Elements of Information Theory
Second Edition
Solutions to Problems

Thomas M. Cover
Joy A. Thomas

September 22, 2006

Contents

Preface

The problems in the book, "Elements of Information Theory, Second Edition", were chosen from the problems used during the course at Stanford. Most of the solutions here were prepared by the graders and instructors of the course. We would particularly like to thank Prof. John Gill, David Evans, Jim Roche, Laura Ekroot and Young Han Kim for their help in preparing these solutions.

Most of the problems in the book are straightforward, and we have included hints in the problem statement for the difficult problems. In some cases, the solutions include extra material of interest (for example, the problem on coin weighing on Pg. 12).

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. 415-723-4505
FAX: 415-723-8473
Email: cover@isl.stanford.edu

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

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 $0 = \text{no}$, $1 = \text{yes}$, $X = \text{Source}$, and $Y = \text{Encoded Source}$. 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 $H(p_1, \ldots, p_n) = H(\mathbf{p})$ 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 $X \to Y \to Z$ 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

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

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

   - 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)

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

(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.

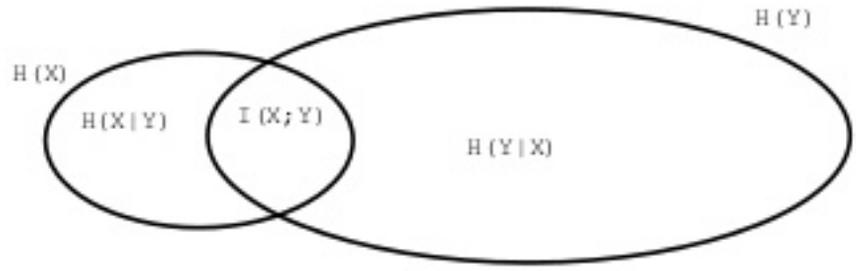
(f) See Figure 1.

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 $Z_i \sim \text{Bernoulli}\left(\frac{1}{2}\right)$, 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

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

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

$$H(X_1, X_2, \ldots, X_n) = H(X_i, \mathbf{R}) \tag{2.36}$$

$$
\begin{aligned}
&= H(\mathbf{R}) + H(X_i|\mathbf{R}) && (2.37)\\
&\leq H(\mathbf{R}) + H(X_i) && (2.38)\\
&\leq H(\mathbf{R}) + 1. && (2.39)
\end{aligned}
$$

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

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

Solution: *Markov inequality applied to entropy.*

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

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

    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 $I(X;Y) \geq 0$ 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 $0 \le p_1 \le 1$, $p_1 \le p_2 \le 1$. 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 $D(p||q) \ge 0$, we will give another proof.

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

(b) Justify the following steps:

$$\begin{aligned} -D(p||q) &= \sum_x p(x) \ln \frac{q(x)}{p(x)} && (2.45) \\ &\leq \sum_x p(x) \left(\frac{q(x)}{p(x)} - 1 \right) && (2.46) \\ &\leq 0 && (2.47) \end{aligned}$$

(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 $p(x) > 0$.

$$\begin{aligned} -D_e(p||q) &= \sum_{x \in A} p(x) ln \frac{q(x)}{p(x)} && (2.50) \\ &\leq \sum_{x \in A} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) && (2.51) \\ &= \sum_{x \in A} q(x) - \sum_{x \in A} p(x) && (2.52) \\ &\leq 0 && (2.53) \end{aligned}$$

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 $q_1 = p_1$, $q_2 = p_2, \ldots, q_{m-2} = p_{m-2}$, 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:

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

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,

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

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

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

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

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

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,

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

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)

X \ Y	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,

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

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

$$H(X) \leq H(P_e) + P_e \log(m-1), \tag{2.66}$$

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

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

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

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

Therefore

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

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:

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

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

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

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

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 $I(X; Q_1, A_1, Q_2, A_2) \leq 2I(X; Q_1, A_1)$.

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 given (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.

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

(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 $H(p) = -\sum p_i log p_i$ and $E \log X = \sum p_i log i = c < \infty$.

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[1].

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

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

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

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

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

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

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

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

By repeatedly applying the extended grouping axiom, we have

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

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

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

Thus $p_2 H_2(1,0) = H_2(1,0)$ 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

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

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

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

and

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

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

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

Therefore summing over n, we have

$$\sum_{n=2}^{N} n b_n = \sum_{n=2}^{N} (n a_n + a_{n-1} + \ldots + a_2) = N \sum_{n=2}^{N} a_i. \quad (2.116)$$

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

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

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

$$f(s) = H_s(\frac{1}{s}, \ldots, \frac{1}{s}) = H(\underbrace{\frac{1}{s}, \ldots, \frac{1}{s}}_{r}, \frac{s-r}{s}) + \frac{s-r}{s} f(s-r) \tag{2.137}$$

$$= H_2(\frac{r}{s}, \frac{s-r}{s}) + \frac{s}{r} f(s) + \frac{s-r}{s} f(s-r) \tag{2.138}$$

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[4]

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)

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

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

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)

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

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 $T_n = C_1 C_2 \ldots C_n = \prod_{i=1}^{n} C_i$. 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 $A^n = \{x^n \in \mathcal{X}^n : |-\frac{1}{n} \log p(x^n) - H| \leq \epsilon\}$. Let $B^n = \{x^n \in \mathcal{X}^n : |\frac{1}{n} \sum_{i=1}^n X_i - \mu| \leq \epsilon\}$.

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

   (d) Show $|A^n \cap B^n| \geq (\frac{1}{2}) 2^{n(H-\epsilon)}$, 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 $Pr(X^n \in B^n) \rightarrow 1$. 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 $Pr(X^n \in A^n \cap B^n) \rightarrow 1$.

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

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

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

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

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

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

   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

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

Solution: *Products.* Let

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

Then

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

with probability 1, by the strong law of large numbers. Thus $P_n \to 2^{E \log X}$ 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 $P_n \to 2^{\frac{1}{4} \log 6} = 1.565$.

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 $-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X)$ 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

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

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

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

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

$$P(A^c \cup B^c) \leq P(A^c) + P(B^c). \quad (3.22)$$

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

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

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

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

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,

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

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

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

(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 k, 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 seqeunces 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 a well defined number.

(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,

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

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

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

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

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

   Expanding the mutual informations in terms of entropies, we have

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

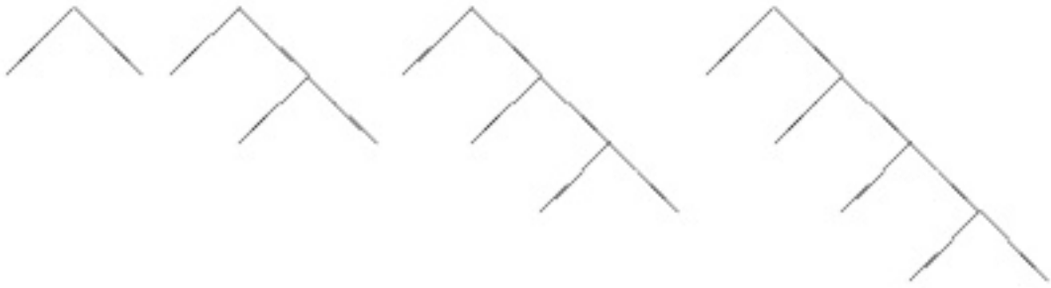
   By stationarity, $H(X_{n-1}) = H(X_n)$ and hence we have

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

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

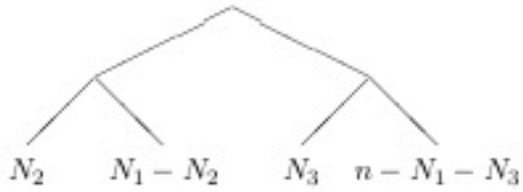
   Then expand one of the two terminal nodes at random:

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 $H_n = H(T_n)$,

$$(n-1)H_n = (n-1)\log(n-1) + 2\sum_{k=1}^{n-1} H_k \tag{4.37}$$

$$(n-2)H_{n-1} = (n-2)\log(n-2) + 2\sum_{k=1}^{n-2} H_k \tag{4.38}$$

$$\tag{4.39}$$

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

$$\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + \frac{\log(n-1)}{n} - \frac{(n-2)\log(n-2)}{n(n-1)} \tag{4.41}$$

$$= \frac{H_{n-1}}{n-1} + C_n \tag{4.42}$$

where

$$C_n = \frac{\log(n-1)}{n} - \frac{(n-2)\log(n-2)}{n(n-1)} \tag{4.43}$$

$$= \frac{\log(n-1)}{n} - \frac{\log(n-2)}{(n-1)} + \frac{2\log(n-2)}{n(n-1)} \tag{4.44}$$

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

$$\frac{H_n}{n} = \sum_{j=3}^{n} C_j + \frac{H_2}{2} \tag{4.45}$$

$$= \sum_{j=3}^{n} \frac{2\log(j-2)}{j(j-1)} + \frac{1}{n}\log(n-1). \tag{4.46}$$

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

$$\lim_{n\to\infty} \frac{H_n}{n} = \sum_{j=3}^{\infty} \frac{2}{j(j-1)} \log j - 2 \tag{4.47}$$

$$\leq \sum_{j=3}^{\infty} \frac{2}{(j-1)^2} \log(j-1) \tag{4.48}$$

$$= \sum_{j=2}^{\infty} \frac{2}{j^2} \log j \tag{4.49}$$

For sufficiently large j, $\log j \leq \sqrt{j}$ 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)
$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. \tag{4.51}$$

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

Solution: *Monotonicity of entropy per element.*

(a) By the chain rule for entropy,

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} = \frac{\sum_{i=1}^{n} H(X_i | X^{i-1})}{n} \tag{4.53}$$

$$= \frac{H(X_n | X^{n-1}) + \sum_{i=1}^{n-1} H(X_i | X^{i-1})}{n} \tag{4.54}$$

$$= \frac{H(X_n | X^{n-1}) + H(X_1, X_2, \ldots, X_{n-1})}{n}. \tag{4.55}$$

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

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

which further implies, by averaging both sides, that,

$$H(X_n|X^{n-1}) \leq \frac{\sum_{i=1}^{n-1} H(X_i|X^{i-1})}{n-1} \quad (4.56)$$

$$= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. \quad (4.57)$$

Combining (4.55) and (4.57) yields,

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq \frac{1}{n}\left[\frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1} + H(X_1, X_2, \ldots, X_{n-1})\right] \quad (4.58)$$

$$= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. \quad (4.59)$$

(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,

$$H(X_n|X^{n-1}) = \frac{\sum_{i=1}^{n} H(X_n|X^{n-1})}{n} \quad (4.60)$$

$$\leq \frac{\sum_{i=1}^{n} H(X_i|X^{i-1})}{n} \quad (4.61)$$

$$= \frac{H(X_1, X_2, \ldots, X_n)}{n}. \quad (4.62)$$

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 $\partial f / \partial p_1 = 0$.

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

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(d) of homework #4. 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

$$I(X_0; X_{n-1}) \geq I(X_0; X_n). \tag{4.63}$$

Therefore

$$H(X_0) - H(X_0|X_{n-1}) \geq H(X_0) - H(X_0|X_n) \tag{4.64}$$

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

$$\begin{aligned} P(S_k \text{ odd}) &= P(S_{k-1} \text{ odd})P(X_k = 0) + P(S_{k-1} \text{ even})P(X_k = 1) && (4.65)\\ &= \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} && (4.66)\\ &= \frac{1}{2}. && (4.67) \end{aligned}$$

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

$$P(X_n = 1) = P(X_n = 0) = \frac{1}{2}. \tag{4.68}$$

    (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,

$$\begin{aligned} P(X_1 = 1, X_n = 1) &= P(X_1 = 1, \sum_{i=2}^{n-1} X_i \text{ even}) && (4.69)\\ &= P(X_1 = 1)P(\sum_{i=2}^{n-1} X_i \text{ even}) && (4.70)\\ &= \frac{1}{2}\frac{1}{2} && (4.71)\\ &= P(X_1 = 1)P(X_n = 1) && (4.72) \end{aligned}$$

    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 $\{0,1\}$,

$$H(X_i, X_j) = H(X_i) + H(X_j) = 1 + 1 = 2 \text{ bits}. \tag{4.73}$$

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

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_1, X_2, \ldots, X_{n-1}) + H(X_n | X_{n-1}, \ldots, X_1) && (4.74)\\ &= \sum_{i=1}^{n-1} H(X_i) + 0 && (4.75)\\ &= n - 1, && (4.76) \end{aligned}$$

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

    $$H(X_n|X_0) = H(X_n, X_0) - H(X_0) \tag{4.77}$$
    $$H(X_{-n}|X_0) = H(X_{-n}, X_0) - H(X_0) \tag{4.78}$$

    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

$$\lim_{n\to\infty} \frac{1}{2n} I(X_1, X_2, \ldots, X_n; X_{n+1}, X_{n+2}, \ldots, X_{2n}) = 0. \tag{4.79}$$

Thus the dependence between adjacent n-blocks of a stationary process does not grow linearly with n.

Solution:

$$\begin{aligned} &I(X_1, X_2, \ldots, X_n; X_{n+1}, X_{n+2}, \ldots, X_{2n}) \\ &= H(X_1, X_2, \ldots, X_n) + H(X_{n+1}, X_{n+2}, \ldots, X_{2n}) - H(X_1, X_2, \ldots, X_n, X_{n+1}, X_{n+2}, \ldots, X_{2n}) \\ &= 2H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n, X_{n+1}, X_{n+2}, \ldots, X_{2n}) \end{aligned} \tag{4.80}$$

since $H(X_1, X_2, \ldots, X_n) = H(X_{n+1}, X_{n+2}, \ldots, X_{2n})$ by stationarity.

Thus

$$\lim_{n\to\infty}\frac{1}{2n}I(X_1,X_2,\ldots,X_n;X_{n+1},X_{n+2},\ldots,X_{2n})$$
$$= \lim_{n\to\infty}\frac{1}{2n}2H(X_1,X_2,\ldots,X_n) - \lim_{n\to\infty}\frac{1}{2n}H(X_1,X_2,\ldots,X_n,X_{n+1},X_{n+2},\ldots,X_{2n}) \quad (4.81)$$
$$= \lim_{n\to\infty}\frac{1}{n}H(X_1,X_2,\ldots,X_n) - \lim_{n\to\infty}\frac{1}{2n}H(X_1,X_2,\ldots,X_n,X_{n+1},X_{n+2},\ldots,X_{2n}) \quad (4.82)$$

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

$$\lim_{n\to\infty}\frac{1}{2n}I(X_1,X_2,\ldots,X_n;X_{n+1},X_{n+2},\ldots,X_{2n}) = 0. \quad (4.83)$$

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

$$Y_i = \phi(X_i), \quad ,i = 1,2,\ldots \quad (4.84)$$

for some function ϕ. Prove that

$$H(\mathcal{Y}) \le H(\mathcal{X}) \quad (4.85)$$

(b) What is the relationship between the entropy rates $H(\mathcal{Z})$ and $H(\mathcal{X})$ if

$$Z_i = \psi(X_i,X_{i+1}), \quad i = 1,2,\ldots \quad (4.86)$$

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)

$$H(Y_1,Y_2,\ldots,Y_n) \le H(X_1,X_2,\ldots,X_n) \quad (4.87)$$

Dividing by n, and taking the limit as $n\to\infty$, we have

$$\lim_{n\to\infty}\frac{H(Y_1,Y_2,\ldots,Y_n)}{n} \le \lim_{n\to\infty}\frac{H(X_1,X_2,\ldots,X_n)}{n} \quad (4.88)$$

or

$$\mathcal{H}(\mathcal{Y}) \le \mathcal{H}(\mathcal{X}) \quad (4.89)$$

15. *Entropy rate.* Let $\{X_i\}$ be a discrete stationary stochastic process with entropy rate $H(\mathcal{X})$. Show

$$\frac{1}{n}H(X_n,\ldots,X_1 \mid X_0,X_{-1},\ldots,X_{-k}) \to H(\mathcal{X}), \quad (4.90)$$

for $k = 1,2,\ldots$.

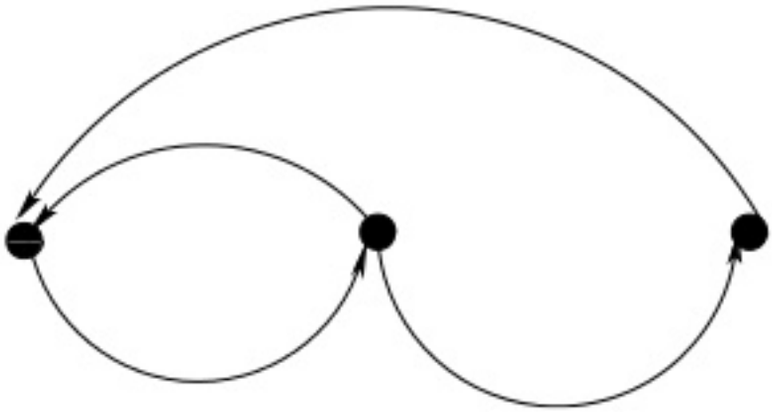


Figure 4.1: Entropy rate of constrained sequence

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

$$\begin{aligned} \frac{1}{n}H(X_1, X_2, \ldots, X_n | X_0, X_{-1}, \ldots, X_{-k}) &= \frac{1}{n}\sum_{i=1}^{n} H(X_i | X_{i-1}, X_{i-2}, \ldots, X_{-k}) && (4.91) \\ &\to \lim H(X_n | X_{n-1}, X_{n-2}, \ldots, X_{-k}) && (4.92) \\ &= \mathcal{H}, && (4.93) \end{aligned}$$

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:

$$\begin{bmatrix} X_1(n) \\ X_2(n) \\ X_3(n) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1(n-1) \\ X_2(n-1) \\ X_3(n-1) \end{bmatrix}, \tag{4.94}$$

with initial conditions $\mathbf{X}(1) = [1\ 1\ 0]^t$.

(c) Let

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{4.95}$$

Then we have by induction

$$\mathbf{X}(n) = A\mathbf{X}(n-1) = A^2\mathbf{X}(n-2) = \cdots = A^{n-1}\mathbf{X}(1). \tag{4.96}$$

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

$$\mathbf{X}(n) = \lambda_1^{n-1}\mathbf{Y}_1 + \lambda_2^{n-1}\mathbf{Y}_2 + \lambda_3^{n-1}\mathbf{Y}_3, \tag{4.97}$$

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

$$\frac{1}{n}\log X_i(n) \to \log\lambda, \tag{4.98}$$

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

$$P = \begin{bmatrix} 0 & 1 & 0 \\ \alpha & 0 & 1-\alpha \\ 1 & 0 & 0 \end{bmatrix}. \tag{4.99}$$

Show that the stationary distribution of this Markov chain is

$$\mu = \left[\frac{1}{3-\alpha}, \frac{1}{3-\alpha}, \frac{1-\alpha}{3-\alpha}\right]. \tag{4.100}$$

(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,

$$X_1(n) = X_2(n-1) + X_3(n-1). \tag{4.101}$$

By similar arguments, we get the other two equations, and we have

$$\begin{bmatrix} X_1(n) \\ X_2(n) \\ X_3(n) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1(n-1) \\ X_2(n-1) \\ X_3(n-1) \end{bmatrix}. \tag{4.102}$$

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

$$\mathbf{X}(n) = A^{n-1}\mathbf{X}(1) = U^{-1}\Lambda^{n-1}U\mathbf{X}(1) \tag{4.103}$$

$$= U^{-1}\begin{bmatrix} \lambda_1^{n-1} & 0 & 0 \\ 0 & \lambda_2^{n-1} & 0 \\ 0 & 0 & \lambda_3^{n-1} \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \tag{4.104}$$

$$= \lambda_1^{n-1}U^{-1}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \lambda_2^{n-1}U^{-1}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$
$$+\lambda_3^{n-1}U^{-1}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \tag{4.105}$$

$$= \lambda_1^{n-1}\mathbf{Y}_1 + \lambda_2^{n-1}\mathbf{Y}_2 + \lambda_3^{n-1}\mathbf{Y}_3, \tag{4.106}$$

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

$$X_1(n) = \lambda_1^{n-1}\mathbf{Y}_{11} + \lambda_2^{n-1}\mathbf{Y}_{21} + \lambda_3^{n-1}\mathbf{Y}_{31} \tag{4.107}$$

$$X_2(n) = \lambda_1^{n-1}\mathbf{Y}_{12} + \lambda_2^{n-1}\mathbf{Y}_{22} + \lambda_3^{n-1}\mathbf{Y}_{32} \tag{4.108}$$

$$X_3(n) = \lambda_1^{n-1}\mathbf{Y}_{13} + \lambda_2^{n-1}\mathbf{Y}_{23} + \lambda_3^{n-1}\mathbf{Y}_{33} \tag{4.109}$$

For large n, this sum is dominated by the largest term. Thus if $\mathbf{Y}_{1i} > 0$, we have

$$\frac{1}{n}\log X_i(n) \to \log \lambda_1. \tag{4.110}$$

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

$$\frac{1}{n}\log X_i(n) \to \log \lambda_1. \tag{4.111}$$

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

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = U^{-1}\Lambda U, \tag{4.112}$$

where

$$\Lambda = \begin{bmatrix} 1.3247 & 0 & 0 \\ 0 & -0.6624 + 0.5623i & 0 \\ 0 & 0 & -0.6624 - 0.5623i \end{bmatrix}, \tag{4.113}$$

and

$$U = \begin{bmatrix} -0.5664 & -0.7503 & -0.4276 \\ 0.6508 - 0.0867i & -0.3823 + 0.4234i & -0.6536 - 0.4087i \\ 0.6508 + 0.0867i & -0.3823i0.4234i & -0.6536 + 0.4087i \end{bmatrix}, \tag{4.114}$$

and therefore

$$\mathbf{Y}_1 = \begin{bmatrix} 0.9566 \\ 0.7221 \\ 0.5451 \end{bmatrix}, \tag{4.115}$$

which has all positive components. Therefore,

$$\frac{1}{n}\log X_i(n) \to \log \lambda_i = \log 1.3247 = 0.4057 \text{ bits}. \tag{4.116}$$

(d) To verify the that

$$\mu = \left[\frac{1}{3-\alpha}, \frac{1}{3-\alpha}, \frac{1-\alpha}{3-\alpha}\right]^T. \tag{4.117}$$

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

$$\mathcal{H} = \sum_i \mu_i \sum_j P_{ij} \ln P_{ij} = \frac{1}{3-\alpha} \left(-\alpha \ln \alpha - (1-\alpha) \ln(1-\alpha)\right), \tag{4.118}$$

and differentiating with respect to α to find the maximum, we find that

$$\frac{d\mathcal{H}}{d\alpha} = \frac{1}{(3-\alpha)^2} \left(-\alpha \ln \alpha - (1-\alpha) \ln(1-\alpha)\right) + \frac{1}{3-\alpha} \left(-1 - \ln \alpha + 1 + \ln(1-\alpha)\right) = 0, \tag{4.119}$$

or

$$(3-\alpha)\left(\ln a - \ln(1-\alpha)\right) = \left(-\alpha \ln \alpha - (1-\alpha) \ln(1-\alpha)\right) \tag{4.120}$$

which reduces to

$$3 \ln \alpha = 2 \ln(1-\alpha), \tag{4.121}$$

i.e.,

$$\alpha^3 = \alpha^2 - 2\alpha + 1, \tag{4.122}$$

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 an 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 $E \log N \leq H(X)$.

    (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,

$$EN = E[E(N|X_0)] = \sum p(X_0 = i)\left(\frac{1}{p(i)}\right) = m. \tag{4.123}$$

    (b) There is a typographical error in the problem. The problem should read $E \log N \leq H(X)$.

    By the same argument, since given $X_0 = i$, N has a geometric distribution with mean $p(i)$ and

$$E(N|X_0 = i) = \frac{1}{p(i)}. \tag{4.124}$$

    Then using Jensen's inequality, we have

$$\begin{aligned} E \log N &= \sum_i p(X_0 = i) E(\log N|X_0 = i) && (4.125)\\ &\leq \sum_i p(X_0 = i) \log E(N|X_0 = i) && (4.126)\\ &= \sum_i p(i) \log \frac{1}{p(i)} && (4.127)\\ &= H(X). && (4.128) \end{aligned}$$

    (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.

    To simplify matters, we will consider $X_1, X_2, \ldots, X_n$ arranged in a circle, so that X_1 follows X_n. Then we can get rid of the edge effects (namely that the waiting time is not defined for X_n, etc) and we can define the waiting time N_k at time

k as $\min\{n > k : X_n = X_k\}$. With this definition, we can write the empirical average of N_k for a particular sample sequence

$$\overline{N} = \frac{1}{n}\sum_{i=1}^{n} N_i \tag{4.129}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=i+1}^{\min\{n>i:x_n=x_i\}} 1\right). \tag{4.130}$$

Now we can rewrite the outer sum by grouping together all the terms which correspond to $x_i = l$. Thus we obtain

$$\overline{N} = \frac{1}{n}\sum_{l=1}^{m}\sum_{i:x_i=l}\left(\sum_{j=i+1}^{\min\{n>i:x_n=l\}} 1\right) \tag{4.131}$$

But the inner two sums correspond to summing 1 over all the n terms, and thus

$$\overline{N} = \frac{1}{n}\sum_{l=1}^{m} n = m \tag{4.132}$$

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

$$\begin{aligned} I(X;Y_1,Y_2) &= H(Y_1,Y_2) - H(Y_1,Y_2|X) && (4.133)\\ &= H(Y_1,Y_2) - H(Y_1|X) - H(Y_2|X) && (4.134)\\ &= H(Y_1,Y_2) - 2H(p) && (4.135)\\ &= H\left(\frac{1}{2}(p^2+(1-p)^2), p(1-p), p(1-p), \frac{1}{2}(p^2+(1-p)^2)\right) - 2H(p) \\ &= H(p(1-p)) + 1 - 2H(p) && (4.136) \end{aligned}$$

where the last step follows from using the grouping rule for entropy.

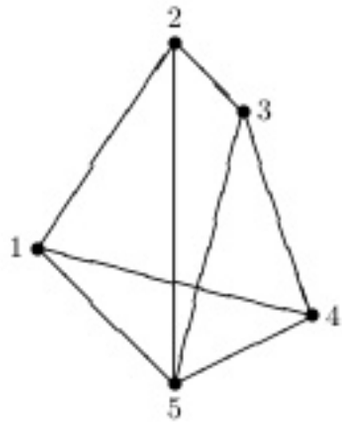
(c)

$$\begin{aligned} \mathcal{H}(\mathcal{Y}) &= \lim \frac{H(Y_1, Y_2, \ldots, Y_n)}{n} && (4.137)\\ &= \lim \frac{H(X, Y_1, Y_2, \ldots, Y_n) - H(X|Y_1, Y_2, \ldots, Y_n)}{n} && (4.138)\\ &= \lim \frac{H(X) + H(Y_1, Y_2, \ldots, Y_n|X) - H(X|Y_1, Y_2, \ldots, Y_n)}{n} && (4.139) \end{aligned}$$

Since $0 \le H(X|Y_1, Y_2, \ldots, Y_n) \le H(X) \le 1$, 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

$$\mathcal{H}(\mathcal{Y}) = \lim \frac{nH(p)}{n} = H(p) \qquad (4.140)$$

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

$$\begin{aligned} I(X_{n+1}; X_n) &= H(X_{n+1}) - H(X_{n+1}|X_n) && (4.141)\\ &= H(3/16, 3/16, 3/16, 3/16, 1/4) - (log16 - H(3/16, 3/16, 3/16, 3/16, 1/4)) && (4.142)\\ &= 2H(3/16, 3/16, 3/16, 3/16, 1/4) - log16 && (4.143)\\ &= 2(\frac{3}{4}\log\frac{16}{3} + \frac{1}{4}\log 4) - log16 && (4.144)\\ &= 3 - \frac{3}{2}\log 3 && (4.145) \end{aligned}$$

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

$$\begin{aligned} \mathcal{H} &= \sum_{i=1}^{9} \mu_i H(X_2|X_1 = i) && (4.146) \\ &= 0.3 \log 3 + 0.5 \log 5 + 0.2 \log 8 && (4.147) \\ &= 2.24 \text{ bits.} && (4.148) \end{aligned}$$

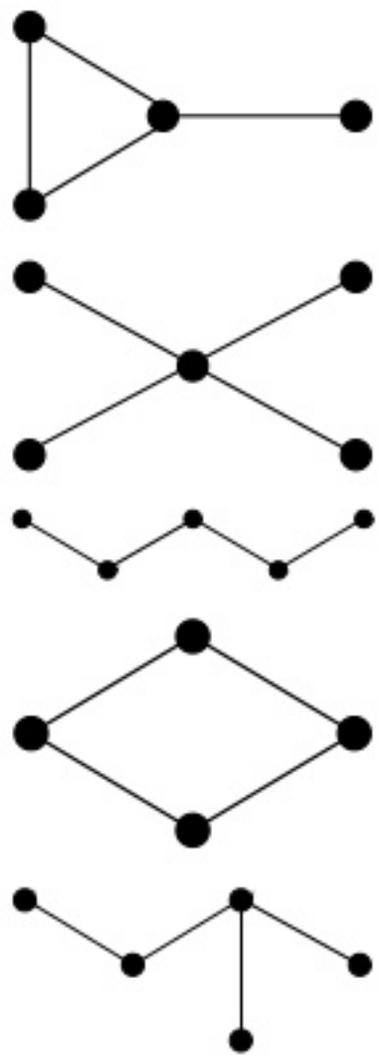
21. *Maximal entropy graphs.* Consider a random walk on a connected graph with 4 edges.

    (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 $H(\mathcal{X}) \leq H(X_1)$.

(b) What are the conditions for equality?

Solution: Entropy Rate

(a) From Theorem 4.2.1

$$H(\mathcal{X}) = H(X_1|X_0, X_{-1}, \ldots) \leq H(X_1) \tag{4.149}$$

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 $\leq, =$, or $\geq$ between each of the pairs listed below:

(a) $H(\mathcal{X}) \overset{?}{\leq} H(\mathcal{Y})$.

(b) $H(\mathcal{X}) \overset{?}{\geq} H(\mathcal{Z})$.

(c) $H(\mathcal{X}) \overset{?}{\geq} H(\mathcal{V})$.

(d) $H(\mathcal{Z}) \overset{?}{\leq} H(\mathcal{X})$.

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 $H(V_1|V_0, \ldots) = H(X_2|X_0, X_{-2}, \ldots) \leq H(X_2|X_1, X_0, X_{-1}, \ldots)$.

(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,

$$H(X|Y_1, Y_2, \ldots, Y_n) \geq H(X|Y_1, Y_2, \ldots, Y_n, Y_{n+1}) \tag{4.150}$$

and hence

$$\begin{aligned} I(X; Y_1, Y_2, \ldots, Y_n) &= H(X) - H(X|Y_1, Y_2, \ldots, Y_n) && (4.151)\\ &\leq H(X) - H(X|Y_1, Y_2, \ldots, Y_n, Y_{n+1}) && (4.152)\\ &= I(X; Y_1, Y_2, \ldots, Y_n, Y_{n+1}) && (4.153) \end{aligned}$$

(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*

$$H(\mathcal{X}) = H(X_{k+1}|X_k, X_{k-1}, \ldots) = H(X_{k+1}|X_k, X_{k-1}) = H(Z_{k+1}) = H(p) \quad (4.154)$$

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_i s 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 $Y_i \sim \text{Bernoulli}((p_1 + p_2)/2)$.

(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.

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.

(a) 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.

(b) It's clear that the variance of S_n grows with n, which again implies that the marginals are not time-invariant.

(c) Process $\{S_n\}$ is an independent increment process. An independent increment process is not stationary (not even wide sense stationary), since $\text{var}(S_n) = \text{var}(X_n) + \text{var}(S_{n-1}) > \text{var}(S_{n-1})$.

(d) We can use chain rule and Markov properties to obtain the following results.

$$\begin{aligned} H(S_1, S_2, \ldots, S_n) &= H(S_1) + \sum_{i=2}^{n} H(S_i|S^{i-1}) \\ &= H(S_1) + \sum_{i=2}^{n} H(S_i|S_{i-1}) \\ &= H(X_1) + \sum_{i=2}^{n} H(X_i) \\ &= \sum_{i=1}^{n} H(X_i) \\ &= 2n \end{aligned}$$

(e) It follows trivially from (e) 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.

(f) 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 success 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

$$\begin{aligned} P(Z_{k+1} | X_k, X_{k-1}) &= P(X_{k+1} - X_k | X_k, X_{k-1}) \\ &= P(X_{k+1} - X_k | X_k) \\ &= P(X_{k+1} - X_k) \\ &= P(Z_{k+1}), \end{aligned}$$

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 $\{X_i\} \sim \text{Bernoulli}(p)$. Consider the associated Markov chain $\{Y_i\}_{i=1}^n$ where $Y_i = (\text{the number of 1's in the current run of 1's})$. 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 counter example. 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

$$I(X_1;X_3) + I(X_2;X_4) \leq I(X_1;X_4) + I(X_2;X_3) \tag{4.155}$$

Solution: *Chain inequality* $X_1 \to X_2 \to X_3 to X_4$

$$\begin{aligned}
I(X_1;X_4) & \quad +I(X_2;X_3) - I(X_1;X_3) - I(X_2;X_4) && (4.156)\\
&= H(X_1) - H(X_1|X_4) + H(X_2) - H(X_2|X_3) - (H(X_1) - H(X_1|X_3)) \\
&\quad -(H(X_2) - H(X_2|X_4)) && (4.157)\\
&= H(X_1|X_3) - H(X_1|X_4) + H(X_2|X_4) - H(X_2|X_3) && (4.158)\\
&= H(X_1,X_2|X_3) - H(X_2|X_1,X_3) - H(X_1,X_2|X_4) + H(X_2|X_1,X_4) && (4.159)\\
&\quad +H(X_1,X_2|X_4) - H(X_1|X_2,X_4) - H(X_1,X_2|X_3) + H(X_1|X_2,X_3) && (4.160)\\
&= -H(X_2|X_1,X_3) + H(X_2|X_1,X_4) && (4.161)\\
&- H(X_2|X_1,X_4) - H(X_2|X_1,X_3,X_4) && (4.162)\\
&= I(X_2;X_3|X_1,X_4) && (4.163)\\
&\geq 0 && (4.164)
\end{aligned}$$

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

$$I(X;Z) + I(X;W) \leq I(X;Y) + I(Z;W) \tag{4.165}$$

Solution: *Broadcast Channel*

$X \to Y \to (Z,W)$, hence by the data processing inequality, $I(X;Y) \geq I(X;(Z,W))$, and hence

$$\begin{aligned}
I(X:Y) & \quad +I(Z;W) - I(X;Z) - I(X;W) && (4.166)\\
&\geq I(X:Z,W) + I(Z;W) - I(X;Z) - I(X;W) && (4.167)\\
&= H(Z,W) + H(X) - H(X,W,Z) + H(W) + H(Z) - H(W,Z) \\
&\quad -H(Z) - H(X) + H(X,Z)) - H(W) - H(X) + H(W,X) && (4.168)\\
&= -H(X,W,Z) + H(X,Z) + H(X,W) - H(X) && (4.169)\\
&= H(W|X) - H(W|X,Z) && (4.170)\\
&= I(W;Z|X) && (4.171)\\
&\geq 0 && (4.172)
\end{aligned}$$

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

$$\begin{aligned}
H(X_n|X_0) - H(X_{n-1}|X_0) - (H(X_{n-1}|X_0) - H(X_{n-2}|X_0)) &= -I(X_1;X_{n-1}|X_0,X_n) && (4.173)\\
&\leq 0 && (4.174)
\end{aligned}$$

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

$$\begin{aligned}
H(X_n|X_0) & -H(X_{n-1}|X_0) - (H(X_{n-1}|X_0) - H(X_{n-2}|X_0) && (4.175)\\
&= H(X_n|X_0) - H(X_{n-1}|X_0, X_{-1}) - (H(X_{n-1}|X_0, X_{-1}) - H(X_{n-2}|X_0, X_{-1})) && (4.176)\\
&= H(X_n|X_0) - H(X_n|X_1, X_0) - (H(X_{n-1}|X_0) - H(X_{n-1}|X_1, X_0) && (4.177)\\
&= I(X_1; X_n|X_0) - I(X_1; X_{n-1}|X_0) && (4.178)\\
&= H(X_1|X_0) - H(X_1|X_n, X_0) - H(X_1|X_0) + H(X_1|X_{n-1}, X_0) && (4.179)\\
&= H(X_1|X_{n-1}, X_0) - H(X_1|X_n, X_0) && (4.180)\\
&= H(X_1, X_{n-1}, X_n, X_0) - H(X_1|X_n, X_0) && (4.181)\\
&= -I(X_1; X_{n-1}|X_n, X_0) && (4.182)\\
&\leq 0 && (4.183)
\end{aligned}$$

where (4.176) and (4.181) follows from Markovity and (4.177) follows from stationarity of the Markov chain.

If we define

$$\Delta_n = H(X_n|X_0) - H(X_{n-1}|X_0) \tag{4.184}$$

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 $a \to 00$, $b \to 01$, $c \to 10$ and $d \to 11$. What is the average length of the binary code for the above random variable constructed by this process?

(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 $a \rightarrow 00$, $b \rightarrow 01$, $c \rightarrow 10$ and $d \rightarrow 11$. What is the average length of the binary code for the above random variable constructed by this process?

Solution: The code constructed by the above process is $A \rightarrow 00$, $B \rightarrow 01$, $C \rightarrow 11$, $D \rightarrow 1000$, $E \rightarrow 1001$, and $F \rightarrow 1010$, 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 p_i's forms 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 $\sum 2^{-l_i} \leq 1$. 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

$$C^* \leq C_{Huffman} \leq C^* + \sum_{i=1}^{m} p_i c_i ?$$

Solution: *Huffman codes with costs.*

(a) We wish to minimize $C = \sum p_i c_i n_i$ subject to $\sum 2^{-n_i} \leq 1$. 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

$$-\log q_i \leq n_i < -\log q_i + 1 \tag{5.37}$$

Multiplying by $p_i c_i$ and summing over i, we get the relationship

$$C^* \leq C_{Huffman} < C^* + Q. \tag{5.38}$$

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

$$\sum_{i=1}^{m-1} D^{-\ell_i} \leq 1 - D^{-\ell_m} < 1 ,$$

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 $011111\ldots1110$. 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 $2^k \leq m \leq 2^{k+1}$, 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, $L' \leq L$. 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 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[3].

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 $(3,3,2,2) \to (4,3,3) \to (6,4) \to (10)$, 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 $W = \sum W_i$, $\tilde{W}_i = W_i/W$, and show that the minimum merge volume V satisfies

$$WH(\tilde{\mathbf{W}}) \leq V \leq WH(\tilde{\mathbf{W}}) + W \tag{5.39}$$

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

$$WH(\tilde{\mathbf{W}}) \leq V \leq WH(\tilde{\mathbf{W}}) + W \tag{5.41}$$

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

|	A_1	|	A_2	|	A_3	$\ldots$	A_m	|		
|	B_1	|	B_2	|	B_3	$\ldots$	B_n	|		

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[5]) 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 $L^*_{1:1} \geq \tilde{L} \stackrel{\triangle}{=} \sum_{i=1}^m p_i \log\left(\frac{i}{2}+1\right)$. Consider the difference

$$F(\mathbf{p}) = H(X) - \tilde{L} = -\sum_{i=1}^m p_i \log p_i - \sum_{i=1}^m p_i \log\left(\frac{i}{2}+1\right). \tag{5.49}$$

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.,

$$H_k \stackrel{\triangle}{=} \sum_{i=1}^k \frac{1}{i} \tag{5.50}$$

(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 $2^k - 1 \leq i \leq 2^k - 2 + 2^k = 2^{k+1} - 2$ use codewords of length k. This corresponds to $2^k < i+2 \leq 2^{k+1}$ or $k < \log(i+2) \leq k+1$ or $k-1 < \log\frac{i+2}{2} \leq k$. 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 $\lceil \log(i/2+1) \rceil \geq \log(i/2+1)$, it follows that $L_{1:1}^* \geq \tilde{L} \stackrel{\triangle}{=} \sum_{i=1}^m p_i \log\left(\frac{i}{2}+1\right)$. 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

$$J(\mathbf{p}) = -\sum_{i=1}^m p_i \log p_i - \sum_{i=1}^m p_i \log\left(\frac{i}{2}+1\right) + \lambda\left(\sum_{i=1}^m p_i - 1\right) \tag{5.57}$$

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

$$H(X) - \log\log|\mathcal{X}| - 2 \leq L_{1:1}^{*} \leq H(X) + 1 \tag{5.77}$$

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

$$EN = 1.\frac{1}{2} + 2\frac{1}{4} + 3\frac{1}{8} + \ldots + \tag{5.91}$$

$$= 2 \tag{5.92}$$

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

$$\begin{array}{lll} 6 & 9 & 25 \\ 6 & 6 & \\ 4 & 6 & \\ 4 & 4 & \\ 2 & & \\ 2 & & \\ 1 & & \end{array}$$

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 $X_1X_2\cdots X_n$ into

$$Z_1Z_2\cdots Z_m = C(X_1)C(X_2)\cdots C(X_n).$$

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

$$nH(\mathcal{X}) \;=\; H(X_1X_2\cdots X_n) \;=\; H(Z_1Z_2\cdots Z_{\sum_1^n L(C(X_i))}) \tag{5.93}$$

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

$$\begin{aligned} H(Z_1Z_2\cdots Z_{\sum_1^n L(C(X_i))}) &= E[\sum_{i=1}^{n} L(C(X_i))]H(\mathcal{Z}) \\ &= nE[L(C(X_1))]H(\mathcal{Z}) \end{aligned} \tag{5.94}$$

(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

$$1_j = \begin{cases} 1, & \text{object } j \text{ yields the same answers for } k \text{ questions as object 1} \\ 0, & \text{otherwise} \end{cases}, \quad \text{for } j = 2, \ldots, m.$$

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

$$K = \sum (\frac{b_0}{o_i} + b_i) = b_0 \sum \frac{1}{o_i} + \sum b_i = b_0 (\sum \frac{1}{o_i} - 1) + 1, \quad (6.19)$$

and

$$r_i = \frac{\frac{b_0}{o_i} + b_i}{K} \quad (6.20)$$

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.,

$$1 - \sum_{i=1}^{m} b'_i = 1 - \sum_{i=1}^{m} \left(b_i - \frac{b_m o_m}{o_i} \right) \tag{6.22}$$

$$= \sum_{i=1}^{m} \frac{b_m o_m}{o_i} \tag{6.23}$$

as cash.

The return on the new portfolio if horse i wins is

$$b'_i o_i = \left(b_i - \frac{b_m o_m}{o_i} \right) o_i + \sum_{i=1}^{m} \frac{b_m o_m}{o_i} \tag{6.24}$$

$$= b_i o_i + b_m o_m \left(\sum_{i=1}^{m} \frac{1}{o_i} - 1 \right) \tag{6.25}$$

$$> b_i o_i, \tag{6.26}$$

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[2], 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 $x_i \geq 0$,

$$\begin{array}{ll} \frac{\partial F}{\partial x_i} \leq 0 & \text{if } x_i = 0 \\ = 0 & \text{if } x_i > 0 \end{array} \tag{6.37}$$

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[2] 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

$$D(\mathbf{p}\|\mathbf{b}) < \frac{1}{4}.$$

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 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 $\sum p_i \log \frac{p_i}{q_i} = D(p||q)$.

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

$$\begin{aligned}
E(Z|J=1) &= P(X=b|J=1)E(Z|X=b,J=1) \\
&\quad + P(X=2b|J=1)E(Z|X=2b,J=1) && (6.72)\\
&= \frac{1}{2}E(Z|X=b,J=1) + \frac{1}{2}E(Z|X=2b,J=1) && (6.73)\\
&= \frac{1}{2}\,(p(J'=1|X=b,J=1)E(Z|J'=1,X=b,J=1) \\
&\quad + p(J'=2|X=b,J=1)E(Z|J'=2,X=b,J=1) \\
&\quad + p(J'=1|X=2b,J=1)E(Z|J'=1,X=2b,J=1) \\
&\quad + p(J'=2|X=2b,J=1)E(Z|J'=2,X=2b,J=1)) && (6.74)\\
&= \frac{1}{2}\,([1-p(b)]2b + p(b)b + p(2b)2b + [1-p(2b)]b) && (6.75)\\
&= \frac{3b}{2} + \frac{1}{2}b(p(2b)-p(b)) && (6.76)\\
&> \frac{3b}{2} && (6.77)
\end{aligned}$$

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

$$\begin{aligned}
W^* &= \sum_i p_i \log p_i o_i && (6.78)\\
&= \sum_i p_i \log \frac{p_i}{\frac{1}{o_i}} && (6.79)\\
&= \sum_i p_i \log \frac{p_i}{q_i} - \sum_i p_i \log \left(\sum_j \frac{1}{o_j}\right) && (6.80)\\
&= \sum_i p_i \log \frac{p_i}{q_i} - \log \left(\sum_j \frac{1}{o_j}\right) && (6.81)
\end{aligned}$$

where

$$q_i = \frac{\frac{1}{o_i}}{\sum_j \frac{1}{o_j}} \qquad (6.82)$$

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 $-\log\left(\sum_j \frac{1}{a_j}\right)$.

(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

$$\begin{aligned}
S^* &= 2^{W^*} = 8.66 \\
S_D &= 2^{W_D} = 7.5
\end{aligned}$$

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 $\bar{b} = 1 - b$ of his money in his pocket and invest the rest in the St. Petersburg game. His wealth at time n is then

$$S_n = \prod_{i=1}^{n} \left(\bar{b} + \frac{bX_i}{c} \right). \tag{6.96}$$

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.

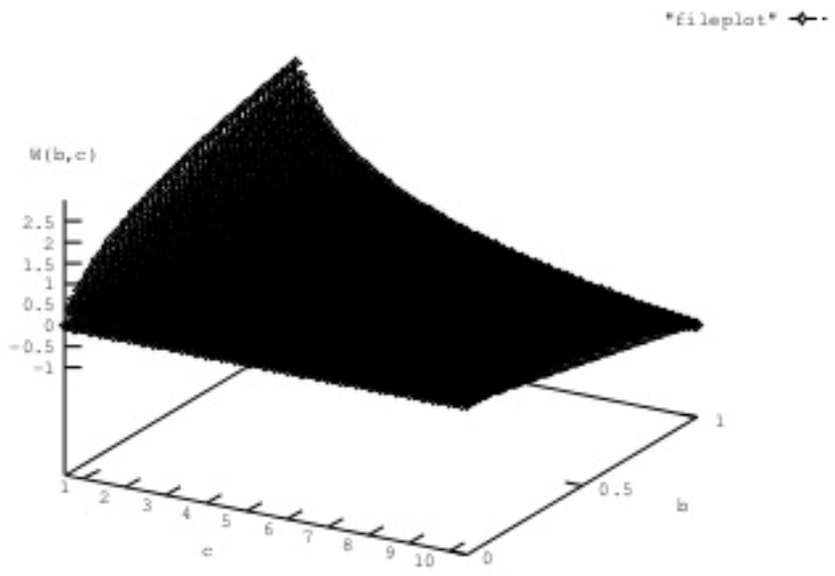


Figure 6.1: St. Petersburg: $W(b,c)$ as a function of b and c.

(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 $b = 1$.

$$\frac{\partial W(b,c)}{\partial b} = \sum_{k=1}^{\infty} 2^{-k} \frac{1}{\left(1 - b + \frac{b2^k}{c}\right)} \left(-1 + \frac{2^k}{c}\right) \tag{6.105}$$

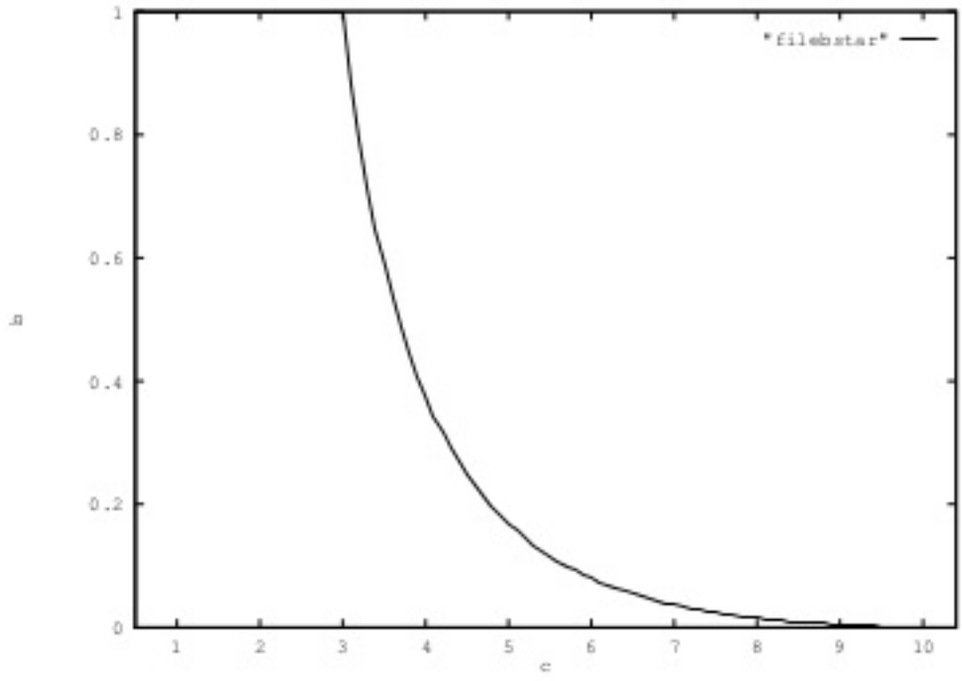


Figure 6.2: St. Petersburg: b^* as a function of c.

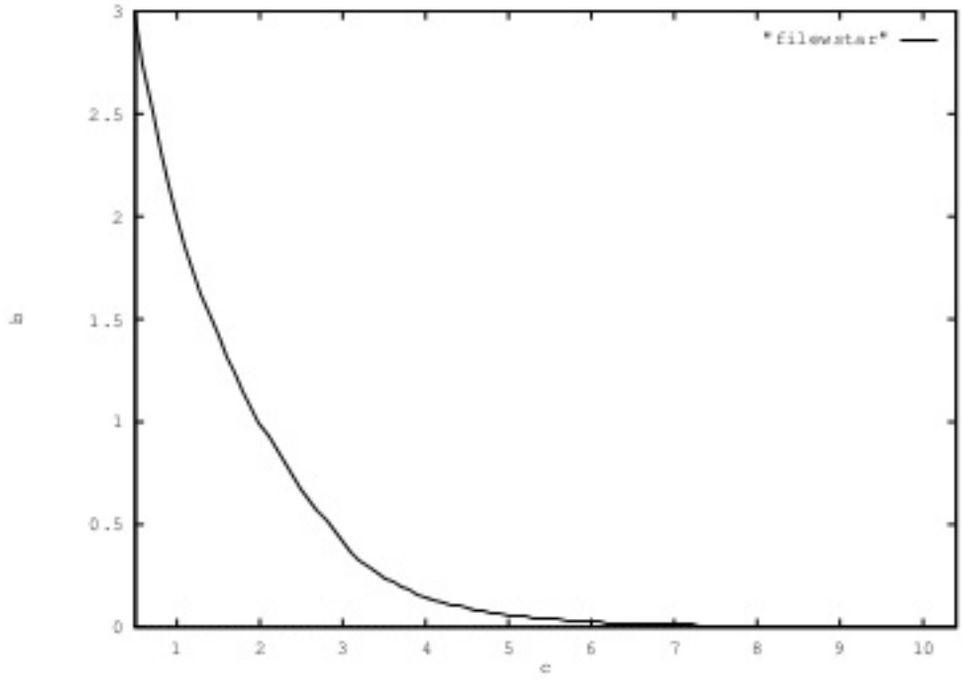


Figure 6.3: St. Petersburg: $W^*(b^*, c)$ as a function of c.

$$= \sum_k \frac{2^{-k}}{\frac{2^k}{c}} \left(\frac{2^k}{d} - 1 \right) \tag{6.106}$$

$$= \sum_{k=1}^{\infty} 2^{-k} - \sum_{k=1}^{\infty} c 2^{-2k} \tag{6.107}$$

$$= 1 - \frac{c}{3} \tag{6.108}$$

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

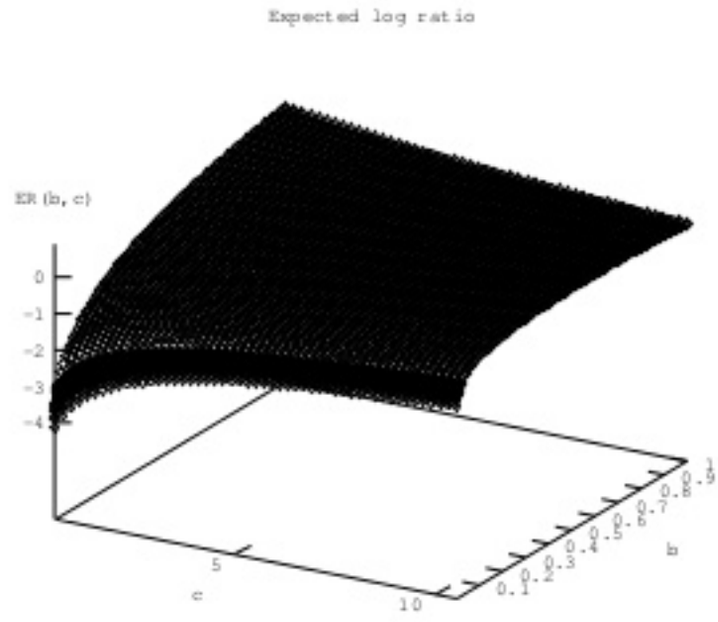


Figure 6.4: Super St. Petersburg: $J(b,c)$ as a function of b and c.

complications. For example, for $k = 6$, 2^{2^k} is outside the normal range of number 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.

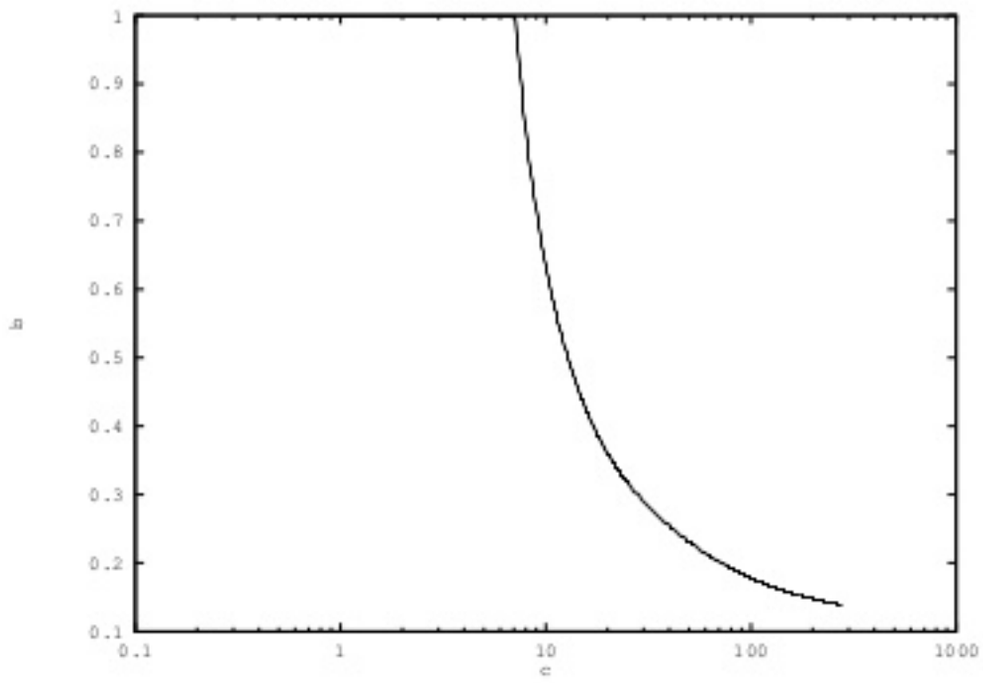


Figure 6.5: Super St. Petersburg: b^* as a function of c.

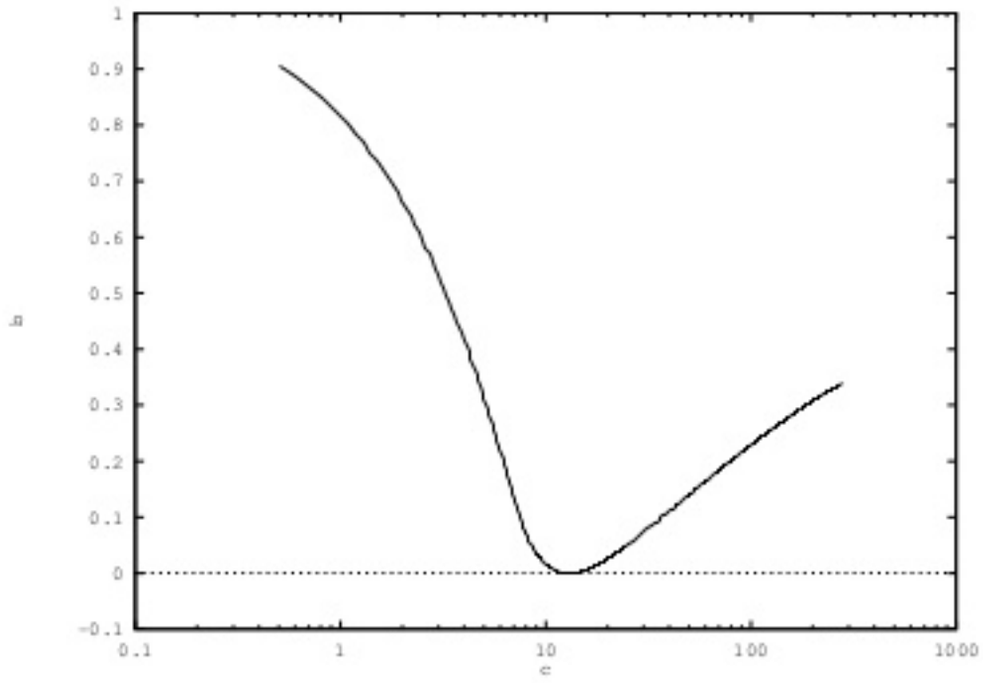


Figure 6.6: Super St. Petersburg: $J^*(b^*, c)$ as a function of c.

Bibliography

[1] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, 1981.

[2] R.G. Gallager. *Information Theory and Reliable Communication*. Wiley, New York, 1968.

[3] R.G. Gallager. Variations on a theme by Huffman. *IEEE Trans. Inform. Theory*, IT-24:668–674, 1978.

[4] A Rényi. *Wahrscheinlichkeitsrechnung, mit einem Anhang über Informationstheorie*. Veb Deutscher Verlag der Wissenschaften, Berlin, 1962.

[5] 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.