}\}}{\sigma_x\sigma_z} \\ &= \frac{\mathbf{E}\{\left(\frac{\sigma_x\rho_{xy}}{\sigma_y}Y\right)\left(\frac{\sigma_z\rho_{zx}}{\sigma_y}Y\right)\}}{\sigma_x\sigma_z} \\ &= \rho_{xy}\rho_{zy} \end{aligned}$$

We can thus conclude that

$$I(X;Y) = -\frac{1}{2}\log(1-\rho_{xy}^2\rho_{zy}^2)$$

10. **The Shape of the Typical Set**

Let X_i be i.i.d. $\sim f(x)$, where

$$f(x) = ce^{-x^4}.$$

Let $h = -\int f \ln f$. Describe the shape (or form) or the typical set $A_\epsilon^{(n)} = \{x^n \in \mathcal{R}^n : f(x^n) \in 2^{-n(h\pm\epsilon)}\}$.

Solution: *The Shape of the Typical Set*

We are interested in the set $\{\, x^n \in \mathcal{R} : f(x^n) \in 2^{-n(h\pm\epsilon)}\,\}$. This is:

$$2^{-n(h-\epsilon)} \leq f(x^n) \leq 2^{-n(h+\epsilon)}$$

Since X_i are i.i.d.,

$$\begin{aligned} f(x^n) &= \prod_{i=1}^n f(x) && (8.49)\\ &= \prod_{i=1}^n ce^{-x_i^4} && (8.50)\\ &= e^{nln(c)-\sum_{i=1}^n x_i^4} && (8.51)\\ & && (8.52) \end{aligned}$$

Plugging this in for $f(x^n)$ in the above inequality and using algebraic manipulation gives:

$$n(ln(c)+(h-\epsilon)ln(2)) \geq \sum_{i=1}^n x_i^4 \geq n(ln(c)+(h+\epsilon)ln(2))$$

So the shape of the typcial set is the shell of a 4-norm ball $\left\{x^n : ||x^n||_4 \in (n(ln(c)+(h\pm\epsilon)ln(2)))^{1/4}\right\}$.

11. **Non ergodic Gaussian process.**

Consider a constant signal V in the presence of iid observational noise $\{Z_i\}$. Thus $X_i = V + Z_i$, where $V \sim N(0,S)$, and Z_i are iid $\sim N(0,N)$. Assume V and $\{Z_i\}$ are independent.

(a) Is $\{X_i\}$ stationary?

(b) Find $\lim_{n\longrightarrow\infty}\frac{1}{n}\sum_{i=1}^{n}X_i$. Is the limit random?

(c) What is the entropy rate h of $\{X_i\}$?

(d) Find the least mean squared error predictor $\hat{X}_{n+1}(X^n)$ and find $\sigma_\infty^2 = \lim_{n\longrightarrow\infty} E(\hat{X}_n - X_n)^2$.

(e) Does $\{X_i\}$ have an AEP? That is, does $-\frac{1}{n}\log f(X^n) \longrightarrow h$?

Solution: *Nonergodic Gaussian process*

(a) Yes. $EX_i = EV + Z_i = 0$ for all i, and

$$EX_iX_j = E(V+Z_i)(V+Z_j) = \begin{cases} S, & i = j \\ S+N. & i \neq j \end{cases} \tag{8.53}$$

Since X_i is Gaussian distributed it is completely characterized by its first and second moments. Since the moments are stationary, X_i is wide sense stationary, which for a Gaussian distribution implies that X_i is stationary.

(b)

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_i &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(Z_i+V) && (8.54)\\
&= V + \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}Z_i && (8.55)\\
&= V + EZ_i(\text{by the strong law of large numbers}) && (8.56)\\
&= V && (8.57)
\end{aligned}$$

The limit is a random variable $\mathcal{N}(0,S)$.

(c) Note that $X^n \sim N(0, K_{X^n})$, where K_{X^n} has diagonal values of $S+N$ and off diagonal values of S. Also observe that the determinant is $|K_{X^n}| = N^n(nS/N+1)$. We now compute the entropy rate as:

$$\begin{aligned}
h(\mathcal{X}) &= \lim \frac{1}{n}h(X_1, X_2, \ldots, X_n) && (8.58)\\
&= \lim \frac{1}{2n}\log((2\pi e)^n |K_{X^n}|) && (8.59)\\
&= \lim \frac{1}{2n}\log\left((2\pi e)^n N^n\left(\frac{nS}{N}+1\right)\right) && (8.60)\\
&= \lim \frac{1}{2n}\log(2\pi e N)^n + \frac{1}{2n}\log\left(\frac{nS}{N}+1\right) && (8.61)\\
&= \frac{1}{2}\log 2\pi e N + \lim \frac{1}{2n}\log\left(\frac{nS}{N}+1\right) && (8.62)\\
&= \frac{1}{2}\log 2\pi e N && (8.63)
\end{aligned}$$

(d) By iterated expectation we can write

$$E\left(X_{n+1}-\hat{X}_{n+1}(X^n)\right)^2 = E\left(E\left(\left(X_{n+1}-\hat{X}_{n+1}(X^n)\right)^2 \middle| X^n\right)\right) \tag{8.64}$$

We note that minimizing the expression is equivalent to minimizing the inner expectation, and that for the inner expectation the predictor is a nonrandom variable. Expanding the inner expectation and taking the derivative with respect to the estimator $\hat{X}_{n+1}(X^n)$, we get

$$\begin{aligned} & E\left((X_{n+1}-\hat{X}_{n+1}(X^n))^2|X^n\right) \\ & \quad = E\left((X_{n+1}^2 - 2X_{n+1}\hat{X}_{n+1}(X^n) + \hat{X}_{n+1}^2(X^n))|X^n\right) \end{aligned} \tag{8.65}$$

so

$$\begin{aligned} \frac{dE\left((X_{n+1}-\hat{X}_{n+1}(X^n))^2|X^n\right)}{d\hat{X}_{n+1}(X^n)} &= E\left(-2X_{n+1}+2\hat{X}_{n+1}(X^n)|X^n\right) & (8.66)\\ &= -2E(X_{n+1}|X^n)+2\hat{X}_{n+1}(X^n) & (8.67) \end{aligned}$$

Setting the derivative equal to 0, we see that the optimal $\hat{X}_{n+1}(X^n) = E(X_{n+1}|X^n)$. To find the limiting squared error for this estimator, we use the fact that V and X^n are normal random variables with known covariance matrix, and therefore the conditional distribution of V given X^n is

$$f(V|X^n) \sim \mathcal{N}\left(\frac{S}{nS+N}\sum_{i=1}^n X_i, \frac{SN}{nS+N}\right) \tag{8.68}$$

Now

$$\begin{aligned} \hat{X}_{n+1}(X^n) &= E(X_{n+1}|X^n) & (8.69)\\ &= E(V|X^n)+E(Z_{n+1}|X^n) & (8.70)\\ &= \frac{S}{nS+N}\sum_{i=1}^n X_i + 0 & (8.71) \end{aligned}$$

and hence the limiting squared error

$$\begin{aligned} e^2 &= \lim_{n\to\infty} E(\hat{X}_n - X_n)^2 & (8.72)\\ &= \lim_{n\to\infty} E\left(\frac{S}{(n-1)S+N}\sum_{i=1}^{n-1} X_i - X_n\right)^2 & (8.73)\\ &= \lim_{n\to\infty} E\left(\frac{S}{(n-1)S+N}\sum_{i=1}^{n-1}(Z_i+V) - Z_n - V\right)^2 & (8.74)\\ &= \lim_{n\to\infty} E\left(\frac{S}{(n-1)S+N}\sum_{i=1}^{n-1} Z_i - Z_n - \frac{N}{(n-1)S+N}V\right)^2 & (8.75) \end{aligned}$$

$$= \lim_{n\to\infty} \left(\frac{S}{(n-1)S+N}\right)^2 \sum_{i=1}^{n-1} EZ_i^2 + EZ_n^2 + \left(\frac{N}{(n-1)S+N}\right)^2 EV^2 \tag{8.76}$$

$$= \lim_{n\to\infty} \left(\frac{S}{(n-1)S+N}\right)^2 (n-1)N + N + \left(\frac{N}{(n-1)S+N}\right)^2 S \tag{8.77}$$

$$= 0 + N + 0 \tag{8.78}$$

$$= N \tag{8.79}$$

(e) Even though the process is not ergodic, it is stationary, and it does have an AEP because

$$-\frac{1}{n}\ln f(X^n) = -\frac{1}{n}\ln \frac{1}{(2\pi)^{n/2}|K_{X^n}|^{\frac{1}{2}}} e^{-X^t K_{X^n}^{-1} X/2} \tag{8.80}$$

$$= \frac{1}{2n}\ln(2\pi)^n + \frac{1}{2n}\ln|K_{X^n}| + \frac{1}{2n} X^t K_{X^n}^{-1} X \tag{8.81}$$

$$= \frac{1}{2n}\ln(2\pi e)^n |K_{X^n}| - \frac{1}{2} + \frac{1}{2n} X^t K_{X^n}^{-1} X \tag{8.82}$$

$$= \frac{1}{n} h(X^n) - \frac{1}{2} + \frac{1}{2n} X^t K_{X^n}^{-1} X \tag{8.83}$$

$$\tag{8.84}$$

Since $X \sim \mathcal{N}(0,K)$, we can write $X = K^{\frac{1}{2}}W$, where $W \sim \mathcal{N}(0,I)$. Then $X^t K^{-1} X = W^t K^{\frac{1}{2}} K^{-1} K^{\frac{1}{2}} W = W^t W = \sum W_i^2$, and therefore $X^t K^{-1} X$ has a χ^2 distribution with n degrees of freedom. The density of the χ^2 distribution is

$$f(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}}} \tag{8.85}$$

The moment generating function of the χ^2 distribution is

$$M(t) = \int f(x) e^{tx}\, dx \tag{8.86}$$

$$= \int \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}}} e^{tx}\, dx \tag{8.87}$$

$$= \int \frac{\frac{1}{(1-2t)^{\frac{n}{2}-1}} ((1-2t)x)^{\frac{n}{2}-1} e^{-(1-2t)x/2}}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}}} (1-2t)\, dx \tag{8.88}$$

$$= \frac{1}{(1-2t)^{\frac{n}{2}}} \tag{8.89}$$

By the Chernoff bound (Lemma 11.19.1)

$$\Pr\left\{\frac{1}{n}\sum W_i^2 > 1+\epsilon\right\} \leq \min_s e^{-s(1+\epsilon)} (1-2s)^{-\frac{n}{2}} \tag{8.90}$$

$$\leq e^{-\frac{n}{2}(\epsilon - \ln(1+\epsilon))} \tag{8.91}$$

setting $s = \frac{\epsilon}{2(1+\epsilon)}$.
Thus

$$\Pr\left\{\left|-\frac{1}{n}\ln f(X^n) - h_n\right| > \epsilon\right\} = \Pr\left\{\left|\frac{1}{2}\left(-1 + \frac{1}{n}\sum W_i^2\right)\right| > \epsilon\right\} \quad (8.92)$$

$$\leq e^{-\frac{n}{2}(\epsilon - \ln(1+\epsilon))} \quad (8.93)$$

and the bound goes to 0 as $n \to \infty$, and therefore by the Borel Cantelli lemma,

$$-\frac{1}{n}\ln f(X^n) - h_n \to 0 \quad (8.94)$$

with probability 1. So X_i satisfies the AEP even though it is not ergodic.

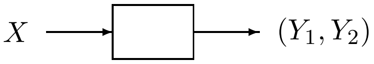
Chapter 9

Gaussian channel

1. **A channel with two independent looks at Y.** Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X.

   (a) Show $I(X;Y_1,Y_2) = 2I(X;Y_1) - I(Y_1;Y_2)$.

   (b) Conclude that the capacity of the channel



   is less than twice the capacity of the channel

Solution: *Channel with two independent looks at* Y.

(a)

$$\begin{aligned}
I(X;Y_1,Y_2) &= H(Y_1,Y_2) - H(Y_1,Y_2|X) && (9.1)\\
&= H(Y_1) + H(Y_2) - I(Y_1;Y_2) - H(Y_1|X) - H(Y_2|X) && (9.2)\\
&\qquad \text{(since } Y_1 \text{ and } Y_2 \text{ are conditionally independent given } X\text{)} && (9.3)\\
&= I(X;Y_1) + I(X;Y_2) - I(Y_1;Y_2) && (9.4)\\
&= 2I(X;Y_1) - I(Y_1;Y_2) \quad \text{(since } Y_1 \text{ and } Y_2 \text{ are conditionally identically distributed)} && (9.5)
\end{aligned}$$

(b) The capacity of the single look channel $X \to Y_1$ is

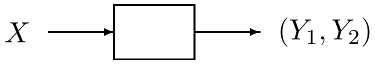
$$C_1 = \max_{p(x)} I(X;Y_1). \tag{9.6}$$

The capacity of the channel $X \rightarrow (Y_1, Y_2)$ is

$$\begin{aligned} C_2 &= \max_{p(x)} I(X;Y_1,Y_2) && (9.7)\\ &= \max_{p(x)} 2I(X;Y_1) - I(Y_1;Y_2) && (9.8)\\ &\leq \max_{p(x)} 2I(X;Y_1) && (9.9)\\ &= 2C_1. && (9.10) \end{aligned}$$

Hence, two independent looks cannot be more than twice as good as one look.

2. **The two-look Gaussian channel.**



Consider the ordinary Gaussian channel with two correlated looks at X, i.e., $Y = (Y_1, Y_2)$, where

$$\begin{aligned} Y_1 &= X + Z_1 && (9.11)\\ Y_2 &= X + Z_2 && (9.12) \end{aligned}$$

with a power constraint P on X, and $(Z_1, Z_2) \sim \mathcal{N}_2(\mathbf{0}, K)$, where

$$K = \begin{bmatrix} N & N\rho \\ N\rho & N \end{bmatrix}. \qquad (9.13)$$

Find the capacity C for

(a) $\rho = 1$

(b) $\rho = 0$

(c) $\rho =$ -1

Solution: *The two look Gaussian channel.*

It is clear that the input distribution that maximizes the capacity is $X \sim \mathcal{N}(0, P)$. Evaluating the mutual information for this distribution,

$$\begin{aligned} C_2 &= \max I(X;Y_1,Y_2) && (9.14)\\ &= h(Y_1,Y_2) - h(Y_1,Y_2|X) && (9.15)\\ &= h(Y_1,Y_2) - h(Z_1,Z_2|X) && (9.16)\\ &= h(Y_1,Y_2) - h(Z_1,Z_2) && (9.17) \end{aligned}$$

Now since

$$(Z_1,Z_2) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} N & N\rho \\ N\rho & N \end{bmatrix}\right), \qquad (9.18)$$

we have

$$h(Z_1, Z_2) = \frac{1}{2}\log(2\pi e)^2|K_Z| = \frac{1}{2}\log(2\pi e)^2 N^2(1-\rho^2). \tag{9.19}$$

Since $Y_1 = X + Z_1$, and $Y_2 = X + Z_2$, we have

$$(Y_1, Y_2) \sim \mathcal{N}\left(\mathbf{0}, \left[\begin{array}{cc} P+N & P+\rho N \\ P+\rho N & P+N \end{array}\right]\right), \tag{9.20}$$

and

$$h(Y_1, Y_2) = \frac{1}{2}\log(2\pi e)^2|K_Y| = \frac{1}{2}\log(2\pi e)^2(N^2(1-\rho^2) + 2PN(1-\rho)). \tag{9.21}$$

Hence the capacity is

$$\begin{aligned} C_2 &= h(Y_1, Y_2) - h(Z_1, Z_2) && (9.22) \\ &= \frac{1}{2}\log\left(1 + \frac{2P}{N(1+\rho)}\right). && (9.23) \end{aligned}$$

(a) $\rho = 1$. In this case, $C = \frac{1}{2}\log(1+\frac{P}{N})$, which is the capacity of a single look channel. This is not surprising, since in this case $Y_1 = Y_2$.

(b) $\rho = 0$. In this case,

$$C = \frac{1}{2}\log\left(1 + \frac{2P}{N}\right), \tag{9.24}$$

which corresponds to using twice the power in a single look. The capacity is the same as the capacity of the channel $X \rightarrow (Y_1 + Y_2)$.

(c) $\rho = -1$. In this case, $C = \infty$, which is not surprising since if we add Y_1 and Y_2, we can recover X exactly.

Note that the capacity of the above channel in all cases is the same as the capacity of the channel $X \rightarrow Y_1 + Y_2$.

3. **Output power constraint**. Consider an additive white Gaussian noise channel with an expected *output* power constraint P. Thus $Y = X + Z$, $Z \sim N(0, \sigma^2)$, Z is independent of X, and $EY^2 \le P$. Find the channel capacity.

Solution: *Output power constraint*

$$\begin{aligned} C &= \max_{f(X):E(X+Z)^2 \le P} I(X;Y) && (9.25) \\ &= \max_{f(X):E(X+Z)^2 \le P} (h(Y) - h(Y|X)) && (9.26) \\ &= \max_{f(X):E(X+Z)^2 \le P} (h(Y) - h(Z)) && (9.27) \\ & && (9.28) \end{aligned}$$

Given a constraint on the output power of Y, the maximum differential entropy is achieved by a normal distribution, and we can achieve this by have $X \sim \mathcal{N}(0, P-N)$, and in this case,

$$C = \frac{1}{2}\log 2\pi eP - \frac{1}{2}\log 2\pi eN = \frac{1}{2}\log\frac{P}{N}. \tag{9.29}$$

4. **Exponential noise channels.** Consider an additive noise channel $Y_i = X_i + Z_i$, where Z_i is i.i.d. exponentially distributed noise with mean μ. Assume that we have a mean constraint on the signal, i.e., $EX_i \leq \lambda$. Show that the capacity of such a channel is $C = \log(1 + \frac{\lambda}{\mu})$.

   Solution: *Exponential noise channels*

   Just as for the Gaussian channel, we can write

$$\begin{aligned} C &= \max_{f(X):EX\leq\lambda} I(X;Y) && (9.30)\\ &= \max_{f(X):EX\leq\lambda} h(Y) - h(Y|X) && (9.31)\\ &= \max_{f(X):EX\leq\lambda} h(Y) - h(Z|X) && (9.32)\\ &= \max_{f(X):EX\leq\lambda} h(Y) - h(Z) && (9.33)\\ &= \max_{f(X):EX\leq\lambda} h(Y) - (1 + \ln\mu) && (9.34)\\ & && (9.35)\end{aligned}$$

   Now $Y = X + Z$, and $EY = EX + EZ \leq \lambda + \mu$. Given a mean constraint, the entropy is maximized by the exponential distribution, and therefore

$$\max_{EY\leq\lambda+\mu} h(Y) = 1 + \ln(\lambda+\mu) \qquad (9.36)$$

   Unlike normal distributions, though, the sum of two exponentially distributed variables is not exponential, so we cannot set X to be an exponential distribution to achive the right distribution of Y. Instead, we can use characterstic functions to find the distribution of X. The characteristic function of an exponential distribution

$$\psi(t) = \int \frac{1}{\mu} e^{-\frac{x}{\mu}} e^{-itx}\,dx = \frac{1}{1 - i\mu t} \qquad (9.37)$$

   The distribution of X that when added to Z will give an exponential distribution for Y is the ratio of the characterstic functions

$$\begin{aligned} \psi_X(t) &= \frac{1 - i\mu t}{1 - i(\lambda+\mu)t} && (9.38)\\ & && (9.39)\end{aligned}$$

   which can been seen to correpond to mixture of a point mass and an exponential distribution. If

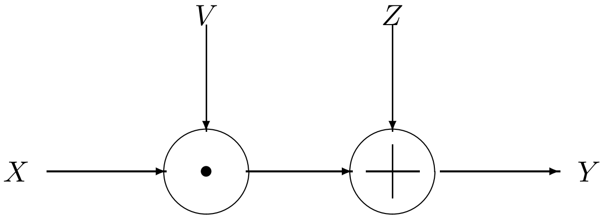
$$X = \begin{cases} 0, & \text{with probability } \frac{\mu}{\lambda+\mu} \\ X_e, & \text{with probability} \frac{\lambda}{\lambda+\mu} \end{cases} \qquad (9.40)$$

   where X_e has an exponential distribution with parameter $\mu + \lambda$, we can verify that the characterstic function of X is correct.

   Using the value of entropy for exponential distributions, we get

$$C = h(Y) - h(Z) = 1 + + \ln(\lambda+\mu) - (1 + \ln\mu) = \ln\left(1 + \frac{\lambda}{\mu}\right) \qquad (9.41)$$

5. **Fading channel.**
Consider an additive noise fading channel



$$Y = XV + Z,$$

where Z is additive noise, V is a random variable representing fading, and Z and V are independent of each other and of X. Argue that knowledge of the fading factor V improves capacity by showing

$$I(X;Y|V) \geq I(X;Y).$$

Solution: *Fading Channel*

Expanding $I(X;Y,V)$ in two ways, we get

$$\begin{aligned} I(X;Y,V) &= I(X;V) + I(X;Y|V) & (9.42)\\ &= I(X;Y) + I(X;V|Y) & (9.43)\end{aligned}$$

i.e.

$$\begin{aligned} I(X;V) + I(X;Y|V) &= I(X;Y) + I(X;V|Y) \\ I(X;Y|V) &= I(X;Y) + I(X;V|Y) & (9.44)\\ I(X;Y|V) &\geq I(X;Y) & (9.45)\end{aligned}$$

where (9.44) follows from the independence of X and V, and (9.45) follows from $I(X;V|Y) \geq 0$.

6. **Parallel channels and waterfilling.** Consider a pair of parallel Gaussian channels, i.e.,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}, \qquad (9.46)$$

where

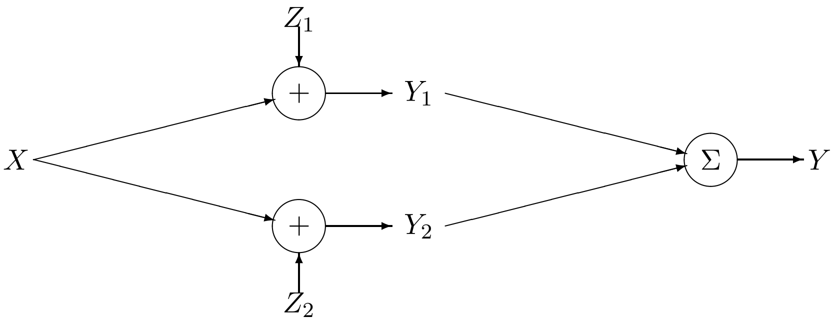
$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right), \qquad (9.47)$$

and there is a power constraint $E(X_1^2 + X_2^2) \leq 2P$. Assume that $\sigma_1^2 > \sigma_2^2$. At what power does the channel stop behaving like a single channel with noise variance σ_2^2, and begin behaving like a pair of channels?

Solution: *Parallel channels and waterfilling.* By the result of Section 10.4, it follows that we will put all the signal power into the channel with less noise until the total power of noise + signal in that channel equals the noise power in the other channel. After that, we will split any additional power evenly between the two channels.

Thus the combined channel begins to behave like a pair of parallel channels when the signal power is equal to the difference of the two noise powers, i.e., when $2P = \sigma_1^2 - \sigma_2^2$.

7. **Multipath Gaussian channel.** Consider a Gaussian noise channel of power contraint P, where the signal takes two different paths and the received noisy signals are added together at the antenna.

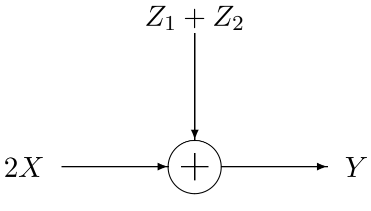


(a) Find the capacity of this channel if Z_1 and Z_2 are jointly normal with covariance matrix $K_Z = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}$.

(b) What is the capacity for $\rho = 0$, $\rho = 1$, $\rho = -1$?

Solution: *Multipath Gaussian channel.*

The channel reduces to the following channel:



The power constraint on the input $2X$ is $4P$. Z_1 and Z_2 are zero mean, and therefore so is $Z_1 + Z_2$. Then

$$\begin{aligned} Var(Z_1 + Z_2) &= E[(Z_1 + Z_2)^2] \\ &= E[Z_1^2 + Z_2^2 + 2Z_1Z_2] \\ &= 2\sigma^2 + 2\rho\sigma^2. \end{aligned}$$

Thus the noise distribution is $\mathcal{N}(0, 2\sigma^2(1+\rho)$.

(a) Plugging the noise and power values into the formula for the one-dimensional (P,N) channel capacity, $C = \frac{1}{2}\log(1+\frac{P}{N})$, we get

$$\begin{aligned} C &= \frac{1}{2}\log\left(1+\frac{4P}{2\sigma^2(1+\rho)}\right) \\ &= \frac{1}{2}\log\left(1+\frac{2P}{\sigma^2(1+\rho)}\right). \end{aligned}$$

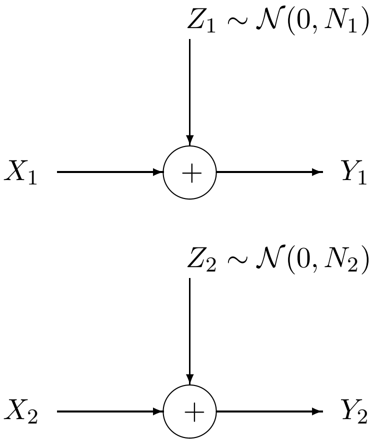
(b) i. When $\rho = 0$, $C = \frac{1}{2}\log(1+\frac{2P}{\sigma^2})$.

ii. When $\rho = 1$, $C = \frac{1}{2}\log(1+\frac{P}{\sigma^2})$.

iii. When $\rho = -1$, $C = \infty$.

8. **Parallel Gaussian channels**

Consider the following parallel Gaussian channel


$Z_1 \sim \mathcal{N}(0, N_1)$
X_1
+
Y_1
$Z_2 \sim \mathcal{N}(0, N_2)$
X_2
+
Y_2

where $Z_1 \sim \mathcal{N}(0,N_1)$ and $Z_2 \sim \mathcal{N}(0,N_2)$ are independent Gaussian random variables and $Y_i = X_i + Z_i$. We wish to allocate power to the two parallel channels. Let β_1 and β_2 be fixed. Consider a total cost constraint $\beta_1 P_1 + \beta_2 P_2 \le \beta$, where P_i is the power allocated to the i^{th} channel and β_i is the cost per unit power in that channel. Thus $P_1 \ge 0$ and $P_2 \ge 0$ can be chosen subject to the cost constraint β.

(a) For what value of β does the channel stop acting like a single channel and start acting like a pair of channels?

(b) Evaluate the capacity and find P_1, P_2 that achieve capacity for $\beta_1 = 1, \beta_2 = 2, N_1 = 3, N_2 = 2$ and $\beta = 10$.

Solution: *Parallel channels*

When we have cost constraints on the power, we need to optimize the total capacity of the two parallel channels

$$C = \frac{1}{2}\log\left(1+\frac{P_1}{N_1}\right) + \frac{1}{2}\log\left(1+\frac{P_2}{N_2}\right) \tag{9.48}$$

subject to the constraint that

$$\beta_1 P_1 + \beta_2 P_2 \le \beta \tag{9.49}$$

Using the methods of Section 9.4, we set

$$J(P_1, P_2) = \sum \frac{1}{2}\log\left(1+\frac{P_i}{N_i}\right) + \lambda(\sum \beta_i P_i) \tag{9.50}$$

and differentiating with respect to P_i, we have

$$\frac{1}{2}\frac{1}{P_i+N_i} + \lambda\beta_i = 0, \tag{9.51}$$

or

$$P_i = (\frac{\nu}{\beta_i} - N_i)^+. \tag{9.52}$$

or

$$\beta_i P_i = (\nu - \beta_i N_i)^+. \tag{9.53}$$

(a) It follows that we will put all the signal power into the channel with less weighted noise ($\beta_i N_i$) until the total weighted power of noise + signal in that channel equals the weighted noise power in the other channel. After that, we will split any additional power between the two channels according to their weights. Thus the combined channel begins to behave like a pair of parallel channels when the signal power is equal to the difference of the two weighted noise powers, i.e., when $\beta_1\beta = \beta_2 N - 2 - \beta_1 N_1$.

(b) In this case, $\beta_1 N_1 < \beta_2 N_2$, so we would put power into channel 1 until $\beta = 1$. After that we would put power according to their weights, i.e. we would divide remaining power of 9 in the ratio 2 is to 1. Thus we would set $P_1 = 6+1$ and $P_2 = 3$, and so that $\nu = 10$ in the equation above. The capacity in this case is

$$C = \frac{1}{2}\log(1+7/3) + \frac{1}{2}\log(1+3/2) = 1.53 \text{ bits.} \tag{9.54}$$

9. **Vector Gaussian channel**

Consider the vector Gaussian noise channel

$$Y = X + Z,$$

where $X = (X_1, X_2, X_3)$, $Z = (Z_1, Z_2, Z_3)$, and $Y = (Y_1, Y_2, Y_3)$, $E\|X\|^2 \le P$, and

$$Z \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}\right).$$

Find the capacity. The answer may be surprising.

Solution: *Vector Gaussian channel*

Normally one would water-fill over the eigenvalues of the noise covariance matrix. Here we have the degenerate case (i.e., one of the eigenvalue is zero), which we can exploit easily.

Musing upon the structure of the noise covariance matrix, one can see $Z_1 + Z_2 = Z_3$. Thus, by processing the output vector as $Y_1+Y_2-Y_3 = (X_1+Z_1)+(X_2+Z_2)-(X_3+Z_3) = X_1+X_2-X_3$, we can get rid of the noise completely. Therefore, we have infinite capacity.

Note that we can reach the conclusion by water-filling on the zero eigenvalue.

10. **The capacity of photographic film.** Here is a problem with a nice answer that takes a little time. We're interested in the capacity of photographic film. The film consists of silver iodide crystals, Poisson distributed, with a density of λ particles per square inch. The film is illuminated without knowledge of the position of the silver iodide particles. It is then developed and the receiver sees only the silver iodide particles that have been illuminated. It is assumed that light incident on a cell exposes the grain if it is there and otherwise results in a blank response. Silver iodide particles that are not illuminated and vacant portions of the film remain blank. The question is, "What is the capacity of this film?"

We make the following assumptions. We grid the film very finely into cells of area dA. It is assumed that there is at most one silver iodide particle per cell and that no silver iodide particle is intersected by the cell boundaries. Thus, the film can be considered to be a large number of parallel binary asymmetric channels with crossover probability $1-\lambda dA$.

By calculating the capacity of this binary asymmetric channel to first order in dA (making the necessary approximations) one can calculate the capacity of the film in bits per square inch. It is, of course, proportional to λ. The question is what is the multiplicative constant?

The answer would be λ bits per unit area if both illuminator and receiver knew the positions of the crystals.

Solution: *Capacity of photographic film*

As argued in the problem, each small cell can be modelled as a binary asymmetric Z-channel with probability transition matrix

$$p(y|x) = \begin{bmatrix} 1 & 0 \\ 1-\lambda\, dA & \lambda\, dA \end{bmatrix} \qquad x,y \in \{0,1\} \tag{9.55}$$

where $x=1$ corresponds to shining light on the cell. Let $\beta = \lambda\, dA$.

First we express $I(X;Y)$, the mutual information between the input and output of the Z-channel, as a function of $\alpha = \Pr(X=1)$:

$$H(Y|X) \;=\; \Pr(X=0)\cdot 0 + \Pr(X=1)\cdot H(\beta) = \alpha H(\beta)$$

$$\begin{aligned} H(Y) &= \mathbf{H}(\Pr(Y=1)) = H(\alpha\beta) \\ I(X;Y) &= H(Y) - H(Y|X) = H(\alpha\beta) - \alpha H(\beta) \end{aligned}$$

Since $I(X;Y) = 0$ when $\alpha = 0$ and $\alpha = 1$, the maximum mutual information is obtained for some value of α such that $0 < \alpha < 1$.

Using elementary calculus, we determine that (converting the equation to nats rather than bits),

$$\frac{d}{d\alpha} I(X;Y) = \beta \ln \frac{1-\alpha\beta}{\alpha\beta} - H_e(\beta)$$

To find the optimal value of α, we set this equal to 0, and solve for α as

$$\alpha = \frac{1}{\beta} \frac{1}{1+e^{\frac{H_e(\beta)}{\beta}}} \tag{9.56}$$

If we let

$$\gamma = \frac{1}{1+e^{\frac{H_e(\beta)}{\beta}}} \tag{9.57}$$

then $\alpha\beta = \gamma$, and

$$\overline{\gamma} = 1-\gamma = \frac{e^{\frac{H_e(\beta)}{\beta}}}{1+e^{\frac{H_e(\beta)}{\beta}}} = \gamma e^{\frac{H_e(\beta)}{\beta}} \tag{9.58}$$

or

$$\ln \overline{\gamma} - \ln \gamma = \frac{H_e(\beta)}{\beta} \tag{9.59}$$

so that

$$\begin{aligned} I(X;Y) &= H_e(\alpha\beta) - \alpha H_e(\beta) && (9.60) \\ &= H_e(\gamma) - \frac{1}{1+e^{\frac{H_e(\beta)}{\beta}}} \frac{H_e(\beta)}{\beta} && (9.61) \\ &= -\gamma \ln \gamma - \overline{\gamma} \ln \overline{\gamma} - \gamma(\ln \overline{\gamma} - \ln \gamma) && (9.62) \\ &= -\ln \overline{\gamma} && (9.63) \\ &= \ln(1+e^{-\frac{H_e(\beta)}{\beta}}) && (9.64) \\ &\approx e^{-\frac{H_e(\beta)}{\beta}} && (9.65) \\ &= e^{-\frac{-\beta\ln\beta-(1-\beta)\ln(1-\beta)}{\beta}} && (9.66) \\ &\approx e^{--\ln\beta} && (9.67) \\ &= \beta && (9.68) \end{aligned}$$

Thus the capacity of this channel is approximately β nats when $\beta \to 0$.

11. **Gaussian mutual information**. Suppose that (X,Y,Z) are jointly Gaussian and that $X \to Y \to Z$ forms a Markov chain. Let X and Y have correlation coefficient ρ_1 and let Y and Z have correlation coefficient ρ_2. Find $I(X;Z)$.

Solution: *Gaussian Mutual Information (Repeat of problem 8.9)*
First note that we may without any loss of generality assume that the means of X, Y and Z are zero. If in fact the means are not zero one can subtract the vector of means without affecting the mutual information or the conditional independence of X, Z given Y. Let

$$\Lambda = \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_z\rho_{xz} \\ \sigma_x\sigma_z\rho_{xz} & \sigma_z^2 \end{pmatrix},$$

be the covariance matrix of X and Z. We can now use Eq. (9.93) and Eq. (9.94) to compute

$$\begin{aligned} I(X;Z) &= h(X) + h(Z) - h(X,Z) \\ &= \frac{1}{2}\log\left(2\pi e\sigma_x^2\right) + \frac{1}{2}\log\left(2\pi e\sigma_x^2\right) - \frac{1}{2}\log\left(2\pi e|\Lambda|\right) \\ &= -\frac{1}{2}\log(1-\rho_{xz}^2) \end{aligned}$$

Now,

$$\begin{aligned} \rho_{xz} &= \frac{\mathbf{E}\{XZ\}}{\sigma_x\sigma_z} \\ &= \frac{\mathbf{E}\{\mathbf{E}\{XZ|Y\}\}}{\sigma_x\sigma_z} \\ &= \frac{\mathbf{E}\{\mathbf{E}\{X|Y\}\mathbf{E}\{Z|Y\}\}}{\sigma_x\sigma_z} \\ &= \frac{\mathbf{E}\{\left(\frac{\sigma_x\rho_{xy}}{\sigma_y}Y\right)\left(\frac{\sigma_z\rho_{zx}}{\sigma_y}Y\right)\}}{\sigma_x\sigma_z} \\ &= \rho_{xy}\rho_{zy} \end{aligned}$$

We can thus conclude that

$$I(X;Y) = -\frac{1}{2}\log(1-\rho_{xy}^2\rho_{zy}^2)$$

12. **Time varying channel.** A train pulls out of the station at constant velocity. The received signal energy thus falls off with time as $1/i^2$. The total received signal at time i is

$$Y_i = \left(\frac{1}{i}\right)X_i + Z_i,$$

where $Z_1, Z_2, \ldots$ are i.i.d. $\sim N(0,N)$. The transmitter constraint for block length n is

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2(w) \le P, \quad w \in \{1,2,\ldots,2^{nR}\}.$$

Using Fano's inequality, show that the capacity C is equal to zero for this channel.

Solution: *Time Varying Channel*

Just as in the proof of the converse for the Gaussian channel

$$\begin{aligned}
nR = H(W) &= I(W;\hat{W}) + H(W|\hat{W}) && (9.69)\\
&\le I(W;\hat{W}) + n\epsilon_n && (9.70)\\
&\le I(X^n;Y^n) + n\epsilon_n && (9.71)\\
&= h(Y^n) - h(Y^n|X^n) + n\epsilon_n && (9.72)\\
&= h(Y^n) - h(Z^n) + n\epsilon_n && (9.73)\\
&\le \sum_{i=1}^n h(Y_i) - h(Z^n) + n\epsilon_n && (9.74)\\
&= \sum_{i=1}^n h(Y_i) - \sum_{i=1}^n h(Z_i) + n\epsilon_n && (9.75)\\
&= \sum_{i=1}^n I(X_i;Y_i) + n\epsilon_n. && (9.76)
\end{aligned}$$

Now let P_i be the average power of the ith column of the codebook, i.e.,

$$P_i = \frac{1}{2^{nR}} \sum_w x_i^2(w). \qquad (9.77)$$

Then, since $Y_i = \frac{1}{i}X_i + Z_i$ and since X_i and Z_i are independent, the average power of Y_i is $\frac{1}{i^2}P_i + N$. Hence, since entropy is maximized by the normal distribution,

$$h(Y_i) \le \frac{1}{2}\log 2\pi e(\frac{1}{i^2}P_i + N). \qquad (9.78)$$

Continuing with the inequalities of the converse, we obtain

$$\begin{aligned}
nR &\le \sum (h(Y_i) - h(Z_i)) + n\epsilon_n && (9.79)\\
&\le \sum \left(\frac{1}{2}\log(2\pi e(\frac{1}{i^2}P_i + N)) - \frac{1}{2}\log 2\pi eN\right) + n\epsilon_n && (9.80)\\
&= \sum \frac{1}{2}\log\left(1 + \frac{P_i}{i^2 N}\right) + n\epsilon_n. && (9.81)
\end{aligned}$$

Since each of the codewords satisfies the power constraint, so does their average, and hence

$$\frac{1}{n}\sum_i P_i \le P. \qquad (9.82)$$

This corresponds to a set of parallel channels with increasing noise powers. Using waterfilling, the optimal solution is to put power into the first few channels which have the lowest noise power. Since the noise power in the channel i is $N_i = i^2 N$, we will put power into channels only where $P_i + N_i \le \lambda$. The height of the water level in the water filling is less than $N+nP$, and hence the for all channels we put power, $i^2N < nP+N$, or only $o(\sqrt{n})$ channels. The average rate is less than $\frac{1}{n}\sqrt{n}\frac{1}{2}\log(1+nP/N)$ and the capacity per transmission goes to 0. Hence there capacity of this channel is 0.

13. **Feedback capacity for** $n = 2$**.** Let $(Z_1, Z_2) \sim N(0, K)$, $\quad K = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. Find the maximum of $\frac{1}{2}\log\frac{|K_{X+Z}|}{|K_Z|}$ with and without feedback given a trace (power) constraint $tr(K_X) \le 2P$.

Solution: *Feedback capacity*

Without feedback, the solution is based on waterfilling. The eigenvalues of the matrix are $1 \pm \rho$, and therefore if $P < \rho$, we would use only one of the channels, and achieve capacity $C = \frac{1}{2}\log(1 + \frac{2P}{1-\rho})$. For $P \ge \rho$, we would use both eigenvalues and the waterlevel for water filling would be obtained by distributing the remaining power equally across both eigenvalues. Thus the water level would be $(1+\rho)+(2P-2\rho)/2 = 1 + P$, and the capacity would be $C = \frac{1}{2}\log(\frac{1+P}{1+\rho}) + \frac{1}{2}\log(\frac{1+P}{1-\rho})$.

With feedback, the solution is a a little more complex. From (9.102), we have

$$C_{n,FB} = \max \frac{1}{2n} \log \frac{|(B+I)K_Z^{(n)}(B+I)^t + K_V|}{|K_Z^{(n)}|} \tag{9.83}$$

where the maximum is taken over all nonnegative definite K_V and strictly lower triangular B such that

$$tr(BK_Z^{(n)}B^t + K_V) \le nP. \tag{9.84}$$

In the case when $n = 2$,

$$\begin{aligned}(B+I)K_Z^{(n)}(B+I)^t + K_V &= \begin{pmatrix} 1 & 0 \\ b & 1 \end{pmatrix}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} P_1 & 0 \\ 0 & P_2 \end{pmatrix} & (9.85)\\ &= \begin{pmatrix} 1+P_1 & \rho+b \\ \rho+b & 1+P_2+2\rho b+b^2 \end{pmatrix} & (9.86)\end{aligned}$$

subject to the constraint that

$$\text{tr}\begin{pmatrix} P_1 & 0 \\ 0 & P_2+b^2 \end{pmatrix} \le 2P \tag{9.87}$$

Expanding this, we obtain the mutual information as

$$I(X;Y) = 1 + P_1 + P_2 + P_1P_2 + P_1b^2 + 2P_1b\rho - \rho^2 \tag{9.88}$$

subject to

$$P_1 + P_2 + b^2 = 2P \tag{9.89}$$

Setting up the functional and differentiating with respect to the variables, we obtain the following relationships

$$P_1 = P_2 + b^2 + 2b\rho \tag{9.90}$$

and

$$b = \rho P_1 \tag{9.91}$$

14. **Additive noise channel.** Consider the channel $Y = X + Z$, where X is the transmitted signal with power constraint P, Z is independent additive noise, and Y is the received signal. Let

$$Z = \begin{cases} 0, & \text{with prob. } \frac{1}{10} \\ Z^*, & \text{with prob. } \frac{9}{10} \end{cases},$$

where $Z^* \sim N(0, N)$. Thus Z has a mixture distribution which is the mixture of a Gaussian distribution and a degenerate distribution with mass 1 at 0.

(a) What is the capacity of this channel? This should be a pleasant surprise.

(b) How would you signal in order to achieve capacity?

Solution: *Additive Noise channel*

The capacity of this channel is infinite, since at the times the noise is 0 the output is exactly equal to the input, and we can send an infinite number of bits.

To send information through this channel, just repeat the same real number at the input. When we have three or four outputs that agree, that should correspond to the points where the noise is 0, and we can decode an infinite number of bits.

15. **Discrete input continuous output channel.** Let $\Pr\{X = 1\} = p$, $\Pr\{X = 0\} = 1 - p$, and let $Y = X + Z$, where Z is uniform over the interval $[0, a]$, $a > 1$, and Z is independent of X.

(a) Calculate

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

(b) Now calculate $I(X;Y)$ the other way by

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

(c) Calculate the capacity of this channel by maximizing over p

Solution: *Discrete input Continuous Output channel*

(a) Since

$$f(Y|X=0) = \begin{cases} \frac{1}{a} & 0 \le y < a \\ 0 & \text{otherwise} \end{cases} \tag{9.92}$$

and

$$f(Y|X=1) = \begin{cases} (1-p)\frac{1}{a} & 0 \le y < 1 \\ \frac{1}{a} & 1 \le y \le a \\ p\frac{1}{a} & a < y < 1+a \end{cases} \tag{9.93}$$

Therefore,

$$f(y) = \begin{cases} (1-p)\frac{1}{a} & 0 \le y < 1 \\ \frac{1}{a} & 1 \le y \le a \\ p\frac{1}{a} & a < y < 1+a \end{cases} \tag{9.94}$$

(b) $H(X) = H(p)$. $H(X|Y = y)$ is nonzero only for $1 \leq y \leq a$, and by Bayes rule, conditioned on Y, the probabilty that $X = 1$ is

$$P(X = 1|Y = y) = \frac{P(X = 1)f(y|X = 1)}{P(X = 1)f(y|X = 1) + P(X = 0)f(y|X = 0)} = p \qquad (9.95)$$

and hence $H(X|Y) = P(1 \leq Y \leq a)H(p) = \frac{a-1}{a}H(p)$. Therefore $I(X;Y) = H(X) - H(X|Y) = \frac{1}{a}H(p)$.

(c) $f(Y|X = 0) \sim U(0, a)$, and hence $h(Y|X = 0) = \log a$, and similarly for $X = 1$, so that $h(Y|X) = \log a$.

The differential entropy $h(Y)$ can be calculated from (9.94) as

$$\begin{aligned} h(Y) &= -\int_0^1 (1-p)\frac{1}{a}\log\frac{1-p}{a}dy - \int_1^a \frac{1}{a}\log\frac{1}{a}dy - \int_a^{1+a}\frac{p}{a}\log\frac{p}{a}dy && (9.96) \\ &= \frac{1}{a}(-p\log p - (1-p)\log(1-p)) + \log a && (9.97) \\ &= \frac{1}{a}H(p) + \log a && (9.98) \end{aligned}$$

and again $I(X;Y) = h(Y) - h(Y|X) = \frac{1}{a}H(p)$.

(d) The mutual information is maximized for $p = 0.5$, and the corresponding capacity of the channel is $\frac{1}{a}$.

16. **Gaussian mutual information**
Suppose that (X, Y, Z) are jointly Gaussian and that $X \to Y \to Z$ forms a Markov chain. Let X and Y have correlation coefficient ρ_1 and let Y and Z have correlation coefficient ρ_2. Find $I(X;Z)$.

Solution: *Gaussian Mutual Information (Repeat of problem 8.9)*
First note that we may without any loss of generality assume that the means of X, Y and Z are zero. If in fact the means are not zero one can subtract the vector of means without affecting the mutual information or the conditional independence of X, Z given Y. Let

$$\Lambda = \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_z\rho_{xz} \\ \sigma_x\sigma_z\rho_{xz} & \sigma_z^2 \end{pmatrix},$$

be the covariance matrix of X and Z. We can now use Eq. (9.93) and Eq. (9.94) to compute

$$\begin{aligned} I(X;Z) &= h(X) + h(Z) - h(X,Z) \\ &= \frac{1}{2}\log(2\pi e\sigma_x^2) + \frac{1}{2}\log(2\pi e\sigma_x^2) - \frac{1}{2}\log(2\pi e|\Lambda|) \\ &= -\frac{1}{2}\log(1-\rho_{xz}^2) \end{aligned}$$

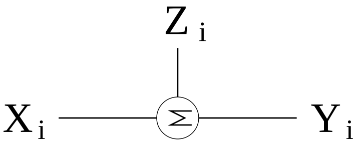
Now,

$$\begin{aligned}
\rho_{xz} &= \frac{\mathbf{E}\{XZ\}}{\sigma_x\sigma_z} \\
&= \frac{\mathbf{E}\{\mathbf{E}\{XZ|Y\}\}}{\sigma_x\sigma_z} \\
&= \frac{\mathbf{E}\{\mathbf{E}\{X|Y\}\mathbf{E}\{Z|Y\}\}}{\sigma_x\sigma_z} \\
&= \frac{\mathbf{E}\{\left(\frac{\sigma_x\rho_{xy}}{\sigma_y}Y\right)\left(\frac{\sigma_z\rho_{zx}}{\sigma_y}Y\right)\}}{\sigma_x\sigma_z} \\
&= \rho_{xy}\rho_{zy}
\end{aligned}$$

We can thus conclude that

$$I(X;Y) = -\frac{1}{2}\log(1-\rho_{xy}^2\rho_{zy}^2)$$

17. **Impulse power.**
Consider the additive white Gaussian channel



where $Z_i \sim N(0,N)$, and the input signal has average power constraint P.

(a) Suppose we use all our power at time 1, i.e. $EX_1^2 = nP$ and $EX_i^2 = 0$, for $i = 2,3,\ldots,n$. Find

$$\max_{f(x^n)} \frac{I(X^n;Y^n)}{n}$$

where the maximization is over all distributions $f(x^n)$ subject to the constraint $EX_1^2 = nP$ and $EX_i^2 = 0$, for $i = 2,3,\ldots,n$.

(b) Find

$$\max_{f(x^n):\ E\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)\le P} \frac{1}{n}I(X^n;Y^n)$$

and compare to part (a).

Solution: *Impulse power.*

(a)

$$\begin{aligned} \max \frac{I(X^n;Y^n)}{n} &\overset{(a)}{=} \max \frac{I(X_1;Y_1)}{n} \\ &\overset{(b)}{=} \frac{\frac{1}{2}\log\left(1+\frac{nP}{N}\right)}{n} \end{aligned}$$

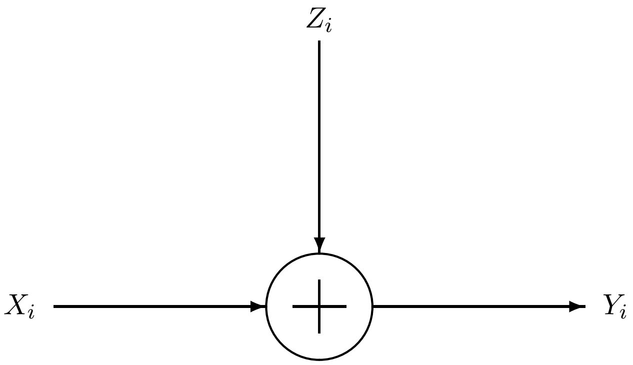
where (a) comes from the constraint that all our power, nP, be used at time 1 and (b) comes from that fact that given Gaussian noise and a power constraint nP, $I(X;Y) \le \frac{1}{2}\log(1+\frac{nP}{N})$.

(b)

$$\begin{aligned} \max \frac{I(X^n;Y^n)}{n} &\overset{(a)}{=} \max \frac{nI(X;Y)}{n} \\ &= \max I(X;Y) \\ &= \frac{1}{2}\log\left(1+\frac{P}{N}\right) \end{aligned}$$

where (a) comes from the fact that the channel is memoryless. Notice that the quantity in part (a) goes to zero as $n \to \infty$ while the quantity in part (b) stays constant. Hence the impulse scheme is suboptimal.

18. **Gaussian channel with time-varying mean.** Find the capacity of the following Gaussian channels.



Let $Z_1, Z_2, \ldots$ be independent and let there be a power contraint P on $x^n(W)$. Find the capacity when

(a) $\mu_i = 0$, for all i.

(b) $\mu_i = e^i, \quad i = 1, 2, \ldots$ Assume μ_i known to the transmitter and receiver.

(c) μ_i unknown, but μ_i i.i.d. $\sim N(0,N_1)$ for all i.

Solution: *Gaussian Noise with time-varying mean*

(a) This is the classical Gaussian channel capacity problem with

$$C = \frac{1}{2}\log\left(1+\frac{P}{N}\right).$$

(b) Since the transmitter and the receiver both know the means, the receiver can simply subtract the mean while decoding. Thus, we are back in case (a). Hence the capacity is

$$C = \frac{1}{2}\log\left(1+\frac{P}{N}\right).$$

(c) Let p_i be the density of Z_i. Clearly p_i is independent of the time index i. Also

$$\begin{aligned} p(y) &= \int \frac{1}{\sqrt{2\pi N1}} e^{-\frac{\mu^2}{2N1}} \frac{1}{\sqrt{2\pi N}} e^{-\frac{(y-\mu)^2}{2N}} \\ &= N(0,N) * N(0,N_1) \\ &= N(0,N+N_1), \end{aligned}$$

where $*$ represents convolution. From the distribution of Z_i it is obvious that the optimal input distribution X_i is $N(0,P)$ and the capacity is

$$C = \frac{1}{2}\log\left(1+\frac{P}{N+N_1}\right).$$

19. **A parametric form for channel capacity**

Consider m parallel Gaussian channels, $Y_i = X_i + Z_i$, where $Z_i \sim N(0,\lambda_i)$ and the noises X_i are independent r.v.'s. Thus $C = \sum_{i=1}^m \frac{1}{2}\log(1+\frac{(\lambda-\lambda_i)^+}{\lambda_i})$ where λ is chosen to satisfy $\sum_{i=1}^m (\lambda-\lambda_i)^+ = P$. Show that this can be rewritten in the form

$$\begin{aligned} P(\lambda) &= \textstyle\sum_{i:\lambda_i\le\lambda}(\lambda-\lambda_i) \\ C(\lambda) &= \textstyle\sum_{i:\lambda_i\le\lambda}\frac{1}{2}\log\frac{\lambda}{\lambda_i}. \end{aligned}$$

Here $P(\lambda)$ is piecewise linear and $C(\lambda)$ is piecewise logarithmic in λ.

Solution: *Parametric form of channel capacity*

The optimal strategy for parallel Gaussian channels is given by water-filling. Here, λ represents the maximum received power in any channel which is being used; i.e. any channel i for which $\lambda_i < \lambda$ will act as a single Gaussian channel with noise $N_i = \lambda_i$ and will communicate a signal with power $P_i = \lambda - N_i$. The $(\cdot)^+$ notation ensures that channels with $\lambda_i > \lambda$ will not be used. Thus, the the total transmitted power, as a function of λ, is given by

$$P(\lambda) = \sum_{i:\lambda_i<\lambda} P_i = \sum_{i:\lambda_i<\lambda} (\lambda - \lambda_i) = \sum_i (\lambda - \lambda_i)^+ \tag{9.99}$$

Now, if we consider the capacity of channel i,

$$\begin{aligned} C_i &= \frac{1}{2}\log\left(1+\frac{P_i}{N_i}\right) && (9.100)\\ &= \frac{1}{2}\log\left(1+\frac{\lambda-\lambda_i}{\lambda_i}\right) && (9.101)\\ &= \frac{1}{2}\log\frac{\lambda}{\lambda_i} && (9.102)\end{aligned}$$

and we obtain

$$C(\lambda) = \sum_{i:\lambda_i<\lambda} C_i = \sum_{i:\lambda_i<\lambda} \frac{1}{2}\log\frac{\lambda}{\lambda_i} \tag{9.103}$$

20. **Robust decoding**. Consider an additive noise channel whose output Y is given by

$$Y = X + Z,$$

where the channel input X is average power limited,

$$EX^2 \le P,$$

and the noise process $\{Z_k\}_{k=-\infty}^{\infty}$ is iid with marginal distribution $p_Z(z)$ (not necessarily Gaussian) of power N,

$$EZ^2 = N.$$

(a) Show that the channel capacity, $C = \max_{EX^2\le P} I(X;Y)$, is lower bounded by C_G where

$$C_G = \frac{1}{2}\log\left(1+\frac{P}{N}\right),$$

i.e., the capacity C_G corresponding to white Gaussian noise.

(b) Decoding the received vector to the codeword that is closest to it in Euclidean distance is in general sub-optimal, if the noise is non-Gaussian. Show, however, that the rate C_G is achievable even if one insists on performing nearest neighbor decoding (minimum Euclidean distance decoding) rather than the optimal maximum-likelihood or joint typicality decoding (with respect to the true noise distribution).

(c) Extend the result to the case where the noise is not iid but is stationary and ergodic with power N.

Hint for b and c: Consider a size 2^{nR} random codebook whose codewords are drawn independently of each other according to a uniform distribution over the n dimensional sphere of radius $\sqrt{nP}$.

- Using a symmetry argument show that, conditioned on the noise vector, the ensemble average probability of error depends on the noise vector only via its Euclidean norm $\|\mathbf{z}\|$.
- Use a geometric argument to show that this dependence is monotonic.
- Given a rate $R < C_G$ choose some $N' > N$ such that

  $$R < \frac{1}{2}\log\left(1 + \frac{P}{N'}\right).$$

  Compare the case where the noise is iid $\mathcal{N}(0, N')$ to the case at hand.
- Conclude the proof using the fact that the above ensemble of codebooks can achieve the capacity of the Gaussian channel (no need to prove that).

Solution: *Robust decoding*

(a) The fact that the worst noise is Gaussian is a consequence of the entropy power inequality, and is proved in problem 9.21. Since C_G is the capacity of the Gaussian, it is the lower bound on the capacity of the channel for all noise distributions.

(b) As suggested in the hint, we will draw codewords at random according to a uniform distribution on a sphere of radius $\sqrt{nP}$. We will send a codeword over the channel, and given the received sequence, find the codeword that is closest (in Euclidean distance) to the received sequence.

First, by the symmetry of the code construction, the probability of error does not depend on which message was sent, so without loss of generality, we can assume that message 1 (i.e., codeword 1 was sent).

The probability of error then depends only on whether the noise sequence Z^n is such that the received vector the closer to some other codeword. However, given any transmitted codeword, all the other codewords are randomly distributed in all directions, and therefore the probability of error does not depend on the direction of the error, only on the norm $||\mathbf{X}(1) + \mathbf{Z}||$ and $||\mathbf{Z}||$. By the spherical symmetry of the choice of $\mathbf{X}(1)$, the probability of error depends only on $||\mathbf{Z}||$.

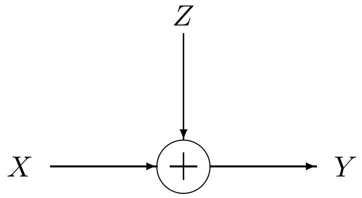
To show monotonocity of the error rate with the norm of the noise, consider an error where the received sequence $\mathbf{X}(1)+\mathbf{Z}$ is closer to some other codeword $\mathbf{X}(2)$ say. Now if increase the norm of the error a little, we have

$$||\mathbf{X}(1) + \mathbf{Z}(1+\Delta) - \mathbf{X}(2)|| = ||\mathbf{X}(1) + \mathbf{Z} - \mathbf{X}(2)|| + \Delta||\mathbf{Z}|| \quad (9.104)$$

by the triangle inequality, and hence if the output is closer to $\mathbf{X}(2)$, then increasing the norm of the noise will not reduce the probability that it is closer to $\mathbf{X}(2)$. Thus the error probability is monotonically decreasing the the norm of the error.

Finally we consider using this code on a Gaussian channel with noise $N' > N$, such the $R < \frac{1}{2}\log(1+P/N')$. Since this is a Gaussian channel, the standard results show that we can achieve arbitrarily low probability of error for this code. Now comparing the non Gaussian channel with the Gaussian channel, we can see that probability close to 1 that the norm of the error in the Gaussian channel is less than the norm of the error for the non Gaussian channel. By the monotonicity of the probability of error with respect to the norm of the noise, we can see that the probability of error for the non-Gaussian channel is less than the probability of error for the Gaussian channel, and hence goes to 0 as the block length goes to ∞.

21. **A mutual information game.** Consider the following channel:



Throughout this problem we shall constrain the signal power

$$EX = 0, \qquad EX^2 = P, \tag{9.105}$$

and the noise power

$$EZ = 0, \qquad EZ^2 = N, \tag{9.106}$$

and assume that X and Z are independent. The channel capacity is given by $I(X;X+Z)$.

Now for the game. The noise player chooses a distribution on Z to minimize $I(X;X+Z)$, while the signal player chooses a distribution on X to maximize $I(X;X+Z)$.

Letting $X^* \sim \mathcal{N}(0,P), Z^* \sim \mathcal{N}(0,N)$, show that Gaussian X^* and Z^* satisfy the saddlepoint conditions

$$I(X;X+Z^*) \le I(X^*;X^*+Z^*) \le I(X^*;X^*+Z). \tag{9.107}$$

Thus

$$\begin{aligned}\min_Z \max_X I(X;X+Z) &= \max_X \min_Z I(X;X+Z) && (9.108)\\ &= \frac{1}{2}\log\left(1+\frac{P}{N}\right), && (9.109)\end{aligned}$$

and the game has a value. In particular, a deviation from normal for either player worsens the mutual information from that player's standpoint. Can you discuss the implications of this?

Note: Part of the proof hinges on the entropy power inequality from Chapter 17, which states that if $\mathbf{X}$ and $\mathbf{Y}$ are independent random n-vectors with densities, then

$$2^{\frac{2}{n}h(\mathbf{X}+\mathbf{Y})} \geq 2^{\frac{2}{n}h(\mathbf{X})} + 2^{\frac{2}{n}h(\mathbf{Y})}. \tag{9.110}$$

Solution: *A mutual information game.*

Let X and Z be random variables with $EX = 0$, $EX^2 = P$, $EZ = 0$ and $EZ^2 = N$. Let $X^* \sim \mathcal{N}(0, P)$ and $Z^* \sim \mathcal{N}(0, N)$. Then as proved in class,

$$\begin{aligned}
I(X; X+Z^*) &= h(X+Z^*) - h(X+Z^*|X) & (9.111)\\
&= h(X+Z^*) - h(Z^*) & (9.112)\\
&\leq h(X^*+Z^*) - h(Z^*) & (9.113)\\
&= I(X^*; X^*+Z^*), & (9.114)
\end{aligned}$$

where the inequality follows from the fact that given the variance, the entropy is maximized by the normal.

To prove the other inequality, we use the entropy power inequality,

$$2^{2h(X+Z)} \leq 2^{2h(X)} + 2^{2h(Z)}. \tag{9.115}$$

Let

$$g(Z) = \frac{2^{2h(Z)}}{2\pi e}. \tag{9.116}$$

Then

$$\begin{aligned}
I(X^*; X^*+Z) &= h(X^*+Z) - h(X^*+Z|X^*) & (9.117)\\
&= h(X^*+Z) - h(Z) & (9.118)\\
&\geq \frac{1}{2}\log\left(2^{2h(X^*)} + 2^{2h(Z)}\right) - h(Z) & (9.119)\\
&= \frac{1}{2}\log\left((2\pi e)P + (2\pi e)g(Z)\right) - \frac{1}{2}\log(2\pi e)g(Z) & (9.120)\\
&= \frac{1}{2}\log\left(1 + \frac{P}{g(Z)}\right), & (9.121)
\end{aligned}$$

where the inequality follows from the entropy power inequality. Now $1 + \frac{P}{g(Z)}$ is a decreasing function of $g(Z)$, it is minimized when $g(Z)$ is maximum, which occurs when $h(Z)$ is maximized, i.e., when Z is normal. In this case, $g(Z^*) = N$ and we have the following inequality,

$$I(X^*; X^*+Z) \geq I(X^*; X^*+Z^*). \tag{9.122}$$

Combining the two inequalities, we have

$$I(X; X+Z^*) \leq I(X^*; X^*+Z^*) \leq I(X^*; X^*+Z). \tag{9.123}$$

Hence, using these inequalities, it follows directly that

$$\begin{aligned}\min_Z \max_X I(X;X+Z) &\le \max_X I(X;X+Z^*) && (9.124)\\ &= I(X^*;X^*+Z^*) && (9.125)\\ &= \min_Z I(X^*;X^*+Z) && (9.126)\\ &\le \max_X \min_Z I(X^*;X^*+Z). && (9.127)\end{aligned}$$

We have shown an inequality relationship in one direction between $\min_Z \max_X I(X;X+Z)$ and $\max_X \min_Z I(X;X+Z)$. We will now prove the inequality in the other direction is a general result for all functions of two variables.

For any function $f(a,b)$ of two variables, for all b, for any a_0,

$$f(a_0,b) \ge \min_a f(a,b). \qquad (9.128)$$

Hence

$$\max_b f(a_0,b) \ge \max_b \min_a f(a,b). \qquad (9.129)$$

Taking the minimum over a_0, we have

$$\min_{a_0} \max_b f(a_0,b) \ge \min_{a_0} \max_b \min_a f(a,b). \qquad (9.130)$$

or

$$\min_a \max_b f(a,b) \ge \max_b \min_a f(a,b). \qquad (9.131)$$

From this result,

$$\min_Z \max_X I(X;X+Z) \ge \max_X \min_Z I(X;X+Z). \qquad (9.132)$$

From (9.127) and (9.132), we have

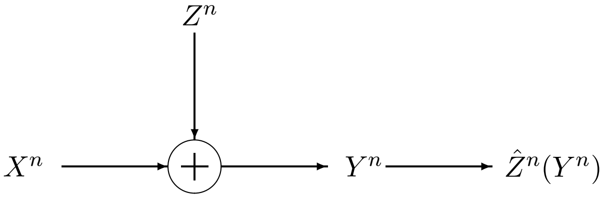
$$\begin{aligned}\min_Z \max_X I(X;X+Z) &= \max_X \min_Z I(X;X+Z) && (9.133)\\ &= \frac{1}{2}\log\left(1+\frac{P}{N}\right). && (9.134)\end{aligned}$$

This inequality implies that we have a saddlepoint in the game, which is the value of the game. If signal player chooses X^*, the noise player cannot do any better than choosing Z^*. Similarly, any deviation by the signal player from X^* will make him do worse, if the noise player has chosen Z^*. Any deviation by either player will make him do worse.

Another implication of this result is that not only is the normal the best possible signal distribution, it is the worst possible noise distribution.

22. **Recovering the noise**
Consider a standard Gaussian channel $Y^n = X^n + Z^n$, where Z_i is i.i.d. $\sim \mathcal{N}(0, N)$, $i = 1, 2, \ldots, n$, and $\frac{1}{n}\sum_{i=1}^n X_i^2 \le P$.

Here we are interested in recovering the noise Z^n and we don't care about the signal X^n. By sending $X^n = (0, 0, \ldots, 0)$, the receiver gets $Y^n = Z^n$ and can fully determine the value of Z^n. We wonder how much variability there can be in X^n and still recover the Gaussian noise Z^n. The use of the channel looks like



Argue that, for some $R > 0$, the transmitter can arbitrarily send one of 2^{nR} different sequences of x^n without affecting the recovery of the noise in the sense that

$$\Pr\{\hat{Z}^n \neq Z^n\} \to 0 \qquad \text{as } n \to \infty .$$

For what R is this possible?

Solution: *Recovering the noise*
We prove that $\sup R = C = C(P/N)$.

If $R < C$, from the achievability proof of the channel coding theorem, 2^{nR} different X^n sequences can be decoded correctly with arbitrarily small error for n large enough. Once X^n is determined, Z^n can be easily computed as $Y^n - X^n$.

We show that this is optimal by using proof by contradiction. Assume that there is some $R > C$ such that Z^n can be recovered with $\Pr\{\hat{Z}^n \neq Z^n\} \to 0$ as $n \to \infty$. But this implies that $X^n = Y^n - Z^n$ can be determined with arbitrary precision; that is, there is a codebook $X^n(W)$, $W = 1, \ldots, 2^{nR}$ with $R > C$ and $\Pr\{\hat{X}^n \neq X^n\} = \Pr\{W \le \hat{W}\} \to 0$ as $n \to \infty$. As we saw in the converse proof of the channel coding theorem, this is impossible. Hence, we have the contradiction and R cannot be greater than C.

Chapter 10

Rate Distortion Theory

1. **One bit quantization of a single Gaussian random variable.** Let $X \sim \mathcal{N}(0,\sigma^2)$ and let the distortion measure be squared error. Here we do not allow block descriptions. Show that the optimum reproduction points for 1 bit quantization are $\pm\sqrt{\frac{2}{\pi}}\sigma$, and that the expected distortion for 1 bit quantization is $\frac{\pi-2}{\pi}\sigma^2$.

   Compare this with the distortion rate bound $D = \sigma^2 2^{-2R}$ for $R = 1$.

   Solution: *One bit quantization of a Gaussian random variable.* Let $X \sim \mathcal{N}(0,\sigma^2)$ and let the distortion measure be squared error. With one bit quantization, the obvious reconstruction regions are the positive and negative real axes. The reconstruction point is the centroid of each region. For example, for the positive real line, the centroid a is

$$a = \int_0^\infty x \frac{2}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\,dx \tag{10.1}$$

$$= \int_0^\infty \sigma\sqrt{\frac{2}{\pi}} e^{-y}\,dy \tag{10.2}$$

$$= \sigma\sqrt{\frac{2}{\pi}}, \tag{10.3}$$

   using the substitution $y = x^2/2\sigma^2$. The expected distortion for one bit quantization is

$$D = \int_{-\infty}^0 \left(x+\sigma\sqrt{\frac{2}{\pi}}\right)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\,dx \tag{10.4}$$

$$+\int_0^\infty \left(x-\sigma\sqrt{\frac{2}{\pi}}\right)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\,dx \tag{10.5}$$

$$= 2\int_{-\infty}^\infty \left(x^2+\sigma^2\frac{2}{\pi}\right) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\,dx \tag{10.6}$$

$$-2\int_0^\infty \left(-2x\sigma\sqrt{\frac{2}{\pi}}\right) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\,dx \tag{10.7}$$

$$= \sigma^2 + \frac{2}{\pi}\sigma^2 - 4\frac{1}{\sqrt{2\pi}}\sigma^2\sqrt{\frac{2}{\pi}} \tag{10.8}$$

$$= \sigma^2\frac{\pi-2}{\pi}. \tag{10.9}$$

2. **Rate distortion function with infinite distortion**. Find the rate distortion function $R(D) = \min I(X;\hat{X})$ for $X \sim$ Bernoulli $(\frac{1}{2})$ and distortion

$$d(x,\hat{x}) = \begin{cases} 0, & x = \hat{x}, \\ 1, & x = 1, \hat{x} = 0, \\ \infty, & x = 0, \hat{x} = 1. \end{cases}$$

Solution: *Rate Distortion.* We wish to evaluate the rate distortion function

$$R(D) = \min_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x})\le D} I(X;\hat{X}). \tag{10.10}$$

Since $d(0,1) = \infty$, we must have $p(0,1) = 0$ for a finite distortion. Thus, the distortion $D = p(1,0)$, and hence we have the following joint distribution for $(X,\hat{X})$ (assuming $D \le \frac{1}{2}$).

$$p(x,\hat{x}) = \begin{bmatrix} \frac{1}{2} & 0 \\ D & \frac{1}{2} - D \end{bmatrix} \tag{10.11}$$

The mutual information for this joint distribution is

$$R(D) = I(X;\hat{X}) = H(X) - H(X|\hat{X}) \tag{10.12}$$

$$= H(\frac{1}{2},\frac{1}{2}) - (\frac{1}{2}+D)H\left(\frac{\frac{1}{2}}{\frac{1}{2}+D}, \frac{D}{\frac{1}{2}+D}\right) \tag{10.13}$$

$$= 1 + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}+D} + D\log\frac{D}{\frac{1}{2}+D}, \tag{10.14}$$

which is the rate distortion function for this binary source if $0 \le D \le \frac{1}{2}$. Since we can achieve $D = \frac{1}{2}$ with zero rate (use $p(\hat{x}=0) = 1$), we have $R(D) = 0$ for $D \ge \frac{1}{2}$.

3. **Rate distortion for binary source with asymmetric distortion.** Fix $p(\hat{x}|x)$ and evaluate $I(X;\hat{X})$ and D for

$$X \sim \text{Bern}(1/2),$$

$$d(x,\hat{x}) = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}$$

(The rate distortion function cannot be expressed in closed form.)

Solution: *Binary source with asymmetric distortion.* $X \sim$ Bern($\frac{1}{2}$), and the distortion measure is

$$d(x,\hat{x}) = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}. \tag{10.15}$$

Proceeding with the minimization to calculate $R(D)$ as

$$R(D) = \min_{p(\hat{x}|x): \sum p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X;\hat{X}), \tag{10.16}$$

we must choose the conditional distribution $p(\hat{x}|x)$. Setting $p(0|0) = \alpha$ and $p(1|1) = \beta$, we get the joint distribution

$$p(x,\hat{x}) = \begin{bmatrix} \frac{\alpha}{2} & \frac{1-\alpha}{2} \\ \frac{1-\beta}{2} & \frac{\beta}{2} \end{bmatrix}. \tag{10.17}$$

Hence the distortion constraint can be written as

$$\frac{1-\alpha}{2}a + \frac{1-\beta}{2}b \leq D. \tag{10.18}$$

The function to be minimized, $I(X;\hat{X})$, can be written

$$I(X;\hat{X}) = H(\hat{X}) - H(\hat{X}|X) = H(\frac{\alpha+1-\beta}{2}) - \frac{1}{2}H(\alpha) - \frac{1}{2}H(\beta). \tag{10.19}$$

Using the method of Lagrange multipliers, we have

$$J(\alpha,\beta,\lambda) = H(\frac{\alpha+1-\beta}{2}) - \frac{1}{2}H(\alpha) - \frac{1}{2}H(\beta) + \lambda(\frac{1-\alpha}{2}a + \frac{1-\beta}{2}b) \tag{10.20}$$

and differentiating to find the maximum, we have the following equations:

$$\begin{aligned}
\frac{1}{2}\left(\log\frac{\frac{1-\alpha+\beta}{2}}{\frac{\alpha+1-\beta}{2}}\right) - \frac{1}{2}\left(\log\frac{1-\alpha}{\alpha}\right) - \frac{\lambda a}{2} &= 0 && (10.21)\\
-\frac{1}{2}\left(\log\frac{\frac{1-\alpha+\beta}{2}}{\frac{\alpha+1-\beta}{2}}\right) - \frac{1}{2}\left(\log\frac{1-\beta}{\beta}\right) - \frac{\lambda b}{2} &= 0 && (10.22)\\
\frac{1-\alpha}{2}a + \frac{1-\beta}{2}b = D & && (10.23)
\end{aligned}$$

In principle, these equations can be solved for α, β, and λ and substituted back in the definition to find the rate distortion function. This problem unfortunately does not have an explicit solution.

4. **Properties of $R(D)$.** Consider a discrete source $X \in \mathcal{X} = \{1,2,\ldots,m\}$ with distribution $p_1, p_2, \ldots, p_m$ and a distortion measure $d(i,j)$. Let $R(D)$ be the rate distortion function for this source and distortion measure. Let $d'(i,j) = d(i,j) - w_i$ be a new distortion measure and let $R'(D)$ be the corresponding rate distortion function. Show that $R'(D) = R(D+\bar{w})$, where $\bar{w} = \sum p_i w_i$, and use this to show that there is no essential loss of generality in assuming that $\min_{\hat{x}} d(i,\hat{x}) = 0$, i.e., for each $x \in \mathcal{X}$, there is one symbol $\hat{x}$ which reproduces the source with zero distortion.

   This result is due to Pinkston[9].

Solution: *Properties of the rate distortion function.* By definition,

$$R'(D') = \min_{p(\hat{x}|x): \sum p(\hat{x}|x)p(x)d'(x,\hat{x}) \le D'} I(X;\hat{X}). \tag{10.24}$$

For any conditional distribution $p(\hat{x}|x)$, we have

$$\begin{aligned} D' &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d'(x,\hat{x}) && (10.25)\\ &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x)(d(x,\hat{x}) - w_x) && (10.26)\\ &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) - \sum_x p(x)w_x \sum_{\hat{x}} p(\hat{x}|x) && (10.27)\\ &= D - \sum_x p(x)w_x && (10.28)\\ &= D - \bar{w}, && (10.29) \end{aligned}$$

or $D = D' + \bar{w}$. Hence

$$\begin{aligned} R'(D') &= \min_{p(\hat{x}|x): \sum p(\hat{x}|x)p(x)d'(x,\hat{x}) \le D'} I(X;\hat{X}) && (10.30)\\ &= \min_{p(\hat{x}|x): \sum p(\hat{x}|x)p(x)d(x,\hat{x}) \le D'+\bar{w}} I(X;\hat{X}) && (10.31)\\ &= R(D' + \bar{w}). && (10.32) \end{aligned}$$

For any distortion matrix, we can set $w_i = \min_{\hat{x}} d(i,\hat{x})$, hence ensuring that $\min_{\hat{x}} d'(x,\hat{x}) = 0$ for every x. This produces only a shift in the rate distortion function and does not change the essential theory. Hence, there is no essential loss of generality in assuming that for each $x \in \mathcal{X}$, there is one symbol $\hat{x}$ which reproduces it with zero distortion.

5. **Rate distortion for uniform source with Hamming distortion.** Consider a source X uniformly distributed on the set $\{1, 2, \ldots, m\}$. Find the rate distortion function for this source with Hamming distortion, i.e.,

$$d(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x}, \\ 1 & \text{if } x \ne \hat{x}. \end{cases}$$

Solution: *Rate distortion for uniform source with Hamming distortion.* X is uniformly distributed on the set $\{1, 2, \ldots, m\}$. The distortion measure is

$$d(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \ne \hat{x} \end{cases}$$

Consider any joint distribution that satisfies the distortion constraint D. Since $D = \Pr(X \ne \hat{X})$, we have by Fano's inequality

$$H(X|\hat{X}) \le H(D) + D\log(m-1), \tag{10.33}$$

and hence

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) && (10.34)\\ &\geq \log m - H(D) - D\log(m-1). && (10.35) \end{aligned}$$

We can achieve this lower bound by choosing $p(\hat{x})$ to be the uniform distribution, and the conditional distribution of $p(x|\hat{x})$ to be

$$p(\hat{x}|x) \begin{cases} = 1 - D & \text{if } \hat{x} = x \\ = D/(m-1) & \text{if } \hat{x} \neq x. \end{cases} \qquad (10.36)$$

It is easy to verify that this gives the right distribution on X and satisfies the bound with equality for $D < 1 - \frac{1}{m}$. Hence

$$R(D) \begin{cases} = \log m - H(D) - D\log(m-1) & \text{if } 0 \leq D \leq 1 - \frac{1}{m} \\ 0 & \text{if } D > 1 - \frac{1}{m}. \end{cases} \qquad (10.37)$$

6. **Shannon lower bound for the rate distortion function.** Consider a source X with a distortion measure $d(x,\hat{x})$ that satisfies the following property: all columns of the distortion matrix are permutations of the set $\{d_1, d_2, \ldots, d_m\}$. Define the function

$$\phi(D) = \max_{\mathbf{p}:\sum_{i=1}^m p_i d_i \leq D} H(\mathbf{p}). \qquad (10.38)$$

The Shannon lower bound on the rate distortion function[13] is proved by the following steps:

(a) Show that $\phi(D)$ is a concave function of D.

(b) Justify the following series of inequalities for $I(X;\hat{X})$ if $Ed(X,\hat{X}) \leq D$,

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) && (10.39)\\ &= H(X) - \sum_{\hat{x}} p(\hat{x}) H(X|\hat{X} = \hat{x}) && (10.40)\\ &\geq H(X) - \sum_{\hat{x}} p(\hat{x}) \phi(D_{\hat{x}}) && (10.41)\\ &\geq H(X) - \phi\left(\sum_{\hat{x}} p(\hat{x}) D_{\hat{x}}\right) && (10.42)\\ &\geq H(X) - \phi(D), && (10.43) \end{aligned}$$

where $D_{\hat{x}} = \sum_x p(x|\hat{x}) d(x,\hat{x})$.

(c) Argue that

$$R(D) \geq H(X) - \phi(D), \qquad (10.44)$$

which is the Shannon lower bound on the rate distortion function.

(d) If in addition, we assume that the source has a uniform distribution and that the rows of the distortion matrix are permutations of each other, then $R(D) = H(X) - \phi(D)$, i.e., the lower bound is tight.

Solution: *Shannon lower bound on the rate distortion function.*

(a) We define

$$\phi(D) = \max_{\mathbf{p}:\sum_{i=1}^m p_i d_i \le D} H(\mathbf{p}). \tag{10.45}$$

From the definition, if $D_1 \ge D_2$, then $\phi(D_1) \ge \phi(D_2)$ since the maximization is over a larger set. Hence $\phi(D)$ is a monotonic increasing function.

To prove concavity of $\phi(D)$, consider two levels of distortion D_1 and D_2 and let $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ achieve the maxima in the definition of $\phi(D_1)$ and $\phi(D_2)$. Let $\mathbf{p}^{(\lambda)}$ be the mixture of the two distributions, i.e.,

$$\mathbf{p}^{(\lambda)} = \lambda \mathbf{p}^{(1)} + (1-\lambda)\mathbf{p}^{(2)}. \tag{10.46}$$

Then the distortion is a mixture of the two distortions

$$D_\lambda = \sum_i p_i^{(\lambda)} d_i = \lambda D_1 + (1-\lambda) D_2. \tag{10.47}$$

Since entropy is a concave function, we have

$$H(\mathbf{p}^{(\lambda)}) \ge \lambda H(\mathbf{p}^{(1)}) + (1-\lambda) H(\mathbf{p}^{(2)}). \tag{10.48}$$

Hence

$$\begin{aligned}
\phi(D_\lambda) &= \max_{\mathbf{p}:\sum p_i d_i = D_\lambda} H(\mathbf{p}) && (10.49)\\
&\ge H(\mathbf{p}^{(\lambda)}) && (10.50)\\
&\ge \lambda H(\mathbf{p}^{(1)}) + (1-\lambda) H(\mathbf{p}^{(2)}) && (10.51)\\
&= \lambda \phi(D_1) + (1-\lambda)\phi(D_2), && (10.52)
\end{aligned}$$

proving that $\phi(D)$ is a concave function of D.

(b) For any $(X, \hat{X})$ that satisfy the distortion constraint, we have

$$\begin{aligned}
I(X;\hat{X}) &\overset{(a)}{=} H(X) - H(X|\hat{X}) && (10.53)\\
&\overset{(b)}{=} H(X) - \sum_{\hat{x}} p(\hat{x}) H(X|\hat{X} = \hat{x}) && (10.54)\\
&\overset{(c)}{\ge} H(X) - \sum_{\hat{x}} p(\hat{x}) \phi(D_{\hat{x}}) && (10.55)\\
&\overset{(d)}{\ge} H(X) - \phi(\sum_{\hat{x}} p(\hat{x}) D_{\hat{x}}) && (10.56)\\
&\overset{(e)}{\ge} H(X) - \phi(D), && (10.57)
\end{aligned}$$

where
(a) follows from the definition of mutual information,
(b) from the definition of conditional entropy,
(c) follows from the definition of $\phi(D_{\hat{x}})$ where $D_{\hat{x}} = \sum p(x|\hat{x})d(x,\hat{x}) = \sum p(x|\hat{x})d_{i'}$ that $H(p(x|\hat{x})) \leq \phi(D_{\hat{x}})$
(d) follows from Jensen's inequality and the concavity of ϕ, and
(e) follows from the monotonicity of ϕ and the fact that $\sum p(\hat{x})D_{\hat{x}} = \sum p(x,\hat{x})d(x,\hat{x}) \leq D$.
Hence, from the definition of the rate distortion function, we have

$$\begin{aligned} R(D) &= \min_{p(\hat{x}|x):\sum p(x,\hat{x})d(x,\hat{x})\leq D} I(X;\hat{X}) && (10.58) \\ &\geq H(X) - \phi(D), && (10.59) \end{aligned}$$

which is the Shannon lower bound on the rate distortion function.

(c) Let $\mathbf{p}^* = (p_1^*, p_2^*, \ldots, p_m^*)$ be the distribution that achieves the maximum in the definition of the $\phi(D)$. Assume that the source has a uniform distribution and that the rows of the distortion matrix are permutations of each other. Let the distortion matrix be $[a_{ij}]$. We can then choose $p(\hat{x})$ to have a uniform distribution and choose $p(x=i|\hat{x}=j) = p_k^*$, if $a_{ij} = d_k$. For this joint distribution,

$$\begin{aligned} p_x(i) &= \sum_j p_{\hat{x}}(j)p_{x|\hat{x}}(i|j) && (10.60) \\ &= \sum_j \frac{1}{m}p_k^* && (10.61) \\ &= \frac{1}{m} && (10.62) \end{aligned}$$

since the rows of the distortion matrix are permutations of each other and therefore each element p_k^*, $k = 1,2,\ldots,m$ occurs once in the above sum. Hence the distribution of x has the desired source distribution. For this joint distribution, we have

$$\begin{aligned} \sum_{i,j} p_{x,\hat{x}}(i,j)a_{ij} &= \sum_j \frac{1}{m}\sum_i p_{x|\hat{x}}(i|j)a_{ij} && (10.63) \\ &= \sum_j \frac{1}{m}\sum_k p_k^* d_k && (10.64) \\ &= \sum_j \frac{1}{m}D && (10.65) \\ &= D, && (10.66) \end{aligned}$$

the desired distortion. The mutual information

$$I(X;\hat{X}) = H(X) - H(X|\hat{X}) \qquad (10.67)$$

$$= H(X) - \sum_j \frac{1}{m} H(X|\hat{X} = j) \tag{10.68}$$

$$= H(X) - \sum_j \frac{1}{m} H(\mathbf{p}^*) \tag{10.69}$$

$$= H(X) - \sum_j \frac{1}{m} \phi(D) \tag{10.70}$$

$$= H(X) - \phi(D). \tag{10.71}$$

Hence using this joint distribution in the definition of the rate distortion function

$$R(D) = \min_{p(\hat{x}|x): \sum p(x,\hat{x})d(x,\hat{x}) \leq D} I(X;\hat{X}) \tag{10.72}$$

$$\leq I(X;\hat{X}) \tag{10.73}$$

$$= H(X) - \phi(D). \tag{10.74}$$

Combining this with the Shannon lower bound on the rate distortion function, we must have equality in the above equation and hence we have equality in the Shannon lower bound.

7. **Erasure distortion.** Consider $X \sim$ Bernoulli($\frac{1}{2}$), and let the distortion measure be given by the matrix

$$d(x,\hat{x}) = \begin{bmatrix} 0 & 1 & \infty \\ \infty & 1 & 0 \end{bmatrix}. \tag{10.75}$$

Calculate the rate distortion function for this source. Can you suggest a simple scheme to achieve any value of the rate distortion function for this source?

Solution: *Erasure distortion.* Consider $X \sim$ Bernoulli($\frac{1}{2}$), and the distortion measure

$$d(x,\hat{x}) = \begin{bmatrix} 0 & 1 & \infty \\ \infty & 1 & 0 \end{bmatrix}. \tag{10.76}$$

The infinite distortion constrains $p(0,1) = p(1,0) = 0$. Hence by symmetry the joint distribution of $(X,\hat{X})$ is of the form shown in Figure 10.1.

For this joint distribution, it is easy to calculate the distortion $D = \alpha$ and that $I(X;\hat{X}) = H(X) - H(X|\hat{X}) = 1 - \alpha$. Hence we have $R(D) = 1 - D$ for $0 \leq D \leq 1$. For $D > 1$, $R(D) = 0$.

It is very see how we could achieve this rate distortion function. If D is rational, say k/n, then we send only the first $n-k$ of any block of n bits. We reproduce these bits exactly and reproduce the remaining bits as erasures. Hence we can send information at rate $1 - D$ and achieve a distortion D. If D is irrational, we can get arbitrarily close to D by using longer and longer block lengths.

8. **Bounds on the rate distortion function for squared error distortion.** For the case of a continuous random variable X with mean zero and variance σ^2 and squared

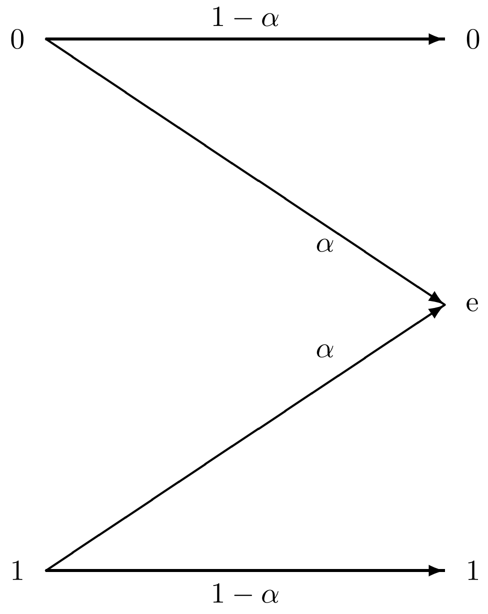


Figure 10.1: Joint distribution for erasure rate distortion of a binary source

error distortion, show that

$$h(X) - \frac{1}{2}\log(2\pi e D) \le R(D) \le \frac{1}{2}\log\frac{\sigma^2}{D}. \qquad (10.77)$$

For the upper bound, consider the joint distribution shown in Figure 10.2. Are Gaussian random variables harder or easier to describe than other random variables with the same variance?

Solution: *Bounds on the rate distortion function for squared error distortion.*

We assume that X has zero mean and variance σ^2. To prove the lower bound, we use the same techniques as used for the Guassian rate distortion function. Let $(X, \hat{X})$ be random variables such that $E(X-\hat{X})^2 \le D$. Then

$$\begin{aligned}
I(X;\hat{X}) &= h(X) - h(X|\hat{X}) && (10.78)\\
&= h(X) - h(X-\hat{X}|\hat{X}) && (10.79)\\
&\ge h(X) - h(X-\hat{X}) && (10.80)\\
&\ge h(X) - h(\mathcal{N}(0, E(X-\hat{X})^2)) && (10.81)\\
&= h(X) - \frac{1}{2}\log(2\pi e)E(X-\hat{X})^2 && (10.82)\\
&\ge h(X) - \frac{1}{2}\log(2\pi e)D. && (10.83)
\end{aligned}$$

To prove the upper bound, we consider the joint distribution as shown in Figure 10.3,

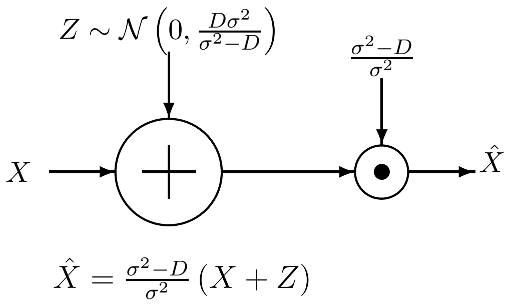

$Z \sim \mathcal{N}\left(0, \frac{D\sigma^2}{\sigma^2-D}\right)$
$\frac{\sigma^2-D}{\sigma^2}$
X
$\hat{X}$
$\hat{X} = \frac{\sigma^2-D}{\sigma^2}(X+Z)$

Figure 10.2: Joint distribution for upper bound on rate distortion function.

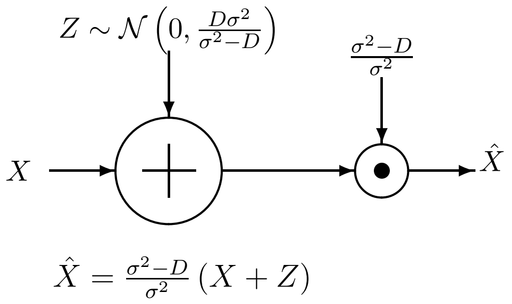

$Z \sim \mathcal{N}\left(0, \frac{D\sigma^2}{\sigma^2-D}\right)$
$\frac{\sigma^2-D}{\sigma^2}$
X
$\hat{X}$
$\hat{X} = \frac{\sigma^2-D}{\sigma^2}(X+Z)$

Figure 10.3: Joint distribution for upper bound on rate distortion function

and calculate the distortion and the mutual information between X and $\hat{X}$. Since

$$\hat{X} = \frac{\sigma^2 - D}{\sigma^2}(X+Z), \tag{10.84}$$

we have

$$\begin{aligned} E(X-\hat{X})^2 &= E\left(\frac{D}{\sigma^2}X - \frac{\sigma^2-D}{\sigma^2}Z\right)^2 &(10.85)\\ &= \left(\frac{D}{\sigma^2}\right)^2 EX^2 + \left(\frac{\sigma^2-D}{\sigma^2}\right)^2 EZ^2 &(10.86)\\ &= \left(\frac{D}{\sigma^2}\right)^2 \sigma^2 + \left(\frac{\sigma^2-D}{\sigma^2}\right)^2 \frac{D\sigma^2}{\sigma^2-D} &(10.87)\\ &= D, &(10.88)\end{aligned}$$

since X and Z are independent and zero mean. Also the mutual information is

$$\begin{aligned} I(X;\hat{X}) &= h(\hat{X}) - h(\hat{X}|X) &(10.89)\\ &= h(\hat{X}) - h(\frac{\sigma^2-D}{\sigma^2}Z). &(10.90)\end{aligned}$$

Now

$$\begin{aligned} E\hat{X}^2 &= \left(\frac{\sigma^2-D}{\sigma^2}\right)^2 E(X+Z)^2 &(10.91)\\ &= \left(\frac{\sigma^2-D}{\sigma^2}\right)^2 (EX^2+EZ^2) &(10.92)\\ &= \left(\frac{\sigma^2-D}{\sigma^2}\right)^2 \left(\sigma^2 + \frac{D\sigma^2}{\sigma^2-D}\right) &(10.93)\\ &= \sigma^2 - D. &(10.94)\end{aligned}$$

Hence, we have

$$\begin{aligned} I(X;\hat{X}) &= h(\hat{X}) - h(\frac{\sigma^2-D}{\sigma^2}Z) &(10.95)\\ &= h(\hat{X}) - h(Z) - \log\frac{\sigma^2-D}{\sigma^2} &(10.96)\\ &\leq h(\mathcal{N}(0,\sigma^2-D)) - \frac{1}{2}\log(2\pi e)\frac{D\sigma^2}{\sigma^2-D} - \log\frac{\sigma^2-D}{\sigma^2} &(10.97)\\ &= \frac{1}{2}\log(2\pi e)(\sigma^2-D) - \frac{1}{2}\log(2\pi e)\frac{D\sigma^2}{\sigma^2-D} - \frac{1}{2}\log\left(\frac{\sigma^2-D}{\sigma^2}\right)^2 &(10.98)\\ &= \frac{1}{2}\log\frac{\sigma^2}{D}, &(10.99)\end{aligned}$$

which combined with the definition of the rate distortion function gives us the required upper bound.

For a Gaussian random variable, $h(X) = \frac{1}{2}\log(2\pi e)\sigma^2$ and the lower bound is equal to the upper bound. For any other random variable, the lower bound is strictly less than the upper bound and hence non-Gaussian random variables cannot require more bits to describe to the same accuracy than the corresponding Gaussian random variables. This is not surprising, since the Gaussian random variable has the maximum entropy and we would expect that it would be the most difficult to describe.

9. **Properties of optimal rate distortion code**. A good (R, D) rate distortion code with $R \approx R(D)$ puts severe constraints on the relationship of the source X^n and the representations $\hat{X}^n$. Examine the chain of inequalities (10.100–10.112) considering the conditions for equality and interpret as properties of a good code. For example, equality in (10.101) implies that $\hat{X}^n$ is a deterministic function of X^n.

   Solution: *Properties of optimal rate distortion code.* The converse of the rate distortion theorem relies on the following chain of inequalities

$$
\begin{aligned}
nR &\overset{(a)}{\geq} H(\hat{X}^n) && (10.100)\\
&\overset{(b)}{\geq} H(\hat{X}^n) - H(\hat{X}^n|X^n) && (10.101)\\
&\overset{(c)}{=} I(\hat{X}^n; X^n) && (10.102)\\
&= H(X^n) - H(X^n|\hat{X}^n) && (10.103)\\
&\overset{(d)}{=} \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) && (10.104)\\
&\overset{(e)}{=} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{i-1}, \ldots, X_1) && (10.105)\\
&\overset{(f)}{\geq} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) && (10.106)\\
&= \sum_{i=1}^{n} I(X_i; \hat{X}_i) && (10.107)\\
&\overset{(g)}{\geq} \sum_{i=1}^{n} R(Ed(X_i, \hat{X}_i)) && (10.108)\\
&= n\sum_{i=1}^{n} \frac{1}{n} R(Ed(X_i, \hat{X}_i)) && (10.109)\\
&\overset{(h)}{\geq} nR\left(\frac{1}{n}\sum_{i=1}^{n} Ed(X_i; \hat{X}_i)\right) && (10.110)\\
&\overset{(i)}{=} nR(Ed(X^n, \hat{X}^n)) && (10.111)\\
&= nR(D). && (10.112)
\end{aligned}
$$

We will have equality in
(a) if $\hat{X}^n$ is uniformly distributed over the set of codewords -i.e., if all the codewords were equally likely,
(b) if $\hat{X}^n$ is a deterministic function of X^n,
(f) if each X_i depends only on the corresponding $\hat{X}_i$ and is conditionally independent of every other $\hat{X}_j$,
(g) if the joint distribution of X_i and $\hat{X}_i$ is the one achieving the minimum in the definition of the rate distortion function, and
(h) if either the rate distortion curve is a straight line or if all the distortions (at each i) are equal.

Thus the optimal rate distortion code would be deterministic, and the joint distribution between the source symbol and the codeword at each instant of time would be independent and equal to the joint distribution that achieves the minimum of the rate distortion function. The distortion would be the same for each time instant.

10. **Rate distortion.** Find and verify the rate distortion function $R(D)$ for X uniform on $\mathcal{X} = \{1, 2, \ldots, 2m\}$ and

$$d(x,\hat{x}) = \begin{cases} 1 & \text{for } x-\hat{x} \text{ odd,} \\ 0 & \text{for } x-\hat{x} \text{ even,} \end{cases}$$

where $\hat{X}$ is defined on $\hat{\mathcal{X}} = \{1, 2, \ldots, 2m\}$.

(You may wish to use the Shannon lower bound in your argument.)

Solution: *Rate distortion*

Since the columns of the distortion measure are alternate 0 and 1, they are all permutations of each other, and we can apply the Shannon lower bound on the rate distortion function. The Shannon lower bound says that

$$R(D) \geq H(X) - \phi(D), \tag{10.113}$$

where

$$\phi(D) = \max_{\mathbf{p}:\sum_{i=1}^m p_i d_i \leq D} H(\mathbf{p}). \tag{10.114}$$

In Problem 6, it was shown that if the input probability distribution is uniform, the bound is tight, and the Shannon lower bound is equal to the rate distortion function.

Therefore to calculate the $R(D)$, we only need to compute $\phi(D)$ for the distortion measure of the problem. Each row of the distortion matrix is a permutation (actually a cyclic shift) of the first row $[010101\ldots01]$. Let Y be random variable with distribution $p_1, \ldots, p_{2m}$, and let Z be the value of the $d(0, Y)$. Thus Z is 0 on the even values of Y and 1 on the odd values. Then

$$\phi(D) = \max_{\mathbf{p}:\sum_{i=1}^m p_i d_i \leq D} H(\mathbf{p}) \tag{10.115}$$

$$= \max_{\mathbf{p}:\sum_{i=1}^m p_i d_i \le D} H(Y) \tag{10.116}$$

$$= \max_{\mathbf{p}:\sum_{i=1}^m p_i d_i \le D} H(Y,Z) \tag{10.117}$$

$$= \max_{\mathbf{p}:\sum_{i=1}^m p_i d_i \le D} H(Z) + H(Y|Z) \tag{10.118}$$

$$= \max_{\mathbf{p}:\sum_{i=1}^m p_i d_i \le D} -p\log p - (1-p)\log(1-p) + (1-p)H(Y|Z=0) + pH(Y|Z=1) \tag{10.119}$$

$$\tag{10.120}$$

where $p = \Pr(Z=1)$. Since $\sum_i p_i d_i = \sum_{i:Z=1} p_i = p$, we have $p \le D$. Given $Z=0$, there are m possible values of Y, and the entropy is maximized by a uniform over these values. Similarly, conditioned on $Z=1$, $H(Y|Z=1)$ is maximized by an uniform distribution on the m values of Y where $Z=1$. Thus

$$\phi(D) = \max_{p:p\le D} H(p) + p\log m + (1-p)\log m = H(D) + \log m \tag{10.121}$$

and hence

$$R(D) = H(X) - \phi(D) = \log 2m - \log m - H(D) = 1 - H(D) \tag{10.122}$$

11. **Lower bound**

Let

$$X \sim \frac{e^{-x^4}}{\int_{-\infty}^{\infty} e^{-x^4}dx}$$

and

$$\frac{\int x^4 e^{-x^4}dx}{\int e^{-x^4}dx} = c.$$

Define $g(a) = \max h(X)$ over all densities such that $EX^4 \le a$. Let $R(D)$ be the rate distortion function for X with the above density and with distortion criterion $d(x,\hat{x}) = (x-\hat{x})^4$. Show $R(D) \ge g(c) - g(D)$.

Solution: *Lower bound*

This is a continuous analog of the Shannon lower bound for the rate distortion function. By similar arguments

$$R(D) = \min_{Ed(X,\hat{X})\le D} I(X;\hat{X}) \tag{10.123}$$

$$= \min_{Ed(X,\hat{X})\le D} h(X) - h(X|\hat{X}) \tag{10.124}$$

$$\tag{10.125}$$

The maximum entropy distribution given the expected fourth power constraint is of the form

$$X \sim \frac{e^{-x^4}}{\int_{-\infty}^{\infty} e^{-x^4}dx} \tag{10.126}$$

and hence $h(X) = g(c)$.

Now $h(X|\hat{X}) = h(X - \hat{X}|\hat{X}) \leq h(X - \hat{X}) \leq g(D)$ from the definition of $g(a) = \max_{EX^4=a} h(X)$. Therefore

$$R(D) \geq g(c) - g(D) \tag{10.127}$$

12. **Adding a column to the distortion matrix**. Let $R(D)$ be the rate distortion function for an i.i.d. process with probability mass function $p(x)$ and distortion function $d(x, \hat{x})$, $x \in \mathcal{X}$, $\hat{x} \in \hat{\mathcal{X}}$. Now suppose that we add a new reproduction symbol $\hat{x}_0$ to $\hat{\mathcal{X}}$ with associated distortion $d(x, \hat{x}_0)$, $x \in \mathcal{X}$. Does this increase or decrease $R(D)$ and why?

Solution: *Adding a column*

Let the new rate distortion function be denoted as $\tilde{R}(D)$, and note that we can still achieve $R(D)$ by restricting the support of $p(x, \hat{x})$, i.e., by simply ignoring the new symbol. Thus, $\tilde{R}(D) \leq R(D)$.

Finally note the duality to the problem in which we added a row to the channel transition matrix to have no smaller capacity (Problem 7.22).

13. **Simplification**. Suppose $\mathcal{X} = \{1, 2, 3, 4\}$, $\hat{\mathcal{X}} = \{1, 2, 3, 4\}$, $p(i) = \frac{1}{4}$, $i = 1, 2, 3, 4$, and $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. The distortion matrix $d(x, \hat{x})$ is given by

$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 0 & 1 & 1 \\ 2 & 0 & 0 & 1 & 1 \\ 3 & 1 & 1 & 0 & 0 \\ 4 & 1 & 1 & 0 & 0 \end{array}$$

(a) Find $R(0)$, the rate necessary to describe the process with zero distortion.

(b) Find the rate distortion function $R(D)$. There are some irrelevant distinctions in alphabets $\mathcal{X}$ and $\hat{\mathcal{X}}$, which allow the problem to be collapsed.

(c) Suppose we have a nonuniform distribution $p(i) = p_i$, $i = 1, 2, 3, 4$. What is $R(D)$?

Solution: *Simplification*

(a) We can achieve 0 distortion if we output $\hat{X} = 1$ if $X = 1$ or 2, and $\hat{X} = 3$ if $X = 3$ or 4. Thus if we set $Y = 1$ if $X = 1$ or 2, and $Y = 2$ if $X = 3$ or 4, we can recover Y exactly if the rate is greater than $H(Y) = 1$ bit. It is also not hard to see that any 0 distortion code would be able to recover Y exactly, and thus $R(0) = 1$.

(b) If we define Y as in the previous part, and $\hat{Y}$ similarly from $\hat{X}$, we can see that the distortion between X and $\hat{X}$ is equal to the Hamming distortion between Y and $\hat{Y}$. Therefore if the rate is greater than the Hamming rate distortion function $R(D)$ for Y, we can recover X to distortion D. Thus $R(D) = 1 - H(D)$.

(c) If the distribution of X is not uniform, the same arguments hold and Y has a distribution $(p_1 + p_2, p_3 + p_4)$, and the rate distortion function is $R(D) = H(p_1 + p_2) - H(D)$,

14. **Rate distortion for two independent sources.** Can one simultaneously compress two independent sources better than by compressing the sources individually? The following problem addresses this question. Let $\{X_i\}$ be iid $\sim p(x)$ with distortion $d(x, \hat{x})$ and rate distortion function $R_X(D)$. Similarly, let $\{Y_i\}$ be iid $\sim p(y)$ with distortion $d(y, \hat{y})$ and rate distortion function $R_Y(D)$.

Suppose we now wish to describe the process $\{(X_i, Y_i)\}$ subject to distortions $Ed(X, \hat{X}) \leq D_1$ and $Ed(Y, \hat{Y}) \leq D_2$. Thus a rate $R_{X,Y}(D_1, D_2)$ is sufficient, where

$$R_{X,Y}(D_1, D_2) = \min_{p(\hat{x},\hat{y}|x,y): Ed(X,\hat{X}) \leq D_1, Ed(Y,\hat{Y}) \leq D_2} I(X, Y; \hat{X}, \hat{Y})$$

Now suppose the $\{X_i\}$ process and the $\{Y_i\}$ process are independent of each other.

(a) Show

$$R_{X,Y}(D_1, D_2) \geq R_X(D_1) + R_Y(D_2).$$

(b) Does equality hold?

Now answer the question.

Solution: *Rate distortion for two independent sources*

(a) Given that X and Y are independent, we have

$$p(x, y, \hat{x}, \hat{y}) = p(x)p(y)p(\hat{x}, \hat{y}|x, y) \tag{10.128}$$

Then

$$\begin{aligned}
I(X, Y; \hat{X}, \hat{Y}) &= H(X, Y) - H(X, Y|\hat{X}, \hat{Y}) &(10.129)\\
&= H(X) + H(Y) - H(X|\hat{X}, \hat{Y}) - H(Y|X, \hat{X}, \hat{Y}) &(10.130)\\
&\geq H(X) + H(Y) - H(X|\hat{X}) - H(Y|\hat{Y}) &(10.131)\\
&= I(X; \hat{X}) + I(Y; \hat{Y}) &(10.132)
\end{aligned}$$

where the inequality follows from the fact that conditioning reduces entropy. Therefore

$$\begin{aligned}
R_{X,Y}(D_1, D_2) &= \min_{p(\hat{x},\hat{y}|x,y): Ed(X,\hat{X}) \leq D_1, Ed(Y,\hat{Y}) \leq D_2} I(X, Y; \hat{X}, \hat{Y}) &(10.133)\\
&\geq \min_{p(\hat{x},\hat{y}|x,y): Ed(X,\hat{X}) \leq D_1, Ed(Y,\hat{Y}) \leq D_2} \left(I(X; \hat{X}) + I(Y; \hat{Y}) \right) &(10.134)\\
&= \min_{p(\hat{x}|x): Ed(X,\hat{X}) \leq D_1} I(X; \hat{X}) + \min_{p(\hat{y}|y): Ed(Y,\hat{Y}) \leq D_2} I(Y; \hat{Y}) &(10.135)\\
&= R_X(D_1) + R_Y(D_2) &(10.136)
\end{aligned}$$

(b) If

$$p(x,y,\hat{x},\hat{y}) = p(x)p(y)p(\hat{x}|x)p(\hat{y}|y), \tag{10.137}$$

then

$$\begin{aligned} I(X,Y;\hat{X},\hat{Y}) &= H(X,Y) - H(X,Y|\hat{X},\hat{Y}) && (10.138) \\ &= H(X) + H(Y) - H(X|\hat{X},\hat{Y}) - H(Y|X,\hat{X},\hat{Y}) && (10.139) \\ &= H(X) + H(Y) - H(X|\hat{X}) - H(Y|\hat{Y}) && (10.140) \\ &= I(X;\hat{X}) + I(Y;\hat{Y}) && (10.141) \end{aligned}$$

Let $p(x,\hat{x})$ be a distribution that achieves the rate distortion $R_X(D_1)$ at distortion D_1 and let $p(y,\hat{y})$ be a distribution that achieves the rate distortion $R_Y(D_2)$ at distortion D_2. Then for the product distribution $p(x,y,\hat{x},\hat{y}) = p(x,\hat{x})p(y,\hat{y})$, where the component distributions achieve rates $(D_1, R_X(D_1))$ and $(D_2, R_X(D_2))$, the mutual information corresponding to the product distribution is $R_X(D_1) + R_Y(D_2)$. Thus

$$R_{X,Y}(D_1,D_2) = \min_{p(\hat{x},\hat{y}|x,y): Ed(X,\hat{X}) \le D_1, Ed(Y,\hat{Y}) \le D_2} I(X,Y;\hat{X},\hat{Y}) = R_X(D_1) + R_Y(D_2) \tag{10.142}$$

Thus by using the product distribution, we can achieve the sum of the rates.

Therefore the total rate at which we encode two independent sources together with distortions D_1 and D_2 is the same as if we encoded each of them separately.

15. **Distortion-rate function.** Let

$$D(R) = \min_{p(\hat{x}|x): I(X;\hat{X}) \le R} Ed(X,\hat{X}) \tag{10.143}$$

be the distortion rate function.

(a) Is $D(R)$ increasing or decreasing in R?

(b) Is $D(R)$ convex or concave in R?

(c) Converse for distortion rate functions: We now wish to prove the converse by focusing on $D(R)$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. Suppose one is given a $(2^{nR}, n)$ rate distortion code $X^n \to i(X^n) \to \hat{X}^n(i(X^n))$, with $i(X^n) \in 2^{nR}$. And suppose that the resulting distortion is $D = Ed(X^n, \hat{X}^n(i(X^n)))$. We must show that $D \ge D(R)$. Give reasons for the following steps in the proof:

$$\begin{aligned} D &= Ed(X^n, \hat{X}^n(i(X^n))) && (10.144) \\ &\overset{(a)}{=} E\frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i) && (10.145) \\ &\overset{(b)}{=} \frac{1}{n}\sum_{i=1}^{n} Ed(X_i, \hat{X}_i) && (10.146) \end{aligned}$$

$$\begin{aligned}
&\overset{(c)}{\geq} \frac{1}{n}\sum_{i=1}^{n} D\left(I(X_i;\hat{X}_i)\right) && (10.147)\\
&\overset{(d)}{\geq} D\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i;\hat{X}_i)\right) && (10.148)\\
&\overset{(e)}{\geq} D\left(\frac{1}{n}I(X^n;\hat{X}^n)\right) && (10.149)\\
&\overset{(f)}{\geq} D(R) && (10.150)
\end{aligned}$$

Solution: *Distortion rate function.*

(a) Since for larger values of R, the minimization in

$$D(R) = \min_{p(\hat{x}|x):I(X;\hat{X})\leq R} Ed(X,\hat{X}) \qquad (10.151)$$

is over a larger set of possible distributions, the minimum has to be at least as small as the minimum over the smaller set. Thus $D(R)$ is a nonincreasing function of R.

(b) By similar arguments as in Lemma 10.4.1, we can show that $D(R)$ is a convex function of R. Consider two rate distortion pairs (R_1, D_1) and (R_2, D_2) which lie on the distortion-rate curve. Let the joint distributions that achieve these pairs be $p_1(x,\hat{x}) = p(x)p_1(\hat{x}|x)$ and $p_2(x,\hat{x}) = p(x)p_2(\hat{x}|x)$. Consider the distribution $p_\lambda = \lambda p_1 + (1-\lambda)p_2$. Since the distortion is a linear function of the distribution, we have $D(p_\lambda) = \lambda D_1 + (1-\lambda)D_2$. Mutual information, on the other hand, is a convex function of the conditional distribution (Theorem 2.7.4) and hence

$$I_{p_\lambda}(X;\hat{X}) \leq \lambda I_{p_1}(X;\hat{X}) + (1-\lambda)I_{p_2}(X;\hat{X}) = \lambda R_1 + (1-\lambda)R_2 \qquad (10.152)$$

Therefore we can achieve a distortion $\lambda D_1 + (1-\lambda)D_2$ with a rate less than $\lambda R_1 + (1-\lambda)R_2$ and hence

$$\begin{aligned}
D(R_\lambda) &\leq D_{p_\lambda}(X;\hat{X}) && (10.153)\\
&= \lambda D(R_1) + (1-\lambda)D(R_2), && (10.154)
\end{aligned}$$

which proves that $D(R)$ is a convex function of R.

(c)

$$\begin{aligned}
D &= Ed(X^n,\hat{X}^n(i(X^n))) && (10.155)\\
&\overset{(a)}{=} E\frac{1}{n}\sum_{i=1}^{n} d(X_i,\hat{X}_i) && (10.156)\\
&\overset{(b)}{=} \frac{1}{n}\sum_{i=1}^{n} Ed(X_i,\hat{X}_i) && (10.157)
\end{aligned}$$

$$\overset{(c)}{\geq} \frac{1}{n}\sum_{i=1}^{n} D\left(I(X_i;\hat{X}_i)\right) \tag{10.158}$$

$$\overset{(d)}{\geq} D\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i;\hat{X}_i)\right) \tag{10.159}$$

$$\overset{(e)}{\geq} D\left(\frac{1}{n}I(X^n;\hat{X}^n)\right) \tag{10.160}$$

$$\overset{(f)}{\geq} D(R) \tag{10.161}$$

(a) follows from the definition of distortion for sequences
(b) from exchanging summation and expectation
(c) from the definition of the distortion rate function based on the joint distribution $p(x_i, \hat{x}_i)$,
(d) from Jensen's inequality and the convexity of $D(R)$
(e) from the fact that

$$\begin{aligned}
I(X^n;\hat{X}^n) &= H(X^n) - H(X^n|\hat{X}^n) && (10.162)\\
&= \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) && (10.163)\\
&= \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{i-1},\ldots,X_1) && (10.164)\\
&\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) && (10.165)\\
&= \sum_{i=1}^{n} I(X_i;\hat{X}_i) && (10.166)
\end{aligned}$$

and
(f) follows from the definition of the distortion rate function.

16. **Probability of conditionally typical sequences.** In Chapter 7, we calculated the probability that two independently drawn sequences X^n and Y^n are weakly jointly typical. To prove the rate distortion theorem, however, we need to calculate this probability when one of the sequences is fixed and the other is random.

The techniques of weak typicality allow us only to calculate the average set size of the conditionally typical set. Using the ideas of strong typicality on the other hand provides us with stronger bounds which work for all typical x^n sequences. We will outline the proof that $\Pr\{(x^n, Y^n) \in A_\epsilon^{*(n)}\} \approx 2^{-nI(X;Y)}$ for all typical x^n. This approach was introduced by Berger[1] and is fully developed in the book by Csiszár and Körner[3].

Let (X_i, Y_i) be drawn i.i.d. $\sim p(x,y)$. Let the marginals of X and Y be $p(x)$ and $p(y)$ respectively.

(a) Let $A_\epsilon^{*(n)}$ be the strongly typical set for X. Show that

$$|A_\epsilon^{*(n)}| \doteq 2^{nH(X)} \tag{10.167}$$

Hint: Theorem 11.1.1 and 11.1.3.

(b) The *joint type* of a pair of sequences (x^n, y^n) is the proportion of times $(x_i, y_i) = (a, b)$ in the pair of sequences, i.e.,

$$p_{x^n,y^n}(a,b) = \frac{1}{n} N(a,b|x^n,y^n) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = a, y_i = b). \tag{10.168}$$

The *conditional type* of a sequence y^n given x^n is a stochastic matrix that gives the proportion of times a particular element of $\mathcal{Y}$ occurred with each element of $\mathcal{X}$ in the pair of sequences. Specifically, the conditional type $V_{y^n|x^n}(b|a)$ is defined as

$$V_{y^n|x^n}(b|a) = \frac{N(a,b|x^n,y^n)}{N(a|x^n)}. \tag{10.169}$$

Show that the number of conditional types is bounded by $(n+1)^{|\mathcal{X}||\mathcal{Y}|}$.

(c) The set of sequences $y^n \in \mathcal{Y}^n$ with conditional type V with respect to a sequence x^n is called the conditional type class $T_V(x^n)$. Show that

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{nH(Y|X)} \le |T_V(x^n)| \le 2^{nH(Y|X)}. \tag{10.170}$$

(d) The sequence $y^n \in \mathcal{Y}^n$ is said to be ϵ*-strongly conditionally typical* with the sequence x^n with respect to the conditional distribution $V(\cdot|\cdot)$ if the conditional type is close to V. The conditional type should satisfy the following two conditions:

i. For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $V(b|a) > 0$,

$$\frac{1}{n} \left| N(a,b|x^n,y^n) - V(b|a)N(a|x^n) \right| \le \frac{\epsilon}{|\mathcal{Y}|+1}. \tag{10.171}$$

ii. $N(a,b|x^n,y^n) = 0$ for all (a,b) such that $V(b|a) = 0$.

The set of such sequences is called the conditionally typical set and is denoted $A_\epsilon^{*(n)}(Y|x^n)$. Show that the number of sequences y^n that are conditionally typical with a given $x^n \in \mathcal{X}^n$ is bounded by

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(Y|X)-\epsilon_1)} \le |A_\epsilon^{*(n)}(Y|x^n)| \le (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(Y|X)+\epsilon_1)}, \tag{10.172}$$

where $\epsilon_1 \to 0$ as $\epsilon \to 0$.

(e) For a pair of random variables (X, Y) with joint distribution $p(x, y)$, the ϵ-*strongly typical* set $A_\epsilon^{*(n)}$ is the set of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ satisfying

i.

$$\left| \frac{1}{n} N(a,b|x^n,y^n) - p(a,b) \right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|} \tag{10.173}$$

for every pair $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a,b) > 0$.

ii. $N(a,b|x^n,y^n) = 0$ for all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a,b) = 0$.

The set of ϵ-strongly jointly typical sequences is called the ϵ-strongly jointly typical set and is denoted $A_\epsilon^{*(n)}(X,Y)$.

Let (X,Y) be drawn i.i.d. $\sim p(x,y)$. For any x^n such that there exists at least one pair $(x^n,y^n) \in A_\epsilon^{*(n)}(X,Y)$, the set of sequences y^n such that $(x^n,y^n) \in A_\epsilon^{*(n)}$ satisfies

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(Y|X)-\delta(\epsilon))} \leq |\{y^n : (x^n,y^n) \in A_\epsilon^{*(n)}\}| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(Y|X)+\delta(\epsilon))}, \tag{10.174}$$

where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. In particular, we can write

$$2^{n(H(Y|X)-\epsilon_2)} \leq |\{y^n : (x^n,y^n) \in A_\epsilon^{*(n)}\}| \leq 2^{n(H(Y|X)+\epsilon_2)}, \tag{10.175}$$

where we can make ϵ_2 arbitrarily small with an appropriate choice of ϵ and n.

(f) Let $Y_1, Y_2, \ldots, Y_n$ be drawn i.i.d. $\sim \prod p(y_i)$. For $x^n \in A_\epsilon^{*(n)}$, the probability that $(x^n, Y^n) \in A_\epsilon^{*(n)}$ is bounded by

$$2^{-n(I(X;Y)+\epsilon_3)} \leq \Pr((x^n,Y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon_3)}, \tag{10.176}$$

where ϵ_3 goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

Solution:

Probability of conditionally typical sequences.

(a) The set of strongly typical sequences is the set of sequence whose type is close the distribution p. We have two conditions - that the proportion of any symbol a in the sequence is close to $p(a)$ and that no symbol with $p(a) = 0$ occurs in the sequence. The second condition may seem a technical one, but is essential in the proof of the strong equipartition theorem below.

By the strong law of large numbers, for a sequence drawn i.i.d. $\sim p(x)$, the asymptotic proportion of any letter a is close to $p(a)$ with high probability. So for appropriately large n, the proportion of every letter is within ϵ of $p(a)$ with probability close to 1, i.e., the strongly typical set has a probability close to 1. We will show that

$$2^{n(H(p)-\epsilon')} \leq |A_\epsilon^{*(n)}| \leq 2^{n(H(p)+\epsilon')}, \tag{10.177}$$

where ϵ' goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

For sequences in the strongly typical set,

$$\begin{aligned} -H(p) - \frac{1}{n}\log p(x^n) &= \sum_{a\in\mathcal{X}} p(a)\log p(a) - \frac{1}{n}\sum_{a\in\mathcal{X}} N(a|x^n)\log p(a) \\ &= -\sum_{a\in\mathcal{X}} \left(\frac{1}{n}N(a|x^n) - p(a)\right)\log p(a), \end{aligned} \tag{10.178}$$

and since $|\frac{1}{n}N(a|x^n) - p(a)| < \epsilon$ if $p(a) > 0$, and $|\frac{1}{n}N(a|x^n) - p(a)| = 0$ if $p(a) = 0$, we have

$$| - H(p) - \frac{1}{n}\log p(x^n)| < \epsilon_1. \tag{10.179}$$

where $\epsilon_1 = \epsilon \sum_{a:p(a)>0} \log \frac{1}{p(a)}$. It follows that $\epsilon_1 \to 0$ as $\epsilon \to 0$.

Recall the definition of weakly typical sequences in Chapter 3. A sequence was defined as ϵ_1-weakly typical if $| - \log p(x^n) - H(p)| \le \epsilon_1$. Hence a sequence that is ϵ-strongly typical is also ϵ_1-weakly typical. Hence the strongly typical set is a subset of the corresponding weakly typical set, i.e., $A_\epsilon^{*(n)} \subset A_{\epsilon_1}^{(n)}$.

Similarly, by the continuity of the entropy function, it follows that for all types in the typical set, the entropy of the type is close to $H(p)$. Specifically, for all $x^n \in A_\epsilon^{*(n)}$, $|p_{x^n}(a) - p(a)| < \epsilon$ and hence by Lemma 10.0.5, we have

$$|H(p_{x^n}) - H(p)| < \epsilon_2, \tag{10.180}$$

where $\epsilon_2 = -|\mathcal{X}|\epsilon \log \epsilon \to 0$ as $\epsilon \to 0$.

There are only a polynomial number of types altogether and hence there are only a polynomial number of types in the strongly typical set. The type class of any type $q \in A_\epsilon^{*(n)}$, by Theorem 12.1.3, has a size bounded by

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(q)} \le |T(q)| \le 2^{nH(q)}. \tag{10.181}$$

By the previous part of this theorem, for $q \in A_\epsilon^{*(n)}$, $|H(q) - H(p)| \le \epsilon_2$, and hence

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{n(H(p)-\epsilon_2)} \le |T(q)| \le 2^{n(H(p)+\epsilon_2)}. \tag{10.182}$$

Since the number of elements in the strongly typical set is the sum of the sizes of the type classes in the strongly typical set, and there are only a polynomial number of them, we have

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{n(H(p)-\epsilon_2)} \le |A_\epsilon^{*(n)}| \le (n+1)^{|\mathcal{X}|} 2^{n(H(p)+\epsilon_2)}, \tag{10.183}$$

i.e., $|\frac{1}{n}\log |A_\epsilon^{*(n)}| - H(p)| \le \epsilon'$, where $\epsilon' = \epsilon_2 + \frac{|\mathcal{X}|}{n}\log(n+1)$ which goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

It is instructive to compare the proofs of the strong AEP with the AEP for weakly typical sequences. The results are similar, but there is one important difference. The lower bound on size of the strongly typical set does not depend on the probability of the set—instead, the bound is derived directly in terms of the size of type classes. This enables the lower bound in the strong AEP to be extended to conditionally typical sequences and sets; the weak AEP cannot be extended similarly. We will consider the extensions of the AEP to conditional distributions in the next part.

(b) The concept of types for single sequences can be extended to pairs of sequences for which we can define the concept of the joint type and the conditional type.

Definition: The *joint type* of a pair of sequences (x^n, y^n) is the proportion of times a pair of symbols (a, b) occurs jointly the the pair of sequences, i.e.,

$$p_{x^n,y^n}(a,b) = \frac{1}{n} N(a,b|x^n,y^n). \tag{10.184}$$

Definition: The *conditional type* of a sequence y^n given x^n is a stochastic matrix that gives the proportion of times a particular element of $\mathcal{Y}$ occurred with each element of $\mathcal{X}$ in the pair of sequences. Specifically, the conditional type $V_{y^n|x^n}(b|a)$ is defined as

$$V_{y^n|x^n}(b|a) = \frac{N(a,b|x^n,y^n)}{N(a|x^n)}. \tag{10.185}$$

The set of sequences $y^n \in \mathcal{Y}^n$ with conditional type V with respect to a sequence x^n is called the conditional type class $T_V(x^n)$.

Lemma 10.0.2 *The number of conditional types for sequences of length n from the alphabet $\mathcal{X}$ and $\mathcal{Y}$ is bounded by $(n+1)^{|\mathcal{X}||\mathcal{Y}|}$.*

Proof: By Theorem 12.1.1, the number of ways of choosing a row of the matrix $V(\cdot|a)$ is bounded by $(n+1)^{|\mathcal{Y}|}$ and there are $|\mathcal{X}|$ different choices of rows. So the total number of different conditional types is bounded by $(n+1)^{|\mathcal{X}||\mathcal{Y}|}$. □

(c) Since $V_{y^n|x^n}$ is a stochastic matrix, we can multiply it with p_{x^n} to find the joint type of (x^n, y^n). We will denote the conditional entropy of Y given X for this joint distribution as $H(V_{y^n|x^n}|p_{x^n})$.

Lemma 10.0.3 *For $x^n \in \mathcal{X}^n$, let $T_V(x^n)$ denote the set of sequences $y^n \in \mathcal{Y}^n$ with conditional type V with respect to x^n. Then*

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{nH(V|p_{x^n})} \le |T_V(x^n)| \le 2^{nH(V|p_{x^n})}. \tag{10.186}$$

Proof: This is a direct consequence of the corresponding lemma about the size of unconditional type classes. We can consider the subsequences of the pair corresponding each element of $\mathcal{X}$. For any particular element $a \in \mathcal{X}$, the number of conditionally typical sequences depends only the conditional type $V(\cdot|a)$, and hence the number of conditionally typical sequences is bounded by

$$\prod_{a\in\mathcal{X}} \frac{1}{(N(a|x^n)+1)^{|\mathcal{Y}|}} 2^{N(a|x^n)H(V|p_{x^n})} \le |T_V(x^n)| \le \prod_{a\in\mathcal{X}} 2^{N(a|x^n)H(V|p_{x^n})} \tag{10.187}$$

which proves the lemma. □

The above two lemmas generalize the corresponding lemmas for unconditional types. We can use these to extend the strong AEP to conditionally typical sets.

(d) We begin with the definition of strongly conditionally typical sequences.

Definition: The sequence $y^n \in \mathcal{Y}^n$ is said to be ϵ*-strongly conditionally typical* with the sequence x^n with respect to the conditional distribution $V(\cdot|\cdot)$ if the conditional type is close to V. The conditional type should satisfy the following two conditions:

i. For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $V(b|a) > 0$,

$$\frac{1}{n}\left|N(a,b|x^n,y^n) - V(b|a)N(a|x^n)\right| \le \epsilon. \tag{10.188}$$

ii. $N(a,b|x^n,y^n) = 0$ for all (a,b) such that $V(b|a) = 0$.

The set of such sequences is called the conditionally typical set and is denoted $A_\epsilon^{*(n)}(Y|x^n)$.

Essentially, a sequence y^n is conditionally typical with x^n if the subsequence of y^n corresponding to the occurrences of a particular symbol a in x^n is typical with respect to the conditional distribution $V(\cdot|a)$. Since the number of such conditionally typical sequences is just the product of the number of subsequences conditionally typically corresponding to each choice of $a \in \mathcal{X}$, we can now extend the strong AEP to derive a bound on the size of the conditionally typical set.

Lemma 10.0.4 *The number of sequences y^n that are conditionally typical with a given $x^n \in \mathcal{X}^n$ is bounded by*

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(V|p_{x^n})-\epsilon_4)} \le |A_\epsilon^{*(n)}(Y|x^n)| \le (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(V|p_{x^n})+\epsilon_4)}, \tag{10.189}$$

where $\epsilon_4 = -|\mathcal{X}||\mathcal{Y}|\epsilon\log\epsilon \to 0$ as $\epsilon \to 0$.

Proof: Just as in the proof of the strong AEP (Theorem 12.2.1), we will derive the bounds using purely combinatorial arguments. The size of the conditional type class is bounded in Lemma 10.0.3 in terms of the entropy of the conditional type. By Lemma 10.0.5 and the definition of the conditionally typical set, we have

$$\left|H(p_{y^n|x^n}|p_{x^n}) - H(V|p_{x^n})\right| \le -|\mathcal{X}||\mathcal{Y}|\epsilon\log\epsilon \tag{10.190}$$

Combining this with the bound on the number of conditional types (Lemma 10.0.2), we have the theorem. □

(e) We now extend the definition of strongly typical sequences to pairs of sequences. The joint type of a pair of sequences is the proportion of occurrences of a pair of symbols together in the pair. A pair of sequences (x^n,y^n) is called jointly strongly typical with respect to a distribution $p(x,y)$ if the joint type is close to $p(x,y)$.

Definition: For a pair of random variables (X,Y) with joint distribution $p(x,y)$, the ϵ*-strongly typical* set $A_\epsilon^{*(n)}$ is the set of sequences $(x^n,y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ satisfying

i.
$$\left|\frac{1}{n}N(a,b|x^n,y^n)-p(a,b)\right|<\epsilon \tag{10.191}$$
for every pair $(a,b)\in\mathcal{X}\times\mathcal{Y}$ with $p(a,b)>0$.

ii. $N(a,b|x^n,y^n)=0$ for all $(a,b)\in\mathcal{X}\times\mathcal{Y}$ with $p(a,b)=0$.

The set of ϵ-strongly jointly typical sequences is called the ϵ-strongly jointly typical set and is denoted $A_\epsilon^{*(n)}(X,Y)$.

Theorem 10.0.1 *(Joint AEP.) Let (X^n,Y^n) be sequences of length n drawn i.i.d. according to $p(x^n,y^n)=\prod_{i=1}^n p(x_i,y_i)$. Then*
$$P(A_\epsilon^{*(n)})\to 1, \quad \textit{as } n\to\infty. \tag{10.192}$$

Proof: Follows directly from the weak law of large numbers. □

From the definition, it is clear that strongly jointly typical sequences are also individually typical, i.e., for x^n such that $(x^n,y^n)\in A_\epsilon^{*(n)}(X,Y)$,
$$\begin{aligned}|p_{x^n}(a)-p(a)| &\le \sum_{b\in\mathcal{Y}}|p_{x^n,y^n}(a,b)-p(a,b)| && (10.193)\\ &\le \epsilon|\mathcal{Y}|, \quad \text{for all } a\in\mathcal{X}. && (10.194)\end{aligned}$$

Hence $x^n\in A_{\epsilon|\mathcal{Y}|}^{*(n)}$. This in turn implies that the pair is also conditionally typical for the conditional distribution $p(y|x)$, i.e., for $(x^n,y^n)\in A_\epsilon^{*(n)}(X,Y)$,
$$|p_{x^n,y^n}(a,b)-p(b|a)p_{x^n}(a)|<\epsilon(|\mathcal{Y}|+1)<\epsilon|\mathcal{X}||\mathcal{Y}|. \tag{10.195}$$

Since conditional entropy is also a continuous function of the distribution, the conditional entropy of the type of a jointly strongly typical sequence, p_{x^n,y^n}, is close to conditional entropy for $p(x,y)$. Hence we can also extend Lemma 10.0.3 for elements of the typical set as follows:

Theorem 10.0.2 *(Size of conditionally typical set)*
Let (X,Y) be drawn i.i.d. $\sim p(x,y)$. For any x^n such that there exists at least one pair $(x^n,y^n)\in A_\epsilon^{*(n)}(X,Y)$, the set of sequences y^n such that $(x^n,y^n)\in A_\epsilon^{*(n)}$ satisfies
$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}}2^{n(H(Y|X)-\delta(\epsilon))}\le|\{y^n:(x^n,y^n)\in A_\epsilon^{*(n)}\}|\le(n+1)^{|\mathcal{X}||\mathcal{Y}|}2^{n(H(Y|X)+\delta(\epsilon))}, \tag{10.196}$$
where $\delta(\epsilon)\to 0$ as $\epsilon\to 0$. In particular, we can write
$$2^{n(H(Y|X)-\epsilon_5)}\le|\{y^n:(x^n,y^n)\in A_\epsilon^{*(n)}\}|\le 2^{n(H(Y|X)+\epsilon_5)}, \tag{10.197}$$
where we can make ϵ_5 arbitrarily small with an appropriate choice of ϵ and n.

Proof: The theorem follows from Theorem 10.0.2 and the continuity of conditional entropy as a function of the joint distribution. Now the set of sequences that are jointly typical with a given x^n are also $\epsilon|\mathcal{X}||\mathcal{Y}|$-strongly conditionally typical, and hence from the upper bound of Theorem 10.0.2, we have

$$|\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(p(b|a)|p_{x^n})+\delta(\epsilon))}, \tag{10.198}$$

where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. Now since

$$H(Y|X) = -\sum_x p(x) \sum p(y|x) \log p(y|x) \tag{10.199}$$

is a linear function of the distribution $p(x)$, we have

$$|H(p(b|a)|p_{x^n}) - H(Y|X)| \leq \epsilon|\mathcal{Y}| \max_{a\in\mathcal{X}} H(Y|X=a) \leq \epsilon|\mathcal{Y}| \log|\mathcal{Y}|, \tag{10.200}$$

which gives us the upper bound of the theorem.
For the lower bound, assume that $(x^n, y^n) \in A_\epsilon^{*(n)}(X,Y)$. Then since the joint type of a pair of sequences is determined by the type of x^n and the conditional type of y^n given x^n, all sequences y^n with this conditional type will also be in $A_\epsilon^{*(n)}(X,Y)$. Hence the number of sequences $|\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}|$ is at least as much as the number of sequences of this conditional type, which by the lower bound of Lemma 10.0.4, and the continuity of conditional entropy as a function of the joint distribution (Lemma 10.0.5 and (10.200)), we have

$$|\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}| \geq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(p(b|a)|p_{x^n})-\delta(\epsilon))}, \tag{10.201}$$

where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. This gives us the theorem with

$$\epsilon_5 = \frac{|\mathcal{X}||\mathcal{Y}|}{n} \log(n+1) + \epsilon|\mathcal{Y}| \log|\mathcal{Y}| - \epsilon|\mathcal{X}|^2|\mathcal{Y}|^2 \log \epsilon|\mathcal{X}||\mathcal{Y}|. \tag{10.202}$$

□

To use this result, we have to assume that there is at least one y^n such that $(x^n, y^n) \in A_\epsilon^{*(n)}(X,Y)$. From the definitions of the strongly typical sets, it is clear that if $|p_{x^n}(a) - p(a)| < \epsilon$, there exists at least one conditional distribution $\hat{p}(b|a)$ such that $|\hat{p}(b|a)p_{x^n}(a) - p(a,b)| < \epsilon$ and hence for large enough n, we have at least one conditional type such that $|p_{x^n,y^n}(a,b) - p(a,b)| \leq \epsilon$ and hence if x^n is ϵ-strongly typical, then there exists a conditional type such the joint type is jointly typical. For such an x^n sequence, we can always find a y^n such that (x^n, y^n) is jointly typical.

(f) Notice that for the results of Theorems 10.0.2, we have used purely combinatorial arguments to bound the size of the conditionally type class and the conditionally typical set. These theorems illustrate the power of the method of types. We will now use the last theorem to bound the probability that a randomly chosen Y^n will be conditionally typical with a given $x^n \in A_\epsilon^{*(n)}$.

Theorem 10.0.3 *Let $Y_1, Y_2, \ldots, Y_n$ be drawn i.i.d. $\sim \prod p(y)$. For $x^n \in A^{*(n)}_{\frac{\epsilon}{3}}$, the probability that $(x^n, Y^n) \in A^{*(n)}_\epsilon$ is bounded by*

$$2^{-n(I(X;Y)+\epsilon_7)} \leq Pr((x^n, Y^n) \in A^{*(n)}_\epsilon) \leq 2^{-n(I(X;Y)-\epsilon_7)} \tag{10.203}$$

where $\epsilon_7 = \epsilon_5 - \epsilon|\mathcal{X}||\mathcal{Y}| \log \epsilon|\mathcal{X}|$ which goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

Proof: If $Y^n \in A^{*(n)}_\epsilon$, then $p(Y^n) \doteq 2^{-n(H(Y))}$, and hence

$$\begin{aligned}
P((x^n, Y^n) \in A^{*(n)}_\epsilon) &= \sum_{y^n:(x^n,y^n)\in A^{*(n)}_\epsilon} p(y^n) && (10.204)\\
&\leq \sum_{y^n:(x^n,y^n)\in A^{*(n)}_\epsilon} 2^{-n(H(Y)-\epsilon_6)} && (10.205)\\
&= |A^{*(n)}_\epsilon(Y|x^n)| 2^{-n(H(Y)-\epsilon_6)} && (10.206)\\
&\leq 2^{n(H(Y|X)+\epsilon_5)} 2^{-n(H(Y)-\epsilon_6)} && (10.207)\\
&= 2^{-n(I(X;Y)-\epsilon_7)}. && (10.208)
\end{aligned}$$

where $\epsilon_6 = -\epsilon|\mathcal{X}||\mathcal{Y}| \log \epsilon|\mathcal{X}|$ since $|p_{y^n} - p| \leq \epsilon|\mathcal{X}|$ if $(x^n, y^n) \in A^{*(n)}_\epsilon$. Also

$$\begin{aligned}
P((x^n, Y^n) \in A^{*(n)}_\epsilon) &= \sum_{y^n:(x^n,y^n)\in A^{*(n)}_\epsilon} p(y^n) && (10.209)\\
&\geq \sum_{y^n:(x^n,y^n)\in A^{*(n)}_\epsilon} 2^{-n(H(Y)+\epsilon_6)} && (10.210)\\
&= |A^{*(n)}_\epsilon(Y|x^n)| 2^{-n(H(Y)+\epsilon_6)} && (10.211)\\
&\geq 2^{n(H(Y|X)-\epsilon_5)} 2^{-n(H(Y)+\epsilon_6)} && (10.212)\\
&= 2^{-n(I(X;Y)+\epsilon_7)}. && (10.213)
\end{aligned}$$

Hence

$$2^{-n(I(X;Y)+\epsilon_7)} \leq Pr((x^n, Y^n) \in A^{*(n)}_\epsilon) \leq 2^{-n(I(X;Y)-\epsilon_7)}. \tag{10.214}$$

□

The main result of this problem is the last theorem, which gives upper and lower bounds on the probability that a randomly chosen sequence y^n will be jointly typical with a given x^n. This was used in the proof of the rate distortion theorem. To end this solution, we will prove a theorem on the continuity of entropy:

Lemma 10.0.5 *If $|p(x) - q(x)| \leq \epsilon$ for all x, then $|H(p) - H(q)| \leq -\epsilon|\mathcal{X}| \log \epsilon$.*

Proof: We will use some simple properties of the function

$$f(x) = -x \ln x \quad \text{for } 0 \leq x \leq \frac{1}{e}. \tag{10.215}$$

Since $f'(x) = -1 - \ln x > 0$ and $f''(x) = -\frac{1}{x}$, $f(x)$ is an increasing concave function. Now consider

$$g(x) = f(x+\epsilon) - f(x) = x\ln x - (x+\epsilon)\ln(x+\epsilon). \qquad (10.216)$$

Then again by differentiation, it is clear that $g'(x) < 0$ so the function is strictly decreasing. Hence $g(x) < g(0) = -\epsilon \ln \epsilon$ for all x.
For any $a \in \mathcal{X}$, assume $p(a) > q(a)$, and hence we have

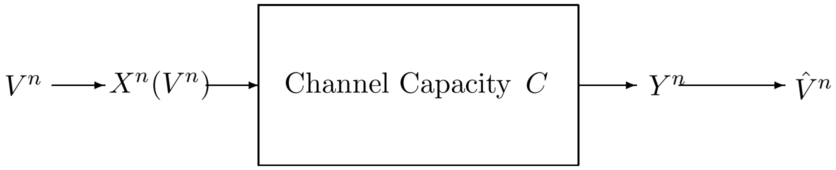
$$p(a) - q(a) \leq \epsilon \qquad (10.217)$$

Hence by the fact that f is an increasing function, we have

$$\begin{aligned} |-p(a)\ln p(a) + q(a)\ln q(a)| &= -p(a)\ln p(a) + q(a)\ln q(a) \qquad (10.218) \\ &\leq -(q(a)+\epsilon)\ln(q(a)+\epsilon) + q(a)\ln q(a) \\ &\leq -\epsilon \ln \epsilon. \qquad (10.219) \end{aligned}$$

Summing this over all $a \in \mathcal{X}$, we have the lemma. □

17. **The source-channel separation theorem with distortion:** Let $V_1, V_2, \ldots, V_n$ be a finite alphabet i.i.d. source which is encoded as a sequence of n input symbols X^n of a discrete memoryless channel. The output of the channel Y^n is mapped onto the reconstruction alphabet $\hat{V}^n = g(Y^n)$. Let $D = Ed(V^n, \hat{V}^n) = \frac{1}{n}\sum_{i=1}^n Ed(V_i, \hat{V}_i)$ be the average distortion achieved by this combined source and channel coding scheme.



(a) Show that if $C > R(D)$, where $R(D)$ is the rate distortion function for V, then it is possible to find encoders and decoders that achieve a average distortion arbitrarily close to D.

(b) (Converse.) Show that if the average distortion is equal to D, then the capacity of the channel C must be greater than $R(D)$.

Solution: *Source channel separation theorem with distortion*

(a) To show achievability, we consider two codes at rate R, where $C > R > R(D)$. The first code is a rate distortion code that achieves distortion D at rate R. The second code is a channel code that allows transmission over the channel at rate R with probability of error going to 0. Using the rate distortion code to encode the source into one of the 2^{nR} messages, and the channel code to send this message over the channel. Since the probability of error is exponentially small, the received message is the same as the transmitted message with probability close to 1. In that case, the results of the achieveability of rate distortion show that the decoded sequence is within distortion D of the input sequence with high probability.

To complete the analysis, we need to consider the case wther the channel code produces an error—however, even in this case, the distortion produced by the error is bounded, and hence the total distortion is essentially the same as achieved without the errors.

(b) To prove the converse, we need to prove that for any encoding system that achieves distortion D, the capacity of the channel should be greater than $R(D)$. Mimicking the steps for the converse for the rate distortion function, we can define coding function f_n and decoding function g_n, Let $\hat{V}^n = \hat{V}^n(Y^n) = g_n(Y^n)$ be the reproduced sequence corresponding to V^n. Assume that $Ed(V^n, \hat{V}^n) \geq D$ for this code. Then we have the following chain of inequalities:

$$\begin{aligned}
I(V^n;\hat{V}^n) &= H(V^n) - H(V^n|\hat{V}^n) && (10.220)\\
&= \sum_{i=1}^{n} H(V_i) - H(V^n|\hat{V}^n) && (10.221)\\
&= \sum_{i=1}^{n} H(V_i) - \sum_{i=1}^{n} H(V_i|\hat{V}^n, V_{i-1}, \ldots, V_1) && (10.222)\\
&\overset{(a)}{\geq} \sum_{i=1}^{n} H(V_i) - \sum_{i=1}^{n} H(V_i|\hat{V}_i) && (10.223)\\
&= \sum_{i=1}^{n} I(V_i;\hat{V}_i) && (10.224)\\
&\geq \sum_{i=1}^{n} R(Ed(V_i,\hat{V}_i)) && (10.225)\\
&= n\left(\frac{1}{n}\sum_{i=1}^{n} R(Ed(V_i,\hat{V}_i))\right) && (10.226)\\
&\overset{(b)}{\geq} nR\left(\frac{1}{n}\sum_{i=1}^{n} Ed(V_i,\hat{V}_i)\right) && (10.227)\\
&= nR(Ed(V^n,\hat{V}^n)) && (10.228)\\
&= nR(D), && (10.229)
\end{aligned}$$

where
(a) follows from the fact that conditioning reduces entropy,
(b) from the convexity of the rate distortion function. Also by the data processing inequality,

$$I(V^n;\hat{V}^n) \geq I(X^n;Y^n) \leq nC \qquad (10.230)$$

where the last inequality follows from Lemma 7.9.2.

18. **Rate distortion.**
Let $d(x,\hat{x})$ be a distortion function. We have a source $X \sim p(x)$. Let $R(D)$ be the associated rate distortion function.

(a) Find $\tilde{R}(D)$ in terms of $R(D)$, where $\tilde{R}(D)$ is the rate distortion function associated with the distortion $\tilde{d}(x,\hat{x}) = d(x,\hat{x}) + a$ for some constant $a > 0$. (They are not equal)

(b) Now suppose that $d(x,\hat{x}) \geq 0$ for all $x, \hat{x}$ and define a new distortion function $d^*(x,\hat{x}) = bd(x,\hat{x})$, where b is some number ≥ 0. Find the associated rate distortion function $R^*(D)$ in terms of $R(D)$.

(c) Let $X \sim N(0,\sigma^2)$ and $d(x,\hat{x}) = 5(x-\hat{x})^2 + 3$. What is $R(D)$?

Solution: *Rate distortion.*

(a)

$$\begin{aligned}\tilde{R}(D) &= \inf_{p(\hat{x}|x):\ E(\tilde{d}(x,\hat{x}))\leq D} I(X;\hat{X}) \\ &= \inf_{p(\hat{x}|x):\ E(d(x,\hat{x}))+a\leq D} I(X;\hat{X}) \\ &= \inf_{p(\hat{x}|x):\ E(d(x,\hat{x}))\leq D-a} I(X;\hat{X}) \\ &= R(D-a)\end{aligned}$$

(b) If $b > 0$

$$\begin{aligned}R^*(D) &= \inf_{p(\hat{x}|x):\ E(d^*(x,\hat{x}))\leq D} I(X;\hat{X}) \\ &= \inf_{p(\hat{x}|x):\ E(bd(x,\hat{x}))\leq D} I(X;\hat{X}) \\ &= \inf_{p(\hat{x}|x):\ E(d(x,\hat{x}))\leq \frac{D}{b}} I(X;\hat{X}) \\ &= R\left(\frac{D}{b}\right),\end{aligned}$$

else if $b = 0$ then $d^* = 0$ and $R^*(D) = 0$.

(c) Let $R_{se}(D)$ be the rate distortion function associate with the distortion $d_{se}(x,\hat{x}) = (x-\hat{x})^2$. Then from parts (a) and (b) we have

$$R(D) = R_{se}\left(\frac{D-3}{5}\right).$$

We know that

$$R_{se}(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D} & 0 \le D \le \sigma^2 \\ 0 & D < \sigma^2 \end{cases}$$

Therefore we have

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{5\sigma^2}{D-3} & 3 \le D \le 5\sigma^2 + 3 \\ 0 & D > 5\sigma^2 + 3 \end{cases}$$

19. **Rate distortion with two constraints**
Let X_i be iid $\sim p(x)$. We are given two distortion functions $d_1(x,\hat{x})$ and $d_2(x,\hat{x})$. We wish to describe X^n at rate R and reconstruct it with distortions $Ed_1(X^n,\hat{X}_1^n) \le D_1$, and $Ed_2(X^n,\hat{X}_2^n) \le D_2$, as shown here:

$$X^n \longrightarrow i(X^n) \longrightarrow (\hat{X}_1^n(i), \hat{X}_2^n(i))$$

$$\begin{aligned} D_1 &= ED(X_1^n, \hat{X}_1^n) \\ D_2 &= ED(X_1^n, \hat{X}_2^n). \end{aligned}$$

Here $i(\cdot)$ takes on 2^{nR} values. What is the rate distortion function $R(D_1,D_2)$?

Solution: *Rate distortion with two constraints*

$$\begin{aligned} R(D_1,D_2) \quad = \quad & \min_{p(\hat{x}_1,\hat{x}_2|x)} I(X;\hat{X}_1,\hat{X}_2) \\ & \text{subject to:} \\ & \mathbf{E}d_1(\hat{X}_1,X) \le D_1 \\ & \mathbf{E}d_2(\hat{X}_2,X) \le D_2 \end{aligned}$$

Some interesting things to note about $R(D_1,D_2)$ are the following. First, $\max(R(D_1),R(D_2)) \le R(D_1,D_2) \le R(D_1)+R(D_2)$. The upper bound occurs when the mutual information is minimized with $\hat{X}_1$ independent of $\hat{X}_2$ which is always allowed. The lower bound occurs because the best rate achieved in the more constrained problem can not be lower than the best rate acheived in either less constrained problem. Note that the optimization is over the set of distributions of the form $p(\hat{x}_1,\hat{x}_2|x)$ which is a larger set than if conditional independence $p(\hat{x}_1|x)p(\hat{x}_2|x)$ were required, and the minimum rate achieving distribution may not have conditional independence. As a simple example of where the optimal solution is conditionally dependent consider a Gaussian source where both distortion measures are square error and the distortion bounds are the same as well. In this case the minimum rate is achieved when $\hat{x}_1 = \hat{x}_2$ almost surely which gives $R(D_1,D_2) = R(D_1) = R(D_2)$. So for $I(X;\hat{X}_1,\hat{X}_2) = I(X;\hat{X}_1) + I(X;\hat{X}_2|\hat{X}_1)$ the second term is zero and the first term is minimized which is not possible of conditional independence is required.

20. **Rate distortion**
Consider the standard rate distortion problem, X_i i.i.d. $\sim p(x)$, $X^n \to i(X^n) \to \hat{X}^n$, $|i(\cdot)| = 2^{nR}$. Consider two distortion criteria $d_1(x,\hat{x})$ and $d_2(x,\hat{x})$.

Suppose $d_1(x,\hat{x}) \le d_2(x,\hat{x})$ for all $x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}$.

Let $R_1(D)$ and $R_2(D)$ be the corresponding rate distortion functions.

(a) Find the inequality relationship between $R_1(D)$ and $R_2(D)$.

(b) Suppose we must describe the source $\{X_i\}$ at the minimum rate R achieving $d_1(X^n, \hat{X}_1^n) \le D$ and $d_2(X^n, \hat{X}_2^n) \le D$. Thus

$$X^n \to i(X^n) \to \begin{cases} \hat{X}_1^n(i(X^n)) \\ \hat{X}_2^n(i(X^n)) \end{cases}$$

and $|i(\cdot)| = 2^{nR}$.
Find the minimum rate R.

Solution: *Rate distortion*

(a) Any rate (and coding scheme) satisfying $d_2(X^n, \hat{X^n}) \le D$ automatically satisfy $d_1(X^n, \hat{X^n}) \le D$. Hence

$$R_1(D) \le R_2(D).$$

(b) As in Problem 10.19,

$$R_2(D) = \max(R_1(D), R_2(D)) \le R(D,D),$$

where $R(D,D)$ is the minimum rate distortion function achieving both distortion criteria.

For the other direction of inequality, repeat the argument we used in part (a): If we use the rate $R_2(D)$ and the optimal coding scheme with $\hat{X}_1 = \hat{X}_2 = \hat{X}$, we satisfy both distortion constraints since $d_1(X^n, \hat{X}^n) \le d_2(X^n, \hat{X}^n) \le D$. This implies $R_2(D)$ is achievable so that $R_2(D) \ge R(D,D)$.

Chapter 11

Information Theory and Statistics

1. **Chernoff-Stein lemma.** Consider the two hypothesis test

$$H_1 : f = f_1 \quad \text{vs.} \quad H_2 : f = f_2$$

Find $D(f_1 \parallel f_2)$ if

(a) $f_i(x) = N(0, \sigma_i^2), i = 1, 2$

(b) $f_i(x) = \lambda_i e^{-\lambda_i x}, x \geq 0, i = 1, 2$

(c) $f_1(x)$ is the uniform density over the interval [0,1] and $f_2(x)$ is the uniform density over $[a, a+1]$. Assume $0 < a < 1$.

(d) f_1 corresponds to a fair coin and f_2 corresponds to a two-headed coin.

Solution: *Stein's lemma.*

(a) $f_1 = \mathcal{N}(0, \sigma_1^2)$, $f_2 = \mathcal{N}(0, \sigma_2^2)$,

$$D(f_1 || f_2) = \int_{-\infty}^{\infty} f_1(x) \left[\frac{1}{2} \ln \frac{\sigma_2^2}{\sigma_1^2} - \left(\frac{x^2}{2\sigma_1^2} - \frac{x^2}{2\sigma_2^2} \right) \right] dx \tag{11.1}$$

$$= \frac{1}{2} \left[\ln \frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1 \right]. \tag{11.2}$$

(b) $f_1 = \lambda_1 e^{-\lambda_1 x}$, $f_2 = \lambda_2 e^{-\lambda_2 x}$,

$$D(f_1 || f_2) = \int_0^{\infty} f_1(x) \left[\ln \frac{\lambda_1}{\lambda_2} - \lambda_1 x + \lambda_2 x \right] dx \tag{11.3}$$

$$= \ln \frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1} - 1. \tag{11.4}$$

(c) $f_1 = U[0,1]$, $f_2 = U[a,a+1]$,

$$\begin{aligned} D(f_1||f_2) &= \int_0^1 f_1 \ln \frac{f_1}{f_2} && (11.5)\\ &= \int_0^a f_1 \ln \infty + \int_a^1 f_1 \ln 1 && (11.6)\\ &= \infty. && (11.7) \end{aligned}$$

In this case, the Kullback Leibler distance of ∞ implies that in a hypothesis test, the two distributions will be distinguished with probability 1 for large samples.

(d) $f_1 = \text{Bern}\left(\frac{1}{2}\right)$ and $f_2 = \text{Bern}(1)$,

$$D(f_1||f_2) = \frac{1}{2}\ln\frac{\frac{1}{2}}{1} + \frac{1}{2}\ln\frac{\frac{1}{2}}{0} = \infty. \qquad (11.8)$$

The implication is the same as in part (c).

2. **A relation between $D(P \parallel Q)$ and Chi-square.** Show that the χ^2 statistic

$$\chi^2 = \Sigma_x \frac{(P(x) - Q(x))^2}{Q(x)}$$

is (twice) the first term in the Taylor series expansion of $D(P \parallel Q)$ about Q. Thus $D(P \parallel Q) = \frac{1}{2}\chi^2 + \ldots$

Suggestion: Write $\frac{P}{Q} = 1 + \frac{P-Q}{Q}$ and expand the log.

Solution: *A relation between $D(P \parallel Q)$ and Chi-square.*

There are many ways to expand $D(P||Q)$ in a Taylor series, but when we are expanding about $P = Q$, we must get a series in $P - Q$, whose coefficients depend on Q only. It is easy to get misled into forming another series expansion, so we will provide two alternative proofs of this result.

- Expanding the log.
  Writing $\frac{P}{Q} = 1 + \frac{P-Q}{Q} = 1 + \frac{\Delta}{Q}$, and $P = Q + \Delta$, we get

$$\begin{aligned} D(P||Q) &= \int P \ln \frac{P}{Q} && (11.9)\\ &= \int (Q+\Delta)\ln\left(1 + \frac{\Delta}{Q}\right) && (11.10)\\ &= \int (Q+\Delta)\left(\frac{\Delta}{Q} - \frac{\Delta^2}{2Q^2} + \ldots\right) && (11.11)\\ &= \int \Delta + \frac{\Delta^2}{Q} - \frac{\Delta^2}{2Q} + \ldots. && (11.12) \end{aligned}$$

The integral of the first term $\int \Delta = \int P - \int Q = 0$, and hence the first non-zero term in the expansion is

$$\frac{\Delta^2}{2Q} = \frac{\chi^2}{2}, \tag{11.13}$$

which shows that locally around Q, $D(P||Q)$ behaves quadratically like χ^2.

- By differentiation.

  If we construct the Taylor series expansion for f, we can write

$$f(x) = f(c) + f'(c)(x-c) + f''(c)\frac{(x-c)^2}{2} + \ldots \tag{11.14}$$

Doing the same expansion for $D(P||Q)$ around the point Q, we get

$$D(P||Q)_{P=Q} = 0, \tag{11.15}$$

$$D'(P||Q)_{P=Q} = (\ln\frac{P}{Q} + 1)_{P=Q} = 1, \tag{11.16}$$

and

$$D''(P||Q)_{P=Q} = \left(\frac{1}{P}\right)_{P=Q} = \frac{1}{Q}. \tag{11.17}$$

Hence the Taylor series is

$$\begin{aligned} D(P||Q) &= 0 + \int 1(P-Q) + \int \frac{1}{Q}\frac{(P-Q)^2}{2} + \ldots & (11.18)\\ &= \frac{1}{2}\chi^2 + \ldots . & (11.19) \end{aligned}$$

and we get $\frac{\chi^2}{2}$ as the first non-zero term in the expansion.

3. **Error exponent for universal codes.** A universal source code of rate R achieves a probability of error $P_e^{(n)} \doteq e^{-nD(P^*\|Q)}$, where Q is the true distribution and P^* achieves $\min D(P \parallel Q)$ over all P such that $H(P) \geq R$.

   (a) Find P^* in terms of Q and R.

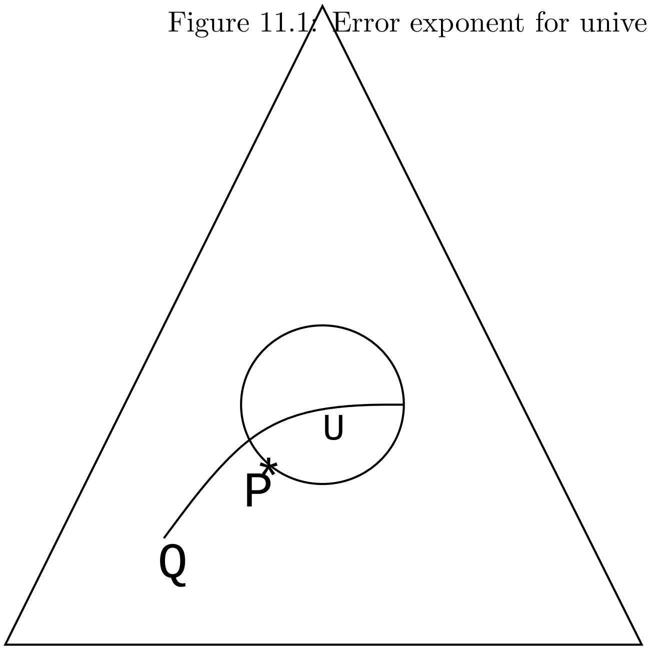
   (b) Now let X be binary. Find the region of source probabilities $Q(x)$, $x \in \{0,1\}$, for which rate R is sufficient for the universal source code to achieve $P_e^{(n)} \to 0$.

Solution: *Error exponent for universal codes.*

(a) We have to minimize $D(p||q)$ subject to the constraint that $H(p) \geq R$. Rewriting this problem using Lagrange multipliers, we get

$$J(p) = \sum p \log\frac{p}{q} + \lambda \sum p \log p + \nu \sum p. \tag{11.20}$$

Figure 11.1: Error exponent for universal codes



Differentiating with respect to $p(x)$ and setting the derivative to 0, we obtain

$$\log\frac{p}{q}+1+\lambda\log p+\lambda+\nu=0, \tag{11.21}$$

which implies that

$$p^*(x)=\frac{q^\mu(x)}{\sum_a q^\mu(a)}. \tag{11.22}$$

where $\mu=\frac{\lambda}{1-\lambda}$ is chosen to satisfy the constraint $H(p^*)=R$. We have to first check that the constraint is active, i.e., that we really need equality in the constraint. For this we set $\lambda=0$ or $\mu=1$, and we get $p^*=q$. Hence if q is such that $H(q)\geq R$, then the maximizing p^* is q. On the other hand, if $H(q)<R$, then $\lambda\neq 0$, and the constraint must be satisfied with equality.

Geometrically it is clear that there will be two solutions for λ of the form (11.22) which have $H(p^*)=R$, corresponding to the minimum and maximum distance to q on the manifold $H(p)=R$. It is easy to see that for $0\leq\mu\leq 1$, $p^*_\mu(x)$ lies on the geodesic from q to the uniform distribution. Hence, the minimum will lie in this region of μ. The maximum will correspond to negative μ, which lies on the other side of the uniform distribution as in the figure.

(b) For a universal code with rate R, any source can be transmitted by the code if $H(p) < R$. In the binary case, this corresponds to $p \in [0, h^{-1}(R))$ or $p \in (1-h^{-1}(R), 1]$, where h is the binary entropy function.

4. **Sequential projection.** We wish to show that projecting Q onto P_1 and then projecting the projection $\hat{Q}$ onto $P_1 \bigcap P_2$ is the same as projecting Q directly onto $P_1 \bigcap P_2$. Let $\mathcal{P}_1$ be the set of probability mass functions on $\mathcal{X}$ satisfying

$$\begin{aligned} \sum_x p(x) &= 1, & (11.23) \\ \sum_x p(x) h_i(x) &\geq \alpha_i, \ i = 1, 2, \ldots, r. & (11.24) \end{aligned}$$

Let $\mathcal{P}_2$ be the set of probability mass functions on $\mathcal{X}$ satisfying

$$\begin{aligned} \sum_x p(x) &= 1, & (11.25) \\ \sum_x p(x) g_j(x) &\geq \beta_j, \ j = 1, 2, \ldots, s. & (11.26) \end{aligned}$$

Suppose $Q \notin P_1 \bigcup P_2$. Let P^* minimize $D(P \parallel Q)$ over all $P \in \mathcal{P}_1$. Let R^* minimize $D(R \parallel Q)$ over all $R \in \mathcal{P}_1 \bigcap \mathcal{P}_2$. Argue that R^* minimizes $D(R \parallel P^*)$ over all $R \in P_1 \bigcap P_2$.

Solution: *Sequential Projection.*

$\mathcal{P}_1$ is defined by the constraints $\{h_i\}$ and $\mathcal{P}_2$ by the constraints $\{g_i\}$. Hence $\mathcal{P}_1 \cap \mathcal{P}_2$ is defined by the union of the constraints.

We will assume that all the constraints are active. In this case, from the parametric form of the distribution that minimizes $D(p||q)$ subject to equality constraints as derived in the first homework, we have

$$\begin{aligned} p^*(x) &= \arg\min_{p \in \mathcal{P}_1} D(p||q) & (11.27) \\ &= c_1 q(x) e^{\sum_{i=1}^r \lambda_i h_i(x)}. & (11.28) \end{aligned}$$

$$\begin{aligned} r^*(x) &= \arg\min_{p \in \mathcal{P}_1 \cap \mathcal{P}_2} D(p||q) & (11.29) \\ &= c_2 q(x) e^{\sum_{i=1}^r \lambda_i' h_i(x) + \sum_{j=1}^s \nu_j' g_j(x)}. & (11.30) \end{aligned}$$

where the constants are chosen so as to satisfy the constraints. Now when we project p^* onto $\mathcal{P}_1 \cap \mathcal{P}_2$, we get

$$\begin{aligned} p^{**}(x) &= \arg\min_{p \in \mathcal{P}_1 \cap \mathcal{P}_2} D(p||p^*) & (11.31) \\ &= c_3 p^*(x) e^{\sum \nu_i g_i(x)} & (11.32) \\ &= c_3 c_1 q(x) e^{\sum \nu_i g_i(x) + \sum \lambda_i h_i(x)}, & (11.33) \end{aligned}$$

which is of the same form as r^*. Since the constants are chosen to satisfy the constraints in both cases, we will obtain the same constants, and hence

$$r^*(x) = p^{**}(x) \quad \text{for all } x. \tag{11.34}$$

Hence sequential projection is equivalent to direct projection, and r^* minimizes $D(r||p^*)$ over all $r \in \mathcal{P}_1 \cap \mathcal{P}_2$.

An alternative proof is to use the fact (proved in the first homework) that for any set E determined by constraints of the type in the problem,

$$D(p||p^*) + D(p^*||q) = D(p||q), \quad \text{for all } p \in E. \tag{11.35}$$

where p^* is the distribution in E that is closest to q. Let p^{**} be the projection of p^* on $\mathcal{P}_1 \cap \mathcal{P}_2$. Then for every element of $\mathcal{P}_1 \cap \mathcal{P}_2$,

$$D(p||p^*) + D(p^*||q) = D(p||q). \tag{11.36}$$

Taking the minimum of both sides over $p \in \mathcal{P}_1 \cap \mathcal{P}_2$, we see that the same p must simultaneously minimize both sides, i.e.,

$$p^{**} = r^*. \tag{11.37}$$

5. **Counting.** Let $\mathcal{X} = \{1, 2, \ldots, m\}$. Show that the number of sequences $x^n \in \mathcal{X}^n$ satisfying $\frac{1}{n}\sum_{i=1}^n g(x_i) \geq \alpha$ is approximately equal to 2^{nH^*}, to first order in the exponent, for n sufficiently large, where

$$H^* = \max_{P: \sum_{i=1}^m P(i)g(i) \geq \alpha} H(P). \tag{11.38}$$

Solution: *Counting.* We wish to count the number of sequences satisfying a certain property. Instead of directly counting the sequences, we will calculate the probability of the set under an uniform distribution. Since the uniform distribution puts a probability of $\frac{1}{m^n}$ on every sequence of length n, we can count the sequences by multiplying the probability of the set by m^n.

The probability of the set can be calculated easily from Sanov's theorem. Let Q be the uniform distribution, and let E be the set of sequences of length n satisfying $\frac{1}{n}\sum g(x_i) \geq \alpha$. Then by Sanov's theorem, we have

$$Q^n(E) \doteq 2^{-nD(P^*||Q)}, \tag{11.39}$$

where P^* is the type in E that is closest to Q. Since Q is the uniform distribution, $D(P||Q) = \log m - H(P)$, and therefore P^* is the type in E that has maximum entropy. Therefore, if we let

$$H^* = \max_{P: \sum_{i=1}^m P(i)g(i) \geq \alpha} H(P), \tag{11.40}$$

we have

$$Q^n(E) \doteq 2^{-n(\log m - H^*)}. \tag{11.41}$$

Multiplying this by m^n to find the number of sequences in this set, we obtain

$$|E| \doteq 2^{-n \log m} 2^{nH^*} m^n = 2^{nH^*}. \tag{11.42}$$

6. **Biased estimates may be better.** Consider the problem of estimating μ and σ^2 from n samples of data drawn i.i.d. from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

(a) Show that $\overline{X_n}$ is an unbiased estimator of μ.

(b) Show that the estimator

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^2 \tag{11.43}$$

is biased and the estimator

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2 \tag{11.44}$$

is unbiased.

(c) Show that S_n^2 has a lower mean squared error than S_{n-1}^2. This illustrates the idea that a biased estimator may be "better" than an unbiased estimator for the same parameter.

Solution: *Biased estimates may be better.*

(a) Let $\overline{X_n} = \frac{1}{n} \sum_{i=1}^n X_i$. Then $E\overline{X_n} = \frac{1}{n} \sum EX_i = \mu$. Thus $\overline{X_n}$ is an unbiased estimator of μ.

(b) Before we compute the expected value of S_n^2, we will first compute the variance of $\overline{X_n}$. By the independence of the X_i's, we have

$$\operatorname{var}(\overline{X_n}) = \frac{1}{n^2} \sum_i \operatorname{var}(X_i) = \frac{\sigma^2}{n}. \tag{11.45}$$

Also, we will need to calculate

$$\begin{aligned} E(X_i - \mu)(\overline{X_n} - \mu) &= E(X_i - \mu) \left(\frac{1}{n} \sum_j (X_j - \mu) \right) && (11.46) \\ &= \frac{1}{n} E(X_i - \mu)^2 + \frac{1}{n} \sum_{j \neq i} E(X_i - \mu)(X_j - \mu) && (11.47) \\ &= \frac{\sigma^2}{n}, && (11.48) \end{aligned}$$

since by independence, X_i and X_j are uncorrelated and therefore $E(X_i - \mu)(X_j - \mu) = 0$.

Therefore, if we let

$$W = \sum_i (X_i - \overline{X_n})^2 = \sum_i \Big((X_i - \mu) - (\overline{X_n} - \mu) \Big)^2 , \tag{11.49}$$

we have

$$\begin{aligned} EW &= \sum_i E(X_i - \mu)^2 - 2 \sum_i E(X_i - \mu)(\overline{X_n} - \mu) + nE(\overline{X_n} - \mu)^2 && (11.50) \\ &= n\sigma^2 - 2n\frac{\sigma^2}{n} + n\frac{\sigma^2}{n} && (11.51) \\ &= (n-1)\sigma^2 && (11.52) \end{aligned}$$

Thus,

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \frac{W}{n} \tag{11.53}$$

has $ES_n^2 = \frac{n-1}{n}\sigma^2$, and it is therefore a biased estimator of σ^2, and

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \frac{W}{n-1} \tag{11.54}$$

has expected value σ^2 and is therefore an unbiased estimator of σ^2.

(c) This involves a lot of algebra. We will need the following properties of the Normal distribution - the third central moment is 0 and the fourth central moment is $3\sigma^4$, and therefore

$$\begin{aligned} EX_i &= \mu && (11.55) \\ EX_i^2 &= \mu^2 + \sigma^2 && (11.56) \\ E(X_i - \mu)^3 &= 0 && (11.57) \\ EX_i^3 &= \mu^3 + 3\sigma^2\mu && (11.58) \\ E(X_i - \mu)^4 &= 3\sigma^4 && (11.59) \\ EX_i^4 &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4. && (11.60) \end{aligned}$$

We also know that $T = \overline{X_n} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, and we have the corresponding results for T:

$$\begin{aligned} ET &= \mu && (11.61) \\ ET^2 &= \mu^2 + \frac{\sigma^2}{n} && (11.62) \\ E(T - \mu)^3 &= 0 && (11.63) \\ ET^3 &= \mu^3 + 3\frac{\sigma^2}{n}\mu && (11.64) \\ E(T - \mu)^4 &= 3\frac{\sigma^4}{n^2} && (11.65) \\ ET^4 &= \mu^4 + 6\mu^2\frac{\sigma^2}{n} + 3\frac{\sigma^4}{n^2}. && (11.66) \end{aligned}$$

Also,

$$\begin{aligned}
EX_iT &= \frac{1}{n}EX_i^2 + \frac{1}{n}\sum_{j\neq i} EX_iX_j = \mu^2 + \frac{\sigma^2}{n} && (11.67)\\
EX_i^2T^2 &= E\frac{1}{n^2}X_i^2\left(\sum_j X_j^2 + 2\sum_{j,k:j<k} X_jX_k\right) && (11.68)\\
&= E\frac{1}{n^2}\Bigg(X_i^4 + \sum_{j\neq i} X_i^2X_j^2 + 2\sum_{j,k:j\neq i,k\neq i,j<k} X_i^2X_jX_k \\
&\qquad +2\sum_{k:k\neq i} X_i^3X_k\Bigg) && (11.69)\\
&= \frac{1}{n^2}\Big(\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 + (n-1)(\mu^2+\sigma^2)^2 \\
&\qquad +2\frac{(n-1)(n-2)}{2}(\mu^2+\sigma^2)\mu^2 + 2(n-1)\mu(\mu^3+3\sigma^2\mu)\Big)\\
&= \frac{n^2\mu^4 + (n^2+5n)\mu^2\sigma^2 + (n+2)\sigma^4}{n^2} && (11.70)
\end{aligned}$$

Now we are in a position to calculate EW^2. We first rewrite

$$\begin{aligned}
W &= \sum_i (X_i - T)^2 && (11.71)\\
&= \sum_i X_i^2 - 2\sum_i X_iT + nT^2 && (11.72)\\
&= \sum_i X_i^2 - 2nTT + nT^2 && (11.73)\\
&= \sum_i X_i^2 - nT^2. && (11.74)
\end{aligned}$$

Thus

$$\begin{aligned}
W^2 &= \left(\sum_i X_i^2 - nT^2\right)^2 && (11.75)\\
&= \sum_i X_i^4 + 2\sum_{i<j} X_i^2X_j^2 + n^2T^4 - 2n\sum_i X_i^2T^2 && (11.76)
\end{aligned}$$

and therefore

$$\begin{aligned}
EW^2 &= n(\mu^4 + 6\sigma^2\mu^2 + 3\sigma^4) + 2\frac{n(n-1)}{2}(\mu^2+\sigma^2)^2 + n^2(\mu^4 + 6\frac{\sigma^2}{n}\mu^2 + 3\frac{\sigma^4}{n^2})\\
&\qquad -2n^2\frac{1}{n^2}(n^2\mu^4 + (n^2+5n)\mu^2\sigma^2 + (n+2)\sigma^4) && (11.77)\\
&= (n^2-1)\sigma^4. && (11.78)
\end{aligned}$$

Now we can calculate the mean squared error of the two estimators. The error of $S_{n-1} = \frac{W}{n-1}$ is

$$\begin{aligned} E(S_{n-1} - \sigma^2)^2 &= E\left(\frac{W^2}{(n-1)^2} - 2\frac{W}{n-1}\sigma^2 + \sigma^4\right) && (11.79) \\ &= \frac{(n^2-1)\sigma^4}{(n-1)^2} - 2\frac{(n-1)\sigma^2}{n-1}\sigma^2 + \sigma^4 && (11.80) \\ &= \frac{2}{n-1}\sigma^4. && (11.81) \end{aligned}$$

The error of $S_{n-1} = \frac{W}{n-1}$ is

$$\begin{aligned} E(S_n - \sigma^2)^2 &= E\left(\frac{W^2}{n^2} - 2\frac{W}{n}\sigma^2 + \sigma^4\right) && (11.82) \\ &= \frac{(n^2-1)\sigma^4}{n^2} - 2\frac{(n-1)\sigma^2}{n}\sigma^2 + \sigma^4 && (11.83) \\ &= \frac{2n-1}{n^2}\sigma^4. && (11.84) \end{aligned}$$

Since $\frac{2n-1}{n^2}$ is less than $\frac{2}{n-1}$ for all positive n, we see that S_n has a lower expected error than S_{n-1}.
In fact, if we let the estimator of σ^2 be cW, then we can easily calculate the expected error of the estimator to be

$$E(cW - \sigma^2)^2 = \sigma^4\left((n^2-1)c^2 - 2(n-1)c + 1\right), \tag{11.85}$$

which is minimized for $c = \frac{1}{n+1}$. Thus neither best unbiased estimator ($c = \frac{1}{n-1}$) or the maximum likelihood estimator ($c = \frac{1}{n}$) produces the minimum mean squared error.

7. **Fisher information and relative entropy**. Show for a parametric family $\{p_\theta(x)\}$ that

$$\lim_{\theta' \to \theta} \frac{1}{(\theta - \theta')^2} D(p_\theta || p_{\theta'}) = \frac{1}{\ln 4} J(\theta). \tag{11.86}$$

Solution: *Fisher information and relative entropy.* Let $t = \theta' - \theta$. Then

$$\frac{1}{(\theta - \theta')^2} D(p_\theta || p_{\theta'}) = \frac{1}{t^2} D(p_\theta || p_{\theta+t}) = \frac{1}{t^2 \ln 2} \sum_x p_\theta(x) \ln \frac{p_\theta(x)}{p_{\theta+t}(x)}. \tag{11.87}$$

Let

$$f(t) = p_\theta(x) \ln \frac{p_\theta(x)}{p_{\theta+t}(x)}. \tag{11.88}$$

We will suppress the dependence on x and expand $f(t)$ in a Taylor series in t. Thus

$$f'(t) = -\frac{p_\theta}{p_{\theta+t}} \frac{dp_{\theta+t}}{dt}, \tag{11.89}$$

and

$$f''(t) = \frac{p_\theta}{p_{\theta+t}^2}\left(\frac{dp_{\theta+t}}{dt}\right)^2 + \frac{p_\theta}{p_{\theta+t}}\frac{d^2 p_{\theta+t}}{dt^2}. \tag{11.90}$$

Thus expanding in the Taylor series around $t = 0$, we obtain

$$f(t) = f(0) + f'(0)t + f''(0)\frac{t^2}{2} + O(t^3), \tag{11.91}$$

where $f(0) = 0$,

$$f'(0) = -\frac{p_\theta}{p_\theta}\left.\frac{dp_{\theta+t}}{dt}\right|_{t=0} = \frac{dp_\theta}{d\theta} \tag{11.92}$$

and

$$f''(0) = \frac{1}{p_\theta}\left(\frac{dp_\theta}{d\theta}\right)^2 + \frac{d^2 p_\theta}{d\theta^2} \tag{11.93}$$

Now $\sum_x p_\theta(x) = 1$, and therefore

$$\sum_x \frac{dp_\theta(x)}{d\theta} = \frac{d}{dt}1 = 0, \tag{11.94}$$

and

$$\sum_x \frac{d^2 p_\theta(x)}{d\theta^2} = \frac{d}{d\theta}0 = 0. \tag{11.95}$$

Therefore the sum of the terms of (11.92) sum to 0 and the sum of the second terms in (11.93) is 0.

Thus substituting the Taylor expansions in the sum, we obtain

$$\frac{1}{(\theta-\theta')^2}D(p_\theta||p_{\theta'}) = \frac{1}{t^2\ln 2}\sum_x p_\theta(x)\ln\frac{p_\theta(x)}{p_{\theta+t}(x)} \tag{11.96}$$

$$= \frac{1}{t^2\ln 2}\left(0 + \sum_x \frac{dp_\theta(x)}{d\theta}t + \sum_x\left(\frac{1}{p_\theta}\left(\frac{dp_\theta}{d\theta}\right)^2 + \frac{d^2p_\theta}{d\theta^2}\right)\frac{t^2}{2} + O(t^3)\right) \tag{11.97}$$

$$= \frac{1}{2\ln 2}\sum_x \frac{1}{p_\theta(x)}\left(\frac{dp_{\theta(x)}}{d\theta}\right)^2 + O(t) \tag{11.98}$$

$$= \frac{1}{\ln 4}J(\theta) + O(t) \tag{11.99}$$

and therefore

$$\lim_{\theta'\to\theta}\frac{1}{(\theta-\theta')^2}D(p_\theta||p_{\theta'}) = \frac{1}{\ln 4}J(\theta). \tag{11.100}$$

8. **Examples of Fisher information.** The Fisher information $J(\Theta)$ for the family $f_\theta(x), \theta \in \mathbf{R}$ is defined by

$$J(\theta) = E_\theta\left(\frac{\partial f_\theta(X)/\partial\theta}{f_\theta(X)}\right)^2 = \int\frac{(f'_\theta)^2}{f_\theta}$$

Find the Fisher information for the following families:

(a) $f_\theta(x) = N(0,\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}$

(b) $f_\theta(x) = \theta e^{-\theta x}, x \ge 0$

(c) What is the Cramèr Rao lower bound on $E_\theta(\hat{\theta}(X) - \theta)^2$, where $\hat{\theta}(X)$ is an unbiased estimator of θ for (a) and (b)?

Solution: *Examples of Fisher information.*

(a) $f_\theta(x) = N(0,\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}$, and therefore

$$f'_\theta = -\frac{1}{2}\frac{1}{\sqrt{2\pi\theta^3}} e^{-\frac{x^2}{2\theta}} + \frac{x^2}{2\theta^2}\frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}, \tag{11.101}$$

and

$$\frac{f'_\theta}{f_\theta} = \left(-\frac{1}{2\theta} + \frac{x^2}{2\theta^2}\right). \tag{11.102}$$

Therefore the Fisher information,

$$\begin{aligned} J(th) &= E_\theta \left(\frac{f'_t h}{f_t h}\right)^2 && (11.103) \\ &= E_\theta \left(\frac{1}{4\theta^2} - 2\frac{1}{2\theta}\frac{x^2}{2\theta^2} + \frac{x^4}{4\theta^4}\right) && (11.104) \\ &= \frac{1}{4\theta^2} - 2\frac{1}{\theta}\frac{\theta}{2\theta^2} + \frac{3\theta^2}{4\theta^4} && (11.105) \\ &= \frac{1}{2\theta^2}, && (11.106) \end{aligned}$$

using the "well-known" or easily verified fact that for a normal $\mathcal{N}(0,\theta)$ distribution, the fourth moment is $3\theta^2$.

(b) $f_\theta(x) = \theta e^{-\theta x}, x \ge 0$, and therefore $\ln f_\theta = \ln\theta - \theta x$, and

$$\frac{d \ln f_\theta}{d\theta} = \frac{1}{\theta} - x, \tag{11.107}$$

and therefore

$$\begin{aligned} J(\theta) &= E_t h \left(\frac{d \ln f_\theta}{d\theta}\right)^2 && (11.108) \\ &= E_\theta \left(\frac{1}{\theta^2} - 2\frac{1}{\theta}x + x^2\right) && (11.109) \\ &= \frac{1}{\theta^2} - 2\frac{1}{\theta}\frac{1}{\theta} + \frac{1}{\theta} + \frac{1}{\theta^2} && (11.110) \\ &= \frac{1}{\theta} && (11.111) \end{aligned}$$

using the fact that for an exponential distribution, $EX = \frac{1}{\theta}$, and $EX^2 = \frac{1}{\theta} + \frac{1}{\theta^2}$.

(c) The Cramer-Rao lower bound is the reciprocal of the Fisher information, and is therefore $2\theta^2$ and θ for parts (a) and (b) respectively.

9. **Two conditionally independent looks double the Fisher information.** Let $g_\theta(x_1, x_2) = f_\theta(x_1) f_\theta(x_2)$. Show $J_g(\theta) = 2J_f(\theta)$.

Solution: *Two conditionally independent looks double the Fisher information.* We can simply use the same arguments as in Section 12.11 in the text. We define the score function $V(X_i) = \frac{\partial}{\partial\theta} \ln f_\theta(x_i)$. Then the score functions are independent mean zero random variables and since the Fisher information of g is the variance of the sum of the score functions, it is the sum of the individual variances. Thus the Fisher information of g is twice the Fisher information of f.

10. **Joint distributions and product distributions.** Consider a joint distribution $Q(x, y)$ with marginals $Q(x)$ and $Q(y)$. Let E be the set of types that look jointly typical with respect to Q, i.e.,

$$\begin{aligned} E = \{P(x,y) \quad : \quad & -\sum_{x,y} P(x,y) \log Q(x) - H(X) = 0, \\ & -\sum_{x,y} P(x,y) \log Q(y) - H(Y) = 0, \\ & -\sum_{x,y} P(x,y) \log Q(x,y) - H(X,Y) = 0\}. \end{aligned} \tag{11.112}$$

(a) Let $Q_0(x, y)$ be another distribution on $\mathcal{X} \times \mathcal{Y}$. Argue that the distribution P^* in E that is closest to Q_0 is of the form

$$P^*(x,y) = Q_0(x,y) e^{\lambda_0 + \lambda_1 \log Q(x) + \lambda_2 \log Q(y) + \lambda_3 \log Q(x,y)}, \tag{11.113}$$

where λ_0, λ_1, λ_2 and λ_3 are chosen to satisfy the constraints. Argue that this distribution is unique.

(b) Now let $Q_0(x, y) = Q(x)Q(y)$. Verify that $Q(x, y)$ is of the form (11.113) and satisfies the constraints. Thus $P^*(x, y) = Q(x, y)$, i.e., the distribution in E closest to the product distribution is the joint distribution.

Solution: *Joint distributions and product distributions.*

(a) This result follows directly from Problem 2 in Chapter 11. We will not repeat the arguments.

(b) If we let $\lambda_0 = 0$, $\lambda_i = -1$, $\lambda_2 = -1$, and $\lambda_3 = 1$, then

$$\begin{aligned} P^*(x,y) &= Q_0(x,y) e^{\lambda_0 + \lambda_1 \log Q(x) + \lambda_2 \log Q(y) + \lambda_3 \log Q(x,y)} && (11.114) \\ &= Q(x)Q(y) \frac{1}{Q(x)} \frac{1}{Q(y)} Q(x,y) && (11.115) \\ &= Q(x,y) && (11.116) \end{aligned}$$

and therefore $Q(x, y)$ is of the form that minimizes the relative entropy. It is easy to verify that $Q(x, y)$ trivially satisfies the constraints involved in the definition of the set E. Therefore the joint distribution is the distribution that looks jointly typical and is closest to the product distribution.

11. **Cramer-Rao inequality with a bias term.** Let $X \sim f(x;\theta)$ and let $T(X)$ be an estimator for θ. Let $b_T(\theta) = E_\theta T - \theta$ be the bias of the estimator. Show that

$$E(T-\theta)^2 \geq \frac{[1+b_T'(\theta)]^2}{J(\theta)} + b_T^2(\theta). \tag{11.117}$$

Solution: *Cramer-Rao inequality with a bias term.* The proof parallels the proof without the bias term (Theorem 12.11.1). We will begin with the calculation of $E_\theta(VT)$, where V is the score function and T is the estimator.

$$\begin{aligned}
E(VT) &= \int \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} T(x) f(x;\theta)\, dx & (11.118)\\
&= \int \frac{\partial}{\partial\theta} f(x;\theta) T(x)\, dx & (11.119)\\
&= \frac{\partial}{\partial\theta} \int f(x;\theta) T(x)\, dx & (11.120)\\
&= \frac{\partial}{\partial\theta} E_\theta T & (11.121)\\
&= \frac{\partial}{\partial\theta} (b_T(\theta)+\theta) & (11.122)\\
&= b_T'(\theta) + 1 & (11.123)
\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$(E\left[(V-EV)(T-ET)\right])^2 \leq E(V-EV)^2 E(T-ET)^2. \tag{11.124}$$

Also, $EV=0$ and therefore $E(V-EV)(T-ET) = E(VT)$. Also, by definition, $\text{var}(V) = J(\theta)$. Thus we have

$$\left(b_T'(\theta)+1\right)^2 \leq J(\theta) E(T-\theta-b_T(\theta))^2. \tag{11.125}$$

Now

$$\begin{aligned}
E(T-\theta-b_T(\theta))^2 &= E(T-\theta)^2 + b_T^2(\theta) - 2E(T-\theta)b_T(\theta) & (11.126)\\
&= E(T-\theta)^2 + b_T^2(\theta) - 2b_T^2(\theta) & (11.127)\\
&= E(T-\theta)^2 - b_T^2(\theta). & (11.128)
\end{aligned}$$

Substituting this in the Cauchy Schwarz inequality, we have the desired result

$$E(T-\theta)^2 \geq \frac{[1+b_T'(\theta)]^2}{J(\theta)} + b_T^2(\theta). \tag{11.129}$$

12. **Hypothesis testing.** Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. Consider the hypothesis test $H_1 : p = p_1$ versus $H_2 : p = p_2$. Let

$$p_1(x) = \begin{cases} \frac{1}{2}, & x = -1 \\ \frac{1}{4}, & x = 0 \\ \frac{1}{4}, & x = 1 \end{cases}$$

and

$$p_2(x) = \begin{cases} \frac{1}{4}, & x = -1 \\ \frac{1}{4}, & x = 0 \\ \frac{1}{2}, & x = 1 \ . \end{cases}$$

(a) Find the error exponent for Pr{ Decide $H_2|H_1$ true } in the best hypothesis test of H_1 vs. H_2 subject to Pr{ Decide $H_1|H_2$ true } $\leq \frac{1}{2}$.

Solution: *Hypothesis testing*

By the Chernoff-Stein lemma, the error exponent in this hypothesis test is the exponent for probability of the acceptance region for H_2 given P_1, which is

$$D(P_2||P_1) = \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{1}{2}} + \frac{1}{4}\log\frac{\frac{1}{4}}{\frac{1}{4}} + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{4}} = 0.25 \tag{11.130}$$

Thus the probability of error will go to 0 as $2^{-\frac{n}{4}}$.

13. **Sanov's theorem:** Prove the simple version of Sanov's theorem for the binary random variables, i.e., let $X_1, X_2, \ldots, X_n$ be a sequence of binary random variables, drawn i.i.d. according to the distribution:

$$\Pr(X = 1) = q, \quad \Pr(X = 0) = 1 - q. \tag{11.131}$$

Let the proportion of 1's in the sequence $X_1, X_2, \ldots, X_n$ be $p_{\mathbf{X}}$, i.e.,

$$p_{X^n} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{11.132}$$

By the law of large numbers, we would expect $p_{\mathbf{X}}$ to be close to q for large n. Sanov's theorem deals with the probability that p_{X^n} is far away from q. In particular, for concreteness, if we take $p > q > \frac{1}{2}$, Sanov's theorem states that

$$-\frac{1}{n}\log\Pr\left\{(X_1, X_2, \ldots, X_n) : p_{X^n} \geq p\right\} \to p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} = D((p, 1-p)||(q, 1-q)). \tag{11.133}$$

Justify the following steps:

- $$\Pr\left\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\right\} \leq \sum_{i=\lfloor np \rfloor}^{n} \binom{n}{i} q^i (1-q)^{n-i} \tag{11.134}$$

- Argue that the term corresponding to $i = \lfloor np \rfloor$ is the largest term in the sum on the right hand side of the last equation.
- Show that this term is approximately 2^{-nD}.
- Prove an upper bound on the probability in Sanov's theorem using the above steps. Use similar arguments to prove a lower bound and complete the proof of Sanov's theorem.

Solution: *Sanov's theorem*

- Since $n\overline{X}_n$ has a binomial distribution, we have

$$\Pr(n\overline{X}_n = i) = \binom{n}{i} q^i (1-q)^{n-i} \tag{11.135}$$

  and therefore

$$\Pr\left\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\right\} \leq \sum_{i=\lfloor np \rfloor}^{n} \binom{n}{i} q^i (1-q)^{n-i} \tag{11.136}$$

- $$\frac{\Pr(n\overline{X}_n = i+1)}{\Pr(n\overline{X}_n = i)} = \frac{\binom{n}{i+1} q^{i+1}(1-q)^{n-i-1}}{\binom{n}{i} q^i (1-q)^{n-i}} = \frac{n-i}{i+1}\frac{q}{1-q} \tag{11.137}$$

  This ratio is less than 1 if $\frac{n-i}{i+1} < \frac{1-q}{q}$,i.e., if $i > nq - (1-q)$. Thus the maximum of the terms occurs when $i = \lfloor np \rfloor$.
- From Example 11.1.3,

$$\binom{n}{\lfloor np \rfloor} \doteq 2^{nH(p)} \tag{11.138}$$

  and hence the largest term in the sum is

$$\binom{n}{\lfloor np \rfloor} q^{\lfloor np \rfloor}(1-q)^{n-\lfloor np \rfloor} = 2^{n(-p\log p-(1-p)\log(1-p))+np\log q+n(1-p)\log(1-q)} = 2^{-nD(p||q)} \tag{11.139}$$

- From the above results, it follows that

$$\begin{aligned}
\Pr\left\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\right\} &\leq \sum_{i=\lfloor np \rfloor}^{n} \binom{n}{i} q^i (1-q)^{n-i} &(11.140)\\
&\leq (n - \lfloor np \rfloor)\binom{n}{\lfloor np \rfloor} q^i (1-q)^{n-i} &(11.141)\\
&\leq (n(1-p)+1) 2^{-nD(p||q)} &(11.142)
\end{aligned}$$

  where the second inequality follows from the fact that the sum is less than the largest term times the number of terms. Taking the logarithm and dividing by n and taking the limit as $n \to \infty$, we obtain

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\left\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\right\} \leq -D(p||q) \tag{11.143}$$

Similarly, using the fact the sum of the terms is larger than the largest term, we obtain

$$\begin{aligned} \Pr\left\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\right\} &\geq \sum_{i=\lceil np \rceil}^{n} \binom{n}{i} q^i (1-q)^{n-i} && (11.144) \\ &\geq \binom{n}{\lceil np \rceil} q^i (1-q)^{n-i} && (11.145) \\ &\geq 2^{-nD(p||q)} && (11.146) \end{aligned}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\left\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\right\} \geq -D(p||q) \qquad (11.147)$$

Combining these two results, we obtain the special case of Sanov's theorem

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\left\{(X_1, X_2, \ldots, X_n) : p_{\mathbf{X}} \geq p\right\} = -D(p||q) \qquad (11.148)$$

14. **Sanov**. Let X_i be i.i.d. $\sim N(0, \sigma^2)$.

   (a) Find the exponent in the behavior of $\Pr\{\frac{1}{n}\sum_{i=1}^n X_i^2 \geq \alpha^2\}$. This can be done from first principles (since the normal distribution is nice) or by using Sanov's theorem.

   (b) What does the data look like if $\frac{1}{n}\sum_{i=1}^n X_i^2 \geq \alpha$? That is, what is the P^* that minimizes $D(P \parallel Q)$?

Solution: *Sanov*

(a) From the properties of the normal distribution, we know that $\sum X_i^2$ has a χ^2 distribution with n degrees of freedom, and we can directly calculate

$$\begin{aligned} \Pr\left(\frac{1}{n}\sum X_i^2 \geq \alpha^2\right) &= \Pr\left(\chi_n^2 \geq n\alpha^2\right) && (11.149) \\ &= \frac{\Gamma(\frac{n}{2}, \frac{n\alpha^2}{2})}{\Gamma(\frac{n}{2})} && (11.150) \end{aligned}$$

However, using Sanov's theorem, we know that the probability of the set

$$-\frac{1}{n} \log \Pr\left(\frac{1}{n}\sum X_i^2 \geq \alpha^2\right) = D(P^*||Q), \qquad (11.151)$$

where P^* is the distribution that satisfies the constraint that is closest to Q. In this case,

$$\begin{aligned} D(P||Q) &= \int f(x) \ln \frac{f(x)}{\frac{1}{\sqrt{2\pi e\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}} && (11.152) \\ &= -H(f) + \int f(x) \ln \sqrt{2\pi e\sigma^2} + \int f(x) \frac{x^2}{2\sigma^2} && (11.153) \\ &= -H(f) + \ln \sqrt{2\pi e\sigma^2} + \frac{E[X^2]}{2\sigma^2} && (11.154) \end{aligned}$$

and hence the distribution that minimizes the relative entropy is the distribution that maximizes the entropy subject to the expected square constraint. Using the results of the maximum entropy chapter (Chapter 12), we can see the the maximum entropy distribution is of the form $f(x) \sim Ce^{-\beta x^2}$, i.e., the maximum entropy distribution is a normal distribution. Thus the P^* that minimizes relative entropy subject to the constraint $\frac{1}{n}\sum X_i^2 \geq \alpha^2$ is the $\mathcal{N}(0,\alpha^2)$ distribution. Substituting this distribution back into the expression for relative entropy, we obtain

$$\begin{aligned} D(P^*||Q) &= -H(\mathcal{N}(0,\alpha^2)) - \ln\sqrt{2\pi e\sigma^2} + \frac{E[X^2]}{2\sigma^2} && (11.155)\\ &= \frac{1}{2}\log(2\pi e\alpha^2) - \frac{1}{2}\log(2\pi e\sigma^2) + \frac{\alpha^2}{2\sigma^2} && (11.156)\\ &= \frac{1}{2}\log\frac{\alpha^2}{\sigma^2} + \frac{1}{2}\frac{\alpha^2}{\sigma^2} && (11.157)\end{aligned}$$

(b) From the above calculation and the conditional limit theorem, the distribution of the data conditional on the constraint is P^*, which is $\mathcal{N}(0,\alpha^2)$.

15. **Counting states.**
Suppose an atom is equally likely to be in each of 6 states, $X \in \{s_1, s_2, s_3, \ldots, s_6\}$. One observes n atoms $X_1, X_2, \ldots, X_n$ independently drawn according to this uniform distribution. It is observed that the frequency of occurrence of state s_1 is twice the frequency of occurrence of state s_2.

(a) To first order in the exponent, what is the probability of observing this event?

(b) Assuming n large, find the conditional distribution of the state of the first atom X_1, given this observation.

Solution: *Counting states*

(a) Using Sanov's theorem, we need to determine the distribution P^* that is closest to the uniform with $p_1 = 2p_2$, which is the empirical constraint. We need to minimize

$$D(P||Q) = \sum p_i \log 6p_i \tag{11.158}$$

subject to the constraints $\sum p_i = 1$ and $p_1 - 2p_2 = 0$. Setting up the functional

$$J(P) = \sum p_i \log 6p_i + \lambda_1 \sum p_i + \lambda_2(p_1 - 2p_2) \tag{11.159}$$

Differentiating with respect to p_i and setting to 0, we obtain

$$\begin{aligned} \log 6p_1 + 1 + \lambda_1 + \lambda_2 &= 0 && (11.160)\\ \log 6p_2 + 1 + \lambda_1 - 2\lambda_2 &= 0 && (11.161)\\ \log 6p_i + 1 + \lambda_1 &= 0, \quad i = 3,4,5,6 && (11.162)\end{aligned}$$

Thus

$$\begin{aligned} p_1 &= c_1 2^{-\lambda_2} && (11.163)\\ p_2 &= c_1 2^{2\lambda_2} && (11.164)\\ p_i &= c_1, \quad i = 3,4,5,6 && (11.165)\end{aligned}$$

where $c_1 = \frac{1}{6}2^{-(1+\lambda_1)}$. Since $p_1 = 2p_2$, we obtain $\lambda_2 = -\frac{1}{3}\log 2$. c_1 should be chosen so that $\sum p_i = 1$, which in turn implies that $c_1 = 1/(2^{\frac{1}{3}} + 2^{-\frac{2}{3}} + 4) = 1/5.889$, and the corresponding distribution is (0.213,0.107,0.17,0.17,0.17,0.17), and the relative entropy distance is 0.0175. Thus the first order probability that this event happens is $2^{-0.0175n}$.

(b) When the event ($p_1 = 2p_2$) happens, the conditional distribution is close to $P^* = (0.213, 0.107, 0.17, 0.17, 0.17, 0.17)$.

16. **Hypothesis testing**

Let $\{X_i\}$ be i.i.d. $\sim p(x)$, $x \in \{1, 2, \ldots\}$. Consider two hypotheses $H_0 : p(x) = p_0(x)$ vs. $H_1 : p(x) = p_1(x)$, where $p_0(x) = \left(\frac{1}{2}\right)^x$, and $p_1(x) = qp^{x-1}$, $x = 1, 2, 3, \ldots$

(a) Find $D(p_0 \parallel p_1)$.

(b) Let $\Pr\{H_0\} = \frac{1}{2}$. Find the minimal probability of error test for H_0 vs. H_1 given data $X_1, X_2, \ldots, X_n \sim p(x)$.

Solution: *Hypothesis testing*

(a)

$$\begin{aligned} D(p_0||p_1) &= \sum_x p_0(x) \log \frac{p_0(x)}{p_1(x)} && (11.166)\\ &= \sum_x \left(\frac{1}{2}\right)^x \log \frac{\left(\frac{1}{2}\right)^x}{qp^{x-1}} && (11.167)\\ &= \sum_x \left(\frac{1}{2}\right)^x \log \left(\left(\frac{1}{2p}\right)^x \frac{p}{q} \right) && (11.168)\\ &= \sum_x \left(\frac{1}{2}\right)^x x \log \left(\frac{1}{2p}\right) + \sum_x \left(\frac{1}{2}\right)^x \log \frac{p}{q} && (11.169)\\ &= \log \left(\frac{1}{2p}\right) \sum_x \left(\frac{1}{2}\right)^x x + \log \frac{p}{q} \sum_x \left(\frac{1}{2}\right)^x && (11.170)\\ &= 2 \log \left(\frac{1}{2p}\right) + \log \frac{p}{q} && (11.171)\\ &= -\log(4pq) && (11.172)\end{aligned}$$

(b) In the Bayesian setting, the minimum probability of error exponent is given by the Chernoff information

$$C(P_1,P_2) \stackrel{\triangle}{=} -\min_{0\le\lambda\le1} \log\left(\sum_x P_1^\lambda(x)P_2^{1-\lambda}(x)\right). \tag{11.173}$$

Now

$$\begin{aligned}\sum_x P_1^\lambda(x)P_2^{1-\lambda}(x) &= \sum_x \left(\frac{1}{2}\right)^{x\lambda}\left(\frac{q}{p}\right)^{1-\lambda} p^{x(1-\lambda)} &(11.174)\\ &= \left(\frac{q}{p}\right)^{1-\lambda}\sum_x\left(\frac{p^{1-\lambda}}{2^\lambda}\right)^x &(11.175)\\ &= \left(\frac{q}{p}\right)^{1-\lambda}\frac{\frac{p^{1-\lambda}}{2^\lambda}}{1-\frac{p^{1-\lambda}}{2^\lambda}} &(11.176)\\ &= \frac{q^{1-\lambda}p^\lambda}{2^\lambda p^\lambda - p} &(11.177)\end{aligned}$$

To find the mimimum of this over λ, we differentiate the logarithm of this with respect to λ, and obtain

$$-\log q + \log p - \frac{1}{(2p)^\lambda - p}(2p)^\lambda \log 2p = 0 \tag{11.178}$$

Solving for λ from this equation and substituting this into the definition of Chernoff information will provide us the answer.

17. **Maximum likelihood estimation.** Let $\{f_\theta(x)\}$ denote a parametric family of densities with parameter $\theta\epsilon\mathcal{R}$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim f_\theta(x)$. The function

$$l_\theta(x^n) = \ln\left(\prod_{i=1}^n f_\theta(x_i)\right)$$

is known as the log likelihood function. Let θ_0 denote the true parameter value.

(a) Let the expected log likelihood be

$$E_{\theta_0} l_\theta(X^n) = \int (\ln \prod_{i=1}^n f_\theta(x_i)) \prod_{i=1}^n f_{\theta_0}(x_i) dx^n \quad ,$$

and show that

$$E_{\theta_0}(l(X^n)) = (-h(f_{\theta_0}) - D(f_{\theta_0}||f_\theta))n \quad .$$

(b) Show that the maximum over θ of the expected log likelihood is achieved by $\theta = \theta_0$.

Solution: *Maximum likelihood*

This problem is the continuous time analog of the cost of miscoding.

(a) Let us denote

$$f_\theta(x^n) = \left(\prod_{i=1}^n f_\theta(x_i)\right) \tag{11.179}$$

Then if

$$\begin{aligned} E_{\theta_0} l_\theta(X^n) &= \int (\log \prod_{i=1}^n f_\theta(x_i)) \prod_{i=1}^n f_{\theta_0}(x_i) dx^n \quad , & (11.180) \\ &- \int f_{\theta_0}(x^n) \log f_\theta(x^n)\, dx^n & (11.181) \\ &= \int \ln f_{\theta_0}(x^n) \log f_{\theta_0}(x^n)\, dx^n + \int f_{\theta_0}(x^n) \log \frac{f_\theta(x^n)}{f_{\theta_0}(x^n)}\, dx^n & (11.182) \\ &= -h(f_{\theta_0}(x^n)) - D(f_{\theta_0}(x^n)||f_\theta(x^n)) & (11.183) \\ &= -n(h(f_{\theta_0}(x)) - D(f_{\theta_0}(x)||f_\theta(x))) & (11.184) \end{aligned}$$

(b) From the non-negativity of relative entropy, it follows from the last equation that the maximum value of the likelihood occurs when

$$D(f_{\theta_0}(x)||f_\theta(x)) = 0 \tag{11.185}$$

or $\theta = \theta_0$.

18. **Large deviations.** Let $X_1, X_2, \ldots$ be i.i.d. random variables drawn according to the geometric distribution

$$Pr\{X = k\} = p^{k-1}(1-p), \quad k = 1, 2, \ldots \quad .$$

Find good estimates (to first order in the exponent) of

(a) $Pr\{\frac{1}{n}\sum_{i=1}^n X_i \geq \alpha\}$.

(b) $Pr\{X_1 = k | \frac{1}{n}\sum_{i=1}^n X_i \geq \alpha\}$.

(c) Evaluate a) and b) for $p = \frac{1}{2}, \alpha = 4$.

Solution: *Large deviations*

By Sanov's theorem, the probability is determined by the relative entropy distance to the closest distribution that satisfies the constraint. Let that distribution be $r_1, r_2, \ldots,$ on the integers 1,2,.... Then the relative entropy distance to the geometric distribution is

$$D(r||p) = \sum r_i \log \frac{r_i}{p^{i-1}(1-p)} \tag{11.186}$$

We need to minimize this subject to the constraints, $\sum r_i = 1$, and $\sum i r_i = \alpha$. We have assumed that the constraint is matched with equality without loss of generality. We set up the functional

$$J(r) = \sum r_i \log \frac{r_i}{p^{i-1}(1-p)} + \lambda_1 \sum r_i + \lambda_2 \sum i r_i \tag{11.187}$$

Differentiating with respect to r_i and setting to 0, we obtain

$$\log r_i - \log(p^{i-1}(1-p)) + \lambda_1 + \lambda_2 i = 0 \tag{11.188}$$

or

$$r_i = p^{i-1}(1-p)c_1 c_2^i \tag{11.189}$$

From the form of the equation for r_i, it is clear that r too is a geometric distribution. Since we need to satisfy the constraint $\sum i r_i = \alpha$, it follows that r_i is a geometric distribution with parameter $1 - \frac{1}{\alpha}$. Therefore

$$r_i = \left(1 - \frac{1}{\alpha}\right)^{i-1} \frac{1}{\alpha} \tag{11.190}$$

and

$$\begin{aligned}
D(r||p) &= \sum_i r_i \log \frac{r_i}{p^{i-1}(1-p)} && (11.191)\\
&= \sum_i r_i \log \frac{p}{1-p} \frac{1-\frac{1}{\alpha}}{\frac{1}{\alpha}} \left(\frac{1}{\alpha p}\right)^i && (11.192)\\
&= \log \frac{p}{(1-p)(\alpha-1)} + \alpha \log \frac{\alpha - 1}{\alpha p} && (11.193)
\end{aligned}$$

(a)

$$-\frac{1}{n} \log \Pr\{\frac{1}{n}\sum_{i=1}^n X_i \geq \alpha\} = D(r||p) = \log \frac{p}{(1-p)(\alpha-1)} + \alpha \log \frac{\alpha-1}{\alpha p} \tag{11.194}$$

(b)

$$\Pr\{X_1 = k | \frac{1}{n}\sum_{i=1}^n X_i \geq \alpha\} = r_k = \left(1 - \frac{1}{\alpha}\right)^{k-1} \frac{1}{\alpha} \tag{11.195}$$

(c) For $\alpha = 4$ and $p = 0.5$, we have $D = \log 27/16 = 0.755$, and the conditional distribution of X_1 is geometric with mean 4, i.e.

$$\Pr\{X_1 = k | \frac{1}{n}\sum_{i=1}^n X_i \geq \alpha\} = r_k = 0.75^{k-1} 0.25 \tag{11.196}$$

19. **Another expression for Fisher information.** Use integration by parts to show that

$$J(\theta) = -E \frac{\partial^2 \ln f_\theta(x)}{\partial \theta^2}.$$

Solution: *Another expression for Fisher information*

From (11.270), we have

$$\begin{aligned}
J(\theta) &= E_\theta \left[\frac{\partial}{\partial\theta} \ln f(X;\theta) \right]^2 && (11.197)\\
&= \int_0^\infty f(x;\theta) \left[\frac{\partial}{\partial\theta} \ln f(x;\theta) \right]^2 \, dx && (11.198)\\
&= \int_0^\infty f(x;\theta) \left[\frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} \right]^2 \, dx && (11.199)\\
&= \int_0^\infty \frac{\left(\frac{\partial}{\partial\theta} f(x;\theta)\right)^2}{f(x;\theta)} \, dx && (11.200)\\
&= \int_0^\infty \left(\frac{\partial}{\partial\theta} f(x;\theta) \right) \left[\frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} \right] \, dx && (11.201)\\
& && (11.202)
\end{aligned}$$

Now integrating by parts, setting $u = \frac{\partial}{\partial\theta} \ln f(x;\theta)$ and $dv = \left(\frac{\partial}{\partial\theta} f(x;\theta)\right) dx$, we have

$$\begin{aligned}
\int_0^\infty dv &= \int_0^\infty \left(\frac{\partial}{\partial\theta} f(x;\theta) \right) dx && (11.203)\\
&= \frac{\partial}{\partial\theta} \int_0^\infty f(x;\theta) \, dx && (11.204)\\
&= \frac{\partial}{\partial\theta} 1 && (11.205)\\
&= 0 && (11.206)
\end{aligned}$$

Therefore since

$$\int u \, dv = uv - \int v \, du \tag{11.207}$$

we have

$$\int_0^\infty \left(\frac{\partial}{\partial\theta} f(x;\theta) \right) \left[\frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} \right] dx \;=\; - \int_0^\infty - \frac{\partial^2}{\partial\theta^2} \ln; f(x;\theta) \, dx \tag{11.208}$$

20. **Stirling's approximation:** Derive a weak form of Stirling's approximation for factorials, i.e., show that

$$\left(\frac{n}{e}\right)^n \leq n! \leq n \left(\frac{n}{e}\right)^n \tag{11.209}$$

using the approximation of integrals by sums. Justify the following steps:

$$\ln(n!) = \sum_{i=2}^{n-1} \ln(i) + \ln(n) \leq \int_2^{n-1} \ln x \, dx + \ln n = ... \tag{11.210}$$

and

$$\ln(n!) = \sum_{i=1}^{n} \ln(i) \geq \int_0^n \ln x \, dx = ... \tag{11.211}$$

Solution: *Stirling's approximation* The basic idea of the proof is find bounds for the sum $\sum_{i=2}^{n} \ln i$. If we plot the sum as a sum of rectangular areas, as shown in Figure 11.2, it is not difficult to see that the total area of the rectangles is bounded above by the integral of the upper curve, and bounded below by the integral of the lower curve.

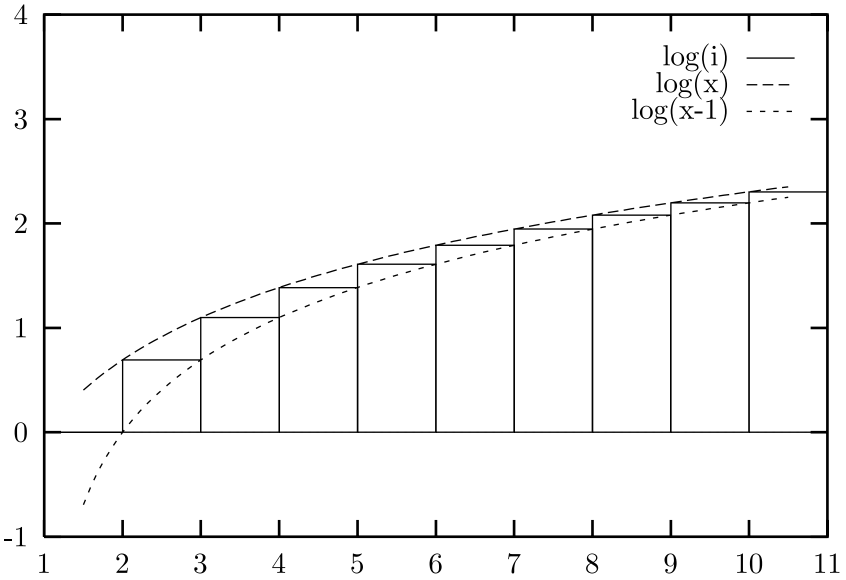


Figure 11.2: Upper and lower bounds on $\log n!$

Now consider the upper bound: From the figure, it follows that the sum of the rectangles starting at 2,3,4,$\ldots, n-1$ is less than the integral of the upper curve from 2 to n. Therefore,

$$\begin{aligned}
\ln(n!) &= \ln n + \sum_{i=2}^{n-1} \ln i + \ln 1 && (11.212)\\
&= \ln n + \sum_{i=2}^{n-1} \ln i && (11.213)\\
&\le \ln n + \int_2^n \ln(x)\, dx && (11.214)\\
&= \ln n + [x \ln x - x]_2^n && (11.215)\\
&= \ln n + n \ln n - n - (2\ln 2 - 2) && (11.216)\\
&= \ln n + n \ln(n/e) - \ln(4/e^2) && (11.217)
\end{aligned}$$

Therefore, exponentiating, we get

$$n! \le n \left(\frac{n}{e}\right)^n \left(\frac{e^2}{4}\right) \le 2n \left(\frac{n}{e}\right)^n, \qquad (11.218)$$

since $e^2/4 = 1.847 < 2$.

For the lower bound, from the figure, it follows that the sum of the areas of the rectanges starting at $1, 2, \ldots, n$ is less than the integral of the lower curve from 1 to $n+1$. Therefore

$$\begin{aligned} \ln(n!) &= \sum_{i=1}^{n} \ln i && (11.219)\\ &\geq \int_1^{n+1} \ln(x-1)\, dx && (11.220)\\ &= \int_0^{n} \ln(x)\, dx && (11.221)\\ &= [x\ln x - x]_0^n && (11.222)\\ &= n\ln n - n - (0\ln 0 - 0) && (11.223)\\ &= n\ln(n/e) && (11.224) \end{aligned}$$

Therefore, exponentiating, we get

$$n! \geq \left(\frac{n}{e}\right)^n. \qquad (11.225)$$

21. **Asymptotic value of $\binom{n}{k}$.** Use the simple approximation of the previous problem to show that, if $0 \leq p \leq 1$, and $k = \lfloor np \rfloor$, i.e., k is the largest integer less than or equal to np, then

$$\lim_{n\to\infty} \frac{1}{n} \log \binom{n}{k} = -p\log p - (1-p)\log(1-p) = H(p). \qquad (11.226)$$

Now let p_i, $i = 1, \ldots, m$ be a probability distribution on m symbols, i.e., $p_i \geq 0$, and $\sum_i p_i = 1$. What is the limiting value of

$$\begin{aligned} &\frac{1}{n}\log \binom{n}{\lfloor np_1\rfloor\ \lfloor np_2\rfloor\ \ldots \lfloor np_{m-1}\rfloor\ n - \sum_{j=0}^{m-1}\lfloor np_j\rfloor} = \\ &\frac{1}{n}\log \frac{n!}{\lfloor np_1\rfloor!\ \lfloor np_2\rfloor!\ \ldots \lfloor np_{m-1}\rfloor!\ (n - \sum_{j=0}^{m-1}\lfloor np_j\rfloor)!} \qquad (11.227) \end{aligned}$$

Solution: *Asymptotic value of* $\binom{n}{k}$

Using the bounds

$$\left(\frac{n}{e}\right)^n \leq n! \leq n\left(\frac{n}{e}\right)^n \qquad (11.228)$$

we obtain

$$\frac{1}{n}\log\binom{n}{k} = \frac{1}{n}\left(\log n! - \log k! - \log(n-k)!\right) \qquad (11.229)$$

$$\begin{aligned}
&\leq \frac{1}{n}\left(\log n\left(\frac{n}{e}\right)^n - \log\left(\frac{k}{e}\right)^k - \log\left(\frac{n-k}{e}\right)^{n-k}\right) && (11.230)\\
&= \frac{1}{n}\log n - \frac{k}{n}\log\frac{k}{n} - \frac{n-k}{n}\log\frac{n-k}{n} && (11.231)\\
&\to H(p) && (11.232)
\end{aligned}$$

Similarly, using the same bounds

$$\begin{aligned}
\frac{1}{n}\log\binom{n}{k} &= \frac{1}{n}\left(\log n! - \log k! - \log(n-k)!\right) && (11.233)\\
&\geq \frac{1}{n}\left(\log\left(\frac{n}{e}\right)^n - \log k\left(\frac{k}{e}\right)^k - \log(n-k)\left(\frac{n-k}{e}\right)^{n-k}\right) && (11.234)\\
&= -\frac{1}{n}\log k(n-k) - \frac{k}{n}\log\frac{k}{n} - \frac{n-k}{n}\log\frac{n-k}{n} && (11.235)\\
&\to H(p) && (11.236)
\end{aligned}$$

and therefore

$$\lim \frac{1}{n}\log\binom{n}{k} = H(p) \qquad (11.237)$$

By the same arguments, it is easy to see that

$$\lim \frac{1}{n}\log\binom{n}{\lfloor np_1\rfloor \ \lfloor np_2\rfloor \ \ldots \lfloor np_{m-1}\rfloor \ n - \sum_{j=0}^{m-1}\lfloor np_j\rfloor} = H(p_1,\ldots,p_m) \qquad (11.238)$$

22. **The running difference..** Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q_1(x)$, and $Y_1, Y_2, \ldots, Y_n$ be i.i.d. $\sim Q_2(y)$. Let X^n and Y^n be independent. Find an expression for $\Pr\{\sum_{i=1}^n X_i - \sum_{i=1}^n Y_i \geq nt\}$, good to first order in the exponent. Again, this answer can be left in parametric form.

    Solution: *Running difference*

    The joint distribution of X and Y is $Q(x,y) = Q_1(x)Q_2(y)$. The constraint that the running difference is greater than nt translates to a constraint on the empirical join distribution, i.e.,

    $$\sum_i \sum_j P_n(i,j)(i-j) \geq t \qquad (11.239)$$

    By Sanov's theorem, the probability of this large deviation is 2^{-nD^*} to the first order in the exponent, where D^* is the minimum relative entropy distance between all distributions P that satisfy the above constraint and $Q(x,y) = Q_1(x)Q_2(y)$, i.e.,

    $$D^* = \min_{\sum_i \sum_j P_n(i,j)(i-j)\geq t} D(P||Q) \qquad (11.240)$$

23. **Large likelihoods.** Let $X_1, X_2, \ldots$ be i.i.d. $\sim Q(x)$, $x \in \{1, 2, \ldots, m\}$. Let $P(x)$ be some other probability mass function. We form the log likelihood ratio

$$\frac{1}{n} \log \frac{P^n(X_1, X_2, \ldots, X_n)}{Q^n(X_1, X_2, \ldots, X_n)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{P(X_i)}{Q(X_i)}$$

of the sequence X^n and ask for the probability that it exceeds a certain threshold. Specifically, find (to first order in the exponent)

$$Q^n \left(\frac{1}{n} \log \frac{P(X_1, X_2, \ldots, X_n)}{Q(X_1, X_2, \ldots, X_n)} > 0 \right).$$

There may be an undetermined parameter in the answer.

Solution:

24. **Fisher information for mixtures.** Let $f_1(x)$ and $f_0(x)$ be two given probability densities. Let Z be Bernoulli(θ), where θ is unknown. Let $X \sim f_1(x)$, if $Z = 1$ and $X \sim f_0(x)$, if $Z = 0$.

    (a) Find the density $f_\theta(x)$ of the observed X.

    (b) Find the Fisher information $J(\theta)$.

    (c) What is the Cramér-Rao lower bound on the mean squared error of an unbiased estimate of θ?

    (d) Can you exhibit an unbiased estimator of θ?

Solution:

25. **Bent coins.** Let $\{X_i\}$ be iid $\sim Q$ where

$$Q(k) = \Pr(X_i = k) = \binom{m}{k} q^k (1-q)^{m-k}, \text{ for } k = 0, 1, 2, \ldots, m.$$

Thus, the X_i's are iid $\sim$ Binomial (m, q).

Show that, as $n \to \infty$,

$$\Pr(X_1 = k | \frac{1}{n} \sum_{i=1}^{n} X_i \geq \alpha) \to P^*(k),$$

where P^* is Binomial (m, λ) (*i.e.* $P^*(k) = \binom{m}{k} \lambda^k (1-\lambda)^{m-k}$ for some $\lambda \in [0, 1]$).

Find λ.

Solution:

26. **Conditional limiting distribution.**

(a) Find the exact value of

$$Pr\{X_1 = 1|\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{4}\} \ , \tag{11.241}$$

if $X_1, X_2, \ldots,$ are Bernoulli($\frac{2}{3}$) and n is a multiple of 4.

(b) Now let $X_i \epsilon \{-1, 0, 1\}$ and let $X_1, X_2 \ldots$ be i.i.d. uniform over $\{-1, 0, +1\}$. Find the limit of

$$Pr\{X_1 = +1|\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{1}{2}\} \tag{11.242}$$

for $n = 2k, \quad k \to \infty$.

Solution:

27. **Variational inequality:** Verify, for positive random variables X, that

$$\log E_P(X) = \sup_Q \left[E_Q(\log X) - D(Q||P)\right] \tag{11.243}$$

where $E_P(X) = \sum_x xP(x)$ and $D(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$, and the supremum is over all $Q(x) \geq 0$, $\sum Q(x) = 1$. It is enough to extremize $J(Q) = E_Q \ln X - D(Q||P) + \lambda(\sum Q(x) - 1)$.

Solution:

28. **Type constraints.**

(a) Find constraints on the type P_{X^n} such that the sample variance $\overline{X_n^2} - (\overline{X}_n)^2 \leq \alpha$, where $\overline{X_n^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ and $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(b) Find the exponent in the probability $Q^n(\overline{X_n^2} - (\overline{X}_n)^2 \leq \alpha)$. You can leave the answer in parametric form.

Solution:

29. **Uniform distribution on the simplex.**
Which of these methods will generate a sample from the uniform distribution on the simplex $\{x \in R^n : x_i \geq 0, \quad \sum_{i=1}^{n} x_i = 1\}$?

(a) Let Y_i be i.i.d. uniform $[0, 1]$, with $X_i = Y_i / \sum_{j=1}^{n} Y_j$.

(b) Let Y_i be i.i.d. exponentially distributed $\sim \lambda e^{-\lambda y}$, $y \geq 0$, with $X_i = Y_i / \sum_{j=1}^{n} Y_j$.

(c) (Break stick into n parts.) Let $Y_1, Y_2, \ldots, Y_{n-1}$ be i.i.d. uniform $[0, 1]$, and let X_i be the length of the i^{th} interval.

Solution:

Chapter 12

Maximum Entropy

1. **Maximum entropy.** Find the maximum entropy density f, defined for $x \geq 0$, satisfying $EX = \alpha_1, E \ln X = \alpha_2$. That is, maximize $-\int f \ln f$ subject to $\int x f(x)\, dx = \alpha_1, \int (\ln x) f(x)\, dx = \alpha_2$, where the integral is over $0 \leq x < \infty$. What family of densities is this?

   Solution: *Maximum entropy.*

   As derived in class, the maximum entropy distribution subject to constraints

   $$\int x f(x)\, dx = \alpha_1 \tag{12.1}$$

   and

   $$\int (\ln x) f(x)\, dx = \alpha_2 \tag{12.2}$$

   is of the form

   $$f(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 \ln x} = c x^{\lambda_2} e^{\lambda_1 x}, \tag{12.3}$$

   which is of the form of a Gamma distribution. The constants should be chosen so as to satisfy the constraints.

2. **Min $D(P \parallel Q)$ under constraints on P.** We wish to find the (parametric form) of the probability mass function $P(x), x \in \{1, 2, \ldots\}$ that minimizes the relative entropy $D(P \parallel Q)$ over all P such that $\sum P(x) g_i(x) = \alpha_i, \quad i = 1, 2, \ldots$.

   (a) Use Lagrange multipliers to guess that

   $$P^*(x) = Q(x) e^{\sum_{i=1}^{\infty} \lambda_i g_i(x) + \lambda_0} \tag{12.4}$$

   achieves this minimum if there exist λ_i's satisfying the α_i constraints. This generalizes the theorem on maximum entropy distributions subject to constraints.

   (b) Verify that P^* minimizes $D(P \parallel Q)$.

   Solution: *Minimize $D(P||Q)$ under constraints on P.*

(a) We construct the functional using Lagrange multipliers

$$J(P) = \int P(x) \ln \frac{P(x)}{Q(x)} + \sum_i \lambda_i \int P(x) h_i(x) + \lambda_0 \int P(x). \tag{12.5}$$

'Differentiating' with respect to $P(x)$, we get

$$\frac{\partial J}{\partial P} = \ln \frac{P(x)}{Q(x)} + 1 + \sum_i \lambda_i h_i(x) + \lambda_0 = 0, \tag{12.6}$$

which indicates that the form of $P(x)$ that minimizes the Kullback Leibler distance is

$$P^*(x) = Q(x) e^{\lambda_0 + \sum_i \lambda_i h_i(x)}. \tag{12.7}$$

(b) Though the Lagrange multiplier method correctly indicates the form of the solution, it is difficult to prove that it is a minimum using calculus. Instead we use the properties of $D(P||Q)$. Let P be any other distribution satisfying the constraints. Then

$$\begin{aligned}
D(P||Q) &- D(P^*||Q) && (12.8)\\
&= \int P(x) \ln \frac{P(x)}{Q(x)} - \int P^*(x) \ln \frac{P^*(x)}{Q(x)} && (12.9)\\
&= \int P(x) \ln \frac{P(x)}{Q(x)} - \int P^*(x)[\lambda_0 + \sum_i \lambda_i h_i(x)] && (12.10)\\
&= \int P(x) \ln \frac{P(x)}{Q(x)} - \int P(x)[\lambda_0 + \sum_i \lambda_i h_i(x)] && \text{(since both } P \text{ and } P^* \text{ satisfy the constraints)}\\
&= \int P(x) \ln \frac{P(x)}{Q(x)} - \int P(x) \ln \frac{P^*(x)}{Q(x)} && (12.11)\\
&= \int P(x) \ln \frac{P(x)}{P^*(x)} && (12.12)\\
&= D(P||P^*) && (12.13)\\
&\geq 0, && (12.14)
\end{aligned}$$

and hence P^* uniquely minimizes $D(P||Q)$.

In the special case when Q is a uniform distribution over a finite set, minimizing $D(P||Q)$ corresponds to maximizing the entropy of P.

3. **Maximum entropy processes.** Find the maximum entropy rate stochastic process $\{X_i\}_{-\infty}^{\infty}$ subject to the constraints:

(a) $EX_i^2 = 1, \quad i = 1, 2, \ldots,$

(b) $EX_i^2 = 1$, $EX_i X_{i+1} = \frac{1}{2}, \quad i = 1, 2, \ldots.$

(c) Find the maximum entropy spectrum for the processes in parts (a) and (b).

Solution: *Maximum Entropy Processes.*

(a) If the only constraint is $EX_i^2 = 1$, then by Burg's theorem, it is clear that the maximum entropy process is a 0-th order Gauss-Markov, i.e., X_i i.i.d. $\sim \mathcal{N}(0,1)$.

(b) If the constraints are $EX_i^2 = 1, \quad EX_iX_{i+1} = \frac{1}{2}$, then by Burg's theorem, the maximum entropy process is a first order Gauss-Markov process of the form

$$X_i = -aX_{i-1} + Z_i, \quad Z_i \sim \mathcal{N}(0,\sigma^2). \tag{12.15}$$

To determine a and σ^2, we use the Yule-Walker equations

$$\begin{aligned} R_0 &= -aR_1 + \sigma^2 && (12.16)\\ R_1 &= -aR_0 && (12.17)\\ & && (12.18) \end{aligned}$$

Substituting $R_0 = 1$ and $R_1 = \frac{1}{2}$, we get $a = -\frac{1}{2}$ and $\sigma^2 = \frac{3}{4}$. Hence the maximum entropy process is

$$X_i = \frac{1}{2}X_{i-1} + Z_i, \quad Z_i \sim \mathcal{N}(0, \frac{3}{4}). \tag{12.19}$$

4. **Maximum entropy with marginals.** What is the maximum entropy distribution $p(x,y)$ that has the following marginals? Hint: You may wish to guess and verify a more general result.

$x\backslash y$	1	2	3	
1	p_{11}	p_{12}	p_{13}	1/2
2	p_{21}	p_{22}	p_{23}	1/4
3	p_{31}	p_{32}	p_{33}	1/4
	2/3	1/6	1/6	

Solution: *Maximum entropy with marginals.*

Given the marginal distributions of X and Y, $H(X)$ and $H(Y)$ are fixed. Since $I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0$, we have

$$H(X,Y) \leq H(X) + H(Y) \tag{12.20}$$

with equality if and only if X and Y are independent. Hence the maximum value of $H(X,Y)$ is $H(X)+H(Y)$, and is attained by choosing the joint distribution to be the product distribution, i.e.,

x y	1	2	3	
1	1/3	1/12	1/12	1/2
2	1/6	1/24	1/24	1/4
3	1/6	1/24	1/24	1/4
	2/3	1/6	1/6	

5. **Processes with fixed marginals.** Consider the set of all densities with fixed pairwise marginals $f_{X_1,X_2}(x_1,x_2), f_{X_2,X_3}(x_2,x_3),\ldots,f_{X_{n-1},X_n}(x_{n-1},x_n)$. Show that the maximum entropy process with these marginals is the first-order (possibly time-varying) Markov process with these marginals. Identify the maximizing $f^*(x_1,x_2,\ldots,x_n)$.

Solution: *Processes with fixed marginals*

By the chain rule,

$$\begin{aligned} h(X_1,X_2,\ldots,X_n) &= h(X_1)+\sum_{i=2}^{n} h(X_i|X_{i-1},\ldots,X_1) && (12.21)\\ &\le h(X_1)+\sum_{i=2}^{n} h(X_i|X_{i-1}), && (12.22)\end{aligned}$$

since conditioning reduces entropy. The quantities $h(X_1)$ and $h(X_i|X_{i-1})$ depend only on the second order marginals of the process and hence the upper bound is true for all processes satisfying the second order marginal constraints.

Define

$$f^*(x_1,x_2,\ldots,x_n)=f_0(x_1)\prod_{i=2}^{n}\frac{f_0(x_{i-1},x_i)}{f_0(x_{i-1})}. \tag{12.23}$$

We will show that f^* maximizes the entropy among all processes with the same second order marginals. To prove this, we just have to show that this process satisfies has the same second order marginals and that this process achieves the upper bound (12.22). The fact that the process satisfies the marginal constraints can be easily proved by induction. Clearly, it is true for $f^*(x_1,x_2)$ and if $f^*(x_{i-1},x_i)=f_0(x_{i-1},x_i)$, then $f^*(x_i)=f_0(x_i)$ and by the definition of f^*, it follows that $f^*(x_i,x_{i+1})=f_0(x_i,x_{i+1})$. Also, since by definition, f^* is first order Markov, $h(X_i|X_{i-1},\ldots,X_1)=h(X_i|X_{i-1})$ and we have equality in (12.22). Hence f^* has the maximum entropy of all processes with the same second order marginals.

6. **Every density is a maximum entropy density.** Let $f_0(x)$ be a given density. Given $r(x)$, let $g_\alpha(x)$ be the density maximizing $h(X)$ over all f satisfying $\int f(x)r(x)\,dx=\alpha$. Now let $r(x)=\ln f_0(x)$. Show that $g_\alpha(x)=f_0(x)$ for an appropriate choice $\alpha=\alpha_0$. Thus $f_0(x)$ is a maximum entropy density under the constraint $\int f\ln f_0=\alpha_0$.

Solution: *Every density is a maximum entropy density.* Given the constraints that

$$\int r(x)f(x)=\alpha \tag{12.24}$$

the maximum entropy density is

$$f^*(x) = e^{\lambda_0+\lambda_1 r(x)} \tag{12.25}$$

With $r(x) = \log f_0(x)$, we have

$$f^*(x) = \frac{f_0^{\lambda_1}(x)}{\int f_0^{\lambda_1}(x)\,dx} \tag{12.26}$$

where λ_1 has to chosen to satisfy the constraint. We can chose the value of the constraint to correspond to the value $\lambda_1 = 1$, in which case $f^* = f_0$. So f_0 is a maximum entropy density under appropriate constraints.

7. **Mean squared error.**
Let $\{X_i\}_{i=1}^n$ satisfy $EX_iX_{i+k} = R_k, \quad k = 0, 1, \ldots, p$. Consider linear predictors for X_n, i.e.

$$\hat{X}_n = \sum_{i=1}^{n-1} b_i X_{n-i}.$$

Assume $n > p$. Find

$$\max_{f(x^n)} \min_b E(X_n - \hat{X}_n)^2,$$

where the minimum is over all linear predictors b and the maximum is over all densities f satisfying $R_0, \ldots, R_p$.

Solution: *Mean squared error.*

8. **Maximum entropy characteristic functions.**
We ask for the maximum entropy density $f(x), \quad 0 \le x \le a$, satisfying a constraint on the characteristic function $\Psi(u) = \int_0^a e^{iux} f(x)dx$. The answers need be given only in parametric form.

   (a) Find the maximum entropy f satisfying $\int_0^a f(x)\cos(u_0x)\,dx = \alpha$, at a specified point u_0.

   (b) Find the maximum entropy f satisfying $\int_0^a f(x)\sin(u_0x)dx = \beta$.

   (c) Find the maximum entropy density $f(x), \quad 0 \le x \le a$, having a given value of the characteristic function $\Psi(u_0)$ at a specified point u_0.

   (d) What problem is encountered if $a = \infty$?

Solution: *Maximum entropy characteristic functions.*

9. **Maximum entropy processes.**

   (a) Find the maximum entropy rate binary stochastic process $\{X_i\}_{i=-\infty}^{\infty}, \quad X_i \in \{0,1\}$, satisfying $\Pr\{X_i = X_{i+1}\} = \frac{1}{3}$, for all i.

   (b) What is the resulting entropy rate?

Solution: *Maximum entropy processes.*

10. **Maximum entropy of sums** Let $Y = X_1 + X_2$ Find the maximum entropy density for Y under the constraint $EX_1^2 = P_1$, $EX_2^2 = P_2$,

    (a) if X_1 and X_2 are independent.

    (b) if X_1 and X_2 are allowed to be dependent.

    (c) Prove part (a).

    Solution: *Maximum entropy of sums*

11. **Maximum entropy Markov chain.**

    Let $\{X_i\}$ be a stationary Markov chain with $X_i \in \{1, 2, 3\}$. Let $I(X_n; X_{n+2}) = 0$ for all n.

    (a) What is the maximum entropy rate process satisfying this constraint?

    (b) What if $I(X_n; X_{n+2}) = \alpha$, for all n for some given value of α, $0 \le \alpha \le \log 3$?

    Solution: *Maximum entropy Markov chain.*

12. **An entropy bound on prediction error.** Let $\{X_n\}$ be an arbitrary real valued stochastic process. Let $\hat{X}_{n+1} = E\{X_{n+1}|X^n\}$. Thus the conditional mean $\hat{X}_{n+1}$ is a random variable depending on the n-past X^n. Here $\hat{X}_{n+1}$ is the minimum mean squared error prediction of X_{n+1} given the past.

    (a) Find a lower bound on the conditional variance $E\{E\{(X_{n+1} - \hat{X}_{n+1})^2|X^n\}\}$ in terms of the conditional differential entropy $h(X_{n+1}|X^n)$.

    (b) Is equality achieved when $\{X_n\}$ is a Gaussian stochastic process?

    Solution: *An entropy bound on prediction error.*

13. **Maximum entropy rate**. What is the maximum entropy rate stochastic process $\{X_i\}$ over the symbol set $\{0, 1\}$ for which the probability that 00 occurs in a sequence is zero?

    Solution: *Maximum entropy rate*

14. **Maximum entropy.**

    (a) What is the parametric form maximum entropy density $f(x)$ satisfying the two conditions

    $$EX^8 = a$$
    $$EX^{16} = b?$$

    (b) What is the maximum entropy density satisfying the condition

    $$E(X^8 + X^{16}) = a + b \qquad ?$$

    (c) Which entropy is higher?

    Solution: *Maximum entropy.*

15. **Maximum entropy.** Find the parametric form of the maximum entropy density f satisfying the Laplace transform condition

$$\int f(x)\mathrm{e}^{-x}dx = \alpha \ ,$$

and give the constraints on the parameter.

Solution: *Maximum entropy.*

16. **Maximum entropy processes**
Consider the set of all stochastic processes with $\{X_i\}, \quad X_i \in \mathcal{R}$, with

$$\begin{array}{l} R_0 = EX_i^2 = 1 \\ R_1 = EX_iX_{i+1} = \frac{1}{2}. \end{array}$$

Find the maximum entropy rate.

Solution: *Maximum entropy processes*

17. **Binary maximum entropy**
Consider a *binary* process $\{X_i\}, \quad X_i \in \{-1, +1\}$, with $R_0 = EX_i^2 = 1$ and $R_1 = EX_iX_{i+1} = \frac{1}{2}$.

(a) Find the maximum entropy process with these constraints.

(b) What is the entropy rate?

(c) Is there a Bernoulli process satisfying these constraints?

Solution: *Binary maximum entropy*

18. **Maximum entropy.**
Maximize $h(Z, V_x, V_y, V_z)$ subject to the energy constraint $E(\frac{1}{2}m \parallel V \parallel^2 + mgZ) = E_0$. Show that the resulting distribution yields

$$E\frac{1}{2}m \parallel V \parallel^2 = \frac{3}{5}E_0$$

$$EmgZ = \frac{2}{5}E_0.$$

Thus $\frac{2}{5}$ of the energy is stored in the potential field, regardless of its strength g.

Solution: *Maximum entropy.*

19. **Maximum entropy discrete processes.**

(a) Find the maximum entropy rate binary stochastic process $\{X_i\}_{i=-\infty}^{\infty}, \quad X_i \in \{0, 1\}$, satisfying $\Pr\{X_i = X_{i+1}\} = \frac{1}{3}$, for all i.

(b) What is the resulting entropy rate?

Solution: *Maximum entropy discrete processes.*

20. **Maximum entropy of sums.**
Let $Y = X_1 + X_2$. Find the maximum entropy of Y under the constraint $EX_1^2 = P_1, \;\; EX_2^2 = P_2$,

    (a) if X_1 and X_2 are independent.

    (b) if X_1 and X_2 are allowed to be dependent.

    Solution: *Maximum entropy of sums.*

21. **Entropy rate**

    (a) Find the maximum entropy rate stochastic process $\{X_i\}$ with $EX_i^2 = 1$, $EX_iX_{i+2} = \alpha, \;\; i = 1, 2, \ldots.$ Be careful.

    (b) What is the maximum entropy rate?

    (c) What is EX_iX_{i+1} for this process?

    Solution: *Entropy rate*

22. **Minimum expected value**

    (a) Find the minimum value of EX over all probability density functions $f(x)$ satisfying the following three constraints:

        (i) $f(x) = 0 \quad$ for $x \le 0$,

        (ii) $\int_{-\infty}^{\infty} f(x)dx = 1$, and

        (iii) $h(f) = h$.

    (b) Solve the same problem if (i) is replaced by

        (i′) $f(x) = 0 \quad$ for $x \le a$.

    Solution: *Minimum expected value*

Chapter 13

Universal Source Coding

1. **Minimax regret data compression and channel capacity.** First consider universal data compression with respect to two source distributions. Let the alphabet $V = \{1, e, 0\}$ and let $p_1(v)$ put mass $1-\alpha$ on $v = 1$ and mass α on $v = e$. Let $p_2(v)$ put mass $1-\alpha$ on 0 and mass α on $v = e$.

   We assign word lengths to V according to $l(v) = \log \frac{1}{p(v)}$, the ideal codeword length with respect to a cleverly chosen probability mass function $p(v)$. The worst case excess description length (above the entropy of the true distribution) is

   $$\max_i \left(E_{p_i} \log \frac{1}{p(V)} - E_{p_i} \log \frac{1}{p_i(V)} \right) = \max_i D(p_i \parallel p). \tag{13.1}$$

   Thus the minimax regret is $R^* = \min_p \max_i D(p_i \parallel p)$.

   (a) Find R^*.

   (b) Find the $p(v)$ achieving R^*.

   (c) Compare R^* to the capacity of the binary erasure channel

   $$\begin{bmatrix} 1-\alpha & \alpha & 0 \\ 0 & \alpha & 1-\alpha \end{bmatrix}$$

   and comment.

   Solution: *Minimax regret data compression and channel capacity.*

2. **Universal data compression.** Consider three possible source distributions on $\mathcal{X}$,

   $$\begin{aligned} P_a &= (.7, .2, .1), \\ P_b &= (.1, .7, .2), \quad \text{and} \\ P_c &= (.2, .1, .7). \end{aligned}$$

(a) Find the minimum incremental cost of compression

$$R^* = \min_P \max_\theta D(P_\theta \| P),$$

and the associated mass function $P = (p_1, p_2, p_3)$, and ideal codeword lengths $l_i = \log(1/p_i)$.

(b) What is the channel capacity of a channel matrix with rows P_a, P_b, P_c?

Solution: *Universal data compression.*

3. **Arithmetic coding**: Let $[X_i]$ be a stationary binary Markov chain with transition matrix

$$p_{ij} = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \tag{13.2}$$

Calculate the first 3 bits of $F(X^\infty) = 0.F_1F_2\ldots$ when $X^\infty = 1010111\ldots$. How many bits of X^∞ does this specify?

Solution: *Arithmetic coding*

4. **Arithmetic coding.** Let X_i be binary stationary Markov with transition matrix $\begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & \frac{1}{3} \end{bmatrix}$.

(a) Find $F(01110) = \Pr\{.X_1X_2X_3X_4X_5 < .01110\}$.

(b) How many bits $.F_1F_2\ldots$ can be known for sure if it is not known how $X = 01110$ continues?

Solution: *Arithmetic coding.*

5. **Lempel-Ziv.** Give the LZ78 parsing and encoding of 00000011010100000110101.

Solution: *Lempel-Ziv.* We first parse the string, looking for strings that we have not seen before. Thus, the parsing yields 0,00,000,1,10,101,0000,01,1010,1. There are 10 phrases, and therefore we need 4 bits to represent the pointer to the prefix. Thus, using the scheme described in the text, we encode the string as (0000,0),(0001,0), (0010,0), (0000,1), (0100,0), (0101,1), (0011,0), (0001,1), (0110,0),(0000,1). (The last phrase, though it is not really a new phrase, is handled like a new phrase).

6. **Lempel Ziv 78**

We are given the constant sequence $x^n = 11111\ldots$

(a) Give the LZ78 parsing for this sequence.

(b) Argue that the number of encoding bits per symbol for this sequence goes to zero as $n \to \infty$.

Solution: *Lempel Ziv 78*

7. **Another idealized version of Lempel-Ziv coding.** An idealized version of LZ was shown to be optimal: The encoder and decoder both have available to them the "infinite past" generated by the process, $\ldots, X_{-1}, X_0$, and the encoder describes the string $(X_1, X_2, \ldots, X_n)$ by telling the decoder the position R_n in the past of the first recurrence of that string. This takes roughly $\log R_n + 2 \log\log R_n$ bits.

   Now consider the following variant: Instead of describing R_n, the encoder describes R_{n-1} plus the last symbol X_n. From these two the decoder can reconstruct the string $(X_1, X_2, \ldots, X_n)$.

   (a) What is the number of bits per symbol used in this case to encode $(X_1, X_2, \ldots, X_n)$?

   (b) Modify the proof given in the text to show that this version is also asymptotically optimal, namely that the expected number of bits-per-symbol converges to the entropy rate.

Solution: *Another idealized version of Lempel-Ziv coding.*

In this version of LZ coding, the encoder and decoder both have available to them the infinite past generated by the process, $\ldots, X_{-1}, X_0$, and the encoder describes the string $X_1^n = (X_1, X_2, \ldots, X_n)$ by telling the decoder the position R_{n-1} in the past of the first recurrence of the string X_1^{n-1}, plus the last symbol X_n.

(a) Let A be the alphabet of the process, and $|A|$ denote its size. Then the number of bits it takes to represent R_{n-1} is roughly $\log R_{n-1} + C \log\log R_{n-1}$, where C is a constant independent of n. To represent X_n, it takes $\lceil \log |A| \rceil$ bits, so the overall number of bits per symbol used for the whole string X_1^n is

$$\frac{L_n(X_1^n)}{n} = \frac{\log R_{n-1} + C \log\log R_{n-1} + \lceil \log |A| \rceil}{n}.$$

(b) To prove that this description is asymptotically optimal it suffices to show that

$$\limsup_{n\to\infty} \mathbf{E}\left(\frac{L_n}{n}\right) \leq H, \tag{13.3}$$

and the optimality will follow since we know that the reverse inequality also holds, by Shannon's Noiseless Coding Theorem.

For the last term in L_n it is immediate that

$$\frac{\lceil \log |A| \rceil}{n} \to 0, \tag{13.4}$$

as $n \to \infty$. Now notice that we always have $R_{n-1} \leq R_n$, so for the first two terms in L_n,

$$\mathbf{E}\left(\frac{\log R_{n-1} + C \log\log R_{n-1}}{n}\right) \leq \mathbf{E}\left(\frac{\log R_n + C \log\log R_n}{n}\right), \tag{13.5}$$

and as was shown in class, the above right-hand-side is asymptotically bounded above by H,

$$\limsup_{n\to\infty} \mathbf{E}\left(\frac{\log R_n + C\log\log R_n}{n}\right) \le H. \tag{13.6}$$

Combining (13.4) with (13.5) and (13.6), yields (13.3), as claimed.

8. **Length of pointers in LZ77**. In the version of LZ77 due of the Storer and Szymanski[15], described in Section 13.4.1, a short match can either be represented by (F, P, L) (flag, pointer, length) or by (F, C) (flag, character). Assume that the window length is W, and assume that the maximum match length is M.

   (a) How many bits are required to represent P? To represent L?

   (b) Assume that C, the representation of a character is 8 bits long. If the representation of P plus L is longer than 8 bits, it would be better to represent a single character match as an uncompressed character rather than as a match within the dictionary. As a function of W and M, what is the shortest match that one should represent as a match rather than as uncompressed characters?

   (c) Let $W = 4096$ and $M = 256$. What is the shortest match that one would represent as a match rather than uncompressed characters?

   Solution: *Length of pointers in LZ77*

   (a) Since P represents the position within the window, $\lceil \log W \rceil$ bits would suffice to represent P. Since L represents the length of the match, which is at most M, $\lceil \log M \rceil$ bits suffice for L.

   (b) Ignoring the integer constraints, we can see that we would use the uncompressed character to represent a match of length 1 if $\log W + \log M > 8$. Similarly, we would use single characters rather than the match representation for matches of length m if $1 + \log W + \log M > m(1 + 8)$, since the representation as a sequence of single characters needs 9 bits per character.

   (c) If $M = 256$, $\log M = 8$, $W = 4096$, $\log W = 14$, and $1 + \log W + \log M = 23$ and we would use the uncompressed representation if $m = 1$ or 2. For $m \ge 3$, the match representation is shorter.

9. **Lempel-Ziv.**

   (a) Continue the Lempel-Ziv parsing of the sequence 0,00,001,00000011010111.

   (b) Give a sequence for which the number of phrases in the LZ parsing grows as fast as possible.

   (c) Give a sequence for which the number of phrases in the LZ parsing grows as slowly as possible.

   Solution: *Lempel-Ziv.*

(a) The Lempel-Ziv parsing is: $0, 00, 001, 000, 0001, 1, 01, 011, 1$

(b) The sequence is: $0, 1, 00, 01, 10, 11, 000, 001, \ldots$ concatenating all binary strings of length 1,2,3, etc. This is the sequence where the phrases are as short as possible.

(c) Clearly the constant sequence will do: $1, 11, 111, 1111, \ldots$

10. **Two versions of fixed-database Lempel-Ziv**. Consider a source $(\mathcal{A}, P)$. For simplicity assume that the alphabet is finite $|\mathcal{A}| = A < \infty$, and the symbols are i.i.d. $\sim P$. A fixed database $\mathcal{D}$ is given, and is revealed to the decoder. The encoder parses the target sequence x_1^n into blocks of length l, and subsequently encodes them by giving the binary description of their last appearance in the database. If a match is not found, the entire block is sent uncompressed, requiring $l \log A$ bits. A flag is used to tell the decoder whether a match location is being described, or the sequence itself. Problems (a) and (b) give some preliminaries you will need in showing the optimality of fixed-database LZ in (c).

(a) Let x_l be a δ-typical sequence of length l starting at 0, and let $R_l(x^l)$ be the corresponding recurrence index in the infinite past $\ldots, X_{-2}, X_{-1}$. Show that

$$E\left[R_l(X^l)|X^l = x^l\right] \leq 2^{l(H+\delta)}$$

where H is the entropy rate of the source.

(b) Prove that for any $\epsilon > 0$, $\Pr\left(R_l(X^l) > 2^{l(H+\epsilon)}\right) \to 0$ as $l \to \infty$.

Hint: Expand the probability by conditioning on strings x^l, and break things up into typical and non-typical. Markov's inequality and the AEP should prove handy as well.

(c) Consider the following two fixed databases (i) $\mathcal{D}_1$ is formed by taking all δ-typical l-vectors; and (ii) $\mathcal{D}_2$ formed by taking the most recent $\tilde{L} = 2^{l(H+\delta)}$ symbols in the infinite past (i.e., $X_{-\tilde{L}}, \ldots, X_{-1}$). Argue that the algorithm described above is asymptotically optimal, namely that the expected number of bits-per-symbol converges to the entropy rate, when used in conjunction with either database $\mathcal{D}_1$ or $\mathcal{D}_2$.

Solution: *Two versions of fixed-database Lempel-Ziv*

(a) Since x^l is δ-typical, the AEP implies that $p(x^l) \geq 2^{-l(H+\delta)}$, and the result follows from Kac's lemma.

(b) Fix $\epsilon > 0$, and $\delta \in (0, \epsilon)$. Let $A_\delta^{(l)}$ be the δ-typical set for $\mathcal{A}^l$. We divide the set of sequences into the typical sequences and the non-typical sequences.

$$\begin{aligned}
\Pr(R_l(X^l) > 2^{l(H+\epsilon)}) &= \sum_{x^l} p(x^l) \Pr(R_l(X^l) > 2^{l(H+\epsilon)}|x^l) \\
&= \sum_{x^l \in A_\delta^{(l)}} p(x^l) \Pr(R_l(X^l) > 2^{l(H+\epsilon)}|x^l)
\end{aligned}$$

$$
\begin{aligned}
& + \sum_{x^l \notin A_\delta^{(l)}} p(x^l) \Pr(R_l(X^l) > 2^{l(H+\epsilon)} | x^l) \\
\overset{\text{(i)}}{\leq} & \sum_{x^l \in A_\delta^{(l)}} p(x^l) \frac{2^{l(H+\delta)}}{2^{l(H+\epsilon)}} + \Pr(X^l \notin A_\delta^{(l)}) \\
\leq & \ 2^{-l(\epsilon-\delta)} + \Pr(X^l \notin A_\delta^{(l)})
\end{aligned}
$$

where (i) follows from Markov's inequality and using the result of part (a). The proof now follows from the AEP, by sending l to infinity.

(c) For $\mathcal{D}_1$ the proof follows trivially from the analysis in §3.2 in Cover and Thomas. For $\mathcal{D}_2$, let $N = n/l$ be the number of blocks in the sequence, and let $L(B_i)$ denote the length of the encoding of the i-th block B_i. To simplify the notation, assume that $N = n/l$ is an integer. We call a block 'good' if we can find a match in $\mathcal{D}_2$, and 'bad' otherwise. Let G be the set of good blocks. If $B_i \in G$, we e ncode it using $\log |\mathcal{D}_2|$ bits, which by our choice of $\mathcal{D}_2$ is equal to $H(l+\delta)$ bits. If $B_i \notin G$ then we encode it using $l \log A$ bits. We throw in one extra bit to distinguish between the two events. Then,

$$
\begin{aligned}
\frac{1}{n} EL(X_1, X_2, \ldots, X_n) &= \frac{1}{n} E \sum_i L(B_i) \\
&= \frac{1}{n} E \sum_{i \in G} (1 + l(H+\delta)) + \frac{1}{n} E \sum_{i \notin G} (1 + l \log A) \\
&\overset{\text{(i)}}{\leq} \frac{(1 + l(H+\delta))}{l} + \Pr\{X^l \notin \mathcal{D}_2\}(\frac{1}{l} + \log A)
\end{aligned}
$$

where step (i) follows from taking the first summation over all N blocks, and using $N^{-1} E \sum_{i \notin G} 1 = \Pr\{G^c\}$. Take l_n to be a sequence of integers such that $l_n \uparrow \infty$ as $n \to \infty$. It now follows from part (b) that $\Pr\{X^l \notin \mathcal{D}_2\} \to 0$ and thus, $\limsup_n n^{-1} EL_n \leq H + \delta$ and since δ is arbitrary, we have $\limsup_n n^{-1} EL_n \leq H$. The proof is now complete since $\liminf_n n^{-1} EL_n \geq H$, by Shannon's source coding theorem.

11. **Tunstall Coding:** The normal setting for source coding maps a symbol (or a block of symbols) from a finite alphabet onto a variable length string. An example of such a code is the Huffman code, which is the optimal (minimal expected length) mapping from a set of symbols to a prefix free set of codewords. Now consider the dual problem of variable-to-fixed length codes, where we map a variable length sequence of source symbols into a fixed length binary (or D-ary) representation. A variable-to-fixed length code for an i.i.d. sequence of random variables $X_1, X_2, \ldots, X_n, X_i \sim p(x), x \in \mathcal{X} = \{0, 1, \ldots, m-1\}$ is defined by a prefix-free set of phrases $A_D \subset \mathcal{X}^*$, where $\mathcal{X}^*$ is the set of finite length strings of symbols of $\mathcal{X}$, and $|A_D| = D$. Given any sequence $X_1, X_2, \ldots, X_n$, the string is parsed into phrases from A_D (unique because of the prefix free property of A_D), and represented by a sequence of symbols from a D-ary alphabet. Define the

efficiency of this coding scheme by

$$R(A_D) = \frac{\log D}{EL(A_D)} \tag{13.7}$$

where $EL(A_D)$ is the expected length of a phrase from A_D.

(a) Prove that $R(A_D) \geq H(X)$.

(b) The process of constructing A_D can be considered as a process of constructing an m-ary tree whose leaves are the phrases in A_D. Assume that $D = 1 + k(m-1)$ for some integer $k \geq 1$.

Consider the following algorithm due to Tunstall:

i. Start with $A = \{0, 1, \ldots, m-1\}$ with probabilities $p_0, p_1, \ldots, p_{m-1}$. This corresponds to a complete m-ary tree of depth 1.

ii. Expand the node with the highest probability. For example, if p_0 is the node with the highest probability, the new set is $A = \{00, 01, \ldots, 0(m-1), 1, \ldots, (m-1)\}$.

iii. Repeat step 2 until the number of leaves (number of phrases) reaches the required value.

Show that the Tunstall algorithm is optimal, in the sense that it constructs a variable to fixed code with the best $R(A_D)$ for a given D, i.e., the largest value of $EL(A_D)$ for a given D.

(c) Show that there exists a D such that $R(A_D^*) < H(X) + 1$.

Solution: *Tunstall Coding:*

(a) We will argue that if $R(A_D) < H(X)$, then it is possible to construct an uniquely decodable code with average length less than the entropy. Consider a long sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n \sim \mathbf{p}$. We can parse this sequence into phrases using the prefix-free set A_D, and these phrases are independent and identically distributed, with the distribution induced by $\mathbf{p}$ on the tree. Thus, using renewal theory, since the expected phrase length is $EL(A_D)$, the number of phrases in the block of length n is $approx n/EL(A_D)$. These phrases can be described with $\log D$ bits each, so that the total description length is $\approx \log D(n/EL(A_D))$. If $R(A_D) < H$, then the total description length is less than nH and we have a contradiction to the fundamental theorem of source coding.

However, making the above argument precise raises issues for which we have two different solutions:

- The algorithm above does not describe how to handle a sequence of random variables that is a prefix of an element of A_D. For example, after parsing a block of length n into phrases from A_D, we might be left with a few symbols of X that are not long enough to make a phrase. We can imagine that these symbols are sent uncompressed, and the overhead is small (the set A_D is finite, and so the maximal length of a phrase in A_D is finite, and so the maximum length of the residue is bounded).

- We could extend the fundamental theorem of source coding directly to the variable to variable case, i.e., we can prove that for any mapping $F : \mathcal{X}^* \rightarrow \{0,1\}^*$, that $EL(A_D)H(X) \leq EL(C)$, where $EL(C)$ is the average length of the binary codewords induced by the mapping. An example of such a mapping for $\mathcal{X} = \{a,b,c\}$ is the mapping $aa \rightarrow 000, ab \rightarrow 001, ac \rightarrow 010, b \rightarrow 011, c \rightarrow 1$. It is easy to see that the above result is a "special" case of such a mapping, where we assume that we can find a code length of $\log D$ for all the elements of A_D.

  Let Y be the random variable whose values are the elements of A_D, and whose distribution is induced by the distribution of X, i.e., if the phrase is ab, the probability $\Pr(Y =' ab') = p_a p_b$. It is easy to see that Y is well defined random variable with total probability 1.

  Now the code C is a code for the random variable Y, and by the standard source coding theorem, we have $EL(C) \geq H(Y)$. We will now prove that $H(Y) = EL(A_D)H(X)$, which will complete the proof.

  There are many ways to prove this result, which is the Wald equation for entropies. We will prove it directly, using a summation over the leaves of a tree. Let $L(X_1, X_2, \ldots, X_n)$ be the stopping defined by set A_D, i.e., $L(X_1, X_2, \ldots, X_n) = l$ if the sequence of length l at the beginning of X is in A_D. Choose n larger than the maximal length in A_D, and let $Y = X_1^L$ be the first phrase in the parsing of $X_1, X_2, \ldots, X_n$. Then

$$\begin{aligned} nH(X) &= H(X_1, X_2, \ldots, X_n) && (13.8)\\ &= H(X_1^L, L, X_{L+1}^n) && (13.9)\\ &= H(X_1^L) + H(L|X_1^L) + H(X_{L+1}^n|L, X_1^L) && (13.10) \end{aligned}$$

  Now since L is fixed given X_1^L, $H(L|X_1^L) = 0$. Also, X_{L+1}^n is indpendent of X_1^L given L, and we can write

$$\begin{aligned} H(X_{L+1}^n|L) &= \sum_{l=1}^{n} \Pr(L=l) H(X_{l+1}^n) && (13.11)\\ &= \sum_{l=1}^{n} \Pr(L=l)(n-l)H(X) && (13.12)\\ &= (n - EL)H(X) && (13.13) \end{aligned}$$

  Substituting this in the equation above, we get

$$nH(X) = H(X_1^L) + (n - EL)H(X) \qquad (13.14)$$

  or

$$H(Y) = H(X_1^L) = ELH(X) \qquad (13.15)$$

  To prove the required result of part (a), we only need to verify that $H(Y) \leq \log D$, which follows directly from the fact that the range of Y is limited to D values.

(b) We will prove the optimality of the Tunstall algorithm by induction in a fashion similar to the proof of Huffman coding optimality. By the statement of the problem, we restrict our attention to complete trees, i.e., trees for which every node is either a leaf (no children) or has m children. Clearly, the algorithm to minimize $R(A_D)$ for a given D has to find the set A_D that maximizes $EL(A_D)$.

We will need some notation for the analysis that follows: nodes in the tree are either leaf nodes (nodes that have no children) or internal nodes (nodes that have m children). The probability of a node is the product of the probability of the symbols that led up to the node. The probability of the root node is 1.

We will assume that the algorithm constructs a tree that is optimal for $D_k = 1 + k(m-1)$. We will show that the algorithm than produces a tree that is optimal for $D_{k+1} = 1 + (k+1)(m-1)$.

Any tree with D_{k+1} nodes consists of tree with D_k nodes with one of the nodes expanded. Let T_k denote a tree with D_k nodes, σ denote a leaf of this tree with probability p_σ, and let T_{k+1} denote the tree with D_{k+1} nodes formed by expanding the node σ. Let $N(T)$ denote the leaf nodes in T. Then

$$\begin{aligned}
EL(T_{k+1}) &= \sum_{i \in N(T_{k+1})} p(i)l(i) && (13.16)\\
&= \sum_{i \in N(T_k), i \neq \sigma} p(i)l(i) + \sum_{j=1}^{m} p(\sigma)p(j)(l(\sigma)+1) && (13.17)\\
&= \sum_{i \in N(T_k), i \neq \sigma} p(i)l(i) + p(\sigma)(l(\sigma)+1) && (13.18)\\
&= \sum_{i \in N(T_k)} p(i)l(i) + p(\sigma) && (13.19)\\
&= EL(T_k) + p(\sigma) && (13.20)
\end{aligned}$$

Thus the expected length for any expanded tree is equal the expected length of the original tree plus the probability of the node that was expanded. This result provides the basic intuition that motivates the algorithm: to maximize $EL(T_{k+1})$ given T_k, we should expand the node with the largest probability. Doing this repeatedly gives us the Tunstall algorithm.

However, using this to prove the optimality of the Tunstall algorithm is surprisingly tricky. This is because there a different sequences of node expansions that give rise to the same final tree. Also, a suboptimal tree of size D_k might have a larger value of $p(\sigma)$ (the steps are not independent, and hence a greedy step early on might not be optimal later) , and thus we cannot directly use the above result for induction.

Instead, we will use another property of the optimal tree constructed by the Tunstall algorithm, that is, the probability of each of the internal nodes is higher than the probability of the leaves. We have the following statement:

Lemma: Any optimal tree T_{k+1} (a tree maximizing EL) has the property that the probability of any of the internal nodes is greater than or equal to the probability of any of the leaves, i.e., if Σ_I is the set of internal nodes and Σ_L is the set of leaf nodes, then

$$\forall \sigma \in \Sigma_I, \forall \sigma' \in \Sigma_L, \quad p(\sigma) \geq p(\sigma') \tag{13.21}$$

Proof: If this were not true, then there exists an internal node with a lower probability than a leaf. Let σ be this internal node, σ_l be the leaf, and since the probability of any descendant node is less than that of the parent, we can work down from σ until we find an internal node σ' just above the leaves that also satisfies this property, i.e., $p(\sigma') < p(\sigma_l)$. Now consider the tree T_k formed by deleting all the leaves coming out σ', and form a tree T'_{k+1} by expanding node σ_l in T_k. We have

$$\begin{aligned} EL(T_{k+1}) &= EL(T_k) + p(\sigma') && (13.22) \\ EL(T'_{k+1}) &= EL(T_k) + p(\sigma_l) && (13.23) \end{aligned}$$

and since $p(\sigma') < p(\sigma_l)$, we have $EL(T_{k+1}) < EL(T'_{k+1})$, contradicting the optimality of T_{k+1}. Thus all optimal trees satisfy the propery above.

We now prove the converse, i.e., that any tree that satisfies this property must be optimal. Again, we prove it by contradiction. Assume that there is a tree T_k satisfying this property that is not optimal, and therefore there is another tree T_k^* which is optimal, i.e.,having larger expected length. By the previous result, this tree also satisfies (13.21).

Now consider the set of nodes that occur in T or T^*. These nodes can be classified into 8 categories.

- S_1: nodes that are internal nodes in both T and T^*.
- S_2: nodes that are leaf nodes in both T and T^*.
- S_3: nodes that are internal nodes in T and leaf nodes in T^*.
- S_4: nodes that leaf nodes in T and internal nodes in T^*.
- S_5: nodes that are internal nodes in T that are not in T^*
- S_6: nodes that are internal nodes in T^* that are not in T
- S_7: nodes that are leaf nodes in T that are not in T^*
- S_8: nodes that are leaf nodes in T^* that are not in T

By assumption, $T \neq T^*$, and therefore there are leaf nodes in T that are not in T^*. Some ancestor of this leaf node σ in T must be a leaf node of T^*, and therefore S_3 is not empty. Similarly, if $T \neq T^*$, S_4 must be non-empty.

We now argue that all nodes in S_3 and S_4 have the same probability. Let $\sigma_3 \in S_3$ and $\sigma_4 \in S_4$ be two nodes in the two sets. By property (13.21) for T^*, it follows that $p(\sigma_3) \leq p(\sigma_4)$. By property (13.21) for T, we have $p(\sigma_4) \leq p(\sigma_3)$. Thus $p(\sigma_3) = p(\sigma_4)$.

We now argue that S_5 and S_6 are empty sets. This follows from the fact that since any node in S_5 has to be a descendant of a node in S_3, and hence $p(\sigma_5) < p(\sigma_3)$. But by the property (13.21) for T, $p(\sigma_5) \geq p(\sigma_4)$, and since $p(\sigma_4) = p(\sigma_3)$, we have a contradiction. Thus there can be no nodes in S_5 or S_6.

Thus the nodes in S_7 are the children of nodes in S_3 and the nodes in S_8 are the children of nodes in S_4. Since T and T^* are equal except for these nodes in S_7 and S_8 and the average length of the trees depends only the probability of the internal nodes, it follows that T and T^* have the same average length.

This finally proves the key result, which is that a tree is optimal if and only if it satisfies property (13.21).

It is now simple to show by induction that the Tunstall algorithm constructs a tree that satisfies (13.21). Initially, the trivial tree of depth 1 satisfies (13.21). Also, if we start with a tree that satisfies (13.21), and expand the leaf with the highest probability, we still satisfy (13.21), since the new internal node has a probability that is at least as high as any other leaf, and the new leaves have a lower probability that the original leaf node that was expanded to form the new internal node. Thus the new tree also satisfies (13.21), and by induction, the tree constructed by the Tunstall algorithm satisfies (13.21). Combining this with the previous result, we see that the tree constructed by the Tunstall algorithm has maximal average length, and therefore minimizes $R(A_D)$.

(c) We will use the familiar Huffman coding procedure and "invert" it to construct the variable to fixed code which achieves a compression ratio within one bit of the entropy.

First, we take a blocks of length 2 for the random variable, ie. $X_1X_2 \in \mathcal{X}^2$ and construct an Huffman code for this pair. By the standard results for Huffman codes, we have

$$2H < EL_2 < 2H + 1 \tag{13.24}$$

Let l_m be the maximal length of any of the Huffman codewords.

Now consider the set of binary sequences of length n, $n >> l_m$. Parse each binary sequence to codewords from Huffman code, and replace the codewords by the corresponding pair of symbols of X. This defines a set of sequences of X, which we will like to use to construct A_D. This set of sequences might not correspond to a complete tree for X. We therefore add to this set by adding the X sequences that correspond to siblings of the X sequences already chosen. This augmented set will be the A_D that we will use in our analysis.

We now show that for an appropriate choice of n large enough, this choice of A_D achieves an compression rate less than $H + 1$.

The number of elements in A_D: It is not difficult to see the code for any sequence in A_D is less than $n + l_m$, and thus the number of sequences in A_D is less than 2^{n+l_m}.

The average length of the sequences in A_D: Using renewal theory, it follows that the expected number of Huffman codewords in the parsing of a binary sequence of length n converges to n/L_2. Thus the average length of the X sequences corresponding to the parsed binary sequences converges to $2n/L_2$, since each Huffman codeword corresponds to a block of two symbols of X.

The fact that we have added sequences to this set to form A_D does not change this result, and we can prove this by considering the parsing of binary sequences of length $n - l_m$ and $n + l_m$. The details of this analysis are omitted.

Thus we have the expected length of sequences of A_D converges to $2n/L_2$ as $n \to \infty$. Thus the compression ratio,

$$R(A_D) = \frac{\log D}{EL(A_D)} \tag{13.25}$$

is upper bounded by $(n + l_m)/(2n/L_2 + \epsilon)$ for n large enough. This converges to $H + 1/2$ as $n \to \infty$, and thus there exists an n such that we can achieve a compression ratio less than $H + 1$. This proves the required result.

Chapter 14

Kolmogorov Complexity

1. **Kolmogorov complexity of two sequences.** Let $x, y \in \{0,1\}^*$. Argue that $K(x,y) \leq K(x) + K(y) + c$.

   Solution: Suppose we are given two sequences, x and y, with Kolmogorov complexity $K(x)$ and $K(y)$ respectively. Then there must exist programs p_x and p_y, of length $K(x)$ and $K(y)$ respectively, which print out x and y. Form the following program:

   ```
   Run the following two programs, not halting after the first;
   Run the program $p_x$, interpreting the halt as a jump to the next step;
   Run the program $p_y$.
   ```

   This program, of length $K(x) + K(y) + c$, prints out x, y. Hence

   $$K(x,y) \leq K(x) + K(y) + c. \tag{14.1}$$

2. **Complexity of the sum.**

   (a) Argue that $K(n) \leq \log n + 2 \log \log n + c$.

   (b) Argue that $K(n_1 + n_2) \leq K(n_1) + K(n_2) + c$.

   (c) Give an example in which n_1 and n_2 are complex but the sum is relatively simple.

   Solution:

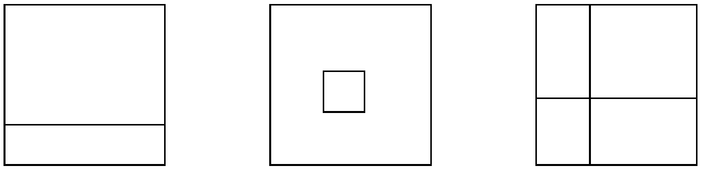
   (a) To describe an integer n, we will tell the computer the length of n, and then tell it the bits of n. Thus the program will be self delimiting. To represent the length of n, i.e., $\log n$, we could use the simple code described in class: repeat each bit of $\log n$ twice, and end the description by 10. This representation requires $2 \log \log n + 2$ bits. It requires $\log n$ bits to represent the bits of n, and hence the total length of the program is $\log n + 2 \log \log n + c$, which is an upper bound on the complexity of n:

   $$K(n) \leq \log n + 2 \log \log n + c. \tag{14.2}$$

(b) Given two programs to print out n_1 and n_2, we can modify them so that they write on the work tape, rather than the output tape. Then we can add an instruction to add the two numbers together and print them out. The length of this program is $K(n_1) + K(n_2) + c$, and hence

$$K(n_1 + n_2) \leq K(n_1) + K(n_2) + c. \tag{14.3}$$

3. **Images.** Consider an $n \times n$ array x of 0's and 1's . Thus x has n^2 bits.



Find the Kolmogorov complexity $K(x \mid n)$ (to first order) if

(a) x is a horizontal line.

(b) x is a square.

(c) x is the union of two lines, each line being vertical or horizontal.

Solution:

(a) The program to print out an image of one horizontal line is of the form

```
For $1 \leq i \leq n$ { Set pixels on row $i$ to 0; }
Set pixels on row $r$ to 1;
Print out image.
```

Since the computer already knows n, the length of this program is $K(r|n) + c$, which is $\leq \log n + c$. Hence, the Kolmogorov complexity of a line image is

$$K(\text{line}|n) \leq \log n + c. \tag{14.4}$$

(b) For a square, we have to tell the program the coordinates of the top left corner, and the length of the side of the square. This requires no more than $3 \log n$ bits, and hence

$$K(\text{square}|n) \leq 3 \log n + c. \tag{14.5}$$

However, we can save some description length by first describing the length of the side of the square and then the coordinates. Knowing the length of the side of the square reduces the range of possible values of the coordinates. Even better, we can count the total number of such squares. There is one $n \times n$ square, four $(n-1) \times (n-1)$ squares, nine $(n-2) \times (n-2)$ squares, etc. The total number of squares is

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6} \approx \frac{n^3}{3}. \tag{14.6}$$

Since we can give the index of a square in a lexicographic ordering,

$$K(\text{square}|n) \leq \log \frac{n^3}{3} + c. \tag{14.7}$$

(c) In this case, we have to tell the program the position of the horizontal line and the position of the vertical line, requiring no more than $2\log n$ bits. Hence

$$K(\text{pair of lines}|n) \leq 2\log n + c. \tag{14.8}$$

In all the above cases, there are many images which are much simpler to describe. For example, in the case of the horizontal line image, the image of the first line or the middle line is much easier to describe. However most of the images have description lengths close to the bounds derived above.

4. **Do computers reduce entropy?** Feed a random program P into an universal computer. What is the entropy of the corresponding output? Specifically, let $X = \mathcal{U}(P)$, where P is a Bernoulli(1/2) sequence. Here the binary sequence X is either undefined or is in $\{0,1\}^*$. Let $H(X)$ be the Shannon entropy of X. Argue that $H(X) = \infty$. Thus although the computer turns nonsense into sense, the output entropy is still infinite.

Solution: *Do computers reduce entropy?* The output probability distribution on strings x is $P_{\mathcal{U}}(x)$, the universal probability of the string x. Thus, by the arguments following equation (7.65), the output distribution includes a mixture of all computable probability distributions.

Consider the following distribution on binary finite length stings:

$$P_1(x) = \begin{cases} \frac{1}{An\log^2 n} & \text{if } x = \underbrace{111\ldots1}_{n \text{ 1's}}0 \\ 0 & \text{otherwise} \end{cases} \tag{14.9}$$

where $A = \sum_{n=1}^{\infty} \frac{1}{n\log^2 n}$ is chosen to ensure that $\sum_x P_1(x) = 1$. Then $P_1(x)$ is a computable probability distribution, and by problem 9 in Chapter 2, $P_1(x)$ has an infinite entropy.

By (7.65) in the text,

$$P_{\mathcal{U}}(x) \geq c_1 P_1(x) \tag{14.10}$$

for some constant c_1 that does not depend on x. Let

$$P_2(c) = \frac{P_{\mathcal{U}}(x) - c_1 P_1(x)}{1 - c_1}. \tag{14.11}$$

It is easy to see that $\sum_x P_2(x) = 1$, and therefore $P_2(x)$ is a probability distribution. Also,

$$P_{\mathcal{U}}(x) = c_1 P_1(x) + (1 - c_1)P_2(x). \tag{14.12}$$

By the results of Chapter 2, $-t\log t$ is a concave function of t and therefore

$$-P_{\mathcal{U}}(x)\log P_{\mathcal{U}}(x) \geq -c_1 P_1(x)\log P_1(x) - (1-c_1)P_2(x)\log P_2(x) \tag{14.13}$$

Summing this over all x, we obtain

$$H(P_{\mathcal{U}}) \geq c_1 H(P_1) + (1-c_1)H(P_2) = \infty \tag{14.14}$$

Thus the entropy at the output of a universal computer fed in Bernoulli(1/2) sequences is infinite.

5. **Monkeys on a computer.** Suppose a random program is typed into a computer. Give a rough estimate of the probability that the computer prints the following sequence:

   (a) 0^n followed by any arbitrary sequence.

   (b) $\pi_1\pi_2\ldots\pi_n$ followed by any arbitrary sequence, where π_i is the i-th bit in the expansion of π.

   (c) $0^n 1$ followed by any arbitrary sequence.

   (d) $\omega_1\omega_2\ldots\omega_n$ followed by any arbitrary sequence.

   (e) A proof of the four color theorem.

   Solution: The probability that a computer with a random input will print will print out the string x followed by any arbitrary sequence is the sum of the probabilities over all sequences starting with the string x.

$$p_{\mathcal{U}}(x\ldots) = \sum_{y\in\{0,1\}^*\cup\{0,1\}^\infty} p_{\mathcal{U}}(xy), \text{ where } p_{\mathcal{U}}(x) = \sum_{p:\mathcal{U}(p)=x} 2^{-\ell(p)}. \tag{14.15}$$

   This sum is lower bounded by the largest term, which corresponds to the simplest concatenated sequence.

   (a) The simplest program to print a sequence that starts with n 0's is

```
Print 0's forever.
```

   This program has constant length c and hence the probability of strings starting with n zeroes is

$$p_{\mathcal{U}}(0^n\ldots) \approx 2^{-c}. \tag{14.16}$$

   (b) Just as in part (a), there is a short program to print the bits of π forever. Hence

$$p_{\mathcal{U}}(\pi_1\pi_2\ldots\pi_n\ldots) \approx 2^{-c}. \tag{14.17}$$

   (c) A program to print out n 0's followed by a 1 must in general specify n. Since most integers n have a complexity $\approx \log^* n$, and given n, the program to print out $0^n 1$ is simple, we have

$$p_{\mathcal{U}}(0^n 1\ldots) \approx 2^{-\log^* n - c}, \tag{14.18}$$

(d) We know that n bits of Ω are essentially incompressible, i.e., their complexity $\geq n-c$. Hence, the shortest program to print out n bits of Ω followed by anything must have a length at least $n-c$, and hence

$$p_{\mathcal{U}}(\omega_1\omega_2\ldots\omega_n\ldots) \approx 2^{-(n-c)}. \tag{14.19}$$

6. **Kolmogorov complexity and ternary programs.**
Suppose that the input programs for a universal computer $\mathcal{U}$ are sequences in $\{0,1,2\}^*$ (ternary inputs). Also, suppose $\mathcal{U}$ prints ternary outputs. Let $K(x|l(x)) = \min_{U(p,l(x))=x} l(p)$. Show that

(a) $K(x^n|n) \leq n+c$.

(b) $|x^n \in \{0,1\}^* : K(x^n|n) < k| < 3^k$.

Solution: *Kolmogorov Complexity and Ternary Programs.*

(a) It is always possible to include a ternary representation of the string to be printed out in the program. This program has a length of $n+c$ ternary digits, and therefore $K(x^n|n) \leq n+c$.

(b) There are less than 3^k ternary programs of length less than k and each of these programs can produce at most one output string and therefore the number of strings with Kolmogorov complexity less than k has to be less than 3^k.

7. **A law of large numbers.** Using ternary inputs and outputs as in Problem 6, outline an argument demonstrating that if a sequence x is algorithmically random, i.e., if $K(x|l(x)) \approx l(x)$, then the proportion of 0's, 1's, and 2's in x must each be near $1/3$.

It may be helpful to use Stirling's approximation $n! \approx (n/e)^n$.

Solution: *A Law of Large Numbers.* The arguments parallel the arguments in the binary case in Theorem 7.5.2. We will only outline the main argument. Let $\theta_0, \theta_1, \theta_2$ be the proportions of 0's, 1's, and 2's in the string x^n. We can construct a two stage description of x^n by first describing $\theta_0, \theta_1, \theta_2$, and then describing the string within the set of all strings with the same proportions of 0,1 and 2. The two stage description has a length bounded by $nH_3(\theta_0,\theta_1,\theta_2) + 6\log n + c$, where H_3 denotes entropy to base 3. If $K(x^n|n) \approx n$, then

$$n - c_n \geq K(x^n|n) \geq nH_3(\theta_0,\theta_1,\theta_2) + 6\log n + c, \tag{14.20}$$

and therefore

$$H_3(\theta_0,\theta_1,\theta_2) \geq 1 - \delta_n, \tag{14.21}$$

where $\delta_n \to 0$. Thus $\theta_0, \theta_1, \theta_2$ must lie in a neighborhood of $(\frac{1}{3},\frac{1}{3},\frac{1}{3})$. This can be seen by considering the behavior of the entropy function—it is close to 1 only in the neighborhood of the center of the three dimensional simplex. Therefore, the proportion of 0's, 1's and 2's must be close to 1/3 for an incompressible ternary sequence.

8. **Image complexity**.
Consider two binary subsets A and B (of an $n \times n$ grid). For example,

Find general upper and lower bounds, in terms of $K(A|n)$ and $K(B|n)$, for

(a) $K(A^c|n)$.
(b) $K(A \cup B|n)$.
(c) $K(A \cap B|n)$.

Solution: *Image Complexity.*

(a) We can describe A^c by first describing A, so

$$K(A^c|n) < K(A|n) + c \tag{14.22}$$

(b) We can describe the union by describing each set separately and taking the union, hence

$$K(A \cup B|n) \leq K(A|n) + K(B|n) + c \tag{14.23}$$

(c) The intersection can also be described similarly, and hence

$$K(A \cap B|n) \leq K(A|n) + K(B|n) + c \tag{14.24}$$

9. **Random program.** Suppose that a random program (symbols i.i.d. uniform over the symbol set) is fed into the nearest available computer.

To our surprise the first n bits of the binary expansion of $1/\sqrt{2}$ are printed out. Roughly what would you say the probability is that the next output bit will agree with the corresponding bit in the expansion of $1/\sqrt{2}$?

Solution: *Random program.* The arguments parallel the argument in Section 7.10, and we will not repeat them. Thus the probability that the next bit printed out will be the next bit of the binary expansion of $\sqrt{2}$ is $\approx \frac{1}{cn+1}$.

10. **The face-vase illusion.**

(a) What is an upper bound on the complexity of a pattern on an $m \times m$ grid that has mirror image symmetry about a vertical axis through the center of the grid *and* consists of horizontal line segments?

(b) What is the complexity K if the image differs in one cell from the pattern described above?

Solution: *The face vase illusion.*

(a) An image with mirror image symmetry has only $m^2/2$ independent pixels. We can describe only one half and ask the computer to construct the other. Therefore the Kolmogorov complexity of the image is less than $m^2/2 + c$.

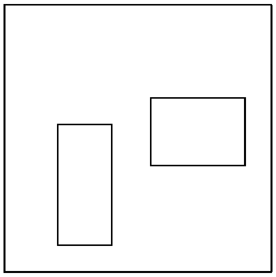
The fact that the image consists of horizontal line segments will not make a difference unless we are given some further restrictions on the line segments. For example, in the image with the face-vase illusion, each half of any horizontal line consists of only two segments, one black and one white. In this case, we can describe the image by a sequence of boundary points between the black and the white. Thus the image will take $m \log(\frac{m}{2}) + c$ bits to describe the m boundary points in one half of the picture (the boundary points on the other half can be calculated from this half). Thus the image with the face-vase illusion has a Kolmogorov complexity less than $m \log m + c$.

(b) We can describe a picture that differs in one pixel from the image above by first describing the above image, and then giving the location of the pixel that is different. Therefore, the Kolmogorov complexity of the new image is less than $m \log m + 2 \log m + c$.

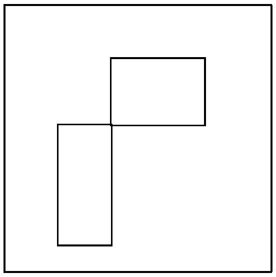
11. **Kolmogorov complexity**

Assume n very large and known. Let all rectangles be parallel to the frame.

(a) What is the (maximal) Kolmogorov complexity of the union of two rectangles on an $n \times n$ grid?



(b) What if the rectangles intersect at a corner?



(c) What if they have the same (unknown) shape?

(d) What if they have the same (unknown) area?

(e) What is the minimum Kolmogorov complexity of the union of two rectangles? That is, what is the simplest union?

(f) What is the (maximal) Kolmogorov complexity over all images (not necessarily rectangles) on an $n \times n$ grid?

Solution: *Kolmogorov complexity* Note that $K(\text{a single point on the screen}|n) \approx 2\log n + c$.

(a) To specify two rectangles, we need to describe the coordinates of two corners (X, Y) and either length and width (L, W) or the opposite corner (X', Y') of the rectangle. Hence we will need to describe 4 numbers, each of which is $\leq n$, and therefore we need $4\log n + c$ bits for each rectangle, for a total of $8\log n + c$ for two rectangles.

With the upper-left corner and the lower-right corner, we can describe a rectangle. Hence, for two rectangles, $K(x|n) \approx K(4 \text{ points}|n) \approx 8\log n + c$.

We have not used the fact that the length and width of the rectangle are not independent of the position of the lower left corner–for example, if the lower left corner is near the NE corner of the square, the length and width of the rectangle have to be small. This will reduce the number of possible $(X, Y, L,)$ combinations to be $(n(n+1)/2)^2$ rather than n^4, but it does not change the key term.

(b) Assuming two rectangles meet at a corner, we need to only describe 3 corners instead of 4. Hence, $K(x|n) \approx K(3 \text{ points}|n) \approx 6\log n + c$.

(c) Assuming two rectangles of the same shape, we need to describe the upper-left and lower right corners of one rectangle and the one corner of the other. Hence, $K(x|n) \approx K(3 \text{ points}|n) \approx 6\log n + c$.

(d) If the rectangles have the same area, then describing one rectangle fully and the other rectangle by one corner and one side (the other side can be calculated). Thus $K(x|n) \approx K(3 \text{ points}|n) + K(1 \text{ side}|n) = 7\log n + c$.

(e) An image is a specification for each pixel whether it is black or white. Since there are n^2 pixels in the image, 1 bit per pixel, the maximal Kolmogorov complexity is $n^2 + c$ bits.

12. **Encrypted text**

Suppose English text x^n is encrypted into y^n by a substitution cypher: a 1 to 1 reassignment of each of the 27 letters of the alphabet (A-Z including the space character) to itself. Suppose the Kolmogorov complexity of the text x^n is $K(x^n) = \frac{n}{4}$. (This is about right for English text. We're now assuming a 27-symbol programming language, instead of a binary symbol-set for the programming language. So, the length of the shortest program, using a 27-ary programming language, that prints out a particular string of English text of length n, is approximately n/4.)

(a) What is the Kolmogorov complexity of the encryption map?

(b) Estimate the Kolmogorov complexity of the encrypted text y^n.

(c) How high must n be before you would expect to be able to decode y^n?

Solution: *Encrypted text*

(a) There are 27! encryption maps. To describe one of them requires in general $\log 27!$ symbols. Note that the question implicitly assumes that we are using a 27-symbol programming language, so the log here is to base 27.

(b) The complexity of the encrypted text cannot be worse than the complexity of the English text plus the complexity of the encryption map (plus some small constant).

(c) The idea here is that in order to be able to decode the encrypted text, the length of the encrypted string, n, must be greater than $n/4 + \log 27!$. Why? Because short strings have short programs that print them out simply by writing including the text verbatim and saying “Print this”. This does not take advantage of the structure of the text, but the text is so short that there isn't really enough structure to take advantage of. *Any* random sequence of symbols of length n can always be printed out by a program of length n (+ c), so if n is less than $\log 27!$ the overhead of expressing it as an encryption of English is higher than including it as verbatim data. It is only as the string grows to length appreciably greater than $log 27!$ that the overhead of expressing it as the encryption of English text becomes neglible. Now the structure starts to dominate.

It should be pointed out that this is only the beginning of an idea about the relationship between encrypted text and the ability to uniquely decipher it. Shannon studied the relationship between encryption and complexity in [12].

13. **Kolmogorov complexity.**

Consider the Kolmogorov complexity $K(n)$ over the integers n. If a specific integer n_1 has a low Kolmogorov complexity $K(n_1)$, by how much can the Kolmogorov complexity $K(n_1 + k)$ for the integer $n_1 + k$ vary from $K(n_1)$?

Solution: *Kolmogorov complexity.*

Since we can describe $n+k$ by describing n and then describing k and then adding them, $K(n+k) \leq K(n) + \log k + c$, since the descriptive complexity of k is less the $\log k$. Similarly, if the complexity of $n+k$ is small, we cand describe n by describing $n+k$ and k, and therefore $K(n) < K(n+k) + \log k + c$. Thus we have $|K(n+k) - K(n)| \leq \log k + c$.

14. **Complexity of large numbers.** Let $A(n)$ be the set of positive integers x for which a terminating program p of length less than or equal to n bits exists that outputs x. Let $B(n)$ be the complement of $A(n)$, i.e., $B(n)$ is the set of integers x for which no program of length less than or equal to n outputs x. Let $M(n)$ be the maximum of $A(n)$ and let $S(n)$ be the minimum of $B(n)$. What is the Kolmogorov complexity $K(M(n))$ (approximately)? What is $K(S(n))$ (approximately)? Which is larger ($M(n)$ or $S(n)$)? Give a reasonable lower bound on $M(n)$ and a reasonable upper bound on $S(n)$.

    Solution: *Complexity of large numbers.*

    Clearly since we can specify the program that printed out $M(n)$ with length less than n, the Kolmogorov complexity of $M(n)$ is less than n. The description "largest number that is printed out by a program of less than n bits" does not give rise to an effective program to compute $M(n)$, because even though we can simulate in parallel all programs of length less than n, we will never know when we have found $M(n)$. Thus a good bound on $K(M(n)) \approx n$.

    While $S(n)$ does not have a program of length less than n to compute it, and therefore $K(S(n)) > n$, we know that since it is the smallest such number, $S(n)-1$ has a short program of length less than n. Therefore we can describe $S(n)$ by describing $S(n)-1$ and the difference, and the complexity $K(S(n)) \approx n$.

    $M(n)$ is likely to be much much larger than $S(n)$ since we can describe very very large numbers with short programs (e.g. iterated exponentials) $S(n)$ on the other hand is a boring small number.

    $M(n)$ could be very large, and a good lower bound is an iterated exponential, i.e., $2^{2^{\cdots}}$, where the iteration is done n times. $S(n)$ on the other hand cannot be less than 2^n since all numbers less than 2^n have description lengths less than n. However since there are not enough short programs, the numbers above 2^n is likely to have complexity greater than n, and so $S(n) \approx 2^n$.

Chapter 15

Network Information Theory

1. **The cooperative capacity of a multiple access channel.** (Figure 15.1)

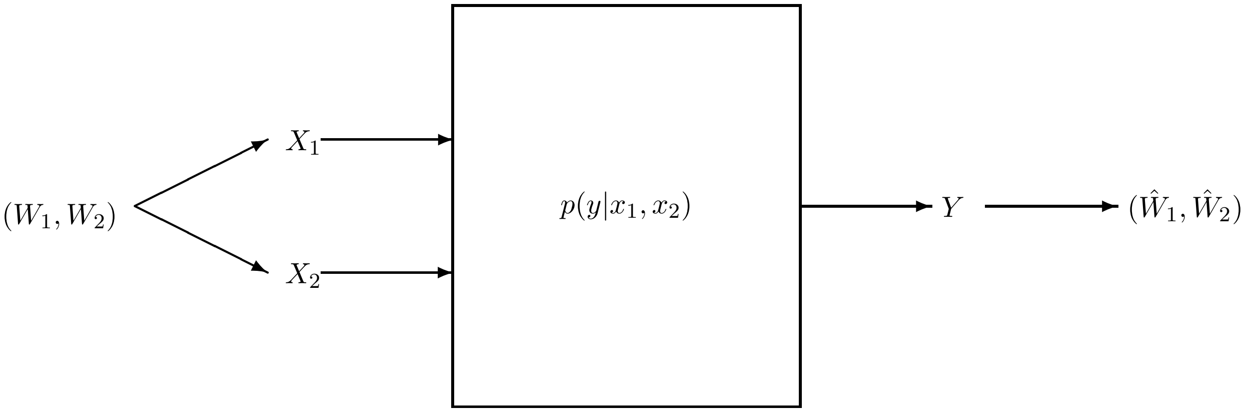


Figure 15.1: Multiple access channel with cooperating senders.

(a) Suppose X_1 and X_2 have access to *both* indices $W_1 \in \{1, 2^{nR}\}, W_2 \in \{1, 2^{nR_2}\}$. Thus the codewords $\mathbf{X}_1(W_1, W_2), \mathbf{X}_2(W_1, W_2)$ depend on both indices. Find the capacity region.

(b) Evaluate this region for the binary erasure multiple access channel $Y = X_1 + X_2, X_i \in \{0, 1\}$. Compare to the non-cooperative region.

Solution: *Cooperative capacity of multiple access channel*

(a) When both senders have access to the pair of messages to be transmitted, they can act in concert. The channel is then equivalent to a single user channel with input alphabet $\mathcal{X}_1 \times \mathcal{X}_2$, and a larger message set $W_1 \times W_2$. The capacity of this single user channel is $C = \max_{p(x)} I(X;Y) = \max_{p(x_1,x_2)} I(X_1,X_2;Y)$. The two senders can send at any combination of rates with the total rate

$$R_1 + R_2 \leq C \tag{15.1}$$

(b) The capacity for the binary erasure multiple access channel was evaluated in class. When the two senders cooperate to send a common message, the capacity is

$$C = \max_{p(x_1,x_2)} I(X_1,X_2;Y) = \max H(Y) = \log 3, \tag{15.2}$$

achieved by (for example) a uniform distribution on the pairs, (0,0), (0,1) and (1,1). The cooperative and non-cooperative regions are illustrated in Figure 15.2.

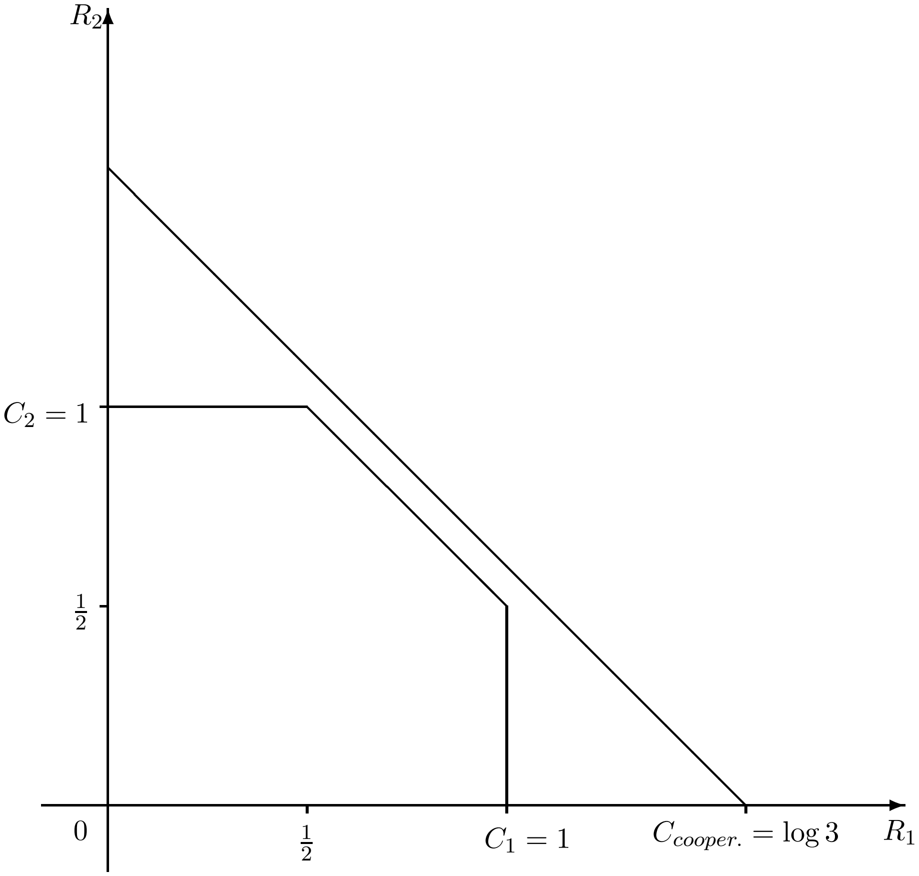


Figure 15.2: Cooperative and non-cooperative capacity for a binary erasure multiple access channel

2. **Capacity of multiple access channels.** Find the capacity region for each of the following multiple access channels:

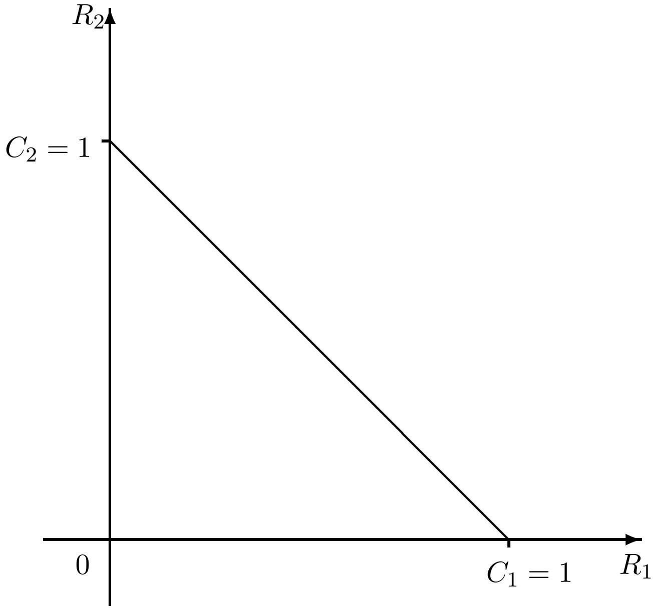


Figure 15.3: Capacity region of additive modulo 2 MAC

(a) Additive modulo 2 multiple access access channel. $X_1 \in \{0,1\}, X_2 \in \{0,1\}, Y = X_1 \oplus X_2$.

(b) Multiplicative multiple access channel. $X_1 \in \{-1,1\}, X_2 \in \{-1,1\}, Y = X_1 \cdot X_2$.

Solution: *Examples of multiple access channels.*

(a) *Additive modulo 2 MAC.*

$Y = X_1 \oplus X_2$. Quite clearly we cannot send at a total rate of more than 1 bit, since $H(Y) \leq 1$. We can achieve a rate of 1 bit from sender 1 by setting $X_2 = 0$, and similarly we can send 1 bit/transmission from sender 2. By simple time sharing we can achieve the entire capacity region which is shown in Figure 15.3.

(b) *Multiplier channel.*

$X_1, X_2 \in \{-1,1\}, Y = X_1.X_2$.

This channel is equivalent to the previous channel with the mapping $-1 \rightarrow 1$ and $1 \rightarrow 0$. Hence the capacity region is the same as the previous channel.

3. **Cut-set interpretation of capacity region of multiple access channel.** For the multiple access channel we know that (R_1, R_2) is achievable if

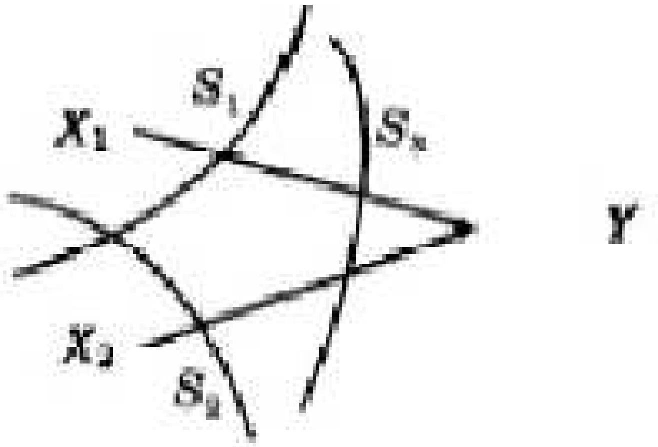
$$R_1 < I(X_1; Y \mid X_2), \quad (15.3)$$

$$R_2 < I(X_2; Y \mid X_1), \quad (15.4)$$

$$R_1 + R_2 < I(X_1, X_2; Y), \quad (15.5)$$

for X_1, X_2 independent. Show, for X_1, X_2 independent, that

$$I(X_1; Y \mid X_2) = I(X_1; Y, X_2).$$



Interpret the information bounds as bounds on the rate of flow across cutsets S_1, S_2 and S_3.

Solution: *Cutset interpretation of the capacity region.*

We can interpret $I(X_1; Y, X_2)$ as the maximum amount of information that could flow across the cutset S_1. This is an upper bound on the rate R_1. Similarly, we can interpret the other bounds.

4. **Gaussian multiple access channel capacity.** For the AWGN multiple access channel, prove, using typical sequences, the achievability of any rate pairs (R_1, R_2) satisfying

$$\begin{aligned} R_1 &< \frac{1}{2}\log(1+\frac{P_1}{N}), && (15.6)\\ R_2 &< \frac{1}{2}\log(1+\frac{P_2}{N}), && (15.7)\\ R_1+R_2 &< \frac{1}{2}\log(1+\frac{P_1+P_2}{N}). && (15.8)\end{aligned}$$

The proof extends the proof for the discrete multiple access channel in the same way as the proof for the single user Gaussian channel extends the proof for the discrete single user channel.

Solution: *Gaussian Multiple Access Channel Capacity.*

The essence of the proof of the achievability of the capacity region for the Gaussian multiple access channel is the same as the discrete multiple access channel. The main difference is the introduction of the power constraint, and the modifications that have to be made to ensure that the codewords satisfy the power constraint with high probability. We will briefly outline the proof of achievability along the lines of the proof in the discrete cases, pausing only to emphasize the differences.

The channel is defined by

$$Y = X_1 + X_2 + Z, \quad Z \sim \mathcal{N}(0, N) \tag{15.9}$$

with power constraints P_1 and P_2 on the inputs. The achievable rates for this channel are

$$\begin{aligned} R_1 &< C\left(\frac{P_1}{N}\right) && (15.10)\\ R_2 &< C\left(\frac{P_2}{N}\right) && (15.11)\\ R_1 + R_2 &< C\left(\frac{P_1 + P_2}{N}\right), && (15.12) \end{aligned}$$

where

$$C(x) = \frac{1}{2}\log(1 + x). \tag{15.13}$$

Codebook generation: Generate 2^{nR_1} independent codewords $\mathbf{X}_1(w_1)$, $w_1 \in \{1, 2, \ldots, 2^{nR_1}\}$, of length n, generating each element i.i.d. $\sim \mathcal{N}(0, P_1 - \epsilon)$ Similarly generate 2^{nR_2} independent codewords $\mathbf{X}_2(w_2)$, $w_2 \in \{1, 2, \ldots, 2^{nR_2}\}$, generating each element i.i.d. $\sim \mathcal{N}(0, P_2 - \epsilon)$. These codewords form the codebook.

Encoding: To send index w_1, sender one sends the codeword $\mathbf{X}_1(w_1)$. Similarly, to send w_2, sender 2 sends $\mathbf{X}_2(w_2)$.

Decoding: The receiver Y^n chooses the pair (i, j) such that

$$(\mathbf{x}_1(i), \mathbf{x}_2(j), \mathbf{y}) \in A_\epsilon^{(n)} \tag{15.14}$$

$$\frac{1}{n}\sum_{k=1}^{n} x_{1k}^2(i) \le P_1 \tag{15.15}$$

and

$$\frac{1}{n}\sum_{k=1}^{n} x_{2k}^2(j) \le P_2 \tag{15.16}$$

if such a pair (i, j) exists and is unique; otherwise, an error is declared.

By the symmetry of the random code construction, the conditional probability of error does not depend on which pair of indices is sent. So, without loss of generality, we can assume that $(w_1, w_2) = (1, 1)$.

An error occurs in the decoding if

- $(\mathbf{x}_1(1), \mathbf{x}_2(1)) \notin A_\epsilon^{(n)}$,
- $(\mathbf{x}_1(i), \mathbf{x}_2(j)) \in A_\epsilon^{(n)}$ for some $i \neq 1$ or $j \neq 1$, or

- $\mathbf{x}_1(1)$ or $\mathbf{x}_2(1)$ do not satisfy the power constraint.

Define the events

$$E_{01} = \{\frac{1}{n}\sum_{k=1}^{n} X_{1k}^2(1) > P_1\} \tag{15.17}$$

and

$$E_{02} = \{\frac{1}{n}\sum_{k=1}^{n} X_{2k}^2(1) > P_2\}. \tag{15.18}$$

For $i \neq 0, j \neq 0$,

$$E_{ij} = \{(\mathbf{X}_1(i), \mathbf{X}_2(j), \mathbf{Y}) \in A_\epsilon^{(n)}\}. \tag{15.19}$$

Then by the union of events bound,

$$\begin{aligned} P_e^{(n)} &= P\left(E_{01}\bigcup E_{02}\bigcup E_{11}^c \bigcup \cup_{(i,j)\neq(1,1)} E_{ij}\right) \quad (15.20)\\ &\leq P(E_{01}) + P(E_{02}) + P(E_{11}^c) + \sum_{i\neq 1,\, j=1} P(E_{i1}) + \sum_{i=1,\, j\neq 1} P(E_{1j}) + \sum_{i\neq 1,\, j\neq 1} P(E_{ij}), \end{aligned}$$

where P is the probability given that $(1,1)$ was sent. Since we choose the codewords according to a normal distribution with mean $P_i - \epsilon$, with very high probability the codeword power will be less than P. Hence, $P(E_{01}) \to 0$ and $P(E_{02}) \to 0$. From the AEP, $P(E_{11}^c) \to 0$. By the AEP, for $i \neq 1$, we have

$$\begin{aligned} P(E_{i1}) &= P((\mathbf{X}_1(i), \mathbf{X}_2(1), \mathbf{Y}) \in A_\epsilon^{(n)}) && (15.21)\\ &= \int_{(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y})\in A_\epsilon^{(n)}} f(\mathbf{x}_1) f(\mathbf{x}_2, \mathbf{y}) && (15.22)\\ &\leq 2^{-n(h(X_1)+h(X_2,Y)-h(X_1,X_2,Y)-3\epsilon)} && (15.23)\\ &= 2^{-n(I(X_1;X_2,Y)-3\epsilon)} && (15.24)\\ &= 2^{-n(I(X_1;Y|X_2)-3\epsilon)} && (15.25)\\ &= 2^{-n(C(\frac{P_1}{N})-3\epsilon)}, && (15.26) \end{aligned}$$

since X_1 and X_2 are independent, and therefore $I(X_1;X_2,Y) = I(X_1;X_2) + I(X_1;Y|X_2) = I(X_1;Y|X_2)$.

Similarly, for $j \neq 1$,

$$P(E_{1j}) \leq 2^{-n(C(\frac{P_2}{N})-3\epsilon)}, \tag{15.27}$$

and for $i \neq 1, j \neq 1$,

$$P(E_{ij}) \leq 2^{-n(C(\frac{P_1+P_2}{N})-4\epsilon)}. \tag{15.28}$$

It follows that

$$\begin{aligned} P_e^{(n)} \leq\ & P(E_{01}) + P(E_{02}) + P(E_{11}^c) + 2^{nR_1} 2^{-n(C(\frac{P_1}{N})-3\epsilon)}\\ & + 2^{nR_2} 2^{-n(C(\frac{P_2}{N})-3\epsilon)} + 2^{n(R_1+R_2)} 2^{-n(C(\frac{P_1+P_2}{N})-4\epsilon)}. \end{aligned} \tag{15.29}$$

Thus $\epsilon > 0$ arbitrary and the conditions of the theorem cause each term to tend to 0 as $n \to \infty$.

The above bound shows that the average probability of error, averaged over all choices of codebooks in the random code construction, is arbitrarily small. Hence there exists at least one code $\mathcal{C}^*$ with arbitrarily small probability of error.

The achievability of the capacity region is proved.

5. **Converse for the Gaussian multiple access channel.** Prove the converse for the Gaussian multiple access channel by extending the converse in the discrete case to take into account the power constraint on the codewords.

Solution: *Converse for the Gaussian multiple access channel.* The proof of the converse for the Gaussian case proceeds on very similar lines to the discrete case. However, for the Gaussian case, the two stages of proof that were required in the discrete case, namely, of finding a new expression for the capacity region and then proving a converse, can be combined into one single step.

By the code construction, it is possible to estimate (W_1, W_2) from the received sequence Y^n with a low probability of error. Hence the conditional entropy of (W_1, W_2) given Y^n must be small. By Fano's inequality,

$$H(W_1, W_2|Y^n) \leq n(R_1 + R_2)P_e^{(n)} + H(P_e^{(n)}) \stackrel{\triangle}{=} n\epsilon_n. \tag{15.30}$$

It is clear that $\epsilon_n \to 0$ as $P_e^{(n)} \to 0$.

Then we have

$$\begin{aligned} H(W_1|Y^n) &\leq H(W_1, W_2|Y^n) \leq n\epsilon_n, & (15.31)\\ H(W_2|Y^n) &\leq H(W_1, W_2|Y^n) \leq n\epsilon_n. & (15.32)\end{aligned}$$

We can now bound the rate R_1 as

$$\begin{aligned}
nR_1 &= H(W_1) & (15.33)\\
&= I(W_1; Y^n) + H(W_1|Y^n) & (15.34)\\
&\stackrel{(a)}{\leq} I(W_1; Y^n) + n\epsilon_n & (15.35)\\
&\stackrel{(b)}{\leq} I(X_1^n(W_1); Y^n) + n\epsilon_n & (15.36)\\
&= H(X_1^n(W_1)) - H(X_1^n(W_1)|Y^n) + n\epsilon_n & (15.37)\\
&\stackrel{(c)}{\leq} H(X_1^n(W_1)|X_2^n(W_2)) - H(X_1^n(W_1)|Y^n, X_2^n(W_2)) + n\epsilon_n & (15.38)\\
&= I(X_1^n(W_1); Y^n|X_2^n(W_2)) + n\epsilon_n & (15.39)\\
&= h(Y^n|X_2^n(W_2)) - h(Y^n|X_1^n(W_1), X_2^n(W_2)) + n\epsilon_n & (15.40)\\
&\stackrel{(d)}{=} h(Y^n|X_2^n(W_2)) - h(Z^n|X_1^n(W_1), X_2^n(W_2)) + n\epsilon_n & (15.41)\\
&\stackrel{(e)}{=} h(Y^n|X_2^n(W_2)) - h(Z^n) + n\epsilon_n & (15.42)
\end{aligned}$$

$$\begin{aligned}
&\overset{(f)}{=} h(Y^n|X_2^n(W_2)) - \sum_{i=1}^n h(Z_i) + n\epsilon_n && (15.43)\\
&\overset{(g)}{\le} \sum_{i=1}^n h(Y_i|X_2^n(W_2)) - \sum_{i=1}^n h(Z_i) + n\epsilon_n && (15.44)\\
&\overset{(h)}{\le} \sum_{i=1}^n h(Y_i|X_{2i}) - \sum_{i=1}^n h(Z_i) + n\epsilon_n && (15.45)\\
&\overset{(i)}{=} \sum_{i=1}^n h(X_{1i}+Z_i|X_{2i}) - \sum_{i=1}^n h(Z_i) + n\epsilon_n && (15.46)\\
&\overset{(j)}{=} \sum_{i=1}^n h(X_{1i}+Z_i) - \sum_{i=1}^n h(Z_i) + n\epsilon_n && (15.47)\\
&\overset{(k)}{\le} \sum_{i=1}^n \frac{1}{2}\log 2\pi e(P_{1i}+N) - \frac{1}{2}\log 2\pi eN + n\epsilon_n && (15.48)\\
&= \sum_{i=1}^n \frac{1}{2}\log\left(1+\frac{P_{1i}}{N}\right) + n\epsilon_n && (15.49)
\end{aligned}$$

where
(a) follows from Fano's inequality,
(b) from the data processing inequality,
(c) from the fact that since W_1 and W_2 are independent, so are $X_1^n(W_1)$ and $X_2^n(W_2)$, and hence it follows that $H(X_1^n(W_1)|X_2^n(W_2)) = H(X_1^n(W_1))$, and $H(X_1^n(W_1)|Y^n, X_2^n(W_2)) \le H(X_1^n(W_1)|Y^n)$ by conditioning,
(d) from the fact that $Y^n = X_1^n + X_2^n + Z^n$,
(e) from the fact that Z^n is independent of X_1^n and X_2^n,
(f) from the fact that the noise is i.i.d.,
(g) from the chain rule and removing conditioning,
(h) from removing conditioning,
(i) from the fact that $Y_i = X_{1i} + X_{2i} + Z_i$,
(j) from the fact that X_{1i} and Z_i are independent of X_{2i}, and
(k) from the entropy maximizing property of the normal (Theorem 9.6.5), after defining $P_{1i} = EX_{1i}^2$.

Hence, we have

$$R_1 \le \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\log\left(1+\frac{P_{1i}}{N}\right) + \epsilon_n. \qquad (15.50)$$

Similarly, we have

$$R_2 \le \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\log\left(1+\frac{P_{2i}}{N}\right) + \epsilon_n. \qquad (15.51)$$

To bound the sum of the rates, we have

$$n(R_1+R_2) \quad = \quad H(W_1, W_2) \qquad (15.52)$$

$$
\begin{aligned}
&= I(W_1, W_2; Y^n) + H(W_1, W_2|Y^n) && (15.53)\\
&\overset{(a)}{\leq} I(W_1, W_2; Y^n) + n\epsilon_n && (15.54)\\
&\overset{(b)}{\leq} I(X_1^n(W_1), X_2^n(W_2); Y^n) + n\epsilon_n && (15.55)\\
&= h(Y^n) - h(Y^n|X_1^n(W_1), X_2^n(W_2)) + n\epsilon_n && (15.56)\\
&\overset{(c)}{=} h(Y^n) - h(Z^n) + n\epsilon_n && (15.57)\\
&\overset{(d)}{=} h(Y^n) - \sum_{i=1}^n h(Z_i) + n\epsilon_n && (15.58)\\
&\overset{(e)}{\leq} \sum_{i=1}^n h(Y_i) - \sum_{i=1}^n h(Z_i) + n\epsilon_n && (15.59)\\
&\overset{(f)}{\leq} \sum_{i=1}^n \frac{1}{2}\log 2\pi e(P_{1i} + P_{2i} + N) - \frac{1}{2}\log 2\pi e N + n\epsilon_n && (15.60)\\
&= \frac{1}{2}\log\left(1 + \frac{P_{1i} + P_{2i}}{N}\right) + n\epsilon_n && (15.61)
\end{aligned}
$$

where
(a) follows from Fano's inequality,
(b) from the data processing inequality,
(c) from the fact that $Y^n = X_1^n + X_2^n + Z^n$, and Z^n is independent of X_1^n and X_2^n,
(d) from the fact that Z_i are i.i.d., (e) follows from the chain rule and removing conditioning, and
(f) from the entropy maximizing property of the normal, and the definitions of P_{1i} and P_{2i}.

Hence we have

$$R_1 + R_2 \leq \frac{1}{n}\sum_{i=1}^n \frac{1}{2}\log\left(1 + \frac{P_{1i} + P_{2i}}{N}\right) + \epsilon_n. \tag{15.62}$$

The power constraint on the codewords imply that

$$\frac{1}{n}\sum_{i=1}^n P_{1i} \leq P_1, \tag{15.63}$$

and

$$\frac{1}{n}\sum_{i=1}^n P_{2i} \leq P_2. \tag{15.64}$$

Now since log is concave function, we can apply Jensens inequality to the expressions in (15.50), (15.51) and (15.62). Thus we obtain

$$R_1 \leq \frac{1}{2}\log\left(1 + \frac{\frac{1}{n}\sum_{i=1}^n P_{1i}}{N}\right) + \epsilon_n \tag{15.65}$$

$$R_2 \leq \frac{1}{2}\log\left(1+\frac{\frac{1}{n}\sum_{i=1}^{n}P_{2i}}{N}\right)+\epsilon_n \tag{15.66}$$

$$R_1+R_2 \leq \frac{1}{2}\log\left(1+\frac{\frac{1}{n}\sum_{i=1}^{n}P_{1i}+P_{2i}}{N}\right)+\epsilon_n. \tag{15.67}$$

which when combined with the power constraints, and taking the limit at $n \rightarrow \infty$, we obtain the desired converse, i.e.,

$$R_1 < \frac{1}{2}\log(1+\frac{P_1}{N}), \tag{15.68}$$

$$R_2 < \frac{1}{2}\log(1+\frac{P_2}{N}), \tag{15.69}$$

$$R_1+R_2 < \frac{1}{2}\log(1+\frac{P_1+P_2}{N}). \tag{15.70}$$

6. **Unusual multiple access channel.** Consider the following multiple access channel: $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y} = \{0,1\}$. If $(X_1,X_2)=(0,0)$, then $Y=0$. If $(X_1,X_2)=(0,1)$, then $Y=1$. If $(X_1,X_2)=(1,0)$, then $Y=1$. If $(X_1,X_2)=(1,1)$, then $Y=0$ with probability $\frac{1}{2}$ and $Y=1$ with probability $\frac{1}{2}$.

   (a) Show that the rate pairs (1,0) and (0,1) are achievable.

   (b) Show that for any non-degenerate distribution $p(x_1)p(x_2)$, we have $I(X_1,X_2;Y) < 1$.

   (c) Argue that there are points in the capacity region of this multiple access channel that can only be achieved by timesharing, i.e., there exist achievable rate pairs (R_1,R_2) which lie in the capacity region for the channel but not in the region defined by

$$R_1 \leq I(X_1;Y|X_2), \tag{15.71}$$

$$R_2 \leq I(X_2;Y|X_1), \tag{15.72}$$

$$R_1+R_2 \leq I(X_1,X_2;Y) \tag{15.73}$$

   for any product distribution $p(x_1)p(x_2)$. Hence the operation of convexification strictly enlarges the capacity region. This channel was introduced independently by Csiszár and Körner[3] and Bierbaum and Wallmeier[2].

Solution:

Unusual multiple access channel.

(a) It is easy to see how we could send 1 bit/transmission from X_1 to Y—simply set $X_2=0$. Then $Y=X_1$, and we can send 1 bit/transmission to from sender 1 to the receiver.

Alternatively, if we evaluate the achievable region for the degenerate product distribution $p(x_1)p(x_2)$ with $p(x_1)=(\frac{1}{2},\frac{1}{2})$, $p(x_2)=(1,0)$, we have $I(X_1;Y|X_2)=1$,

$I(X_2;Y|X_1) = 0$, and $I(X_1,X_2;Y) = 1$. Hence the point $(1,0)$ lies in the achievable region for the multiple access channel corresponding to this product distribution.

By symmetry, the point $(0,1)$ also lies in the achievable region.

(b) Consider any non-degenerate product distribution, and let $p_1 = p(X_1 = 1)$, and let $p_2 = p(X_2 = 1)$. By non-degenerate we mean that $p_1 \neq 0$ or 1, and $p_2 \neq 0$ or 1. In this case, $Y = 0$ when $(X_1,X_2) = (0,0)$ and half the time when $(X_1,X_2) = (1,1)$, i.e., with a probability $(1-p_1)(1-p_2) + \frac{1}{2}p_1p_2$. $Y_1 = 1$ for the other input pairs, i.e., with a probability $p_1(1-p_2) + p_2(1-p_1) + \frac{1}{2}p_1p_2$. We can evaluate the achievable region of the multiple access channel for this product distribution. In particular,

$$R_1+R_2 \leq I(X_1,X_2;Y) = H(Y)-H(Y|X_1,X_2) = H((1-p_1)(1-p_2)+\frac{1}{2}p_1p_2)-p_1p_2. \tag{15.74}$$

Now $H((1-p_1)(1-p_2) + \frac{1}{2}p_1p_2) \leq 1$ (entropy of a binary random variable is at most 1) and $p_1p_2 > 0$ for a non-degenerate distribution. Hence $R_1 + R_2$ is strictly less than 1 for any non-degenerate distribution.

(c) The degenerate distributions have either R_1 or R_2 equal to 0. Hence all the distributions that achieve rate pairs (R_1,R_2) with both rates positive have $R_1+R_2 < 1$. For example the union of the achievable regions over all product distributions does not include the point $(\frac{1}{2},\frac{1}{2})$. But this point is clearly achievable by timesharing between the points $(1,0)$ and $(0,1)$. Or equivalently, the point $(\frac{1}{2},\frac{1}{2})$ lies in the convex hull of the union of the achievable regions, but not the union itself. So the operation of taking the convex hull has strictly increased the capacity region for this multiple access channel.

7. **Convexity of capacity region of broadcast channel.** Let $\mathbf{C} \subseteq \mathbf{R}^2$ be the capacity region of all achievable rate pairs $\mathbf{R} = (R_1,R_2)$ for the broadcast channel. Show that $\mathbf{C}$ is a convex set by using a timesharing argument.

Specifically, show that if $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ are achievable, then $\lambda\mathbf{R}^{(1)} + (1-\lambda)\mathbf{R}^{(2)}$ is achievable for $0 \leq \lambda \leq 1$.

Solution: *Convexity of Capacity Regions.*

Let $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ be two achievable rate pairs. Then there exist a sequence of $((2^{nR_1^{(1)}},2^{nR_2^{(1)}}),n)$ codes and a sequence of $((2^{nR_1^{(2)}},2^{nR_2^{(2)}}),n)$ codes for the channel with $P_e^{(n)}(1) \to 0$ and $P_e^{(n)}(2) \to 0$. We will now construct a code of rate $\lambda\mathbf{R}^{(1)} + (1-\lambda)\mathbf{R}^{(2)}$.

For a code length n, use the concatenation of the codebook of length λn and rate $\mathbf{R}^{(1)}$ and the code of length $(1-\lambda)n$ and rate $\mathbf{R}^{(2)}$. The new codebook consists of all pairs of codewords and hence the number of X_1 codewords is $2^{\lambda nR_1^{(1)}}2^{(1-\lambda)nR_1^{(2)}}$, and hence the rate is $\lambda R_1^{(1)} + (1-\lambda)R_1^{(2)}$. Similarly the rate of the X_2 codeword is $\lambda R_2^{(1)} + (1-\lambda)R_2^{(2)}$.

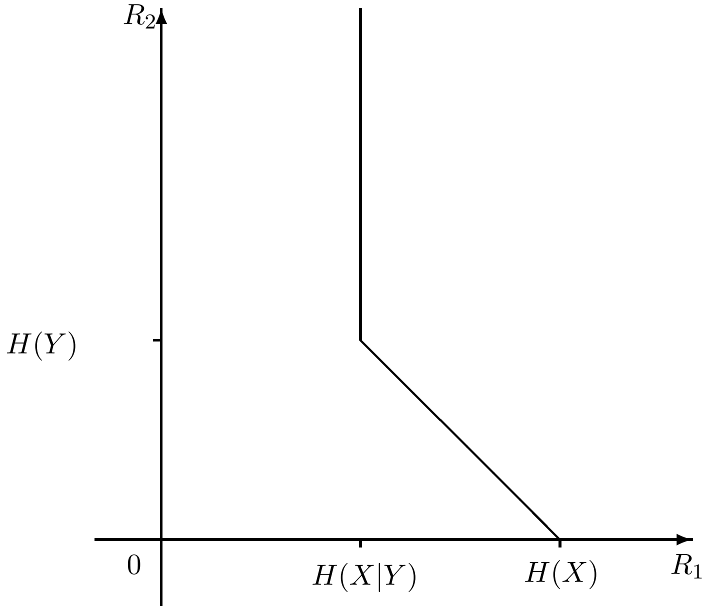


Figure 15.4: Slepian Wolf rate region for $Y = f(X)$.

We will now show that the probability of error for this sequence of codes goes to zero. The decoding rule for the concatenated code is just the combination of the decoding rule for the parts of the code. Hence the probability of error for the combined codeword is less than the sum of the probabilities for each part. For the combined code,

$$P_e^{(n)} \leq P_e^{(\lambda n)}(1) + P_e^{((1-\lambda)n)}(2) \tag{15.75}$$

which goes to 0 as $n \to \infty$. Hence the overall probability of error goes to 0, which implies the $\lambda \mathbf{R}^{(1)} + (1-\lambda)\mathbf{R}^{(2)}$ is achievable.

8. **Slepian-Wolf for deterministically related sources.** Find and sketch the Slepian-Wolf rate region for the simultaneous data compression of (X, Y), where $y = f(x)$ is some deterministic function of x.

   Solution: *Slepian Wolf for* $Y = f(X)$.

   The quantities defining the Slepian Wolf rate region are $H(X,Y) = H(X)$, $H(Y|X) = 0$ and $H(X|Y) \geq 0$. Hence the rate region is as shown in the Figure 15.4.

9. **Slepian-Wolf.** Let X_i be i.i.d. Bernoulli(p). Let Z_i be i.i.d. $\sim$ Bernoulli(r), and let $\mathbf{Z}$ be independent of $\mathbf{X}$. Finally, let $\mathbf{Y} = \mathbf{X} \oplus \mathbf{Z}$ (mod 2 addition). Let $\mathbf{X}$ be described at rate R_1 and $\mathbf{Y}$ be described at rate R_2. What region of rates allows recovery of $\mathbf{X}, \mathbf{Y}$ with probability of error tending to zero?

   Solution: *Slepian Wolf for binary sources.*

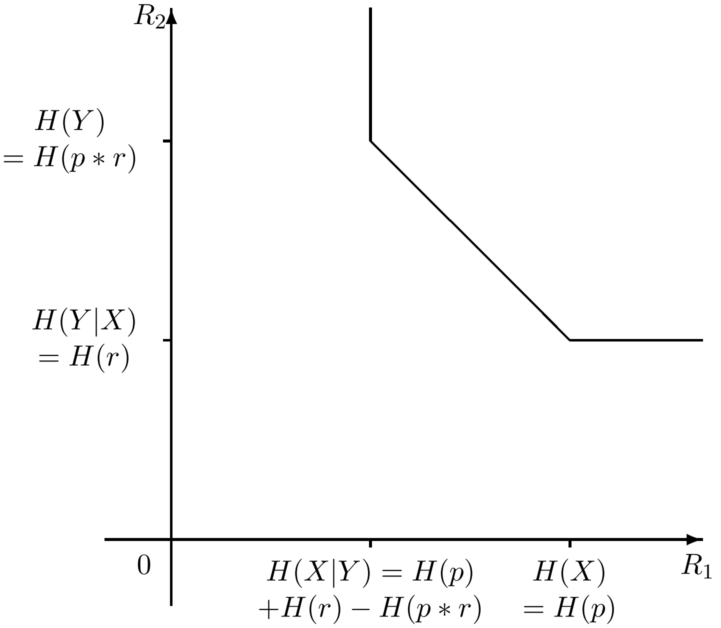


Figure 15.5: Slepian Wolf region for binary sources

$X \sim$ Bern(p). $Y = X \oplus Z$, $Z \sim$ Bern(r). Then $Y \sim$ Bern$(p * r)$, where $p * r = p(1-r) + r(1-p)$. $H(X) = H(p)$. $H(Y) = H(p*r)$, $H(X,Y) = H(X,Z) = H(X) + H(Z) = H(p) + H(r)$. Hence $H(Y|X) = H(r)$ and $H(X|Y) = H(p) + H(r) - H(p*r)$.

The Slepian Wolf region in this case is shown in Figure 15.5.

10. **Broadcast capacity depends only on the conditional marginals.** Consider the general broadcast channel $(X, Y_1 \times Y_2, p(y_1, y_2 \mid x))$. Show that the capacity region depends only on $p(y_1 \mid x)$ and $p(y_2 \mid x)$. To do this, for any given $((2^{nR_1}, 2^{nR_2}), n)$ code, let

$$\begin{aligned} P_1^{(n)} &= P\{\hat{W}_1(\mathbf{Y}_1) \neq W_1\}, & (15.76) \\ P_2^{(n)} &= P\{\hat{W}_2(\mathbf{Y}_2) \neq W_2\}, & (15.77) \\ P^{(n)} &= P\{(\hat{W}_1, \hat{W}_2) \neq (W_1, W_2)\}. & (15.78) \end{aligned}$$

Then show

$$\max\{P_1^{(n)}, P_2^{(n)}\} \leq P^{(n)} \leq P_1^{(n)} + P_2^{(n)}.$$

The result now follows by a simple argument.

Remark: The probability of error $P^{(n)}$ *does* depend on the conditional joint distribution $p(y_1, y_2 \mid x)$. But whether or not $P^{(n)}$ can be driven to zero (at rates (R_1, R_2)) *does not* (except through the conditional marginals $p(y_1 \mid x), p(y_2 \mid x)$).

Solution: *Broadcast channel capacity depends only on conditional marginals*

$$P_1^{(n)} = P(\hat{W}_1(\mathbf{Y}_1) \neq W_1) \tag{15.79}$$

$$P_2^{(n)} = P(\hat{W}_2(\mathbf{Y}_2) \neq W_2) \tag{15.80}$$

$$P^{(n)} = P((\hat{W}_1(\mathbf{Y}_1), \hat{W}_2(\mathbf{Y}_2)) \neq (W_1, W_2)) \tag{15.81}$$

Then by the union of events bound, it is obvious that

$$P^{(n)} \leq P_1^{(n)} + P_2^{(n)}. \tag{15.82}$$

Also since $(\hat{W}_1(\mathbf{Y}_1) \neq W_1)$ or $(\hat{W}_2(\mathbf{Y}_2) \neq W_2)$ implies $((\hat{W}_1(\mathbf{Y}_1), \hat{W}_2(\mathbf{Y}_2)) \neq (W_1, W_2))$, we have

$$P^{(n)} \geq \max\{P_1^{(n)}, P_2^{(n)}\}. \tag{15.83}$$

Hence $P^{(n)} \to 0$ iff $P_1^{(n)} \to 0$ and $P_2^{(n)} \to 0$.

The probability of error, $P^{(n)}$, for a broadcast channel does depend on the joint conditional distribution. However, the individual probabilities of error $P_1^{(n)}$ and $P_2^{(n)}$ however depend only on the conditional marginal distributions $p(y_1|x)$ and $p(y_2|x)$ respectively. Hence if we have a sequence of codes for a particular broadcast channel with $P^{(n)} \to 0$, so that $P_1^{(n)} \to 0$ and $P_2^{(n)} \to 0$, then using the same codes for another broadcast channel with the same conditional marginals will ensure that $P^{(n)}$ for that channel as well, and the corresponding rate pair is achievable for the second channel. Hence the capacity region for a broadcast channel depends only on the conditional marginals.

11. **Converse for the degraded broadcast channel.** The following chain of inequalities proves the converse for the degraded discrete memoryless broadcast channel. Provide reasons for each of the labeled inequalities.

Setup for converse for degraded broadcast channel capacity:

$$(W_1, W_2)_{\text{indep.}} \to X^n(W_1, W_2) \to Y_1^n \to Y_2^n$$

Encoding $f_n : 2^{nR_1} \times 2^{nR_2} \to \mathcal{X}^n$

Decoding: $g_n : \mathcal{Y}_1^n \to 2^{nR_1}, \qquad h_n : \mathcal{Y}_2^n \to 2^{nR_2}$

Let $U_i = (W_2, Y_1^{i-1})$. Then

$$nR_2 \quad \overset{\cdot}{\leq}_{\text{Fano}} \quad I(W_2; Y_2^n) \tag{15.84}$$

$$\overset{(a)}{=} \quad \sum_{i=1}^{n} I(W_2; Y_{2i} \mid Y_2^{i-1}) \tag{15.85}$$

$$\overset{(b)}{=} \quad \sum_i (H(Y_{2i} \mid Y_2^{i-1}) - H(Y_{2i} \mid W_2, Y_2^{i-1})) \tag{15.86}$$

$$\overset{(c)}{\leq} \quad \sum_i (H(Y_{2i}) - H(Y_{2i} \mid W_2, Y_2^{i-1}, Y_1^{i-1})) \tag{15.87}$$

$$\overset{(d)}{=} \quad \sum_i (H(Y_{2i}) - H(Y_{2i} \mid W_2, Y_1^{i-1})) \tag{15.88}$$

$$\overset{(e)}{=} \quad \sum_{i=1}^{n} I(U_i; Y_{2i}). \tag{15.89}$$

Continuation of converse. Give reasons for the labeled inequalities:

$$\begin{aligned}
nR_1 \quad & \dot{\leq}_{\text{Fano}} \quad I(W_1; Y_1^n) & (15.90)\\
& \overset{(f)}{\leq} \quad I(W_1; Y_1^n, W_2) & (15.91)\\
& \overset{(g)}{\leq} \quad I(W_1; Y_1^n \mid W_2) & (15.92)\\
& \overset{(h)}{=} \quad \sum_{i-1}^{n} I(W_1; Y_{1i} \mid Y_1^{i-1}, W_2) & (15.93)\\
& \overset{(i)}{\leq} \quad \sum_{i=1}^{n} I(X_i; Y_{1i} \mid U_i). & (15.94)
\end{aligned}$$

Now let Q be a time sharing random variable with $\Pr(Q = i) = 1/n$, $i = 1, 2, \ldots, n$. Justify the following:

$$\begin{aligned}
R_1 \quad & \leq \quad I(X_Q; Y_{1Q} | U_Q, Q) & (15.95)\\
R_2 \quad & \leq \quad I(U_Q; Y_{2Q} | Q), & (15.96)
\end{aligned}$$

for some distribution $p(q)p(u|q)p(x|u,q)p(y_1, y_2|x)$. By appropriately redefining U, argue that this region is equal to the convex closure of regions of the form

$$\begin{aligned}
R_1 \quad & \leq \quad I(X; Y_1|U) & (15.97)\\
R_2 \quad & \leq \quad I(U; Y_2), & (15.98)
\end{aligned}$$

for some joint distribution $p(u)p(x|u)p(y_1, y_2|x)$.

Solution: *Converse for the degraded broadcast channel.*

$$(W_1, W_2) \to \mathbf{X}(W_1, W_2) \to \mathbf{Y} \to \mathbf{Z} \tag{15.99}$$

We also have

$$(W_1, W_2) \to X_i(W_1, W_2) \to Y_i \to Z_i. \tag{15.100}$$

Let $U_i = (W_2, Y^{i-1})$.

By Fano's inequality,

$$H(W_2|Z^n) \leq P_2^{(n)} nR_2 + H(P_2^{(n)}) = n\epsilon_n \tag{15.101}$$

where $\epsilon_n \to 0$ as $P_2^{(n)} \to 0$.

We then have the following chain of inequalities

$$\begin{aligned}
nR_2 &= H(W_2) && (15.102)\\
&= I(W_2;Z^n) + H(W_2|Z^n) && (15.103)\\
&\le I(W_2;Z^n) + n\epsilon_n && (15.104)\\
&\overset{(a)}{=} \sum_i I(W_2;Z_i|Z^{i-1}) + n\epsilon_n && (15.105)\\
&\overset{(b)}{=} \sum_i (H(Z_i|Z^{i-1}) - H(Z_i|Z^{i-1},W_2)) + n\epsilon_n && (15.106)\\
&\overset{(c)}{\le} \sum_i (H(Z_i) - H(Z_i|Z^{i-1},W_2,Y^{i-1})) + n\epsilon_n && (15.107)\\
&\overset{(d)}{=} \sum_i (H(Z_i) - H(Z_i|W_2,Y^{i-1})) + n\epsilon_n && (15.108)\\
&\overset{(e)}{=} \sum_i I(U_i;Z_i) + n\epsilon_n && (15.109)
\end{aligned}$$

where (15.104) follows from Fano's inequality,
(a) from the chain rule,
(b) from the definition of conditional mutual information,
(c) from the fact that removing conditioning increases entropy and adding conditioning reduces it,
(d) from the fact that since the broadcast channel is degraded, Z^{i-1} depends only on Y^{i-1} and is conditionally independent of everything else, hence Z_i is conditionally independent of Z^{i-1} given Y^{i-1},
(e) follows from the definition of U_i.

Continuation of Converse.

Similarly by Fano's inequality,

$$H(W_1|Y^n) \le P_1^{(n)} nR_1 + H(P_1^{(n)}) = n\epsilon_n \qquad (15.110)$$

and we have the chain of inequalities,

$$\begin{aligned}
nR_1 &= H(W_1) && (15.111)\\
&= I(W_1;Y^n) + H(W_1|Y^n) && (15.112)\\
&\le I(W_1;Y^n) + n\epsilon_n && (15.113)\\
&\overset{(f)}{\le} I(W_1;W_2,Y^n) + n\epsilon_n && (15.114)\\
&\overset{(g)}{=} I(W_1;Y^n|W_2) + n\epsilon_n && (15.115)\\
&\overset{(h)}{\le} I(W_1;Y_i|W_2,Y^{i-1}) + n\epsilon_n && (15.116)\\
&\le I(W_1,X_i;Y_i|W_2,Y^{i-1}) + n\epsilon_n && (15.117)
\end{aligned}$$

$$\overset{(i)}{\leq} \quad I(X_i; Y_i|W_2, Y^{i-1}) + n\epsilon_n \tag{15.118}$$

$$= \quad I(X_i; Y_i|U_i) + n\epsilon_n \tag{15.119}$$

where (15.113) follows from Fano's inequality,
(f) follows from the fact that the difference, $I(W_1; W_2|Y^n) \geq 0$,
(g) follows from the chain rule for I and the fact that W_1 and W_2 are independent,
(h) from the chain rule for mutual information, and
(i) from the data processing inequality.

We can then use standard techniques like the introduction of a time-sharing random variable to complete the proof of the converse for the broadcast channel.

12. **Capacity points.**

(a) For the degraded broadcast channel $X \rightarrow Y_1 \rightarrow Y_2$, find the points a and b where the capacity region hits the R_1 and R_2 axes (Figure 15.6).

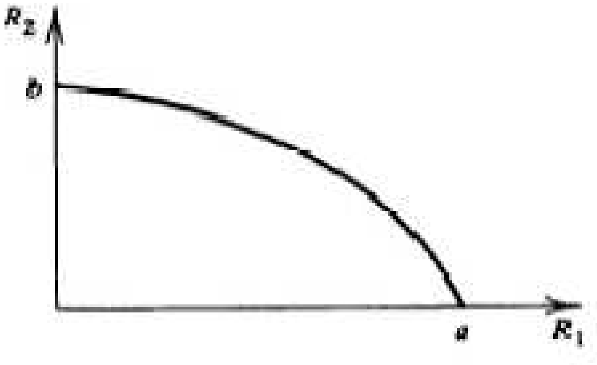


Figure 15.6: Capacity region of a broadcast channel

(b) Show that $b \leq a$.

Solution: *Capacity region of broadcast channel.*

(a) The capacity region of the degraded broadcast channel $X \rightarrow Y_1 \rightarrow Y_2$ is the convex hull of regions of the form

$$R_1 \leq I(X; Y_1|U) \tag{15.120}$$

$$R_2 \leq I(U; Y_2) \tag{15.121}$$

over all choices of auxiliary random variable U and joint distribution of the form $p(u)p(x|u)p(y_1, y_2|x)$.

The region is of the form illustrated in Figure 15.7.

The point a on the figure corresponds to the maximum achievable rate from the sender to receiver 2. From the expression for the capacity region, it is the maximum value of $I(U; Y_2)$ for all auxiliary random variables U.

For any random variable U and $p(u)p(x|u)$, $U \rightarrow X \rightarrow Y_2$ forms a Markov chain, and hence $I(U;Y_2) \leq I(X;Y_2) \leq \max_{p(x)} I(X;Y_2)$. The maximum can be achieved by setting $U = X$ and choosing the distribution of X to be the one that maximizes $I(X;Y_2)$. Hence the point a corresponds to $R_2 = \max_{p(x)} I(X;Y_2), R_1 = I(X;Y_1|U) = I(X;Y_1|X) = 0$.

The point b has a similar interpretation. The point b corresponds to the maximum rate of transmission to receiver 1. From the expression for the capacity region,

$$R_1 \leq I(X;Y_1|U) = H(Y_1|U) - H(Y_1|X,U) = H(Y_1|U) - H(Y_1|X), \qquad (15.122)$$

since $U \rightarrow X \rightarrow Y_1$ forms a Markov chain. Since $H(Y_1|U) \leq H(Y_1)$, we have

$$R_1 \leq H(Y_1) - H(Y_1|X) = I(X;Y_1) \leq \max_{p(x)} I(X;Y_1), \qquad (15.123)$$

and the maximum is attained when we set $U \equiv 0$ and choose $p(x) = p(x|u)$ to be the distribution that maximizes $I(X;Y_1)$. In this case, $R_2 \leq I(U;Y_2) = 0$.

Hence point b corresponds to the rates $R_1 = \max_{p(x)} I(X;Y_1), R_2 = 0$.

These results have a simple single user interpretation. If we not sending any information to receiver 1, then we can treat the channel to receiver 2 as a single user channel and send at capacity for this channel, i.e., $\max I(X;Y_2)$. Similarly, if we are not sending any information to receiver 2, we can send at capacity to receiver 1, which is $\max I(X;Y_1)$.

(b) Since $X \rightarrow Y_1 \rightarrow Y_2$ forms a Markov chain for all distributions $p(x)$, we have by the data processing inequality

$$\begin{aligned} a = \max_{p(x)} I(X;Y_2) &= I(X^*;Y_2) && (15.124)\\ &\leq I(X^*;Y_1) && (15.125)\\ &= \max_{p(x)} I(X;Y_1) = b, && (15.126) \end{aligned}$$

where X^* has the distribution that maximizes $I(X;Y_2)$.

13. **Degraded broadcast channel.** Find the capacity region for the degraded broadcast channel in Figure 15.8.

Solution: *Degraded broadcast channel.* From the expression for the capacity region, it is clear that the only on trivial possibility for the auxiliary random variable U is that it be binary. From the symmetry of the problem, we see that the auxiliary random variable should be connected to X by a binary symmetric channel with parameter β.

Hence we have the setup as shown in Figure 15.9.

We can now evaluate the capacity region for this choice of auxiliary random variable. By symmetry, the best distribution for U is the uniform. Hence

$$R_2 = I(U;Y_2) \qquad (15.127)$$

$$
\begin{aligned}
&= H(Y_2) - H(Y_2|U) && (15.128)\\
&= H\left(\frac{\overline{\alpha}}{2}, \alpha, \frac{\overline{\alpha}}{2}\right) - H((\overline{\beta}\overline{p} + \beta p)\overline{\alpha}, \alpha, (\overline{\beta}p + \beta\overline{p})\overline{\alpha}) && (15.129)\\
&= H(\alpha) + \overline{\alpha}H\left(\frac{1}{2}\right) - H(\alpha) - \overline{\alpha}H(\overline{\beta}p + \beta\overline{p}) && (15.130)\\
&= \overline{\alpha}(1 - H(\overline{\beta}p + \beta\overline{p})). && (15.131)
\end{aligned}
$$

Also

$$
\begin{aligned}
R_1 &= I(X;Y_1|U) && (15.132)\\
&= H(Y_1|U) - H(Y_1|U,X) && (15.133)\\
&= H(\beta\overline{p} + \overline{\beta}p) - H(p). && (15.134)
\end{aligned}
$$

These two equations characterize the boundary of the capacity region as β varies. When $\beta = 0$, then $R_1 = 0$ and $R_2 = \overline{\alpha}(1 - H(p))$. When $\beta = \frac{1}{2}$, we have $R_1 = 1 - H(p)$ and $R_2 = 0$.

The capacity region is sketched in Figure 15.10.

14. **Channels with unknown parameters.** We are given a binary symmetric channel with parameter p. The capacity is $C = 1 - H(p)$.

Now we change the problem slightly. The receiver knows only that $p \in \{p_1, p_2\}$, i.e.,$p = p_1$ or $p = p_2$, where p_1 and p_2 are given real numbers. The transmitter knows the actual value of p. Devise two codes for use by the transmitter, one to be used if $p = p_1$, the other to be used if $p = p_2$, such that transmission to the receiver can take place at rate $\approx C(p_1)$ if $p = p_1$ and at rate $\approx C(p_2)$ if $p = p_2$.

Hint: Devise a method for revealing p to the receiver without affecting the asymptotic rate. Prefixing the codeword by a sequence of 1's of appropriate length should work.

Solution: *Capacity of channels with unknown parameters.*

We have two possiblities; the channel is a BSC with parameter p_1 or a BSC with parameter p_2. If both sender and receiver know that state of channel, then we can achieve the capacity corresponding to which channel is in use, i.e., $1 - H(p_1)$ or $1 - H(p_2)$.

If the receiver does not know the state of the channel, then he cannot know which codebook is being used by the transmitter. He cannot then decode optimally; hence he cannot achieve the rates corresponding to the capacities of the channels.

But the transmitter can inform the receiver of the state of the channel so that the receiver can decode optimally. To do this, the transmitter can precede the codewords by a sequence of 1's or 0's. Let us say we use a string of m 1's to indicate that the channel was in state p_1 and m 0's to indicate state p_2. Then, if $m = o(n)$ and $m \to \infty$, where n is the block length of the code used, we have the probability of error in decoding the state of the channel going to zero. Since the receiver will then use the right code for the rest of the message, it will be decoded correctly with $P_e^{(n)} \to 0$. The effective rate for this code is $\frac{\log 2^{nC(p_i)}}{n+m} \to C(p_i)$, since $m = o(n)$. So we can achieve the same asymptotic rate as if both sender and receiver knew the state of the channel.

15. **Two-way channel**. Consider the two-way channel shown in Figure 15.6. The outputs Y_1 and Y_2 depend only on the current inputs X_1 and X_2.

    (a) By using independently generated codes for the two senders, show that the following rate region is achievable:

$$R_1 < I(X_1;Y_2|X_2), \quad (15.135)$$
$$R_2 < I(X_2;Y_1|X_1) \quad (15.136)$$

    for some product distribution $p(x_1)p(x_2)p(y_1,y_2|x_1,x_2)$.

    (b) Show that the rates for any code for a two-way channel with arbitrarily small probability of error must satisfy

$$R_1 \leq I(X_1;Y_2|X_2), \quad (15.137)$$
$$R_2 \leq I(X_2;Y_1|X_1) \quad (15.138)$$

    for some joint distribution $p(x_1,x_2)p(y_1,y_2|x_1,x_2)$.

The inner and outer bounds on the capacity of the two-way channel are due to Shannon[14]. He also showed that the inner bound and the outer bound do not coincide in the case of the binary multiplying channel $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y}_1 = \mathcal{Y}_2 = \{0,1\}$, $Y_1 = Y_2 = X_1X_2$. The capacity of the two-way channel is still an open problem.

Solution: *Two-way channel.*

(a) We will only outline the proof of achievability. It is quite straightforward compared to the more complex channels considered in the text.

Fix $p(x_1)p(x_2)p(y_1,y_2|x_1,x_2)$.

Code generation: Generate a code of size 2^{nR_1} of codewords $\mathbf{X}_1(w_1)$, where the x_{1i} are generate i.i.d. $\sim p(x_1)$. Similarly generate a codebook $\mathbf{X}_2(w_2)$ of size 2^{nR_2}.

Encoding: To send index w_1 from sender 1, he sends $\mathbf{X}_1(w_1)$. Similarly, sender 2 sends $\mathbf{X}_2(w_2)$.

Decoding: Receiver 1 looks for the unique w_2, such that $(\mathbf{X}_1(w_1), \mathbf{x}_2(w_2), \mathbf{Y}_1) \in A_\epsilon^{(n)}(X_1,X_2,Y_1)$. If there is no such w_2 or more than one such, it declares an error. Similarly, receiver 2 looks for the unique w_1, such that $(\mathbf{x}_1(w_1), \mathbf{X}_2(w_2), \mathbf{Y}_2) \in A_\epsilon^{(n)}(X_1,X_2,Y_2)$.

Analysis of the probability of error: We will only analyze the error at receiver 1. The analysis for receiver 2 is similar.

Without loss of generality, by the symmetry of the random code construction, we can assume that (1,1) was sent. We have an error at receiver 1 if

- $(\mathbf{X}_1(1), \mathbf{X}_2(1), \mathbf{Y}_1) \notin A_\epsilon^{(n)}(X_1, X_2, Y_1)$. The probability of this goes to 0 by the law of large numbers as $n \to \infty$.
- There exists an $j \neq 1$, such that $(\mathbf{X}_1(1), \mathbf{X}_2(j), \mathbf{Y}_1) \in A_\epsilon^{(n)}(X_1, X_2, Y_1)$.

Define the events

$$E_j = \{(\mathbf{X}_1(1), \mathbf{X}_2(j), \mathbf{Y}_1) \in A_\epsilon^{(n)}\}. \tag{15.139}$$

Then by the union of events bound,

$$\begin{aligned} P_e^{(n)} &= P\left(E_1^c \bigcup \cup_{j \neq 1} E_j\right) && (15.140) \\ &\leq P(E_1^c) + \sum_{j \neq 1} P(E_j), && (15.141) \end{aligned}$$

where P is the probability given that $(1,1)$ was sent. From the AEP, $P(E_1^c) \to 0$. By Theorem 14.2.3, for $j \neq 1$, we have

$$\begin{aligned} P(E_j) &= P((\mathbf{X}_1(1), \mathbf{X}_2(j), \mathbf{Y}_1) \in A_\epsilon^{(n)}) && (15.142) \\ &= \sum_{(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1) \in A_\epsilon^{(n)}} p(\mathbf{x}_2) p(\mathbf{x}_1, \mathbf{y}_1) && (15.143) \\ &\leq |A_\epsilon^{(n)}| 2^{-n(H(X_2)-\epsilon)} 2^{-n(H(X_1,Y)-\epsilon)} && (15.144) \\ &\leq 2^{-n(H(X_2)+H(X_1,Y)-H(X_1,X_2,Y)-3\epsilon)} && (15.145) \\ &= 2^{-n(I(X_2;X_1,Y)-3\epsilon)} && (15.146) \\ &= 2^{-n(I(X_2;Y|X_1)-3\epsilon)}, && (15.147) \end{aligned}$$

since X_1 and X_2 are independent, and therefore $I(X_1; X_2, Y) = I(X_1; X_2) + I(X_1; Y|X_2) = I(X_1; Y|X_2)$. Therefore

$$P_e^{(n)} \leq P(E_1^c) + 2^{nR_2} 2^{-n(I(X_2;Y|X_1)-3\epsilon)}, \tag{15.148}$$

Since $\epsilon > 0$ is arbitrary, the conditions of the theorem imply that the probability of error tends to 0 as $n \to \infty$. Similarly, we can show that the probability of error at receiver two goes to 0, and thus we have proved the achievability of the region for the two way channel.

(b) The converse is a simple application of the general Theorem 14.10.1 to this simple case. The sets S can be taken in turn to be each node. We will not go into the details.

16. **Multiple-access channel**

Let the output Y of a multiple-access channel be given by

$$Y = X_1 + \operatorname{sgn}(X_2)$$

where X_1, X_2 are both real and power limited,

$$\begin{aligned} E[X_1^2] &\le P_1, \\ E[X_2^2] &\le P_2, \end{aligned}$$

and $\text{sgn}(x) = \begin{cases} 1, & x > 0 \\ -1, & x \le 0 \end{cases}$.

Note that there is interference but no noise in this channel.

(a) Find the capacity region.

(b) Describe a coding scheme that achieves the capacity region.

Solution: *Multiple-access channel*

(a) This is continuous noiseless multiple access channel, if we let $U_2 = sgn(X_2)$, we can consider a channel from X_1 and U_2 to Y

$$I(X_1;Y|X_2) = h(Y|X_2) - h(Y|X_1,X_2) = h(X_1|X_2) - (-\infty) = \infty \qquad (15.149)$$

since X_1 and X_2 are independent, and similarly

$$I(X_2;Y|X_1) = I(X_2,U_2;Y|X_1) = I(U_2;Y|X_1)+I(X_2;Y|X_1,U_2) = I(U_2;Y|X_1) = H(U_2)-H(U_2|Y \qquad (15.150)$$

$I(X_1,X_2;Y) = \infty$. Thus we can send at infinite rate from X_1 to Y and at a maximum rate of 1 bit/transmission from X_2 to Y.

(b) We can send a 1 for X_2 in the first transmission, and knowing this, Y can recover X_1 perfectly, recovering an infinite number of bits. From then on, X_1 can be 0 and we can send 1 bit per transmission using the sign of X_2.

17. **Slepian Wolf**

Let (X,Y) have the joint pmf $p(x,y)$

p(x,y)	1	2	3
1	α	β	β
2	β	α	β
3	β	β	α

where $\beta = \frac{1}{6} - \frac{\alpha}{2}$. (Note: This is a joint, not a conditional, probability mass function.)

(a) Find the Slepian Wolf rate region for this source.

(b) What is $\Pr\{X = Y\}$ in terms of α?

(c) What is the rate region if $\alpha = \frac{1}{3}$?

(d) What is the rate region if $\alpha = \frac{1}{9}$?

Solution: *Slepian Wolf*

(a) $H(X,Y) = -\sum p(x,y)\log p(x,y) = -3\alpha\log\alpha - 6\beta\log\beta$. Since X and Y are uniformly distributed

$$H(X) = H(Y) = \log 3 \tag{15.151}$$

and

$$H(X|Y) = H(Y|X) = H(3\alpha, 3\beta, 3\beta) \tag{15.152}$$

Hence the Slepian Wolf rate region is

$$R_1 \geq H(X|Y) = H(3\alpha, 3\beta, 3\beta) \tag{15.153}$$
$$R_2 \geq H(Y|X) = H(3\alpha, 3\beta, 3\beta) \tag{15.154}$$
$$R_1 + R_2 \geq H(X,Y) = H(3\alpha, 3\beta, 3\beta) + \log 3 \tag{15.155}$$

(b) From the joint distribution, $\Pr(X = Y) = 3\alpha$.

(c) If $\alpha = \frac{1}{3}$, $\beta = 0$, and $H(X|Y) = H(Y|X) = 0$. The rate region then becomes

$$R_1 \geq 0 \tag{15.156}$$
$$R_2 \geq 0 \tag{15.157}$$
$$R_1 + R_2 \geq \log 3 \tag{15.158}$$

(d) If $\alpha = \frac{1}{9}$, $\beta = \frac{1}{9}$, and $H(X|Y) = H(Y|X) = \log 3$. X and Y are independent, and the rate region then becomes

$$R_1 \geq \log 3 \tag{15.159}$$
$$R_2 \geq \log 3 \tag{15.160}$$
$$R_1 + R_2 \geq 2\log 3 \tag{15.161}$$

18. **Square channel**

What is the capacity of the following multiple access channel?

$$\begin{aligned} X_1 &\in \{-1, 0, 1\} \\ X_2 &\in \{-1, 0, 1\} \\ Y &= X_1^2 + X_2^2 \end{aligned}$$

(a) Find the capacity region.

(b) Describe $p^*(x_1), p^*(x_2)$ achieving a point on the boundary of the capacity region.

Solution: *Square channel*

(a) If we let $U_1 = X_1^2$ and $U_2 = X_2^2$, then the channel is equivalent to a sum multiple access channel $Y = U_1 + U_2$. We could also get the same behaviour by using only two input symbols (0 and 1) for both X_1 and X_2.

Thus the capacity region

$$\begin{aligned} R_1 &< I(X_1;Y|X_2) = H(Y|X_2) && (15.162)\\ R_2 &< I(X_2;Y|X_1) = H(Y|X_1) && (15.163)\\ R_1 + R_2 &< I(X_1,X_2;Y) = H(Y) && (15.164) \end{aligned}$$

By choosing $p(x_1,x_2) = 1/4$ for $(x_1,x_2) = (1,0),(0,0),(0,1),(1,1)$ and 0 otherwise, we obtain $H(Y|X_1) = H(Y|X_2) = 1$, $H(Y) = 1.5$, and by the results for the binary erasure multiple access channel, the capacity of the channel is limited by

$$\begin{aligned} R_1 &< 1 && (15.165)\\ R_2 &< 1 && (15.166)\\ R_1 + R_2 &< 1.5 && (15.167) \end{aligned}$$

(b) One possible distribution that achieves points on the boundary of the rate region is given by the distribution in part (a).

19. **Slepian-Wolf**: Two senders know random variables U_1 and U_2 respectively. Let the random variables (U_1,U_2) have the following joint distribution:

$U_1 \backslash U_2$	0	1	2	$\cdots$	$m-1$
0	α	$\frac{\beta}{m-1}$	$\frac{\beta}{m-1}$	$\cdots$	$\frac{\beta}{m-1}$
1	$\frac{\gamma}{m-1}$	0	0	$\cdots$	0
2	$\frac{\gamma}{m-1}$	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$m-1$	$\frac{\gamma}{m-1}$	0	0	$\cdots$	0

where $\alpha + \beta + \gamma = 1$. Find the region of rates (R_1, R_2) that would allow a common receiver to decode both random variables reliably.

Solution: *Slepian-Wolf*

For this joint distribution,

$$H(U_1) = H(\alpha+\beta, \frac{\gamma}{m-1}, \ldots, \frac{\gamma}{m-1}) = H(\alpha+\beta,\gamma) + \gamma \log(m-1) \qquad (15.168)$$

$$H(U_2) = H(\alpha+\gamma, \frac{\beta}{m-1}, \ldots, \frac{\beta}{m-1}) = H(\alpha+\gamma,\beta) + \beta \log(m-1) \qquad (15.169)$$

$$H(U_1,U_2) = H\left(\alpha, \frac{\beta}{m-1}, \ldots, \frac{\beta}{m-1}, \frac{\gamma}{m-1}, \ldots, \frac{\gamma}{m-1}\right) = H(\alpha,\beta,\gamma) + \beta\log(m-1) + \gamma\log(m-1) \qquad (15.170)$$

$$H(U_1|U_2) = H(\alpha,\beta,\gamma) - H(\alpha+\gamma,\beta) + \gamma\log(m-1) \tag{15.171}$$

$$H(U_2|U_1) = H(\alpha,\beta,\gamma) - H(\alpha+\beta,\gamma) + \beta\log(m-1) \tag{15.172}$$

and hence the Slepian Wolf region is

$$R_1 \geq H(\alpha,\beta,\gamma) - H(\alpha+\gamma,\beta) + \gamma\log(m-1) \tag{15.173}$$

$$R_2 \geq H(\alpha,\beta,\gamma) - H(\alpha+\beta,\gamma) + \beta\log(m-1) \tag{15.174}$$

$$R_1 + R_2 \geq H(\alpha,\beta,\gamma) + \beta\log(m-1) + \gamma\log(m-1) \tag{15.175}$$

20. **Multiple access.**

(a) Find the capacity region for the multiple access channel

$$Y = X_1^{X_2}$$

where

$$X_1 \epsilon \{2,4\}\ ,\quad X_2 \epsilon \{1,2\}\ \ .$$

(b) Suppose the range of X_1 is $\{1,2\}$. Is the capacity region decreased? Why or why not?

Solution: *Multiple access.*

(a) With $X_1 \in \{2,4\}, X_2 \in \{1,2\}$, the channel $Y = X_1^{X_2}$ behaves as:

X_1	X_2	Y
2	1	2
4	1	4
2	2	4
4	2	16

We compute

$$\begin{aligned} R_1 &\leq I(X_1;Y|X_2) = I(X_1;X_1^{X_2}|X_2) = H(X_1) = 1 \text{ bit per trans} \\ R_2 &\leq I(X_2;Y|X_1) = I(X_2;X_1^{X_2}|X_1) = H(X_2) = 1 \text{ bit per trans} \\ R_1 + R_2 &\leq I(X_1,X_2;Y) = H(Y) - H(Y|X_1,X_2) = H(Y) = \frac{3}{2} \text{ bits per trans,} \end{aligned}$$

where the bound on $R_1 + R_2$ is met at the corners in the picture below, where either sender 1 or 2 sends 1 bit per transmission and the other user treats the channel as a binary erasure channel with capacity $1 - p_{erasure} = 1 - \frac{1}{2} = \frac{1}{2}$ bits per use of the channel. Other points on the line are achieved by timesharing.

(b) With $X_1 \in \{1,2\}, X_2 \in \{1,2\}$, the channel $Y = X_1^{X_2}$ behaves as:

X_1	X_2	Y
1	1	1
2	1	2
1	2	1
2	2	4

Note when $X_1 = 1$, X_2 has no effect on Y and can not be recovered given X_1 and Y. If $X_1 \sim Br(\alpha)$ and $X_2 \sim Br(\beta)$ then:

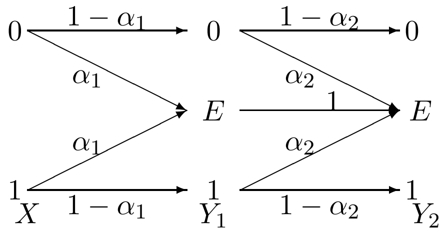
$$\begin{aligned}
R_1 &\leq I(X_1;Y|X_2) = I(X_1;X_1^{X_2}|X_2) = H(\alpha)\\
R_2 &\leq I(X_2;Y|X_1) = H(Y|X_1) - H(Y|X_1,X_2) = H(Y|X_1)\\
&= p(X_1=1)H(Y|X_1=1) + p(X_1=2)H(Y|X_1=2)\\
&= \alpha H(\beta)\\
R_1+R_2 &\leq I(X_1,X_2;Y) = H(Y) - H(Y|X_1,X_2) = H(Y)\\
&= H(\alpha\beta, \overline{\alpha}\overline{\beta}, 1-\alpha\beta-\overline{\alpha}\overline{\beta}) = H(\alpha) + \alpha H(\beta)
\end{aligned}$$

We may choose $\beta = \frac{1}{2}$ to maximize the above bounds, giving

$$\begin{aligned}
R_1 &\leq H(\alpha)\\
R_2 &\leq \alpha\\
R_1+R_2 &\leq H(\alpha)+\alpha
\end{aligned}$$

Above, we plot the region for $X_1 \in \{2,4\}$ (solid line) against that when $X_1 \in \{1,2\}$ (dotted). What we find is that, surprisingly, the rate region from the first case is not reduced in the second. In fact, neither region contains the other, so for each version of this channel, there are achievable rate pairs which are *not* achievable in the other.

21. **Broadcast Channel.** Consider the following degraded broadcast channel.



(a) What is the capacity of the channel from X to Y_1?

(b) From X to Y_2?

(c) What is the capacity region of all (R_1, R_2) achievable for this broadcast channel? Simplify and sketch.

Solution: *Broadcast Channel.*

(a) The channel from X to Y_1 is a standard erasure channel with probability of erasure $= \alpha_1$, and hence the capacity is $1-\alpha_1$

(b) We can show that the effective channel from X to Y_2 is a binary erasure channel with erasure probability $\alpha_1 + \alpha_2 - \alpha_1\alpha_2$, and hence the capacity is $1-\alpha_1-\alpha_2+\alpha_1\alpha_2 = (1-\alpha_1)(1-\alpha_2)$

(c) As in Problem 15.13, the auxiliary random variable U in the capacity region of the broadcast channel has to be binary. Hence we have the following picture

We can now evaluate the capacity region for this choice of auxiliary random variable. By symmetry, the best distribution for U is the uniform. Let $\alpha = \alpha_1 + \alpha_2 - \alpha_1\alpha_2$, and therefore $1 - \alpha = \overline{\alpha} = \overline{\alpha_1\alpha_2}$. Hence

$$\begin{aligned}
R_2 &= I(U;Y_2) && (15.176)\\
&= H(Y_2) - H(Y_2|U) && (15.177)\\
&= H\left(\frac{\overline{\alpha}}{2}, \alpha, \frac{\overline{\alpha}}{2}\right) - H((\overline{\beta}\overline{\alpha_1\alpha_2}, \alpha_1 + \overline{\alpha_1}\alpha_2, \beta\overline{\alpha_1\alpha_2}) && (15.178)\\
&= H(\alpha) + \overline{\alpha}H\left(\frac{1}{2}\right) - H(\alpha) - \overline{\alpha}H(\overline{\beta}, \beta) && (15.179)\\
&= \overline{\alpha}(1 - H(\beta)). && (15.180)
\end{aligned}$$

Also

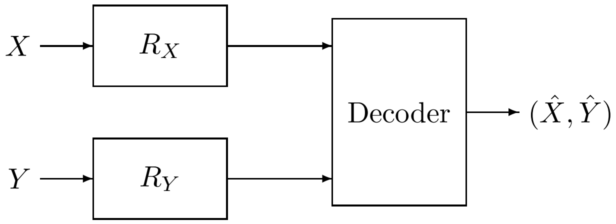
$$\begin{aligned}
R_1 &= I(X;Y_1|U) && (15.181)\\
&= H(Y_1|U) - H(Y_1|U,X) && (15.182)\\
&= H(\overline{\beta}\overline{\alpha_1}, \alpha_1, \beta\overline{\alpha_1}) - H(\alpha_1) && (15.183)\\
&= \overline{\alpha_1}H(\beta) + H(\alpha_1) - H(\alpha_1) && (15.184)\\
&= \overline{\alpha_1}H(\beta) && (15.185)
\end{aligned}$$

These two equations characterize the boundary of the capacity region as β varies. When $\beta = 0$, then $R_1 = 0$ and $R_2 = \overline{\alpha}$. When $\beta = \frac{1}{2}$, we have $R_1 = \overline{\alpha_1}$ and $R_2 = 0$.

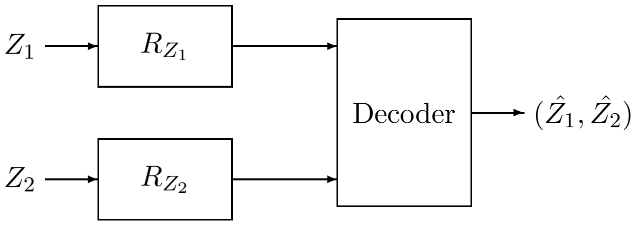
The capacity region is sketched in Figure 15.13.

22. **Stereo.** The sum and the difference of the right and left ear signals are to be individually compressed for a common receiver. Let Z_1 be Bernoulli (p_1) and Z_2 be Bernoulli (p_2) and suppose Z_1 and Z_2 are independent. Let $X = Z_1 + Z_2$, and $Y = Z_1 - Z_2$.

(a) What is the Slepian Wolf rate region of achievable (R_X, R_Y)?



(b) Is this larger or smaller than the rate region of (R_{Z_1}, R_{Z_2})? Why?



There is a simple way to do this part.

Solution: *Stereo.*

The joint distribution of X and Y is shown in following table

Z_1	Z_2	X	Y	probability
0	0	0	0	$(1-p_1)(1-p_2)$
0	1	1	-1	$(1-p_1)p_2$
1	0	1	1	$p_1(1-p_2)$
1	1	2	0	p_1p_2

and hence we can calculate

$$H(X) = H(p_1p_2, p_1+p_2-2p_1p_2, (1-p_1)(1-p_2)) \tag{15.186}$$

$$H(Y) = H(p_1p_2+(1-p_1)(1-p_2), p_1-p_1p_2, p_2-p_1p_2) \tag{15.187}$$

and

$$H(X,Y) = H(Z_1,Z_2) = H(p_1)+H(p_2) \tag{15.188}$$

and therefore

$$H(X|Y) = H(p_1)+H(p_2)-H(p_1p_2+(1-p_1)(1-p_2), p_1-p_1p_2, p_2-p_1p_2) \tag{15.189}$$

$$H(Y|X) = H(p_1)+H(p_2)-H(p_1p_2, p_1+p_2-2p_1p_2, (1-p_1)(1-p_2)) \tag{15.190}$$

The Slepian Wolf region in this case is

$$R_1 \geq H(X|Y) = H(p_1)+H(p_2)-H(p_1p_2+(1-p_1)(1-p_2), p_1-p_1p_2, p_2-p_1p_2) \tag{15.191}$$

$$R_2 \geq H(Y|X) = H(p_1)+H(p_2)-H(p_1p_2, p_1+p_2-2p_1p_2, (1-p_1)(1-p_2)) \tag{15.192}$$

$$R_1+R_2 \geq H(p_1)+H(p_2) \tag{15.193}$$

23. The Slepian Wolf region for (Z_1, Z_2) is

$$R_1 \geq H(Z_1|Z_2) = H(p_1) \tag{15.194}$$

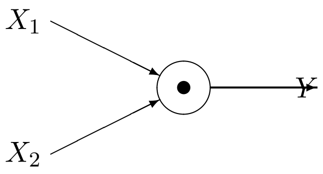
$$R_2 \geq H(Z_2|Z_1) = H(p_2) \tag{15.195}$$

$$R_1+R_2 \geq H(Z_1,Z_2) = H(p_1)+H(p_2) \tag{15.196}$$

which is a rectangular region.

The minimum sum of rates is the same in both cases, since if we knew both X and Y, we could find Z_1 and Z_2 and vice versa. However, the region in part (a) is usually pentagonal in shape, and is larger than the region in (b).

24. **Multiplicative multiple access channel.** Find and sketch the capacity region of the multiplicative multiple access channel



with $X_1 \in \{0,1\}$, $X_2 \in \{1,2,3\}$, and $Y = X_1 X_2$.

Solution: *Multiplicative multiple access channel.*

Since $Y = X_1 X_2$, if $X_1 = 0$, $Y = 0$ and we receive no information about X_2. When $X_1 = 1$, $Y = X_2$, and we can decode X_2 perfectly, thus we can achieve a rate $R_1 = 0, R_2 = \log 3$.

Let α be the probability that $X_1 = 1$. By symmetry, X_2 should have an uniform distribution on $\{1,2,3\}$. The capacity region of the multiple access channel

$$\begin{aligned} I(X_1;Y|X_2) &= H(X_1|X_2) - H(X_1|Y,X_2) = H(X_1) = H(\alpha) && (15.197) \\ I(X_2;Y|X_1) &= H(Y|X_1) = \alpha H(X_2) = \alpha \log 3 && (15.198) \\ I(X_1,X_2;Y) &= H(Y) = H(1-\alpha, \frac{\alpha}{3}, \frac{\alpha}{3}, \frac{\alpha}{3}) = H(\alpha) + \alpha \log 3 && (15.199) \end{aligned}$$

Thus the rate region is characterized by the equations

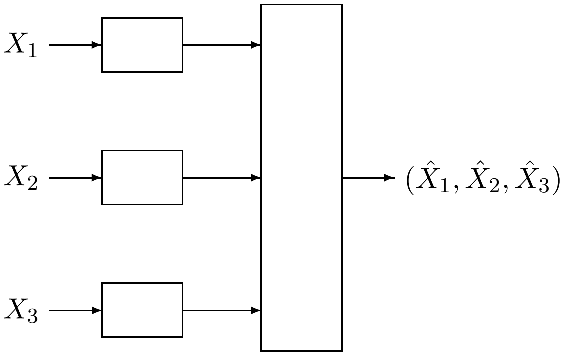
$$\begin{aligned} R_1 &\leq H(\alpha) && (15.200) \\ R_2 &\leq \alpha \log 3 && (15.201) \end{aligned}$$

where α varies from 0 to 1

The maximum value for R_1 occurs for $\alpha = \frac{1}{2}$. The maximum value for the sum of the rates occurs (by calculus) at $\alpha = \frac{3}{4}$.

25. **Distributed data compression.** Let Z_1, Z_2, Z_3 be independent Bernoulli (p). Find the Slepian-Wolf rate region for the description of (X_1, X_2, X_3) where

$$\begin{aligned} X_1 &= Z_1 \\ X_2 &= Z_1 + Z_2 \\ X_3 &= Z_1 + Z_2 + Z_3 \quad . \end{aligned}$$



Solution: *Distributed data compression.*

To establish the rate region, appeal to Theorem 14.4.2 in the text, which generalizes the case with two encoders. The inequalities defining the rate region are given by

$$R(S) > H(X(S)|X(S^c))$$

for all $S \subseteq \{1,2,3\}$, and $R(S) = \sum_{i \in S} R_i$.

The rest is calculating entropies $H(X(S)|X(S^c))$ for each S. We have

$$\begin{aligned}
H_1 &= H(X_1) = H(Z_1) = H(p), \\
H_2 &= H(X_2) = H(Z_1 + Z_2) = H(p^2, 2p(1-p), (1-p)^2), \\
H_3 &= H(X_3) = H(Z_1 + Z_2 + Z_3) \\
&= H(p^3, 3p^2(1-p), 3p(1-p)^2, (1-p)^3), \\
H_{12} &= H(X_1, X_2) = H(Z_1, Z_2) = 2H(p), \\
H_{13} &= H(X_1, X_3) = H(X_1) + H(X_3|X_1) = H(X_1) + H(Z_2 + Z_3) \\
&= H(p^2, 2p(1-p), (1-p)^2) + H(p), \\
H_{23} &= H(X_2, X_3) = H(X_2) + H(X_3|X_2) = H(X_2) + H(Z_3) \\
&= H(p^2, 2p(1-p), (1-p)^2) + H(p), \qquad \text{and} \\
H_{123} &= H(X_1, X_2, X_3) = H(Z_1, Z_2, Z_3) = 3H(p).
\end{aligned}$$

Using the above identities and chain rule, we obtain the rate region as

$$\begin{aligned}
R_1 &> H(X_1|X_2, X_3) = H_{123} - H_{23} \\
&= 2H(p) - H(p^2, 2p(1-p), (1-p)^2) = 2p(1-p), \\
R_2 &> H(X_2|X_1, X_3) = H_{123} - H_{13} = 2p(1-p), \\
R_3 &> H(X_3|X_1, X_2) = H_{123} - H_{12} = H(p), \\
R_1 + R_2 &> H(X_1, X_2|X_3) = H_{123} - H_3 \\
&= 3H(p) - H(p^3, 3p^2(1-p), 3p(1-p)^2, (1-p)^3) = 3p(1-p)\log(3),
\end{aligned}$$

$$\begin{aligned} R_1+R_3 &> H(X_1,X_3|X_2) = H_{123}-H_2 \\ &= 3H(p)-H(p^2,2p(1-p),(1-p)^2) = H(p)+2p(1-p), \\ R_2+R_3 &> H(X_2,X_3|X_1) = H_{123}-H_1 = 2H(p), \qquad \text{and} \\ R_1+R_2+R_3 &> H_{123} = 3H(p). \end{aligned}$$

(Simplifications are contributed by KBS.)

26. **Noiseless multiple access channel** Consider the following multiple access channel with two binary inputs $X_1, X_2 \in \{0,1\}$ and output $Y=(X_1,X_2)$.

(a) Find the capacity region. Note that each sender can send at capacity.

(b) Now consider the cooperative capacity region, $R_1 \geq 0, R_2 \geq 0, R_1+R_2 \leq \max_{p(x_1,x_2)} I(X_1,X_2;Y)$. Argue that the throughput R_1+R_2 does not increase, but the capacity region increases.

Solution: *Noiseless multiple access channel*

(a) Since $Y=(X_1,X_2)$, $I(X_1;Y|X_2)=H(X_1|X_2)=H(X_1)\leq 1$, and $I(X_1,X_2;Y)=H(X_1,X_2)\leq 2$, and hence the capacity region of the MAC becomes $R_1 \leq 1$, $R_2 \leq 1$, $R_1+R_2 \leq 2$.

(b) The cooperative capacity region is $R_1+R_2 \leq \max_{p(x_1,x_2)} I(X_1,X_2;Y)=2$. Thus, the cooperative capacity has the same sum of rates, but with cooperation, one of the senders could send 2 bits (while the other rate is 0). Thus the capacity region increases from the square ($R_1 \leq 1$, $R_2 \leq 1$) to the triangle $R_1+R_2 \leq 2$.

27. **Infinite bandwidth multiple access channel** Find the capacity region for the Gaussian multiple access channel with infinite bandwidth. Argue that all senders can send at their individual capacities, i.e., infinite bandwidth eliminates interference.

Solution: *Infinite bandwidth multiple access channel*

The capacity of a Gaussian multiple access channel with bandwidth W is given by the following rate region

$$R_1 \leq W\log\left(1+\frac{P_1}{NW}\right) \tag{15.202}$$

$$R_2 \leq W\log\left(1+\frac{P_2}{NW}\right) \tag{15.203}$$

$$R_1+R_2 \leq W\log\left(1+\frac{P_1+P_2}{NW}\right) \tag{15.204}$$

A hueristic argument to prove this follows from the single user Gaussian channel capacity with bandwidth W combined with “onion-peeling” and timesharing.

As $W \to \infty$, these bounds reduce to

$$\begin{aligned} R_1 &\leq \frac{P_1}{N} & (15.205) \\ R_2 &\leq \frac{P_2}{N} & (15.206) \\ R_1 + R_2 &\leq \frac{P_1+P_2}{N} & (15.207) \end{aligned}$$

which is a rectangular region corresponding to no interference between the two senders.

28. **A multiple access identity.**
Let $C(x) = \frac{1}{2}\log(1+x)$ denote the channel capacity of a Gaussian channel with signal to noise ratio x. Show

$$C\left(\frac{P_1}{N}\right) + C\left(\frac{P_2}{P_1+N}\right) = C\left(\frac{P_1+P_2}{N}\right).$$

This suggests that 2 independent users can send information as well as if they had pooled their power.

Solution: *A multiple access identity.*

$$\begin{aligned} C(\frac{P_1+P_2}{N}) &= \frac{1}{2}\log(1+\frac{P_1+P_2}{N}) & (15.208) \\ &= \frac{1}{2}\log(\frac{N+P_1+P_2}{N}) & (15.209) \\ &= \frac{1}{2}\log(\frac{N+P_1+P_2}{N+P_1}\cdot\frac{N+P_1}{N}) & (15.210) \\ &= \frac{1}{2}\log(\frac{N+P_1+P_2}{N+P_1}) + \frac{1}{2}\log(\frac{N+P_1}{N}) & (15.211) \\ &= C(\frac{P_2}{P_1+N}) + C(\frac{P}{N_1}) & (15.212) \end{aligned}$$

29. **Frequency Division Multiple Access (FDMA).** Maximize the throughput $R_1 + R_2 = W_1 \log(1+\frac{P_1}{NW_1}) + (W-W_1)\log(1+\frac{P_2}{N(W-W_1)})$ over W_1 to show that bandwidth should be proportional to transmitted power for FDMA.

Solution: *Frequency Division Multiple Access (FDMA).*

Allocating bandwidth W_1 and $W_2 = W - W_1$ to the two senders, we can achieve the following rates

$$R_1 = W_1 \log\left(1+\frac{P_1}{NW_1}\right), \qquad (15.213)$$

$$R_2 = W_2 \log\left(1+\frac{P_2}{NW_2}\right). \qquad (15.214)$$

To maximimize the sum of the rates, we write

$$R = R_1 + R_2 = W_1 \log\left(1 + \frac{P_1}{NW_1}\right) + (W - W_1)\log\left(1 + \frac{P_2}{N(W-W_1)}\right) \qquad (15.215)$$

and differentiating with respect to W_1, we obtain

$$\log\left(1 + \frac{P_1}{NW_1}\right) + \frac{W_1}{1 + \frac{P_1}{NW_1}}\left(-\frac{P_1}{NW_1^2}\right)$$
$$-\log\left(1 + \frac{P_2}{N(W-W_1)}\right) + \frac{W - W_1}{1 + \frac{P_2}{N(W-W_1)}}\left(\frac{P_2}{N(W-W_1)^2}\right) = 0 \qquad (15.216)$$

Instead of solving this equation, we can verify that if we set

$$W_1 = \frac{P_1}{P_1 + P_2} W \qquad (15.217)$$

so that

$$\frac{P_1}{NW_1} = \frac{P_2}{NW_2} = \frac{P_1 + P_2}{NW} \qquad (15.218)$$

that (15.216) is satisfied, and that using bandwidth proportional to the power optimizes the total rate for Frequency Division Multiple Access.

30. **Trilingual speaker broadcast channel**
A speaker of Dutch, Spanish and French wishes to communicate simultaneously to three people: D, S, and F. D knows only Dutch, but can distinguish when a Spanish word is being spoken as distinguished from a French word, similarly for the other two, who know only Spanish and French respectively, but can distinguish when a foreign word is spoken and which language is being spoken.

Suppose each language, Dutch, Spanish, and French, has M words: M words of Dutch, M words of French, and M words of Spanish.

(a) What is the maximum rate at which the trilingual speaker can speak to D?

(b) If he speaks to D at the maximum rate, what is the maximum rate he can simultaneously speak to S?

(c) If he is speaking to D and S at the above joint rate, can he also speak to F at some positive rate? If so, what is it? If not, why not?

Solution: *Trilingual speaker broadcast channel*

(a) Speaking Dutch gives M words, and in addition two words for the distinguishability of French and Spanish from Dutch, thus $\log(M+2)$ bits.

(b) Transmitting $\log M$ bits for a fraction of $1/(M+2)$ of the time gives $R = (\log M)/(M+2)$.

(c) Same reasoning as in (b) gives $R = (\log M)/(M+2)$.

31. **Parallel Gaussian channels from a mobile telephone**
Assume that a sender X is sending to two fixed base stations.

Assume that the sender sends a signal X that is constrained to have average power P. Assume that the two base stations receive signals Y_1 and Y_2, where

$$\begin{aligned} Y_1 &= \alpha_1 X + Z_1 \\ Y_2 &= \alpha_2 X + Z_2 \end{aligned}$$

where $Z_i \sim \mathcal{N}(0, N_1)$, $Z_2 \sim \mathcal{N}(0, N_2)$, and Z_1 and Z_2 are independent. We will assume the α's is constant over a transmitted block.

(a) Assuming that both signals Y_1 and Y_2 are available at a common decoder $Y = (Y_1, Y_2)$, what is the capacity of the channel from the sender to the common receiver?

(b) If instead the two receivers Y_1 and Y_2 each independently decode their signals, this becomes a broadcast channel. Let R_1 be the rate to base station 1 and R_2 be the rate to base station 2. Find the capacity region of this channel.

Solution: *Parallel Gaussian channels from a mobile telephone*

(a) Let $Y = [Y_1, Y_2]^T$. Obviously,

$$I(X;Y) = h(Y_1, Y_2) - h(Z_1, Z_2)$$

thus it is clear that the maximizing distribution on X is Gaussian $\mathcal{N}(0, P)$. Therefore we have

$$h(Y_1, Y_2) = \frac{1}{2} \log 2\pi e |K_Y|$$

and consequently, by independence of the noises

$$C = \frac{1}{2} \log \frac{|K_Y|}{N_1 N_2} \quad .$$

Plugging in $|K_Y| = (1-\alpha)^2 P N_1 + \alpha^2 P N_2 + N_1 N_2$ we have

$$C = \frac{1}{2} \log \left(1 + \frac{(1-\alpha)^2 P}{N_2} + \frac{\alpha^2 P}{N_1} \right) \quad .$$

(b) The problem is equivalent to the degraded broadcast channel with

$$\begin{aligned} Y_1 &= X + Z_1/\alpha \\ Y_2 &= X + Z_2/(1-\alpha) \quad . \end{aligned}$$

Thus, the noise is $\mathcal{N}(0, N_1/\alpha^2)$ and $\mathcal{N}(0, N_2/(1-\alpha)^2)$. Without loss of generality assume that $N_2/(1-\alpha)^2 > N_1/\alpha^2$. Then, referring to Example 14.6.6. in Cover and Thomas, the rate region is

$$\begin{aligned} R_1 &< C\left(\frac{\theta\alpha^2 P}{N_1}\right) \\ R_2 &< C\left(\frac{(1-\theta)(1-\alpha)^2 P}{\theta(1-\alpha)^2 P + N_2}\right) \quad , \quad 0 \le \theta \le 1 \quad . \end{aligned}$$

32. **Gaussian multiple access.**

A group of m users, each with power P, is using a Gaussian multiple access channel at capacity, so that

$$\sum_{i=1}^{m} R_i = C\left(\frac{mP}{N}\right), \tag{15.219}$$

where $C(x) = \frac{1}{2}\log(1+x)$ and N is the receiver noise power.

A new user of power P_0 wishes to join in.

(a) At what rate can he send without disturbing the other users?

(b) What should his power P_0 be so that the new users rate is equal to the combined communication rate $C(mP/N)$ of all the other users?

Solution: *Gaussian multiple access.*

(a) If the new user can be decoded while treating all the other senders as part of the noise, then his signal can be subtracted out before decoding the other senders, and hence will not disturb the rates of the other senders. Therefore if

$$R_0 < \frac{1}{2}\log\left(1 + \frac{P_0}{mP+N}\right), \tag{15.220}$$

the new user will not disturb the other senders.

(b) The new user will have a rate equal to the sum of the existing senders if

$$\frac{1}{2}\log\left(1 + \frac{P_0}{mP+N}\right) = \frac{1}{2}\log\left(1 + \frac{mP}{N}\right) \tag{15.221}$$

or

$$P_0 = (mP+N)\frac{mP}{N} \tag{15.222}$$

33. **Converse for deterministic broadcast channel.**

A deterministic broadcast channel is defined by an input X, two outputs, Y_1 and Y_2 which are functions of the input X. Thus $Y_1 = f_1(X)$ and $Y_2 = f_2(X)$. Let R_1 and R_2 be the rates at which information can be sent to the two receivers. Prove that

$$R_1 \le H(Y_1) \tag{15.223}$$

$$R_2 \le H(Y_2) \tag{15.224}$$

$$R_1 + R_2 \le H(Y_1, Y_2) \tag{15.225}$$

Solution: *Converse for deterministic broadcast channel.*

We can derive this bound from single user arguments. The maximum rate that the sender can send information to receiver 1 is less than

$$R_1 \le I(X;Y_1) = H(Y_1) - H(Y_1|X) = H(Y_1) \tag{15.226}$$

since the channel is deterministic and therefore $H(Y_1|X) = H(Y_2|X) = 0$. Similarly, $R_2 \le H(Y_2)$.

Also, if the receivers cooperated with each other, the capacity

$$R_1 + R_2 \le I(X;Y_1,Y_2) = H(Y_1,Y_2) \tag{15.227}$$

since the sum of rates to the two receivers without cooperation cannot be greater than the single user capacity of a channel from X to (Y_1,Y_2).

34. **Multiple access channel**
Consider the multiple access channel $Y = X_1 + X_2 \pmod 4$, where $X_1 \in \{0,1,2,3\}$, $X_2 \in \{0,1\}$.

(a) Find the capacity region (R_1, R_2).

(b) What is the maximum throughput $R_1 + R_2$?

Solution: *Multiple access channel*

(a) The MAC capacity region is given by the standard set of equations which reduce as follows since there is no noise:

$$\begin{aligned} R_1 &< I(X_1;Y|X_2) = H(Y|X_2) - H(Y|X_1,X_2) = H(Y|X_2) = H(X_1) \\ R_2 &< I(X_2;Y|X_1) = H(Y|X_1) - H(Y|X_1,X_2) = H(Y|X_1) = H(X_2) \\ R_1 + R_2 &< I(X_1,X_2;Y) = H(Y) - H(Y|X_1,X_2) = H(Y) \end{aligned}$$

Since entropy is maximized under a uniform distribution over the finite alphebet, $R_1 < H(X_1) \le 2$, $R_2 < H(X_2) \le 1$, and $R_1 + R_2 < H(Y) \le 2$. Further, if $X_1 \sim unif(0,1,2,3)$, and $X_2 \sim unif(0,1)$ then $Y \sim unif(0,1,2,3)$, so the upper bounds are achieved. This gives the capacity region in Figure 15.14.

(b) The throughput of $R_1 + R_2 \le 2$ by the third constraint above, and is achieved at many points including when $R_1 = 2$ and $R_2 = 0$. So the maximum throughput is $R_1 + R_2 = 2$.

35. **Distributed source compression**
Let

$$Z_1 = \begin{cases} 1, & p \\ 0, & q, \end{cases}$$

$$Z_2 = \begin{cases} 1, & p \\ 0, & q, \end{cases}$$

and let $U = Z_1Z_2$, $V = Z_1 + Z_2$. Assume Z_1 and Z_2 are independent. This induces a joint distribution on (U, V). Let (U_i, V_i) be iid according to this distribution. Sender 1 describes U^n at rate R_1, and sender 2 describes V^n at rate R_2.

(a) Find the Slepian-Wolf rate region for recovering (U^n, V^n) at the receiver.

(b) What is the residual uncertainty (conditional entropy) that the receiver has about (X^n, Y^n).

Solution: *Distributed source compression*

(a) Below is a table listing the possible results and their associated probabilities.

Z_1	Z_2	U	V	Prob
0	0	0	0	q^2
0	1	0	1	pq
1	0	0	1	pq
1	1	1	2	p^2

Evaluating the three standard inequalities for the Slepian-Wolf rate region gives the following:

$$\begin{aligned} R_1 &> H(U|V) = 0 \\ R_2 &> H(V|U) = Pr(U=0)H(V|U=0) = (1-p^2)H\left(\frac{q^2}{1-p^2}\right) \\ R_1 + R_2 &> H(U,V) = H(V) + H(U|V) = H(V) = H(q^2, 2pq, p^2) \end{aligned}$$

Where the first equation comes because U is a deterministic function of V. The second equation comes from the definition of conditional entropy and noting that $H(V|U=1) = 0$. The Slepian-Wolf rate region is depicted in figure 15.15.

(b) The residual uncertainty is given by $H(Z_1^n, Z_2^n|U^n, V^n) = nH(Z_1, Z_2|U, V)$ because everything is iid. Since there is only uncertainty in (Z_1, Z_2) when $(U = 0, V = 1)$, the residual uncertainty simplifies to $nPr(U = 0, V = 1)H(Z_1, Z_2|U = 0, V = 1) = n(2pq)H\left(\frac{1}{2}\right) = 2pqn$.

36. **MAC capacity with costs**

The cost of using symbol x is $r(x)$. The cost of a codeword x^n is $r(x^n) = \frac{1}{n}\sum_{i=1}^n r(x_i)$. A $(2^{nR}, n)$ codebook satisfies cost constraint r if $\frac{1}{n}\sum_{i=1}^n r(x_i(w)) \le r$, for all $w \in 2^{nR}$.

(a) Find an expression for the capacity $C(r)$ of a discrete memoryless channel with cost constraint r.

(b) Find an expression for the multiple access channel capacity region for $(\mathcal{X}_1 \times \mathcal{X}_2, p(y|x_1, x_2), \mathcal{Y})$ if sender X_1 has cost constraint r_1 and sender X_2 has cost constraint r_2.

(c) Prove the converse for (b).

Solution: *MAC capacity with costs*

(a) The capacity of a discrete memoryless channel with cost constraint r is given by

$$C(r) = \max_{p(x):\sum_x p(x)r(x)\leq r} I(X;Y). \tag{15.228}$$

The achievability follows immediately from Shannon's 'average over random codebooks' method and joint typicality decoding. (See Section 9.1 for the power constraint example.)

For the converse, we need to establish following simple properties of the capacity-cost function $C(r)$.

Theorem 15.0.4 *The capacity cost function* $C(r)$ *given in* (15.228) *is a non-decreasing concave function of* r.

Remark: These properties of the capacity cost function $C(r)$ exactly parallel those of the rate distortion function $R(D)$. (See Lemma 10.4.1 of the text.)

Proof: The monotonicity is a direct consequence of the definition of $C(r)$. To prove the concavity, consider two points (C_1, r_1) and (C_2, r_2) which lie on the capacity cost curve. Let the distributions that achieve these pairs be $p_1(x)$ and $p_2(x)$. Consider the distribution $p_\lambda = \lambda p_1 + (1-\lambda)p_2$. Since the cost is a linear function of the distribution, we have $r(p_\lambda) = \lambda r_1 + (1-\lambda)r_2$. Mutual information, on the other hand, is a concave function of the input distribution (Theorem 2.7.4) and hence

$$\begin{aligned}
C(\lambda r_1 + (1-\lambda)r_2) &= C(r(p_\lambda)) && (15.229)\\
&\geq I_{p_\lambda}(X;Y) && (15.230)\\
&\geq \lambda I_{p_1}(X;Y) + (1-\lambda)I_{p_2}(X;Y) && (15.231)\\
&= \lambda C(r_1) + (1-\lambda)C(r_2), && (15.232)
\end{aligned}$$

which proves that $C(r)$ is concave in r. □

Now we are ready to prove the converse. Consider any $(2^{nR}, n)$ code that satisfies the cost constraint

$$\frac{1}{n}\sum_{i=1}^{n} r(x_i(w)) \leq r$$

for $w = 1, 2, \ldots, 2^{nR}$, which in turn implies that

$$\frac{1}{n}\sum_{i=1}^{n} E(r(X_i)) \leq r, \tag{15.233}$$

where the expectation is with respect to the uniformly drawn message index W. As in the case without the cost constraint, we begin with Fano's inequality to obtain the following chain of inequalities:

$$
\begin{aligned}
nR &= H(W) && (15.234)\\
&\le I(W;Y^n)+n\epsilon_n && (15.235)\\
&\le I(X^n;Y^n)+n\epsilon_n && (15.236)\\
&\le H(Y^n)-H(Y^n|X^n)+n\epsilon_n && (15.237)\\
&\le \sum_{i=1}^n H(Y_i)-H(Y_i|X^n,Y^{i-1})+n\epsilon_n && (15.238)\\
&= \sum_{i=1}^n H(Y_i)-H(Y_i|X_i)+n\epsilon_n && (15.239)\\
&= \sum_{i=1}^n I(X_i;Y_i)+n\epsilon_n && (15.240)\\
&\overset{(a)}{\le} \sum_{i=1}^n C(E(r(X_i)))+n\epsilon_n && (15.241)\\
&= n\sum_{i=1}^n \frac{1}{n}C(E(r(X_i)))+n\epsilon_n && (15.242)\\
&\overset{(b)}{\le} nC\left(\frac{1}{n}\sum_{i=1}^n E(r(X_i))\right)+n\epsilon_n && (15.243)\\
&\overset{(c)}{\le} nC(r)+n\epsilon_n, && (15.244)
\end{aligned}
$$

where
(a) follows from the definition of the capacity cost function,
(b) from the concavity of the capacity cost function and Jensen's inequality, and
(c) from Eq. (15.233) and the fact that $C(r)$ is non-decreasing in r.

Note that we cannot jump from (15.240) to (15.244) since $E(r(X_i))$ may be greater than r for some i.

(b) The capacity region under cost constraints r_1 and r_2 is given by the closure of the set of all (R_1,R_2) pairs satisfying

$$
\begin{aligned}
R_1 &< I(X_1;Y|X_2,Q),\\
R_2 &< I(X_2;Y|X_1,Q),\\
R_1+R_2 &< I(X_1,X_2;Y|Q)
\end{aligned}
$$

for some choice of the joint distribution $p(q)p(x_1|q)p(x_2|q)p(y|x_1,x_2)$ with

$$\sum_{x_1} p(x_1)r_1(x_1)\le r_1,$$

$$\sum_{x_2} p(x_2)r_2(x_2)\le r_2,$$

and $|\mathcal{Q}|\le 4$.

(c) Again the achievability proof is an easy extension from the case without cost constraints. For the converse, consider any sequence of $((2^{nR_1}, 2^{nR_2}), n)$ codes with $P_e^{(n)} \to 0$ satisfying

$$\begin{aligned}\frac{1}{n}\sum_i r_1(x_{1i}(w_{1i})) &\leq r_1\\ \frac{1}{n}\sum_i r_2(x_{2i}(w_{2i})) &\leq r_2,\end{aligned}$$

for all $w_{1i} = 1, 2, \ldots, 2^{nR_1}$, $w_{2i} = 1, 2, \ldots, 2^{nR_2}$. By taking expectation with respect to the random message index pair (W_1, W_2), we get

$$\frac{1}{n}\sum_i E(r_1(X_{1i})) \leq r_1 \quad \text{and} \quad \frac{1}{n}\sum_i E(r_2(X_{2i})) \leq r_2. \tag{15.245}$$

By starting from Fano's inequality and taking the exact same steps as in the converse proof for the MAC without constraints (see Section 14.3.4 of the text), we obtain

$$\begin{aligned}nR_1 &\leq \sum_{i=1}^n I(X_{1i}; Y_i|X_{2i}) + n\epsilon_{1n} = nI(X_{1Q}; Y_Q|X_{2Q}, Q) + n\epsilon_{1n},\\ nR_2 &\leq \sum_{i=1}^n I(X_{2i}; Y_i|X_{1i}) + n\epsilon_{2n} = nI(X_{2Q}; Y_Q|X_{1Q}, Q) + n\epsilon_{2n},\\ n(R_1 + R_2) &\leq \sum_{i=1}^n I(X_{1i}, X_{2i}; Y_i) + n\epsilon_n = nI(X_{1Q}, X_{2Q}; Y_Q|Q) + n\epsilon_n,\end{aligned}$$

where the random variable Q is uniform over $\{1, 2, \ldots, n\}$ and independent of (X_{1i}, X_{2i}, Y_i) for all i.

Now define $X_1 \overset{\triangle}{=} X_{1Q}$, $X_2 \overset{\triangle}{=} X_Q$, and $Y \overset{\triangle}{=} Y_Q$. It is easy to check that (Q, X_1, X_2, Y) have a joint distribution of the form $p(q)p(x_1|q)p(x_2|q)p(y|x_1, x_2)$. Moreover, from Eq. (15.245),

$$\begin{aligned}\sum_{x_1} \Pr(X_1 = x_1) r_1(x_1) &= \sum_{x_1} \Pr(X_{1Q} = x_1) r_1(x_1)\\ &= \sum_{x_1}\sum_{i=1}^n \Pr(X_{1Q} = x_1|Q = i) \Pr(Q = i) r_1(x_1)\\ &= \sum_{x_1}\sum_{i=1}^n \frac{1}{n} \Pr(X_{1i} = x_1) r_1(x_1)\\ &= \frac{1}{n}\sum_{i=1}^n \sum_{x_1} \Pr(X_{1i} = x_1) r_1(x_1)\\ &= \frac{1}{n}\sum_{i=1}^n \sum_{x_1} \Pr(X_{1i} = x_1) r_1(x_1)\end{aligned}$$

$$\begin{aligned} &= \frac{1}{n}\sum_{i=1}^{n} E(r_1(X_{1i})) \\ &\leq r_1, \end{aligned}$$

and similarly,

$$\sum_{x_2} \Pr(X_2 = x_2) r_2(x_2) \leq r_2.$$

Therefore, we have shown that any sequence of $((2^{nR_1}, 2^{nR_2}), n)$ codes satisfying cost constraints with $P_e^{(n)} \to 0$ should have the rates satisfying

$$\begin{aligned} R_1 &< I(X_1; Y|X_2, Q), \\ R_2 &< I(X_2; Y|X_1, Q), \\ R_1 + R_2 &< I(X_1, X_2; Y|Q) \end{aligned}$$

for some choice of the joint distribution $p(q)p(x_1|q)p(x_2|q)p(y|x_1, x_2)$ with

$$\sum_{x_1} p(x_1) r_1(x_1) \leq r_1$$

and

$$\sum_{x_2} p(x_2) r_2(x_2) \leq r_2.$$

Finally, from Theorem 14.3.4, the region is unchanged if we limit the cardinality of $\mathcal{Q}$ to 4, which completes the proof of the converse.

Note that, compared to the single user case in part (a), the converse for the MAC with cost constraints is rather straightforward. Here the time sharing random variable Q saves the trouble of dealing with costs at each time index i.

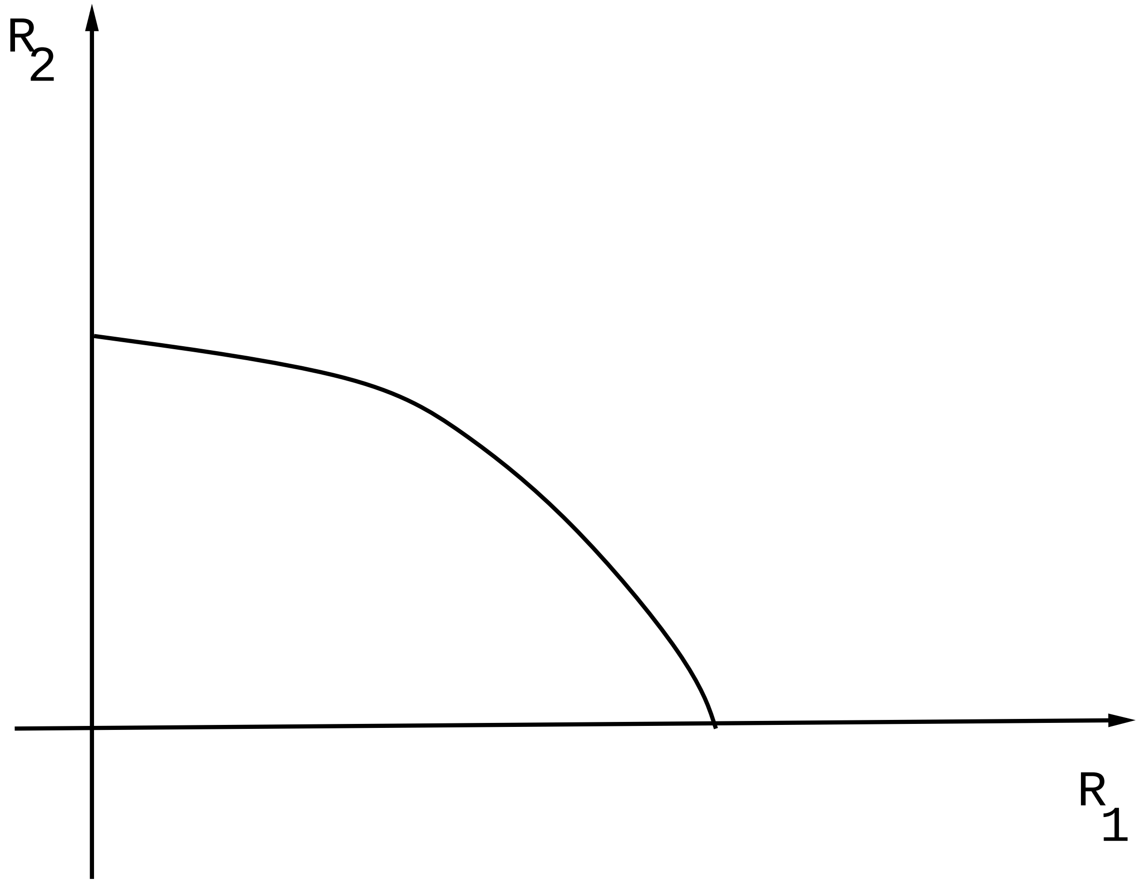


Figure 15.7: Capacity region of degraded broadcast channel

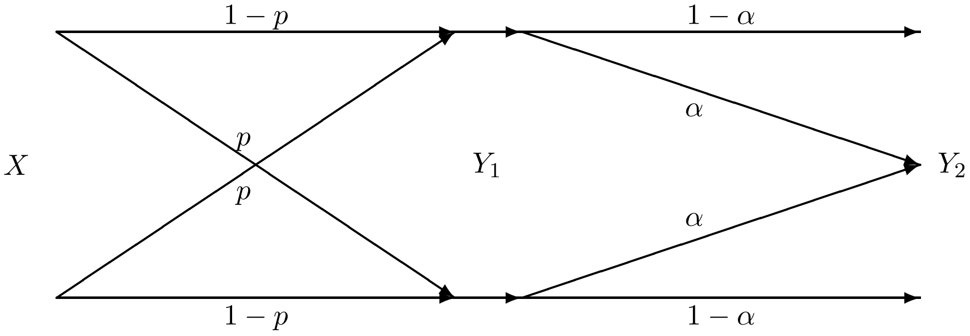


Figure 15.8: Broadcast channel with a binary symmetric channel and an erasure channel

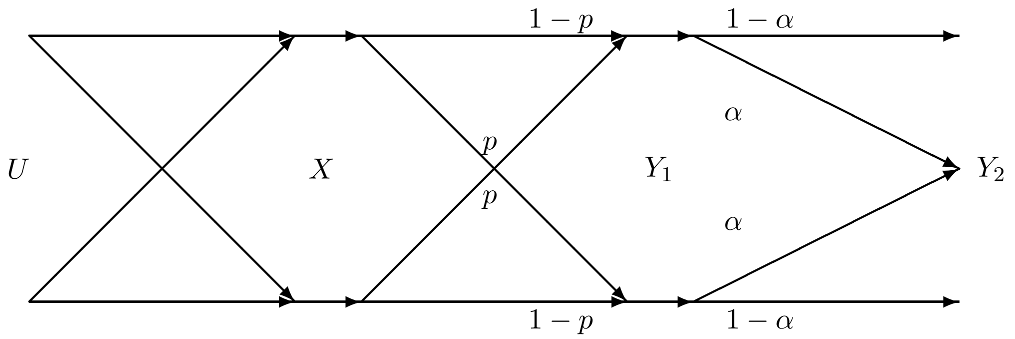
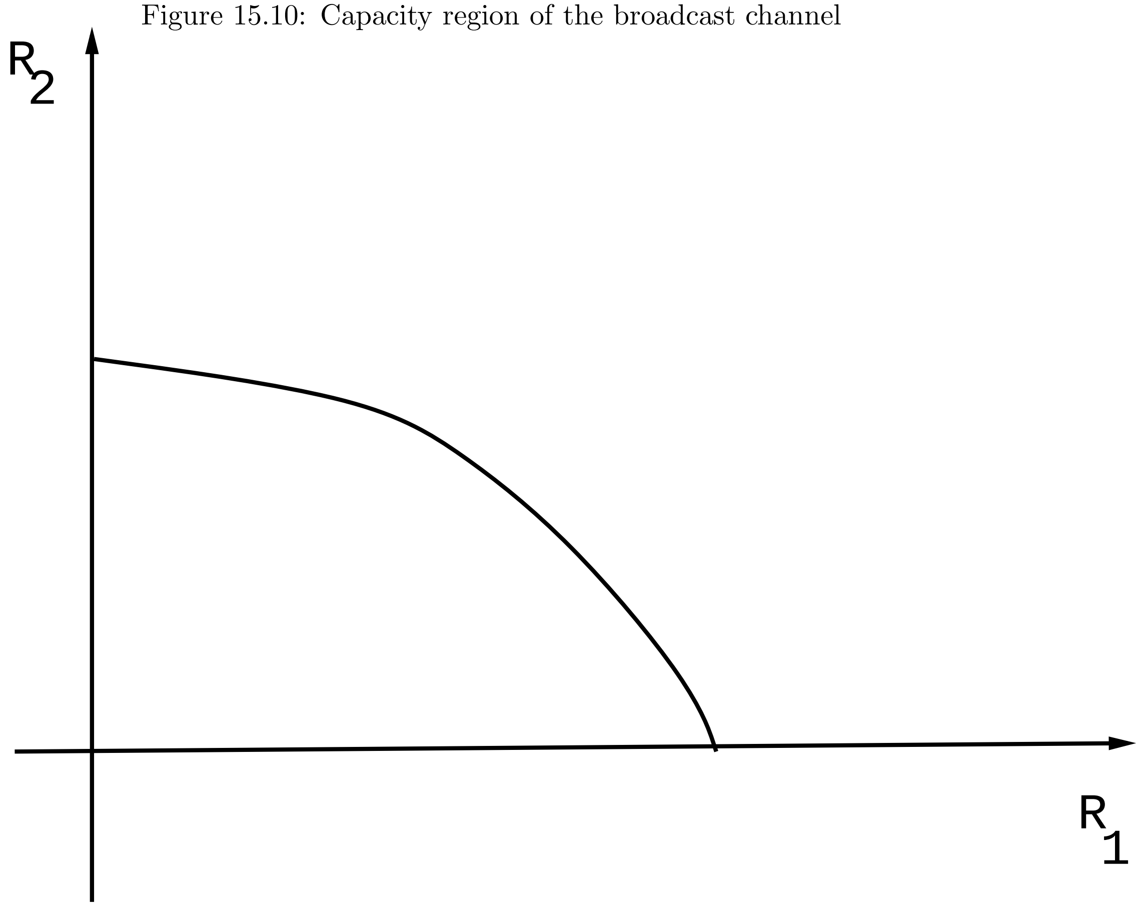


Figure 15.9: Broadcast channel with auxiliary random variable

Figure 15.10: Capacity region of the broadcast channel



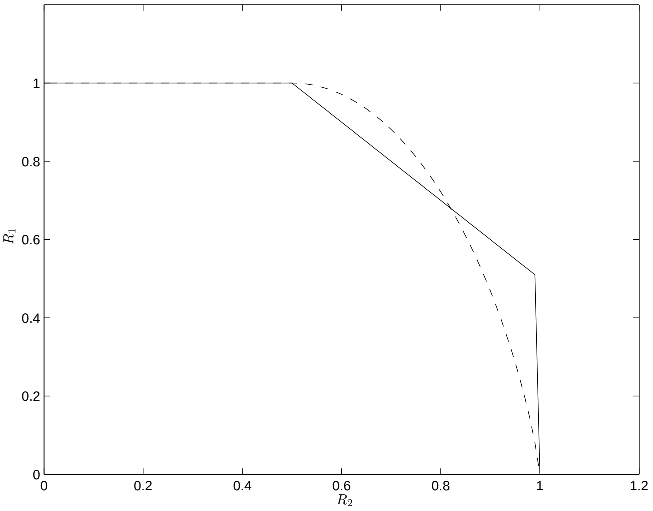


Figure 15.11: Rate regions for $X_1 \in \{2, 4\}$ and $X_1 \in \{1, 2\}$

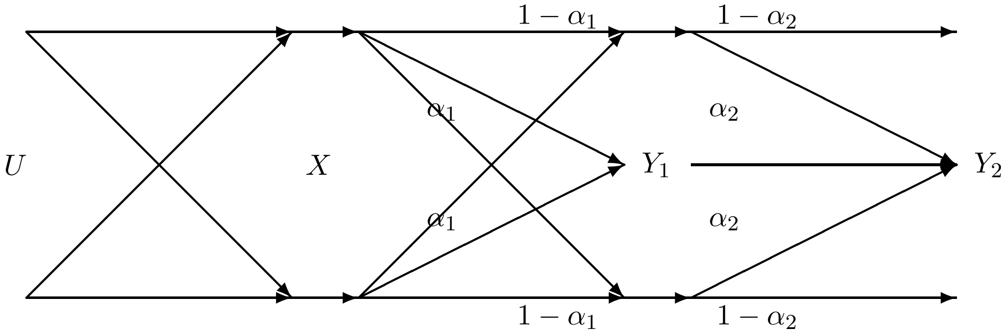
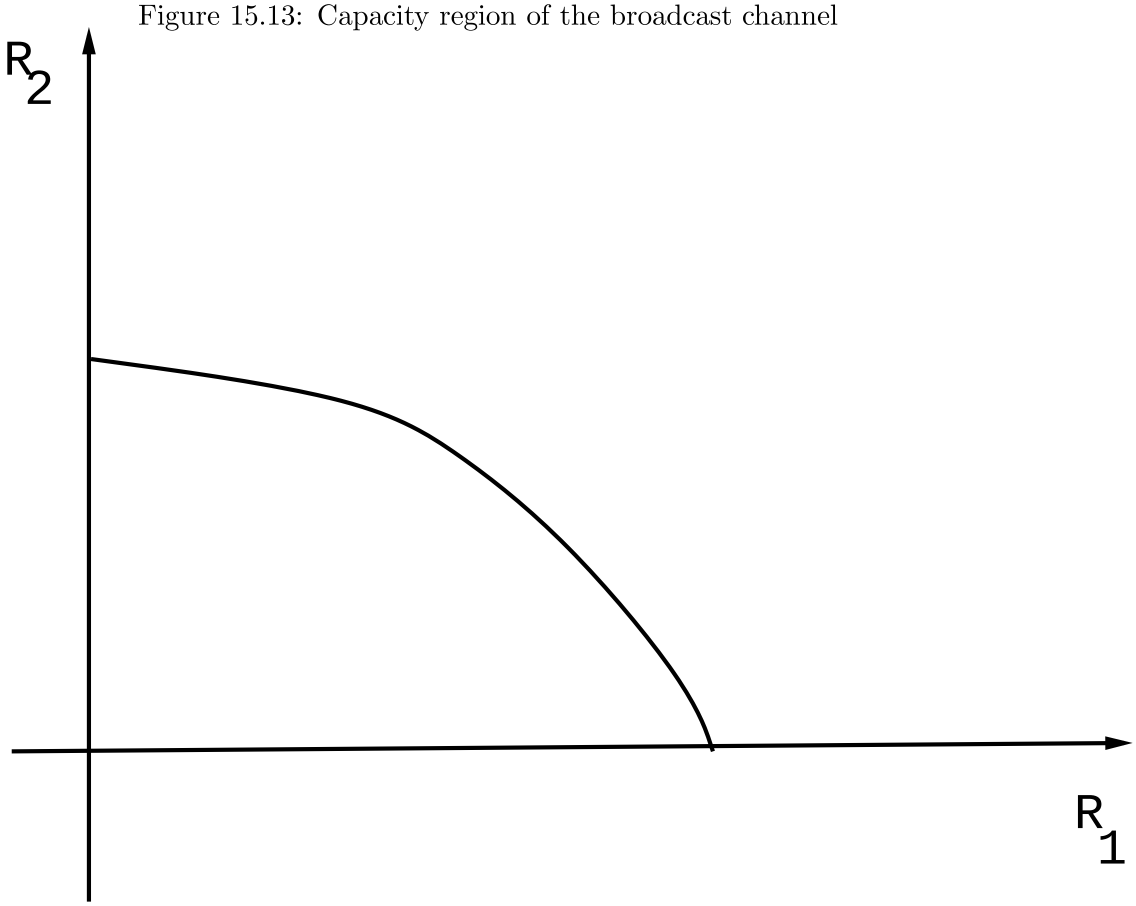


Figure 15.12: Broadcast channel with auxiliary random variable

Figure 15.13: Capacity region of the broadcast channel



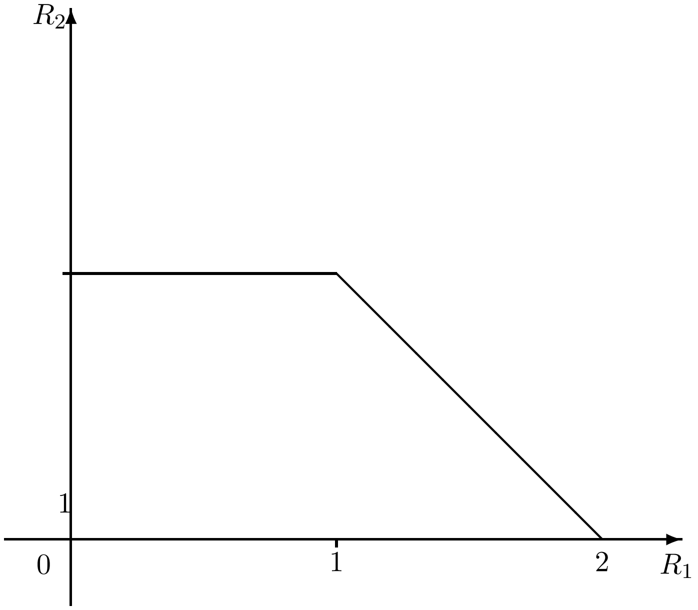


Figure 15.14: MAC Capacity Region

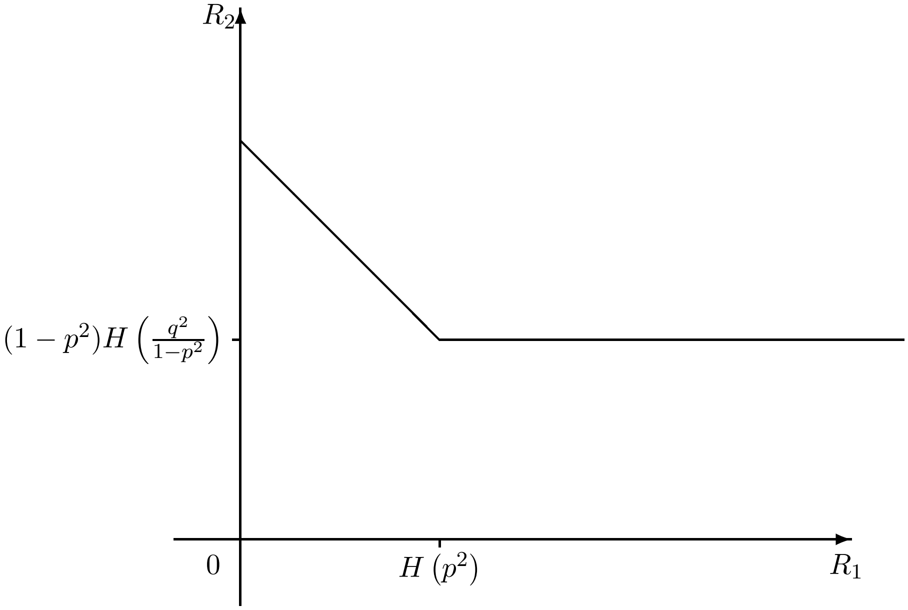


Figure 15.15: SW Rate Region

Chapter 16

Information Theory and Portfolio Theory

1. **Growth rate.** Let

$$X = \begin{cases} (1,a), & \text{with probability } 1/2 \\ (1,1/a), & \text{with probability } 1/2 \end{cases},$$

where $a > 1$. This vector X represents a stock market vector of cash vs. a hot stock. Let

$$W(\mathbf{b},F) = E \log \mathbf{b}^t \mathbf{X},$$

and

$$W^* = \max_{\mathbf{b}} W(\mathbf{b},F)$$

be the growth rate.

(a) Find the log optimal portfolio $\mathbf{b}^*$.

(b) Find the growth rate W^*.

(c) Find the asymptotic behavior of

$$S_n = \prod_{i=1}^{n} \mathbf{b}^t \mathbf{X}_i$$

for all $\mathbf{b}$.

Solution: *Doubling Rate.*

(a) Let the portfolio be $(1-b_2, b_2)$. Then

$$W(\mathbf{b},F) = \frac{1}{2}\ln(1-b_2+ab_2) + \frac{1}{2}\ln(1-b_2+\frac{b_2}{a}). \tag{16.1}$$

Differentiating to find the maximum, we have

$$\frac{dW}{db_2} = \frac{1}{2}\frac{a-1}{1-b_2+ab_2} - \frac{1-\frac{1}{a}}{1-b_2+\frac{b_2}{a}} = 0 \tag{16.2}$$

Solving this equation, we get $b_2^* = \frac{1}{2}$. Hence the log optimal portfolio $\mathbf{b}^*$ is $(\frac{1}{2}, \frac{1}{2})$.

(b) The optimal doubling rate $W^* = W(\mathbf{b}^*, F)$ is

$$\begin{aligned} W^* &= \frac{1}{2}\ln\left(\frac{1}{2}+\frac{a}{2}\right) + \frac{1}{2}\ln\left(\frac{1}{2}+\frac{1}{2a}\right) && (16.3) \\ &= \frac{1}{2}\ln\frac{(1+a)^2}{a} - \ln 2. && (16.4) \end{aligned}$$

(c) The asymptotic behavior of an infinite product of i.i.d. terms is essentially determined by the expected log of the individual terms.

$$\begin{aligned} S_n &= \prod_{i=1}^{n} \mathbf{b}^t\mathbf{X}_i && (16.5) \\ &= e^{n\frac{1}{n}\sum_{i=1}^{n}\ln \mathbf{b}^t\mathbf{X}_i} && (16.6) \\ &\to e^{nE\ln \mathbf{b}^t\mathbf{X}} && (16.7) \\ &= e^{nW(\mathbf{b},F)}, && (16.8) \end{aligned}$$

where the convergence is with probability 1 by the strong law of large numbers. We can substitute for $W(\mathbf{b}, F)$ from (16.1).

2. **Side information.** Suppose, in the previous problem, that

$$\mathbf{Y} = \begin{cases} 1, & \text{if } (X_1, X_2) \geq (1,1), \\ 0, & \text{if } (X_1, X_2) \leq (1,1). \end{cases}$$

Let the portfolio $\mathbf{b}$ depend on $\mathbf{Y}$. Find the new growth rate W^{**} and verify that $\Delta W = W^{**} - W^*$ satisfies

$$\Delta W \leq I(X;Y).$$

Solution: *Side Information.*

In the previous problem, if we knew Y so that we knew which of the two possible stock vectors would occur, then the optimum strategy is clear. In the case when $Y = 1$, we should put all out money in the second stock to maximize the conditional expected log return. Similarly, when $Y = 0$, we should put all the money in the first stock. The average expected log return is

$$W^*(Y) = \frac{1}{2}\ln a + \frac{1}{2}\ln 1 = \frac{1}{2}\ln a. \tag{16.9}$$

The increase in doubling rate due to the side information is

$$\begin{aligned}
\Delta W &= W^*(Y) - W^* && (16.10)\\
&= \frac{1}{2}\ln a - \frac{1}{2}\ln\frac{(1+a)^2}{a} + \ln 2 && (16.11)\\
&= \ln\frac{a}{1+a} + \ln 2 && (16.12)\\
&\le \ln 2, && (16.13)
\end{aligned}$$

since $\frac{a}{1+a} < 1$. Also in this case,

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) = \ln 2, \qquad (16.14)$$

since Y is a function of X and uniformly distributed on $\{0,1\}$.

We can hence verify that

$$\Delta W \le I(X;Y). \qquad (16.15)$$

3. **Stock dominance.** Consider a stock market vector

$$\mathbf{X} = (X_1, X_2).$$

Suppose $X_1 = 2$ with probability 1.

(a) Find necessary and sufficient conditions on the distribution of stock X_2 such that the log optimal portfolio $\mathbf{b}^*$ invests all the wealth in stock X_2, i.e., $\mathbf{b}^* = (0,1)$.

(b) Argue for any distribution on X_2 that the growth rate satisfies $W^* \ge 1$.

Solution: *Stock Market* We have a stock market vector

$$\mathbf{X} = (X_1, X_2)$$

with $X_1 = 2$.

(a) The Kuhn Tucker conditions for the portfolio $\mathbf{b} = (0,1)$ to be optimal is that

$$E\frac{X_2}{X_2} = 1 \qquad (16.16)$$

and

$$E\frac{2}{X_2} \le 1. \qquad (16.17)$$

The first is trivial. So the second condition is the only condition on the distribution for the optimal portfolio to be (0,1).

(b) Since the optimal portfolio does better than the (1,0) portfolio

$$W^* \ge W(\mathbf{b}) = W(1,0) = \log 2 = 1. \qquad (16.18)$$

4. **Including experts and mutual funds.** Let $\mathbf{X} \sim F(\mathbf{x}), \mathbf{x} \in \mathcal{R}_+^m$ be the vector of price relatives for a stock market. Suppose an "expert" suggests a portfolio $\mathbf{b}$. This would result in a wealth factor $\mathbf{b}^t\mathbf{X}$. We add this to the stock alternatives to form $\tilde{\mathbf{X}} = (X_1, X_2, \ldots, X_m, \mathbf{b}^t\mathbf{X})$. Show that the new growth rate

$$\tilde{W}^* = \max_{b_1,\ldots,b_m,b_{m+1}} \int \ln(\mathbf{b}^t\tilde{\mathbf{x}}) dF(\tilde{\mathbf{x}}) \tag{16.19}$$

is equal to the old growth rate

$$W^* = \max_{b_1,\ldots,b_m} \int \ln(\mathbf{b}^t\mathbf{x}) dF(\mathbf{x}). \tag{16.20}$$

Solution: *Including experts and mutual funds.*

This problem asks you to show that the existence of a mutual fund does not fundamentally change the market; that is, it asks you to show that you can make as much money without the presence of the mutual fund as you can make with it. This should be obvious, since, if you thought a particular mutual fund would be a good idea to hold, you could always invest in its constituent stocks directly in exactly the same proportions as the mutual fund did.

(a) **Outline of Proof**
We are asked to compare two quantities, $\hat{W}^*$ and W^*. $\hat{W}^*$ is the maximum doubling rate of the "extended" market. That is, it is the maximimum achievable doubling-rate over the set of extended portfolios: those that include investment in the mutual fund. W^* is the maximum doubling rate of the "non-extended" market. That is, it is the maximum achievable doubling-rate over the set of non-extended portfolios: those without investment in the mutual fund. Our strategy will be to show that the set of achievable doubling rates in the extended market is precisely the same as the set of achievable doubling rates in the non-extended market, and hence that the maximum value on both sets must be the same. In particular, we need to show that for any extended portfolio $\hat{\mathbf{b}}$ that achieves some particular doubling rate $\hat{W}$ on the extended market, there exists a corresponding non-extended portfolio $\mathbf{b}$ that achieves the same doubling rate $W = \hat{W}$ on the non-extended market, and, conversely, that for any non-extended portfolio $\mathbf{b}$ achieving some particular doubling-rate on the non-extended market, we can find an equivalent extended portfolio $\hat{\mathbf{b}}$ that achieves the same doubling-rate on the extended market.

(b) **Converse:** $W^* \le \hat{W}^*$

The converse is easy. Let $\mathbf{b} = (b_1, b_2, \ldots, b_m)$ be any non-extended portfolio. Then clearly the extended portfolio $\hat{\mathbf{b}} = (\mathbf{b}, 0) = (b_1, b_2, \ldots b_m, 0)$ achieves the same doubling rate on the extended market. In particular, then, if $\mathbf{b}^*$ achieves W^* on the non-extended market, then $(\mathbf{b}^*, 0)$ achieves W^* on the extended market, and so the maximum doubling rate on the extended market must be at least as big as W^*, that is: $W^* \le \hat{W}^*$

(c) $\hat{W}^* \leq W^*$

First, some definitions. Let $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ be the non-extended stock-market. Let $\mathbf{c} = (c_1, c_2, \ldots, c_m)$ be the portfolio that generates the mutual fund, X_{m+1}. Thus, $X_{m+1} = \mathbf{c}^T\mathbf{X}$. Let $\hat{\mathbf{X}} = (X_1, X_2, \ldots, X_m, X_{m+1})$ be the extended stock-market.

Now consider any extended portfolio $\hat{\mathbf{b}} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_m, \hat{b}_{m+1})$. The doubling rate $\hat{W}$ associated with the portfolio $\hat{\mathbf{b}}$ is

$$\begin{aligned}
\hat{W} &= \mathbf{E}[\log(\hat{\mathbf{b}}^T\hat{\mathbf{X}})] \\
&= \mathbf{E}[\log(\hat{b}_1X_1 + \hat{b}_2X_2 + \ldots + \hat{b}_mX_m + \hat{b}_{m+1}X_{m+1})] \\
&= \mathbf{E}[\log(\hat{b}_1X_1 + \hat{b}_2X_2 + \ldots + \hat{b}_mX_m + \hat{b}_{m+1}c^T\mathbf{X})] \\
&= \mathbf{E}[\log(\hat{b}_1X_1 + \hat{b}_2X_2 + \ldots + \hat{b}_mX_m + \hat{b}_{m+1}(c_1X_1 + c_2X_2 + \ldots + c_mX_m))] \\
&= \mathbf{E}[\log(\hat{b}_1X_1 + \hat{b}_2X_2 + \ldots + \hat{b}_mX_m + \hat{b}_{m+1}(c_1X_1 + c_2X_2 + \ldots + c_mX_m))] \\
&= \mathbf{E}[\log((\hat{b}_1 + \hat{b}_{m+1}c_1)X_1 + (\hat{b}_2 + \hat{b}_{m+1}c_2)X_2 + \ldots + (\hat{b}_m + \hat{b}_{m+1}c_m)X_m)]
\end{aligned}$$

But this last expression can be re-expressed as the doubling rate W associated with the non-extended portfolio $\mathbf{b}$, where $b_i = \hat{b}_i + \hat{b}_{m+1}c_i$. In particular, then, when $\hat{\mathbf{b}} = \hat{\mathbf{b}}^*$ is the portfolio achieving the optimal doubling rate $\hat{W}^*$, then there is an associated portfolio $\mathbf{b}$, on the non-extended market, given by $b_i = \hat{b}_i^* + \hat{b}_{m+1}^*c_i$ that also achieves doubling-rate $\hat{W}^*$. Hence, $\hat{W}^* \leq W^*$.

Combining the above two inequalities, we must conclude that $\hat{W}^* = W^*$.

5. **Growth rate for symmetric distribution.** Consider a stock vector $\mathbf{X} \sim F(\mathbf{x})$, $\mathbf{X} \in \mathcal{R}^m, \mathbf{X} \geq 0$, where the component stocks are exchangeable. Thus $F(x_1, x_2, \ldots, x_m) = F(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(m)})$, for all permutations σ.

   (a) Find the portfolio $\mathbf{b}^*$ optimizing the growth rate and establish its optimality.
   Now assume that $\mathbf{X}$ has been normalized so that $\frac{1}{m}\sum_{i=1}^m X_i = 1$, and F is symmetric as before.

   (b) Again assuming $\mathbf{X}$ to be normalized, show that all symmetric distributions F have the same growth rate against $\mathbf{b}^*$.

   (c) Find this growth rate.

Solution: *Growth rate for symmetric distribution.*

(a) By the assumption of exchangeability, putting an equal amount in each stock is clearly the best strategy. In fact,

$$E\frac{X_i}{\mathbf{b}_*^T\mathbf{X}} = E\frac{X_i}{m^{-1}\sum X_i} = 1 \quad ,$$

so $\mathbf{b}_*$ satisfies the Kuhn-Tucker conditions.

(b)+(c) Putting an equal amount in each stock we get

$$\begin{aligned} E \log \mathbf{b}_*^T \mathbf{X} &= E \log \frac{1}{m} \sum_{i=1}^{m} X_i \\ &= E \log 1 \end{aligned}$$

Thus the growth rate is 0.

6. **Convexity.** We are interested in the set of stock market densities that yield the same optimal porfolio. Let $P_{\mathbf{b}_0}$ be the set of all probability densities on $\mathcal{R}_+^m$ for which $\mathbf{b}_0$ is optimal. Thus $P_{\mathbf{b}_0} = \{p(x) : \int \ln(\mathbf{b}^t x) p(x) dx$ is maximized by $\mathbf{b} = \mathbf{b}_0\}$. Show that $P_{\mathbf{b}_0}$ is a convex set. It may be helpful to use Theorem 16.2.2.

Solution: *Convexity.*

Let f_1 and f_2 be two stock-market densities in the set $\mathbf{P}_{b_0}$. Since both f_1 and f_2 are in this set, then, by definition, b_0 is the optimal constant-rebalance portfolio when the stock market vector is drawn according to f_1, and it is also the optimal constant-rebalance portfolio when when stock market vector is drawn according to f_2.

In order to show that the set $\mathbf{P}_{b_0}$ is convex, we need to show that any arbitrary mixture distribution, $f = \lambda f_1 + \bar{\lambda} f_2$, is also in the set; that is, we must show that b_0 is also the optimal portfolio for f.

We know that $W(b, f)$ is linear in f. So

$$\begin{aligned} W(b, f) &= W(b, \lambda f_1 + \bar{\lambda} f_2) \\ &= \lambda W(b, f_1) + \bar{\lambda} W(b, f_2) \end{aligned}$$

But by assumption each of the summands in the last expression is maximized when $b = b_0$, so the entire expression is also maximized when $b = b_0$. Hence, f is in $\mathbf{P}_{b_0}$ and the set is convex.

7. **Short selling.** Let

$$X = \begin{cases} (1, 2), p \\ (1, \frac{1}{2}), 1 - p \end{cases}$$

Let $B = \{(b_1, b_2) : b_1 + b_2 = 1\}$

Thus this set of portfolios B does not include the constraint $b_i \geq 0$. (This allows short selling.)

(a) Find the log optimal portfolio $\mathbf{b}^*(p)$.

(b) Relate the growth rate $W^*(p)$ to the entropy rate $H(p)$.

Solution: *Short selling.*

First, some philosophy. What does it mean to allow negative components in our portfolio vector? Suppose at the beginning of a trading day our current wealth is S. We want to invest our wealth S according to the portfolio $\mathbf{b}$. If b_i is positive, then we want to own $b_i S$ dollars worth of stock i. But if b_i is negative, then we want to *owe* $b_i S$ dollars worth of stock i. This is what selling-short means. It means we sell a stock we don't own in exchange for cash, but then we end up owing our broker so many shares of the stock we sold. Instead of owing money, we owe stock. The difference is that if the stock goes down in price by the end of the trading day, then we owe less money! So selling short is equivalent to betting that the stock will go down.

So, this is all well and good, but it seems to me that there may be some problems. First of all, why do we still insist that the components sum to one? It made a lot of sense when we interpreted the components, all positive, as fractions of our wealth, but it makes less sense if we are allowed to borrow money by selling short. Why not have the components sum to zero instead?

Secondly, if you owe money, then it's possible for your wealth to be negative. This is bad for our model because the log of a negative value is undefined. The reason we take logs in the first place is to turn a product into a sum that converges almost surely. But we are only justified in taking the logs in the first place if the product is positive, which it may not be if we allow short-selling.

Now, having gotten all these annoying philosophical worries out of the way, we can solve the problem quite simply by viewing it just as an unconstrained calculus problem and not worrying about what it all means.

(a) We'll represent an arbitrary portfolio as $\mathbf{b} = (b, 1-b)$. The quantity we're trying to maximize is

$$\begin{aligned}
W(b) &= \mathbf{E}[\log(\mathbf{b}^T\mathbf{X})] \\
&= \mathbf{E}[\log(bX_1 + (1-b)X_2)] \\
&= p\log(b + 2(1-b)) + (1-p)\log(b + \frac{1}{2}(1-b)) \\
&= p\log(b + 2 - 2b) + (1-p)\log(b + \frac{1}{2} - \frac{1}{2}b) \\
&= p\log(2-b) + (1-p)\log(\frac{1}{2} + \frac{1}{2}b) \\
&= p\log(2-b) + (1-p)\log\frac{1}{2} + (1-p)\log(1+b)
\end{aligned}$$

We solve for the maximum of $W(b)$ by taking the derivative and solving for zero:

$$\begin{aligned}\frac{dW}{db} &= \frac{-p}{2-b} + \frac{1-p}{1+b} = 0 \\ &\Rightarrow b = 2-3p \\ &\Rightarrow \mathbf{b} = (2-3p, 3p-1)\end{aligned}$$

(b) This questions asks us to relate the growth rate W^* to the entropy rate $H(p)$ of the market. Evidently there is some equality or inequality we should discover, as is the case with the horse race. Our intuition should tell us that low entropy rates correspond to high doubling rates and that high entropy rates correspond to low doubling rates. Quite simply, the more certain we are about what the market is going to do next (low entropy rate), the more money we should be able to make in it.

$$\begin{aligned}W^* &= W(2-3p) \\ &= p\log((2-3p)+2(3p-1)) + (1-p)\log((2-3p)+\frac{1}{2}(3p-1)) \\ &= p\log(2-3p+6p-2) + (1-p)\log(2-3p+\frac{3}{2}p-\frac{1}{2}) \\ &= p\log 3p + (1-p)\log(\frac{3}{2}-\frac{3}{2}p) \\ &= p\log p + p\log 3 + (1-p)\log\frac{3}{2} + (1-p)\log(1-p) \\ &= -H(p) + p\log 3 + (1-p)\log 3 - (1-p)\log 2 \\ &= -H(p) + \log 3 - (1-p)\log 2 \\ &\Rightarrow \\ W^* + H(p) &= \log 3 - (1-p) \\ &\le \log 3\end{aligned}$$

Hence, we can conclude that $W^* + H(p) \le \log 3$

8. **Normalizing x.** Suppose we define the log optimal portfolio $\mathbf{b}^*$ to be the portfolio maximizing the relative growth rate

$$\int \ln \frac{\mathbf{b}^t\mathbf{x}}{\frac{1}{m}\sum_{i=1}^m x_i} dF(x_1,\ldots,x_m).$$

The virtue of the normalization $\frac{1}{m}\sum X_i$, which can be viewed as the wealth associated with a uniform portfolio, is that the relative growth rate is finite, even when the growth rate $\int \ln b^t x dF(x)$ is not. This matters, for example, if X has a St. Petersburg-like distribution. Thus the log optimal portfolio $\mathbf{b}^*$ is defined for all distributions F, even those with infinite growth rates $W^*(F)$.

(a) Show that if $\mathbf{b}$ maximizes $\int \ln(\mathbf{b}^t\mathbf{x})dF(x)$, it also maximizes $\int \ln \frac{\mathbf{b}^t\mathbf{x}}{u^t x}dF(x)$, where $u = (\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m})$.

(b) Find the log optimal portfolio $\mathbf{b}^*$ for

$$\mathbf{X} = \begin{cases} (2^{2^k+1}, 2^{2^k}), & 2^{-(k+1)} \\ (2^{2^k}, 2^{2^k+1}), & 2^{-(k+1)} \end{cases}$$

where $k = 1, 2, \ldots$

(c) Find EX and W^*.

(d) Argue that $\mathbf{b}^*$ is competitively better than any portfolio $\mathbf{b}$ in the sense that $\Pr\{\mathbf{b}^t\mathbf{X} > c\mathbf{b}^{*t}\mathbf{X}\} \leq \frac{1}{c}$.

Solution: *Normalizing* $\mathbf{x}$

(a) $\mathbf{E}[\frac{\log \mathbf{b}^T\mathbf{X}}{\mathbf{u}^T\mathbf{X}}] = \mathbf{E}[\log \mathbf{b}^T\mathbf{X} - \log \mathbf{u}^T\mathbf{X}] = \mathbf{E}[\log \mathbf{b}^T\mathbf{X}] - \mathbf{E}[\log \mathbf{u}^T\mathbf{X}]$
where the second quantity in this last expression is just a number that does not change as the portfolio $\mathbf{b}$ changes. So any portfolio that maximizes the first quantity in the last expression maximizes the entire expression.

(b) Well, you can grunge out all the math here, which is messy but not difficult. But you can also notice that the symmetry of the values that $\mathbf{X}$ can take on demands that, if there is any optimum solution, it must be at $\mathbf{b} = (\frac{1}{2}, \frac{1}{2})$. For every value of the form (a, b) that $\mathbf{X}$ can take on, there is a value of the form (b, a) that $\mathbf{X}$ takes on with equal probability, so there is absolutely no bias in the market between allocating funds to stock 1 vs. stock 2.
Normalizing X by $u^t x = \frac{1}{2}(2^{2^k+1} + 2^{2^k}) = \frac{3}{2}2^{2^k}$, we obtain

$$\hat{X} = \begin{cases} (\frac{4}{3}, \frac{2}{3}), & \text{with probability } 2^{-(k+1)} \\ (\frac{2}{3}, \frac{4}{3}), & \text{with probability } 2^{-(k+1)} \end{cases} \tag{16.21}$$

Since $\hat{X}$ only takes on two values, we can sum over k and obtain

$$\hat{X} = \begin{cases} (\frac{4}{3}, \frac{2}{3}), & \text{with probability } \frac{1}{2} \\ (\frac{2}{3}, \frac{4}{3}), & \text{with probability } \frac{1}{2} \end{cases} \tag{16.22}$$

The doubling rate for a portfolio on this distbribution is

$$W(\mathbf{b}) = \frac{1}{2}\log\left(\frac{4}{3}b_1 + \frac{2}{3}(1-b_1)\right) + \frac{1}{2}\log\left(\frac{2}{3}b_1 + \frac{4}{3}(1-b_1)\right) \tag{16.23}$$

Differentiating and setting to zero and solving gives $\mathbf{b} = (\frac{1}{2}, \frac{1}{2})$.

(c) It is easy to calculate that

$$\begin{aligned} \mathbf{E}[\mathbf{X}] &= \sum_{k=1}^{\infty}\left(2^{2^k+1} + 2^{2^k}\right)2^{-(k+1)} && (16.24) \\ &= \sum_{k=1}^{\infty} 3 \cdot 2^{2^k-k-1} && (16.25) \\ &= \infty && (16.26) \end{aligned}$$

and similarly that

$$W^* = \sum_{k=1}^{\infty}\left[\log\left(\frac{1}{2}2^{2^k+1}+\frac{1}{2}2^{2^k}\right)+\log\left(\frac{1}{2}2^{2^k}+\frac{1}{2}2^{2^k+1}\right)\right]2^{-(k+1)} \tag{16.27}$$

$$= \sum_{k=1}^{\infty} 2^{-k}\log\left(2^{2^k}\frac{3}{2}\right) \tag{16.28}$$

$$= \sum_{k=1}^{\infty} 2^{-k}\left(2^k\log 2+\log\frac{3}{2}\right) \tag{16.29}$$

$$= \infty, \tag{16.30}$$

for the standard definition of W^*. If we use the new definition, then obviously $W^* = 0$, since the maximizing distribution $\mathbf{b}^*$ *is* the uniform distribution, which is the distribution by which we are normalizing.

(d) The inequality can be shown by Markov's inequality and Theorem 16.2.2 as follows

$$\Pr\left\{b^tX > cb^{*t}X\right\} = \Pr\left\{\frac{b^tX}{b^{*t}X} > c\right\} \tag{16.31}$$

$$\leq \frac{\mathbf{E}\frac{b^tX}{b^{*t}X}}{c} \tag{16.32}$$

$$\leq \frac{1}{c} \tag{16.33}$$

and therefore no portfolio exists that almost surely beats b^*. Also the probability that any other portfolio is more than twice the return of b^* is less than $\frac{1}{2}$, etc.

9. **Universal portfolio.** We examine the first $n = 2$ steps of the implementation of the universal portfolio for $m = 2$ stocks. Let the stock vectors for days 1 and 2 be $\mathbf{x}_1 = (1, \frac{1}{2})$, and $\mathbf{x}_2 = (1, 2)$. Let $\mathbf{b} = (b, 1-b)$ denote a portfolio.

(a) Graph $S_2(\mathbf{b}) = \prod_{i=1}^{2}\mathbf{b}^t\mathbf{x}_i, \quad 0 \leq b \leq 1$.

(b) Calculate $S_2^* = \max_{\mathbf{b}} S_2(\mathbf{b})$.

(c) Argue that $\log S_2(\mathbf{b})$ is concave in $\mathbf{b}$.

(d) Calculate the (universal) wealth $\hat{S}_2 = \int_0^1 S_2(\mathbf{b})d\mathbf{b}$.

(e) Calculate the universal portfolio at times $n = 1$ and $n = 2$:

$$\hat{\mathbf{b}}_1 = \int_0^1 \mathbf{b}d\mathbf{b}$$

$$\hat{\mathbf{b}}_2(\mathbf{x}_1) = \int_0^1 \mathbf{b}S_1(\mathbf{b})d\mathbf{b} / \int_0^1 S_1(\mathbf{b})d\mathbf{b}.$$

(f) Which of $S_2(\mathbf{b}), S_2^*, \hat{S}_2, \hat{\mathbf{b}}_2$ are unchanged if we permute the order of appearance of the stock vector outcomes, i.e., if the sequence is now $(1,2), (1,\frac{1}{2})$?

Solution: *Universal portfolio.*

All integrals, unless otherwise stated are over $[0,1]$.

(a) $S_2(b) = (b/2 + 1/2)(2 - b) = 1 + b/2 - b^2/2.$

(b) Maximizing over $S_2(b)$ we have $S_2^* = S_2(1/2) = 9/8.$

(c) $S_2(b)$ is concave and $\log(\cdot)$ is a monotonic increasing concave function so $\log S_2(b)$ is concave as well (check!).

(d) Using (a) we have $\hat{S_2} = \int(1 + b/2 + b^2/2)db = 13/12.$

(e) Clearly $\hat{b}_1 = 1/2$, and

$$\begin{aligned}\hat{b}_2(\mathbf{x_1}) &= \int bS_1(b)db / \int S_1(b)db \\ &= \int 0.5b(b+1)db / \int 0.5(b+1)db \\ &= 5/9.\end{aligned}$$

(f) Only $\hat{b}_2(\mathbf{x_1})$ changes.

10. **Growth optimal.** Let $X_1, X_2 \geq 0$, be price relatives of two independent stocks. Suppose $EX_1 > EX_2$. Do you always want some of X_1 in a growth rate optimal portfolio $S(\mathbf{b}) = bX_1 + \bar{b}X_2$? Prove or provide a counterexample.

Solution: *Growth optimal.*

Yes, we always want some of X_1. The following is a proof by contradiction. Assume that $\mathbf{b}^* = (0,1)^t$ so that X_1 is not active. Then the KKT conditions for this choice of $\mathbf{b}^*$ imply that $\mathbf{E}\frac{X_1}{X_2} \leq 1$ and $E\frac{X_2}{X_2} = 1$, because by assumption stock 1 is inactive and stock 2 is active. The second condition is obviously satisfied, so only the first condition needs to be checked. Since X_1 and X_2 are independent the expectation can be rewritten as $EX_1 E\frac{1}{X_2}$. Since X_2 is nonnegative, $\frac{1}{X_2}$ is convex over the region of interest, so by Jensen's inequality $E\frac{1}{X_2} \geq \frac{1}{EX_2}$. This gives that $E\frac{X_1}{X_2} \geq \frac{EX_1}{EX_2} > 1$ since $EX_1 > EX_2$. But this contradicts the KKT condition, therefore the assumption that $\mathbf{b}^* = (0,1)^t$ must be wrong, and so we must want some of X_1.

Note that we never want to short sell X_1. For any $b < 0$, we have

$$\begin{aligned}E\ln(bX_1 + (1-b)X_2) - E\log X_2 &\leq E\ln\left(b\frac{X_1}{X_2} + (1-b)\right) \\ &\leq \ln\left(bE\frac{X_1}{X_2} + (1-b)\right) \\ &< \ln 1 = 0.\end{aligned}$$

Hence, the short selling on X_1 is always worse than $\mathbf{b} = (0,1)$.

Alternatively, we can prove the same result directly as follows. Let $-\infty < b < \infty$. Consider the growth rate $W(b) = E\ln(bX_1 - (1-b)X_2)$. Differentiating w.r.t. b, we get

$$W'(b) = E\frac{X_1 - X_2}{bX_1 + (1-b)X_2}.$$

Note that $W(b)$ is concave in b. Thus $W'(b)$ is monotonically nonincreasing. Since $W'(b^*) = 0$ and $W'(0) = E\frac{X_1}{X_2} - 1 > 0$, it is immediate that $b^* > 0$.

11. **Cost of universality.** In the discussion of finite horizon universal portfolios, it was shown that the loss factor due to universality is

$$\frac{1}{V_n} = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}. \tag{16.34}$$

Evaluate V_n for $n = 1, 2, 3$.

Solution: *Cost of universality.*

Simple computation of the equation allows us to calculate

n	$\frac{1}{V_n}$	V_n
1	0	∞
2	0.125	8
3	0.197530864197531	5.0625
4	0.251953125	3.96899224806202
5	0.29696	3.36745689655172
6	0.336076817558299	2.97551020408163
7	0.371099019723317	2.69469857599079
8	0.403074979782104	2.48092799146348
9	0.43267543343584	2.31120124398809
10	0.460358496	2.17222014731754

12. **Convex families.** This problem generalizes Theorem 16.2.2. We say that $\mathcal{S}$ is a convex family of random variables if $S_1, S_2 \in \mathcal{S}$ implies $\lambda S_1 + (1-\lambda) S_2 \in \mathcal{S}$. Let $\mathcal{S}$ be a closed convex family of random variables. Show that there is a random variable $S^* \in \mathcal{S}$ such that

$$E \ln\left(\frac{S}{S^*}\right) \le 0 \tag{16.35}$$

for all $S \in \mathcal{S}$ if and only if

$$E\left(\frac{S}{S^*}\right) \le 1 \tag{16.36}$$

for all $S \in \mathcal{S}$.

Solution: *Convex families.*

Define S^* as the random variable that maximizes $E \ln S$ over all $S \in \mathcal{S}$. Since this is a maximization of a concave function over a convex set, there is a global maximum. For this value of S^*, we have

$$E \ln S \le E \ln S^* \tag{16.37}$$

for all $S \in \mathcal{S}$, and therefore

$$E \ln \frac{S}{S^*} \le 0 \tag{16.38}$$

for all $S \in \mathcal{S}$.

We need to show that for this value of S^*, that

$$E\frac{S}{S^*} \leq 1 \tag{16.39}$$

for all $S \in \mathcal{S}$. Let $T \in \mathcal{S}$ be defined as $T = \lambda S + (1-\lambda)S^* = S^* + \lambda(S - S^*)$. Then as $\lambda \to 0$, expanding the logarithm in a Taylor series and taking only the first term, we have

$$\begin{aligned} E\ln T - E\ln S^* &= E\ln S^*\left(1 + \frac{\lambda(S-S^*)}{S^*}\right) - E\ln S^* && (16.40)\\ &= E\frac{\lambda(S-S^*)}{S^*} && (16.41)\\ &= \lambda\left(E\frac{S}{S^*} - 1\right) && (16.42)\\ &\leq 0 && (16.43)\end{aligned}$$

where the last inequality follows from the fact that S^* maximizes the expected logarithm. Therefore if S^* maximizes the expected logarithm over the convex set, then for every S in the set,

$$E\frac{S}{S^*} \leq 1 \tag{16.44}$$

The other direction follows from Jensen's inequality, since if $ES/S^* \leq 1$ for all S, then

$$E\ln\frac{S}{S^*} \leq \ln E\frac{S}{S^*} \leq \ln 1 = 0. \tag{16.45}$$

Chapter 17

Inequalities in Information Theory

1. **Sum of positive definite matrices.** For any two positive definite matrices, K_1 and K_2, show that $|K_1 + K_2| \geq |K_1|$.

   Solution: *Sum of positive definite matrices*

   Let $\mathbf{X}$, $\mathbf{Y}$ be independent random vectors with $\mathbf{X} \sim \phi_{K_1}$ and $\mathbf{Y} \sim \phi_{K_2}$. Then $\mathbf{X}+\mathbf{Y} \sim \phi_{K_1+K_2}$ and hence $\frac{1}{2}\ln(2\pi e)^n|K_1+K_2| = h(\mathbf{X}+\mathbf{Y}) \geq h(\mathbf{X}) = \frac{1}{2}\ln(2\pi e)^n|K_1|$, by Lemma 17.2.1.

2. **Fan's inequality[5] for ratios of determinants.** For all $1 \leq p \leq n$, for a positive definite $K = K(1,2,\ldots,n)$, show that

$$\frac{|K|}{|K(p+1,p+2,\ldots,n)|} \leq \prod_{i=1}^{p} \frac{|K(i,p+1,p+2,\ldots,n)|}{|K(p+1,p+2,\ldots,n)|}. \tag{17.1}$$

   Solution: *Ky Fan's inequality for the ratio of determinants.* We use the same idea as in Theorem 17.9.2, except that we use the conditional form of Theorem 17.1.5.

$$\begin{aligned} \frac{1}{2}\ln(2\pi e)^p\frac{|K|}{|K(p+1,p+2,\ldots,n)|} &= h(X_1,X_2,\ldots,X_p|X_{p+1},X_{p+2},\ldots,X_n) \\ &\leq \sum h(X_i|X_{p+1},X_{p+2},\ldots,X_n) \\ &= \sum_{i=1}^{p}\frac{1}{2}\ln 2\pi e\frac{|K(i,p+1,p+2,\ldots,n)|}{|K(p+1,p+2,\ldots,n)|}. \end{aligned} \tag{17.2}$$

3. **Convexity of determinant ratios**. For positive definite matrices K, K_0, show that $\ln\frac{|K+K_0|}{|K|}$ is convex in K.

   Solution: *Convexity of determinant ratios*

   The form of the expression is related to the capacity of the Gaussian channel, and hence we can use results from the concavity of mutual information to prove this result.

Consider a colored noise Gaussian channel

$$Y_i = X_i + Z_i, \tag{17.3}$$

where $X_1, X_2, \ldots, X_n \sim \mathcal{N}(0, K_0)$ and $Z_1, Z_2, \ldots, Z_n \sim \mathcal{N}(0, K)$, and X and Z are independent

Then

$$\begin{aligned}
I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) &= h(Y_1, Y_2, \ldots, Y_n) - h(Y_1, Y_2, \ldots, Y_n | X_1, X_2, \ldots, X_n) && (17.4)\\
&= h(Y_1, Y_2, \ldots, Y_n) - h(Z_1, Z_2, \ldots, Z_n) && (17.5)\\
&= \frac{1}{2}\log(2\pi e)^n |K + K_0| - \frac{1}{2}\log(2\pi e)^n |K| && (17.6)\\
&= \frac{1}{2}\log\frac{|K_0 + K|}{|K|} && (17.7)
\end{aligned}$$

Now from Theorem 2.7.2, relative entropy is a convex function of the the distributions (The theorem should be extended to the continuous case by replacing probability mass functions by densities and summations by integrations.) Thus if $f_\lambda(x,y) = \lambda f_1(x,y) + (1-\lambda) f_2(x,y)$, $g_\lambda(x,y) = \lambda g_1(x,y) + (1-\lambda) g_2(x,y)$, we have

$$D(f_\lambda(x,y)||g_\lambda(x,y)) \le \lambda D(f_1(x,y)||g_1(x,y)) + (1-\lambda) D(f_2(x,y)||g_2(x,y)) \tag{17.8}$$

Let $Z^n \sim \mathcal{N}(0, K_1)$ with probability λ and $Z^n \sim \mathcal{N}(0, K_2)$ with probability $1-\lambda$. Let $f_1(x^n, y^n)$ be the joint distribution corresponding to $Y^n = X^n + Z^n$ when $Z^n \sim \mathcal{N}(0, K_1)$, and $g_1(x,y) = f_1(x) f_1(y)$ be the corresponding product distribution. Then

$$I(X_1^n; Y_1^n) = D(f_1(x^n, y^n)||f_1(x^n) f_1(y^n)) = D(f_1(x^n, y^n)||g_1(x^n, y^n)) = \frac{1}{2}\log\frac{|K_0 + K_1|}{|K_1|} \tag{17.9}$$

Similarly

$$I(X_2^n; Y_2^n) = D(f_1(x^n, y^n)||f_1(x^n) f_1(y^n)) = D(f_1(x^n, y^n)||g_1(x^n, y^n)) = \frac{1}{2}\log\frac{|K_0 + K_2|}{|K_2|} \tag{17.10}$$

However, the mixture distribution is not Guassian, and cannot write the same expression in terms of determinants. Instead, using the fact that the Gaussian is the worst noise given the moment constraints, we have by convexity of relative entropy

$$\begin{aligned}
\frac{1}{2}\log\frac{|K_0 + K_\lambda|}{|K_\lambda|} &\le I(X_\lambda^n; Y_\lambda^n) && (17.11)\\
&= D(f_\lambda(x^n, y^n)||f_\lambda(x^n) f_\lambda(y^n)) && (17.12)\\
&\le \lambda D(f_1(x,y)||g_1(x,y)) + (1-\lambda) D(f_2(x,y)||g_2(x,y)) && (17.13)\\
&= \lambda I(X_1^n; Y_1^n) + (1-\lambda) I(X_2^n; Y_2^n) && (17.14)\\
&= \lambda \frac{1}{2}\log\frac{|K_0 + K_1|}{|K_1|} + (1-\lambda)\frac{1}{2}\log\frac{|K_0 + K_2|}{|K_2|} && (17.15)
\end{aligned}$$

proving the convexity of the determinant ratio.

4. **Data Processing Inequality:** Let random variable X_1, X_2, X_3 and X_4 form a Markov chain $X_1 \to X_2 \to X_3 \to X_4$. Show that

$$I(X_1; X_3) + I(X_2; X_4) \leq I(X_1; X_4) + I(X_2; X_3). \tag{17.16}$$

Solution: *Data Processing Inequality:* (repeat of Problem 4.33)

$X_1 \to X_2 \to X_3 \to X_4$

$$\begin{aligned}
I(X_1; X_4) & \quad +I(X_2; X_3) - I(X_1; X_3) - I(X_2; X_4) && (17.17)\\
&= H(X_1) - H(X_1|X_4) + H(X_2) - H(X_2|X_3) - (H(X_1) - H(X_1|X_3)) \\
&\qquad -(H(X_2) - H(X_2|X_4)) && (17.18)\\
&= H(X_1|X_3) - H(X_1|X_4) + H(X_2|X_4) - H(X_2|X_3) && (17.19)\\
&= H(X_1, X_2|X_3) - H(X_2|X_1, X_3) - H(X_1, X_2|X_4) + H(X_2|X_1, X_4) && (17.20)\\
&\quad +H(X_1, X_2|X_4) - H(X_1|X_2, X_4) - H(X_1, X_2|X_3) + H(X_1|X_2, X_3) && (17.21)\\
&= -H(X_2|X_1, X_3) + H(X_2|X_1, X_4) && (17.22)\\
&- H(X_2|X_1, X_4) - H(X_2|X_1, X_3, X_4) && (17.23)\\
&= I(X_2; X_3|X_1, X_4) && (17.24)\\
&\geq 0 && (17.25)
\end{aligned}$$

where $H(X_1|X_2, X_3) = H(X_1|X_2, X_4)$ by the Markovity of the random variables.

5. **Markov chains:** Let random variables X, Y, Z and W form a Markov chain so that $X \to Y \to (Z, W)$, i.e., $p(x, y, z, w) = p(x)p(y|x)p(z, w|y)$. Show that

$$I(X; Z) + I(X; W) \leq I(X; Y) + I(Z; W) \tag{17.26}$$

Solution: *Markov chains:* (repeat of Problem 4.34)

$X \to Y \to (Z, W)$, hence by the data processing inequality, $I(X; Y) \geq I(X; (Z, W))$, and hence

$$\begin{aligned}
I(X : Y) & \quad +I(Z; W) - I(X; Z) - I(X; W) && (17.27)\\
&\geq I(X : Z, W) + I(Z; W) - I(X; Z) - I(X; W) && (17.28)\\
&= H(Z, W) + H(X) - H(X, W, Z) + H(W) + H(Z) - H(W, Z) \\
&\qquad -H(Z) - H(X) + H(X, Z)) - H(W) - H(X) + H(W, X) && (17.29)\\
&= -H(X, W, Z) + H(X, Z) + H(X, W) - H(X) && (17.30)\\
&= H(W|X) - H(W|X, Z) && (17.31)\\
&= I(W; Z|X) && (17.32)\\
&\geq 0 && (17.33)
\end{aligned}$$

Bibliography

[1] T. Berger. Multiterminal source coding. In G. Longo, editor, *The Information Theory Approach to Communications.* Springer-Verlag, New York, 1977.

[2] M. Bierbaum and H.M. Wallmeier. A note on the capacity region of the multiple access channel. *IEEE Trans. Inform. Theory*, IT-25:484, 1979.

[3] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic Press, 1981.

[4] Ky Fan. On a theorem of Weyl concerning the eigenvalues of linear transformations II. *Proc. National Acad. Sci. U.S.*, 36:31–35, 1950.

[5] Ky Fan. Some inequalities concerning positive-definite matrices. *Proc. Cambridge Phil. Soc.*, 51:414–421, 1955.

[6] R.G. Gallager. *Information Theory and Reliable Communication.* Wiley, New York, 1968.

[7] R.G. Gallager. Variations on a theme by Huffman. *IEEE Trans. Inform. Theory*, IT-24:668–674, 1978.

[8] L. Lovasz. On the Shannon capacity of a graph. *IEEE Trans. Inform. Theory*, IT-25:1–7, 1979.

[9] J.T. Pinkston. An application of rate-distortion theory to a converse to the coding theorem. *IEEE Trans. Inform. Theory*, IT-15:66–71, 1969.

[10] A Rényi. *Wahrscheinlichkeitsrechnung, mit einem Anhang über Informationstheorie.* Veb Deutscher Verlag der Wissenschaften, Berlin, 1962.

[11] A.A. Sardinas and G.W. Patterson. A necessary and sufficient condition for the unique decomposition of coded messages. In *IRE Convention Record, Part 8*, pages 104–108, 1953.

[12] C.E. Shannon. Communication theory of secrecy systems. *Bell Sys. Tech. Journal*, 28:656–715, 1949.

[13] C.E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record, Part 4*, pages 142–163, 1959.

[14] C.E. Shannon. Two-way communication channels. In *Proc. 4th Berkeley Symp. Math. Stat. Prob.*, volume 1, pages 611–644. Univ. California Press, 1961.

[15] J.A. Storer and T.G. Szymanski. Data compression via textual substitution. *J. ACM*, 29(4):928–951, 1982.