Relative entropy:
$D(p(x, y)||q(x, y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$.

Jensen's inequality. If f is a convex function, then $Ef(X) \geq f(EX)$.

Log sum inequality. For n positive numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{2.165}$$

with equality if and only if $\frac{a_i}{b_i} = \text{constant}$.

Data-processing inequality. If $X \to Y \to Z$ forms a Markov chain, $I(X; Y) \geq I(X; Z)$.

Sufficient statistic. $T(X)$ is sufficient relative to $\{f_\theta(x)\}$ if and only if $I(\theta; X) = I(\theta; T(X))$ for all distributions on θ.

Fano's inequality. Let $P_e = \Pr\{\hat{X}(Y) \neq X\}$. Then

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|Y). \tag{2.166}$$

Inequality. If X and X' are independent and identically distributed, then

$$\Pr(X = X') \geq 2^{-H(X)}, \tag{2.167}$$

PROBLEMS

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

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

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

(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.

2.2 *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) $Y = \cos X$?

2.3 *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 that achieve this minimum.

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

$$H(X, g(X)) \overset{\text{(a)}}{=} H(X) + H(g(X) \mid X) \tag{2.168}$$

$$\overset{\text{(b)}}{=} H(X), \tag{2.169}$$

$$H(X, g(X)) \overset{\text{(c)}}{=} H(g(X)) + H(X \mid g(X)) \tag{2.170}$$

$$\overset{\text{(d)}}{\geq} H(g(X)). \tag{2.171}$$

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

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

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

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

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

2.7 *Coin weighing*. Suppose that 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?

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

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

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

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

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

$$\rho(X, Y) = H(X) + H(Y) - 2I(X; Y) \tag{2.172}$$

$$= H(X, Y) - I(X; Y) \tag{2.173}$$

$$= 2H(X, Y) - H(X) - H(Y). \tag{2.174}$$

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

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

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

2.11 *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 that $\rho = \frac{I(X_1;X_2)}{H(X_1)}$.
(b) Show that $0 \le \rho \le 1$.
(c) When is $\rho = 0$?
(d) When is $\rho = 1$?

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

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

Find:
(a) $H(X), H(Y)$.
(b) $H(X \mid Y), H(Y \mid X)$.
(c) $H(X, Y)$.
(d) $H(Y) - H(Y \mid X)$.
(e) $I(X; Y)$.
(f) Draw a Venn diagram for the quantities in parts (a) through (e).

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

2.14 *Entropy of a sum*. Let X and Y be random variables that take on values $x_1, x_2, \ldots, x_r$ and $y_1, y_2, \ldots, y_s$, respectively. Let $Z = X + Y$.
(a) Show that $H(Z|X) = H(Y|X)$. Argue that if X, Y are independent, then $H(Y) \le H(Z)$ and $H(X) \le H(Z)$. 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 $H(Z) = H(X) + H(Y)$?

2.15 *Data processing*. Let $X_1 \to X_2 \to X_3 \to \cdots \to X_n$ form a Markov chain in this order; that is, 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.

2.16 *Bottleneck*. Suppose that a (nonstationary) Markov chain starts in one of n states, necks down to $k < n$ states, and then fans back to $m > k$ states. Thus, $X_1 \to X_2 \to X_3$, that is,

$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$, for all $x_1 \in \{1, 2, \ldots, n\}$, $x_2 \in \{1, 2, \ldots, k\}$, $x_3 \in \{1, 2, \ldots, m\}$.

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

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

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

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

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

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

the games are independent, calculate $H(X)$, $H(Y)$, $H(Y|X)$, and $H(X|Y)$.

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

2.20 *Run-length coding*. Let $X_1, X_2, \ldots, X_n$ be (possibly dependent) binary random variables. Suppose that 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.

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

$$\Pr\{p(X) \leq d\} \log \frac{1}{d} \leq H(X). \tag{2.175}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(b) Justify the following steps:

$$-D(p||q) = \sum_x p(x) \ln \frac{q(x)}{p(x)} \tag{2.176}$$

$$\leq \sum_x p(x) \left(\frac{q(x)}{p(x)} - 1 \right) \tag{2.177}$$

$$\leq 0. \tag{2.178}$$

(c) What are the conditions for equality?

2.27 *Grouping rule for entropy*. Let $\mathbf{p} = (p_1, p_2, \ldots, p_m)$ be a probability distribution on m elements (i.e., $p_i \geq 0$ and $\sum_{i=1}^{m} p_i = 1$).

Define a new distribution $\mathbf{q}$ on $m-1$ elements as $q_1 = p_1, q_2 = p_2, \ldots, q_{m-2} = p_{m-2}$, and $q_{m-1} = p_{m-1} + p_m$ [i.e., the distribution $\mathbf{q}$ is the same as $\mathbf{p}$ on $\{1, 2, \ldots, m-2\}$, and the probability of the last element in $\mathbf{q}$ is the sum of the last two probabilities of $\mathbf{p}$]. Show that

$$H(\mathbf{p}) = H(\mathbf{q}) + (p_{m-1} + p_m) H\left(\frac{p_{m-1}}{p_{m-1} + p_m}, \frac{p_m}{p_{m-1} + p_m}\right). \tag{2.179}$$

2.28 *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.

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

2.30 *Maximum entropy.* Find the probability mass function $p(x)$ that maximizes the entropy $H(X)$ of a nonnegative 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 $H(X)$.

2.31 *Conditional entropy.* Under what conditions does $H(X|g(Y)) = H(X|Y)$?

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

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

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

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

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

2.33 *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.

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

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

Symbol	$p(x)$	$q(x)$
a	$\frac{1}{2}$	$\frac{1}{3}$
b	$\frac{1}{4}$	$\frac{1}{3}$
c	$\frac{1}{4}$	$\frac{1}{3}$

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

2.36 *Symmetric relative entropy.* Although, as Problem 2.35 shows, $D(p||q) \neq D(q||p)$ in general, there could be distributions for which equality holds. Give an example of two distributions p and q on a binary alphabet such that $D(p||q) = D(q||p)$ (other than the trivial case $p = q$).

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

$$D(p(x, y, z)||p(x)p(y)p(z)) = E\left[\log \frac{p(x, y, z)}{p(x)p(y)p(z)}\right]. \quad (2.180)$$

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

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

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

Suppose that $\Pr\{X \in S\} = \alpha$. Find the decrease in uncertainty $H(X) - H(X|Y)$.
Apparently, any set S with a given α is as good as any other.

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

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

(b) Give an example achieving this minimum.

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

(a) Find $H(X)$.

(b) Find $H(Y)$.

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

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

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

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

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

(a) $H(5X)$ vs. $H(X)$

(b) $I(g(X); Y)$ vs. $I(X; Y)$

(c) $H(X_0|X_{-1})$ vs. $H(X_0|X_{-1}, X_1)$

(d) $H(X, Y)/(H(X) + H(Y))$ vs. 1

2.43 *Mutual information of heads and tails*

(a) Consider a fair coin flip. What is the mutual information between the top and bottom sides of the coin?

(b) A six-sided fair die is rolled. What is the mutual information between the top side and the front face (the side most facing you)?

2.44 *Pure randomness*. We wish to use a three-sided coin to generate a fair coin toss. Let the coin X have probability mass function

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

where p_A, p_B, p_C are unknown.

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

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

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

2.46 *Axiomatic definition of entropy (Difficult)*. If we assume certain axioms for our measure of information, we will be forced to use a logarithmic measure such as entropy. Shannon used this to justify his initial definition of entropy. In this book we 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

$$H_m(p_1, p_2, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i, \qquad m = 2, 3, \ldots. \tag{2.181}$$

There are various other axiomatic formulations which result in the same definition of entropy. See, for example, the book by Csiszár and Körner [149].

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

2.48 *Sequence length*. How much information does the length of a sequence give about the content of a sequence? Suppose that we consider a Bernoulli $(\frac{1}{2})$ process $\{X_i\}$. Stop the process when the first 1 appears. Let N designate this stopping time. Thus, X^N is an element of the set of all finite-length binary sequences $\{0, 1\}^* = \{0, 1, 00, 01, 10, 11, 000, \ldots\}$.

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

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

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

Let's now consider a different stopping time. For this part, again assume that $X_i \sim \text{Bernoulli}(\frac{1}{2})$ but stop at time $N = 6$, with probability $\frac{1}{3}$ and stop at time $N = 12$ with probability $\frac{2}{3}$. Let this stopping time be independent of the sequence $X_1 X_2 \cdots X_{12}$.

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

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

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

HISTORICAL NOTES

The concept of entropy was introduced in thermodynamics, where it was used to provide a statement of the second law of thermodynamics. Later, statistical mechanics provided a connection between thermodynamic entropy and the logarithm of the number of microstates in a macrostate of the system. This work was the crowning achievement of Boltzmann, who had the equation $S = k \ln W$ inscribed as the epitaph on his gravestone [361].

In the 1930s, Hartley introduced a logarithmic measure of information for communication. His measure was essentially the logarithm of the alphabet size. Shannon [472] was the first to define entropy and mutual information as defined in this chapter. Relative entropy was first defined by Kullback and Leibler [339]. It is known under a variety of names, including the Kullback–Leibler distance, cross entropy, information divergence, and information for discrimination, and has been studied in detail by Csiszár [138] and Amari [22].

SUMMARY

AEP. "Almost all events are almost equally surprising." Specifically, if $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$, then

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X) \text{ in probability.} \tag{3.28}$$

Definition. The *typical set* $A_\epsilon^{(n)}$ is the set of sequences $x_1, x_2, \ldots, x_n$ satisfying

$$2^{-n(H(X)+\epsilon)} \le p(x_1, x_2, \ldots, x_n) \le 2^{-n(H(X)-\epsilon)}. \tag{3.29}$$

Properties of the typical set

1. If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then $p(x_1, x_2, \ldots, x_n) = 2^{-n(H\pm\epsilon)}$.
2. $\Pr\left\{A_\epsilon^{(n)}\right\} > 1 - \epsilon$ for n sufficiently large.
3. $\left|A_\epsilon^{(n)}\right| \le 2^{n(H(X)+\epsilon)}$, where $|A|$ denotes the number of elements in set A.

Definition. $a_n \doteq b_n$ means that $\frac{1}{n}\log\frac{a_n}{b_n} \to 0$ as $n \to \infty$.

Smallest probable set. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$, and for $\delta < \frac{1}{2}$, let $B_\delta^{(n)} \subset \mathcal{X}^n$ be the smallest set such that $\Pr\{B_\delta^{(n)}\} \ge 1 - \delta$. Then

$$|B_\delta^{(n)}| \doteq 2^{nH}. \tag{3.30}$$

PROBLEMS

3.1 *Markov's inequality and Chebyshev's inequality*

(a) (*Markov's inequality*) For any nonnegative random variable X and any $t > 0$, show that

$$\Pr\{X \ge t\} \le \frac{EX}{t}. \tag{3.31}$$

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 $\epsilon > 0$,

$$\Pr\{|Y - \mu| > \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}. \tag{3.32}$$

(c) (*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

$$\Pr\left\{\left|\overline{Z}_n - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}. \tag{3.33}$$

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*.

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

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

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

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

Thus, for example, the first cut (and choice of largest piece) may result in a piece of size $\frac{3}{5}$. Cutting and choosing from this piece might reduce it to size $\left(\frac{3}{5}\right)\left(\frac{2}{3}\right)$ at time 2, and so on. How large, to first order in the exponent, is the piece of cake after n cuts?

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

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

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

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

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

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

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

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

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

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

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

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

3.7 *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 1's.

(a) Assuming that all codewords are the same length, find the minimum length required to provide codewords for all sequences with three or fewer 1's.

(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).

3.8 *Products.*
Let

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

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

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

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

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

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

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

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

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

(b) Justify the steps in the chain of inequalities

$$1 - \epsilon - \delta \leq \Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) \tag{3.34}$$

$$= \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \tag{3.35}$$

$$\leq \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\epsilon)} \tag{3.36}$$

$$= |A_\epsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H-\epsilon)} \tag{3.37}$$

$$\leq |B_\delta^{(n)}| 2^{-n(H-\epsilon)}. \tag{3.38}$$

(c) Complete the proof of the theorem.

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

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

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

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

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

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

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

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

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

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

(a) Calculate $H(X)$.

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

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

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

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

HISTORICAL NOTES

The asymptotic equipartition property (AEP) was first stated by Shannon in his original 1948 paper [472], where he proved the result for i.i.d. processes and stated the result for stationary ergodic processes. McMillan [384] and Breiman [74] proved the AEP for ergodic finite alphabet sources. The result is now referred to as the AEP or the Shannon–McMillan–Breiman theorem. Chung [101] extended the theorem to the case of countable alphabets and Moy [392], Perez [417], and Kieffer [312] proved the $\mathcal{L}_1$ convergence when $\{X_i\}$ is continuous valued and ergodic. Barron [34] and Orey [402] proved almost sure convergence for real-valued ergodic processes; a simple sandwich argument (Algoet and Cover [20]) will be used in Section 16.8 to prove the general AEP.

4. The conditional entropy $H(X_n|X_1)$ increases with time for a stationary Markov chain.
5. The conditional entropy $H(X_0|X_n)$ of the initial condition X_0 increases for any Markov chain.

Functions of a Markov chain. If $X_1, X_2, \ldots, X_n$ form a stationary Markov chain and $Y_i = \phi(X_i)$, then

$$H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) \leq H(\mathcal{Y}) \leq H(Y_n|Y_{n-1}, \ldots, Y_1) \tag{4.80}$$

and

$$\lim_{n\to\infty} H(Y_n|Y_{n-1}, \ldots, Y_1, X_1) = H(\mathcal{Y}) = \lim_{n\to\infty} H(Y_n|Y_{n-1}, \ldots, Y_1). \tag{4.81}$$

PROBLEMS

4.1 *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.

4.2 *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, 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.

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

$$H(TX) \geq H(TX|T) \tag{4.82}$$

$$= H(T^{-1}TX|T) \tag{4.83}$$

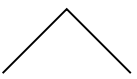
$$= H(X|T) \tag{4.84}$$

$$= H(X) \tag{4.85}$$

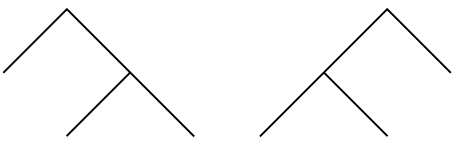
if X and T are independent.

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

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

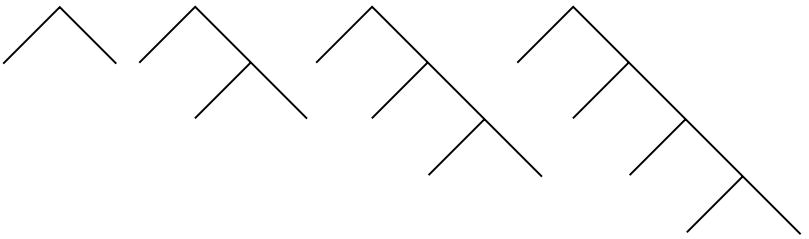


Then expand one of the two terminal nodes at random:

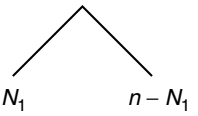


At time k, choose one of the $k-1$ terminal nodes according to a uniform distribution and expand it. Continue until n terminal nodes

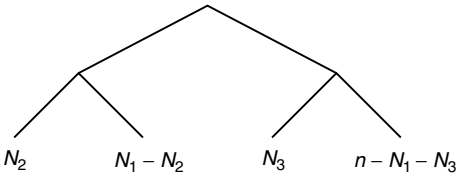
have been generated. Thus, a sequence leading to a five-node tree might look like this:



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



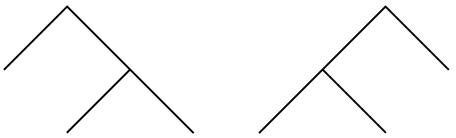
Then choose an integer N_2 uniformly distributed over $\{1, 2, \ldots, N_1 - 1\}$, and independently choose another integer N_3 uniformly over $\{1, 2, \ldots, (n - N_1) - 1\}$. The picture is now



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

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

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



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

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:

$$H(T_n) \overset{\text{(a)}}{=} H(N_1, T_n) \tag{4.86}$$

$$\overset{\text{(b)}}{=} H(N_1) + H(T_n|N_1) \tag{4.87}$$

$$\overset{\text{(c)}}{=} \log(n-1) + H(T_n|N_1) \tag{4.88}$$

$$\overset{\text{(d)}}{=} \log(n-1) + \frac{1}{n-1}\sum_{k=1}^{n-1}(H(T_k) + H(T_{n-k})) \tag{4.89}$$

$$\overset{\text{(e)}}{=} \log(n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H(T_k) \tag{4.90}$$

$$= \log(n-1) + \frac{2}{n-1}\sum_{k=1}^{n-1} H_k. \tag{4.91}$$

(f) Use this to show that

$$(n-1)H_n = nH_{n-1} + (n-1)\log(n-1) - (n-2)\log(n-2) \tag{4.92}$$

or

$$\frac{H_n}{n} = \frac{H_{n-1}}{n-1} + c_n \tag{4.93}$$

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.

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

(a)

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \leq \frac{H(X_1, X_2, \ldots, X_{n-1})}{n-1}. \tag{4.94}$$

(b)

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \geq H(X_n|X_{n-1}, \ldots, X_1). \tag{4.95}$$

4.7 *Entropy rates of Markov chains*

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

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

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

(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 $\frac{1}{2}$, 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).]

4.8 *Maximum entropy process*. A discrete memoryless source has the 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 $H(\mathcal{X}) = \frac{H(X)}{ET}$. What is the maximum value $H(\mathcal{X})$?

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

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

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

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

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

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

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

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

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

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

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

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

4.12 *Entropy rate of a dog looking for a bone.* A dog walks on the integers, possibly reversing direction at each step with probability $p = 0.1$. 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 the dog.

(c) What is the expected number of steps that the dog takes before reversing direction?

4.13 *The past has little to say about the future.* For a stationary stochastic process $X_1, X_2, \ldots, X_n, \ldots$, show that

$$\lim_{n\to\infty} \frac{1}{2n} I(X_1, X_2, \ldots, X_n; X_{n+1}, X_{n+2}, \ldots, X_{2n}) = 0. \quad (4.96)$$

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

4.14 *Functions of a stochastic process*

(a) Consider a stationary stochastic process $X_1, X_2, \ldots, X_n$, and let $Y_1, Y_2, \ldots, Y_n$ be defined by

$$Y_i = \phi(X_i), \quad i = 1, 2, \ldots \quad (4.97)$$

for some function ϕ. Prove that

$$H(\mathcal{Y}) \leq H(\mathcal{X}). \quad (4.98)$$

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

$$Z_i = \psi(X_i, X_{i+1}), \quad i = 1, 2, \ldots \tag{4.99}$$

for some function ψ?

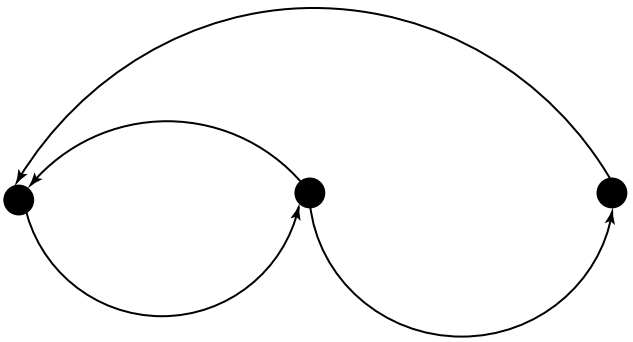
4.15 *Entropy rate*. Let $\{X_i\}$ be a discrete stationary stochastic process with entropy rate $H(\mathcal{X})$. Show that

$$\frac{1}{n} H(X_n, \ldots, X_1 \mid X_0, X_{-1}, \ldots, X_{-k}) \to H(\mathcal{X}) \tag{4.100}$$

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

4.16 *Entropy rate of constrained sequences*. In magnetic recording, the mechanism of recording and reading the bits imposes constraints on the sequences of bits that can be recorded. For example, to ensure proper synchronization, 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 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:

$$\begin{bmatrix} X_1(n) \\ X_2(n) \\ X_3(n) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X_1(n-1) \\ X_2(n-1) \\ X_3(n-1) \end{bmatrix}, \tag{4.101}$$

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

(c) Let

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{4.102}$$

Then we have by induction

$$\mathbf{X}(n) = A\mathbf{X}(n-1) = A^2\mathbf{X}(n-2) = \cdots = A^{n-1}\mathbf{X}(1). \tag{4.103}$$

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

$$\mathbf{X}(n) = \lambda_1^{n-1}\mathbf{Y}_1 + \lambda_2^{n-1}\mathbf{Y}_2 + \lambda_3^{n-1}\mathbf{Y}_3, \tag{4.104}$$

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

$$\frac{1}{n}\log X_i(n) \to \log \lambda, \tag{4.105}$$

where λ is the largest (positive) eigenvalue. Thus, the number of sequences of length n grows as λ^n for large n. Calculate λ for the matrix A above. (The case when the eigenvalues are not distinct can be handled in a similar manner.)

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

$$P = \begin{bmatrix} 0 & 1 & 0 \\ \alpha & 0 & 1-\alpha \\ 1 & 0 & 0 \end{bmatrix}. \tag{4.106}$$

Show that the stationary distribution of this Markov chain is

$$\mu = \left[\frac{1}{3-\alpha}, \frac{1}{3-\alpha}, \frac{1-\alpha}{3-\alpha}\right]. \tag{4.107}$$

(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?

4.17 *Recurrence times are insensitive to distributions.* Let $X_0, X_1, X_2, \ldots$ be drawn i.i.d. $\sim p(x), x \in \mathcal{X} = \{1, 2, \ldots, m\}$, and let N be the waiting time to the next occurrence of X_0. Thus $N = \min_n\{X_n = X_0\}$.

(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.

4.18 *Stationary but not ergodic process.* A bin has two biased coins, one with probability of heads p and the other with probability of heads $1 - p$. One of these coins is chosen at random (i.e., with probability $\frac{1}{2}$) and is then tossed n times. Let X denote the identity of the coin that is picked, and let Y_1 and Y_2 denote the results of the first two tosses.

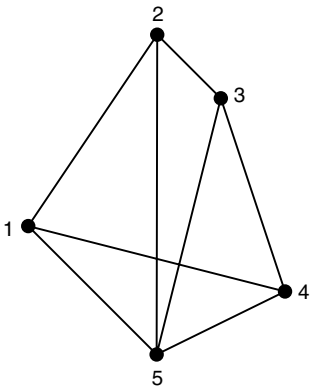
(a) Calculate $I(Y_1; Y_2|X)$.

(b) Calculate $I(X; Y_1, Y_2)$.

(c) Let $\mathcal{H}(\mathcal{Y})$ be the entropy rate of the Y process (the sequence of coin tosses). Calculate $\mathcal{H}(\mathcal{Y})$. [*Hint:* Relate this to $\lim \frac{1}{n} H(X, Y_1, Y_2, \ldots, Y_n)$.]

You can check the answer by considering the behavior as $p \to \frac{1}{2}$.

4.19 *Random walk on graph.* Consider a random walk on the following graph:



(a) Calculate the stationary distribution.

(b) What is the entropy rate?

(c) Find the mutual information $I(X_{n+1}; X_n)$ assuming that the process is stationary.

4.20 *Random walk on chessboard*. Find the entropy rate of the Markov chain associated with a random walk of a king on the 3×3 chessboard

1	2	3
4	5	6
7	8	9

What about the entropy rate of rooks, bishops, and queens? There are two types of bishops.

4.21 *Maximal entropy graphs*. Consider a random walk on a connected graph with four edges.

(a) Which graph has the highest entropy rate?

(b) Which graph has the lowest?

4.22 *Three-dimensional maze*. A bird is lost in a $3 \times 3 \times 3$ cubical maze. The bird flies from room to room going to adjoining rooms with equal probability through each of the walls. For example, the corner rooms have three exits.

(a) What is the stationary distribution?

(b) What is the entropy rate of this random walk?

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

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

(b) What are the conditions for equality?

4.24 *Entropy rates*. Let $\{X_i\}$ be a stationary process. Let $Y_i = (X_i, X_{i+1})$. Let $Z_i = (X_{2i}, X_{2i+1})$. Let $V_i = X_{2i}$. Consider the entropy rates $H(\mathcal{X})$, $H(\mathcal{Y})$, $H(\mathcal{Z})$, and $H(\mathcal{V})$ of the processes $\{X_i\}$,$\{Y_i\}$, $\{Z_i\}$, and $\{V_i\}$. What is the inequality relationship $\leq$, $=$, or $\geq$ between each of the pairs listed below?

(a) $H(\mathcal{X}) \gtreqless H(\mathcal{Y})$.

(b) $H(\mathcal{X}) \gtreqless H(\mathcal{Z})$.

(c) $H(\mathcal{X}) \gtreqless H(\mathcal{V})$.

(d) $H(\mathcal{Z}) \gtreqless H(\mathcal{X})$.

4.25 *Monotonicity*

(a) Show that $I(X; Y_1, Y_2, \ldots, Y_n)$ is nondecreasing in n.

(b) Under what conditions is the mutual information constant for all n?

4.26 *Transitions in Markov chains*. Suppose that $\{X_i\}$ forms an irreducible Markov chain with transition matrix P and stationary distribution μ. Form the associated "edge process" $\{Y_i\}$ by keeping track only of the transitions. Thus, the new process $\{Y_i\}$ takes values in $\mathcal{X} \times \mathcal{X}$, and $Y_i = (X_{i-1}, X_i)$. For example,

$$X^n = 3, 2, 8, 5, 7, \ldots$$

becomes

$$Y^n = (\emptyset, 3), (3, 2), (2, 8), (8, 5), (5, 7), \ldots.$$

Find the entropy rate of the edge process $\{Y_i\}$.

4.27 *Entropy rate*. Let $\{X_i\}$ be a stationary $\{0, 1\}$-valued stochastic process obeying

$$X_{k+1} = X_k \oplus X_{k-1} \oplus Z_{k+1},$$

where $\{Z_i\}$ is Bernoulli(p)and $\oplus$ denotes mod 2 addition. What is the entropy rate $H(\mathcal{X})$?

4.28 *Mixture of processes*. Suppose that we observe one of two stochastic processes but don't know which. What is the entropy rate? Specifically, let $X_{11}, X_{12}, X_{13}, \ldots$ be a Bernoulli process with parameter p_1, and let $X_{21}, X_{22}, X_{23}, \ldots$ be Bernoulli(p_2). Let

$$\theta = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ 2 & \text{with probability } \frac{1}{2} \end{cases}$$

and let $Y_i = X_{\theta i}, i = 1, 2, \ldots$, be the stochastic process observed. Thus, Y observes the process $\{X_{1i}\}$ or $\{X_{2i}\}$. Eventually, Y will know which.

(a) Is $\{Y_i\}$ stationary?

(b) Is $\{Y_i\}$ an i.i.d. process?

(c) What is the entropy rate H of $\{Y_i\}$?

(d) Does

$$-\frac{1}{n} \log p(Y_1, Y_2, \ldots Y_n) \longrightarrow H?$$

(e) Is there a code that achieves an expected per-symbol description length $\frac{1}{n} EL_n \longrightarrow H$?

Now let θ_i be Bern($\frac{1}{2}$). Observe that

$$Z_i = X_{\theta_i i}, \qquad i = 1, 2, \ldots.$$

Thus, θ is not fixed for all time, as it was in the first part, but is chosen i.i.d. each time. Answer parts (a), (b), (c), (d), (e) for the process $\{Z_i\}$, labeling the answers (a′), (b′), (c′), (d′), (e′).

4.29 *Waiting times*. Let X be the waiting time for the first heads to appear in successive flips of a fair coin. For example, $\Pr\{X = 3\} = (\frac{1}{2})^3$. Let S_n be the waiting time for the nth head to appear. Thus,

$$S_0 = 0$$

$$S_{n+1} = S_n + X_{n+1},$$

where $X_1, X_2, X_3, \ldots$ are i.i.d according to the distribution above.

(a) Is the process $\{S_n\}$ stationary?

(b) Calculate $H(S_1, S_2, \ldots, S_n)$.

(c) Does the process $\{S_n\}$ have an entropy rate? If so, what is it? If not, why not?

(d) What is the expected number of fair coin flips required to generate a random variable having the same distribution as S_n?

4.30 *Markov chain transitions*

$$P = [P_{ij}] = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

Let X_1 be distributed uniformly over the states $\{0, 1, 2\}$. Let $\{X_i\}_1^\infty$ be a Markov chain with transition matrix P; thus, $P(X_{n+1} = j | X_n = i) = P_{ij}, i, j \in \{0, 1, 2\}$.

(a) Is $\{X_n\}$ stationary?

(b) Find $\lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n)$.

Now consider the derived process $Z_1, Z_2, \ldots, Z_n$, where

$$Z_1 = X_1$$

$$Z_i = X_i - X_{i-1} \pmod 3, \qquad i = 2, \ldots, n.$$

Thus, Z^n encodes the transitions, not the states.

(c) Find $H(Z_1, Z_2, \ldots, Z_n)$.

(d) Find $H(Z_n)$ and $H(X_n)$ for $n \geq 2$.

(e) Find $H(Z_n|Z_{n-1})$ for $n \geq 2$.

(f) Are Z_{n-1} and Z_n independent for $n \geq 2$?

4.31 *Markov*. Let $\{X_i\} \sim \text{Bernoulli}(p)$. Consider the associated Markov chain $\{Y_i\}_{i=1}^n$, where
$Y_i = (\text{the number of 1's in the current run of 1's})$. For example, if $X^n = 101110\ldots$, we have $Y^n = 101230\ldots$.

(a) Find the entropy rate of X^n.

(b) Find the entropy rate of Y^n.

4.32 *Time symmetry*. Let $\{X_n\}$ be a stationary Markov process. We condition on (X_0, X_1) and look into the past and future. For what index k is

$$H(X_{-n}|X_0, X_1) = H(X_k|X_0, X_1)?$$

Give the argument.

4.33 *Chain inequality*. Let $X_1 \to X_2 \to X_3 \to X_4$ form a Markov chain. Show that

$$I(X_1; X_3) + I(X_2; X_4) \leq I(X_1; X_4) + I(X_2; X_3). \tag{4.108}$$

4.34 *Broadcast channel*. Let $X \to Y \to (Z, W)$ form a Markov chain [i.e., $p(x, y, z, w) = p(x)p(y|x)p(z, w|y)$ for all x, y, z, w]. Show that

$$I(X; Z) + I(X; W) \leq I(X; Y) + I(Z; W). \tag{4.109}$$

4.35 *Concavity of second law*. Let $\{X_n\}_{-\infty}^{\infty}$ be a stationary Markov process. Show that $H(X_n|X_0)$ is concave in n. Specifically, show that

$$\begin{aligned} &H(X_n|X_0) - H(X_{n-1}|X_0) - (H(X_{n-1}|X_0) - H(X_{n-2}|X_0)) \\ &\quad = -I(X_1; X_{n-1}|X_0, X_n) \leq 0. \end{aligned} \tag{4.110}$$

Thus, the second difference is negative, establishing that $H(X_n|X_0)$ is a concave function of n.

HISTORICAL NOTES

The entropy rate of a stochastic process was introduced by Shannon [472], who also explored some of the connections between the entropy rate of the process and the number of possible sequences generated by the process. Since Shannon, there have been a number of results extending the basic

Shannon code

$$l_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil \tag{5.137}$$

$$H_D(X) \le L < H_D(X) + 1. \tag{5.138}$$

Huffman code

$$L^* = \min_{\sum D^{-l_i} \le 1} \sum p_i l_i \tag{5.139}$$

$$H_D(X) \le L^* < H_D(X) + 1. \tag{5.140}$$

Wrong code. $X \sim p(x)$, $l(x) = \left\lceil \log \frac{1}{q(x)} \right\rceil$, $L = \sum p(x) l(x)$:

$$H(p) + D(p||q) \le L < H(p) + D(p||q) + 1. \tag{5.141}$$

Stochastic processes

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} \le L_n < \frac{H(X_1, X_2, \ldots, X_n)}{n} + \frac{1}{n}. \tag{5.142}$$

Stationary processes

$$L_n \to H(\mathcal{X}). \tag{5.143}$$

Competitive optimality. Shannon code $l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil$ versus any other code $l'(x)$:

$$\Pr\left(l(X) \ge l'(X) + c\right) \le \frac{1}{2^{c-1}}. \tag{5.144}$$

PROBLEMS

5.1 *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?

5.2 *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.

5.3 *Slackness in the Kraft inequality*. An instantaneous code has word lengths $l_1, l_2, \ldots, l_m$, which satisfy the strict inequality

$$\sum_{i=1}^{m} D^{-l_i} < 1.$$

The code alphabet is $\mathcal{D} = \{0, 1, 2, \ldots, D-1\}$. Show that there exist arbitrarily long sequences of code symbols in $\mathcal{D}^*$ which cannot be decoded into sequences of codewords.

5.4 *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 code length for this encoding.

(c) Find a ternary Huffman code for X.

5.5 *More Huffman codes*. Find the binary Huffman code for the source with probabilities $(\frac{1}{3}, \frac{1}{5}, \frac{1}{5}, \frac{2}{15}, \frac{2}{15})$. Argue that this code is also optimal for the source with probabilities $(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$.

5.6 *Bad codes*. Which of these codes cannot be Huffman codes for any probability assignment?

(a) $\{0, 10, 11\}$

(b) $\{00, 01, 10, 110\}$

(c) $\{01, 10\}$

5.7 *Huffman 20 questions*. Consider a set of n objects. Let $X_i = 1$ or 0 accordingly as the ith object is good or defective. Let $X_1, X_2, \ldots, X_n$ be independent with $\Pr\{X_i = 1\} = p_i$; and $p_1 > p_2 > \cdots > p_n > \frac{1}{2}$. 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? What two sets are we distinguishing with this question? Assume a compact (minimum average length) sequence of questions.

(c) Give an upper bound (within one question) on the minimum average number of questions required.

5.8 *Simple optimum compression of a Markov source.* Consider the three-state Markov process $U_1, U_2, \ldots$ having transition matrix

U_{n-1} \ U_n	S_1	S_2	S_3
S_1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
S_2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
S_3	0	$\frac{1}{2}$	$\frac{1}{2}$

Thus, the probability that S_1 follows S_3 is equal to zero. Design three codes C_1, C_2, C_3 (one for each state 1, 2 and 3, each code mapping elements of the set of S_i's into sequences of 0's and 1's, such that this Markov process can be sent with maximal compression by the following scheme:

(a) Note the present symbol $X_n = i$.

(b) Select code C_i.

(c) Note the next symbol $X_{n+1} = j$ and send the codeword in C_i corresponding to j.

(d) Repeat for the next symbol. What is the average message length of the next symbol conditioned on the previous state $X_n = i$ using this coding scheme? What is the unconditional average number of bits per source symbol? Relate this to the entropy rate $H(\mathcal{U})$ of the Markov chain.

5.9 *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 $L > H(X) + 1 - \epsilon$].

5.10 *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

$$L = \frac{H(X)}{\log_2 3} = H_3(X). \tag{5.145}$$

(a) Show that each symbol of X has a probability of the form 3^{-i} for some i.

(b) Show that m is odd.

5.11 *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.

5.12 *Shannon codes and Huffman codes*. Consider a random variable X that 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 $\left\lceil \log \frac{1}{p(x)} \right\rceil$.

5.13 *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.

5.14 *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.

5.15 *Huffman codes*

(a) Construct a binary Huffman code for the following distribution on five symbols: $\mathbf{p} = (0.3, 0.3, 0.2, 0.1, 0.1)$. What is the average length of this code?

(b) Construct a probability distribution $\mathbf{p}'$ on five symbols for which the code that you constructed in part (a) has an average length (under $\mathbf{p}'$) equal to its entropy $H(\mathbf{p}')$.

5.16 *Huffman codes*. Consider a random variable X that takes six values $\{A, B, C, D, E, F\}$ with probabilities 0.5, 0.25, 0.1, 0.05, 0.05, and 0.05, respectively.

(a) Construct a binary Huffman code for this random variable. What is its average length?

(b) Construct a quaternary Huffman code for this random variable [i.e., a code over an alphabet of four symbols (call them a, b, c and d)]. What is the average length of this code?

(c) One way to construct a binary code for the random variable is to start with a quaternary code and convert the symbols into binary using the mapping $a \rightarrow 00$, $b \rightarrow 01$, $c \rightarrow 10$, and $d \rightarrow 11$. What is the average length of the binary code for the random variable above constructed by this process?

(d) For any random variable X, let L_H be the average length of the binary Huffman code for the random variable, and let L_{QB} be the average length code constructed by first building a quaternary Huffman code and converting it to binary. Show that

$$L_H \leq L_{QB} < L_H + 2. \tag{5.146}$$

(e) The lower bound in the example is tight. Give an example where the code constructed by converting an optimal quaternary code is also the optimal binary code.

(f) The upper bound (i.e., $L_{QB} < L_H + 2$) is not tight. In fact, a better bound is $L_{QB} \leq L_H + 1$. Prove this bound, and provide an example where this bound is tight.

5.17 *Data compression*. Find an optimal set of binary codeword lengths $l_1, l_2, \ldots$ (minimizing $\sum p_i l_i$) for an instantaneous code for each of the following probability mass functions:

(a) $\mathbf{p} = (\frac{10}{41}, \frac{9}{41}, \frac{8}{41}, \frac{7}{41}, \frac{7}{41})$

(b) $\mathbf{p} = (\frac{9}{10}, (\frac{9}{10})(\frac{1}{10}), (\frac{9}{10})(\frac{1}{10})^2, (\frac{9}{10})(\frac{1}{10})^3, \ldots)$

5.18 *Classes of codes*. Consider the code $\{0, 01\}$.

(a) Is it instantaneous?

(b) Is it uniquely decodable?

(c) Is it nonsingular?

5.19 *The game of Hi-Lo*

(a) A computer generates a number X according to a known probability mass function $p(x), x \in \{1, 2, \ldots, 100\}$. The player asks a question, "Is $X = i$?" and is told "Yes," "You're too high," or "You're too low." He continues for a total of six questions. If he is right (i.e., he receives the answer "Yes") during this sequence, he receives a prize of value $v(X)$. How should the player proceed to maximize his expected winnings?

(b) Part (a) 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 1 unit each. How should the player proceed? What is the expected payoff?

(c) Continuing part (b), what if $v(x)$ is fixed but $p(x)$ can be chosen by the computer (and then announced to the player)? The computer wishes to minimize the player's expected return. What should $p(x)$ be? What is the expected return to the player?

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

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

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

(c) Can you show that

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

5.21 *Conditions for unique decodability.* Prove that a code C is uniquely decodable if (and only if) the extension

$$C^k(x_1, x_2, \ldots, x_k) = C(x_1)C(x_2)\cdots C(x_k)$$

is a one-to-one mapping from $\mathcal{X}^k$ to D^* for every $k \ge 1$. (The "only if" part is obvious.)

5.22 *Average length of an optimal code.* Prove that $L(p_1, \ldots, p_m)$, the average codeword length for an optimal D-ary prefix code for probabilities $\{p_1, \ldots, p_m\}$, is a continuous function of $p_1, \ldots, p_m$. This is true even though the optimal code changes discontinuously as the probabilities vary.

5.23 *Unused code sequences.* Let C be a variable-length code that satisfies the Kraft inequality with an 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.

5.24 *Optimal codes for uniform distributions.* Consider a random variable with m equiprobable outcomes. The entropy of this information source is obviously $\log_2 m$ bits.

(a) Describe the optimal instantaneous binary code for this source and compute the average codeword length L_m.

(b) For what values of m does the average codeword length L_m equal the entropy $H = \log_2 m$?

(c) We know that $L < H + 1$ for any probability distribution. The *redundancy* of a variable-length code is defined to be $\rho = L - H$. For what value(s) of m, where $2^k \le m \le 2^{k+1}$, is the redundancy of the code maximized? What is the limiting value of this worst-case redundancy as $m \to \infty$?

5.25 *Optimal codeword lengths.* Although the codeword lengths of an optimal variable-length code are complicated functions of the message probabilities $\{p_1, p_2, \ldots, p_m\}$, it can be said that less probable

symbols are encoded into longer codewords. Suppose that the message probabilities are given in decreasing order, $p_1 > p_2 \geq \cdots \geq p_m$.

(a) Prove that for any binary Huffman code, if the most probable message symbol has probability $p_1 > \frac{2}{5}$, 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 $p_1 < \frac{1}{3}$, that symbol must be assigned a codeword of length ≥ 2.

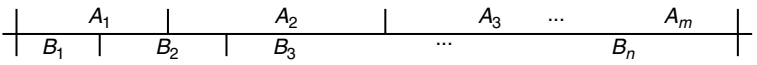
5.26 *Merges*. Companies with values $W_1, W_2, \ldots, W_m$ are merged as follows. The two least valuable companies are merged, thus forming a list of $m-1$ companies. The *value of the merge* is the sum of the values of the two merged companies. This continues until one supercompany remains. Let V equal the sum of the values of the merges. Thus, V represents the total reported dollar volume of the merges. For example, if $\mathbf{W} = (3, 3, 2, 2)$, the merges yield $(3, 3, 2, 2) \rightarrow (4, 3, 3) \rightarrow (6, 4) \rightarrow (10)$ and $V = 4 + 6 + 10 = 20$.

(a) Argue that V is the minimum volume achievable by sequences of pairwise merges terminating in one supercompany. (*Hint:* Compare to Huffman coding.)

(b) Let $W = \sum W_i, \quad \tilde{W}_i = W_i / W$, and show that the minimum merge volume V satisfies

$$W H(\tilde{\mathbf{W}}) \leq V \leq W H(\tilde{\mathbf{W}}) + W. \tag{5.147}$$

5.27 *Sardinas–Patterson test for unique decodability*. A code is not uniquely decodable if and only if there exists a finite sequence of code symbols which can be resolved into sequences of codewords in two different ways. 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 [456]) in

the following way: Construct a set S of all possible dangling suffixes. The code is uniquely decodable if and only if S contains no codeword.

(a) State the precise rules for building the set S.

(b) Suppose that 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.

5.28 *Shannon code*. Consider the following method for generating a code for a random variable X that takes on m values $\{1, 2, \ldots, m\}$ with probabilities $p_1, p_2, \ldots, p_m$. Assume that the probabilities are ordered so that $p_1 \geq p_2 \geq \cdots \geq p_m$. Define

$$F_i = \sum_{k=1}^{i-1} p_k, \tag{5.148}$$

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 that the average length satisfies

$$H(X) \leq L < H(X) + 1. \tag{5.149}$$

(b) Construct the code for the probability distribution (0.5, 0.25, 0.125, 0.125).

5.29 *Optimal codes for dyadic distributions*. For a Huffman code tree, define the probability of a node as the sum of the probabilities of all the leaves under that node. Let the random variable X be drawn from a dyadic distribution [i.e., $p(x) = 2^{-i}$, for some i, for all $x \in \mathcal{X}$]. Now consider a binary Huffman code for this distribution.

(a) Argue that for any node in the tree, the probability of the left child is equal to the probability of the right child.

(b) Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. $\sim p(x)$. Using the Huffman code for $p(x)$, we map $X_1, X_2, \ldots, X_n$ to a sequence of bits $Y_1, Y_2, \ldots, Y_{k(X_1, X_2, \ldots, X_n)}$. (The length of this sequence will depend on the outcome $X_1, X_2, \ldots, X_n$.) Use part (a) to argue that the sequence $Y_1, Y_2, \ldots$ forms a sequence of fair coin flips [i.e., that $\Pr\{Y_i = 0\} = \Pr\{Y_i = 1\} = \frac{1}{2}$, independent of $Y_1, Y_2, \ldots, Y_{i-1}$]. Thus, the entropy rate of the coded sequence is 1 bit per symbol.

(c) Give a heuristic argument why the encoded sequence of bits for any code that achieves the entropy bound cannot be compressible and therefore should have an entropy rate of 1 bit per symbol.

5.30 *Relative entropy is cost of miscoding*. Let the random variable X have five possible outcomes $\{1, 2, 3, 4, 5\}$. Consider two distributions $p(x)$ and $q(x)$ on this random variable.

Symbol	$p(x)$	$q(x)$	$C_1(x)$	$C_2(x)$
1	$\frac{1}{2}$	$\frac{1}{2}$	0	0
2	$\frac{1}{4}$	$\frac{1}{8}$	10	100
3	$\frac{1}{8}$	$\frac{1}{8}$	110	101
4	$\frac{1}{16}$	$\frac{1}{8}$	1110	110
5	$\frac{1}{16}$	$\frac{1}{8}$	1111	111

(a) Calculate $H(p)$, $H(q)$, $D(p||q)$, and $D(q||p)$.

(b) The last two columns represent codes for the random variable. Verify that the average length of C_1 under p is equal to the entropy $H(p)$. Thus, C_1 is optimal for p. Verify that C_2 is optimal for q.

(c) Now assume that we use code C_2 when the distribution is p. What is the average length of the codewords. By how much does it exceed the entropy p?

(d) What is the loss if we use code C_1 when the distribution is q?

5.31 *Nonsingular codes*. The discussion in the text focused on instantaneous codes, with extensions to uniquely decodable codes. Both these are required in cases when the code is to be used repeatedly to encode a sequence of outcomes of a random variable. But if we need to encode only one outcome and we know when we have reached the end of a codeword, we do not need unique decodability—the fact that the code is nonsingular would suffice. For example, if a random variable X takes on three values, a, b, and c, we could encode them by 0, 1, and 00. Such a code is nonsingular but not uniquely decodable.

In the following, assume that we have a random variable X which takes on m values with probabilities $p_1, p_2, \ldots, p_m$ and that the probabilities are ordered so that $p_1 \geq p_2 \geq \cdots \geq p_m$.

(a) By viewing the nonsingular binary code as a ternary code with three symbols, 0, 1, and "STOP," show that the expected length of a nonsingular code $L_{1:1}$ for a random variable X satisfies the following inequality:

$$L_{1:1} \geq \frac{H_2(X)}{\log_2 3} - 1, \tag{5.150}$$

where $H_2(X)$ is the entropy of X in bits. Thus, the average length of a nonsingular code is at least a constant fraction of the average length of an instantaneous code.

(b) Let L_{INST} be the expected length of the best instantaneous code and $L^*_{1:1}$ be the expected length of the best nonsingular code for X. Argue that $L^*_{1:1} \leq L^*_{\text{INST}} \leq H(X) + 1$.

(c) Give a simple example where the average length of the nonsingular code is less than the entropy.

(d) The set of codewords available for a nonsingular code is {0, 1, 00, 01, 10, 11, 000, ...}. Since $L_{1:1} = \sum_{i=1}^{m} p_i l_i$, show that this is minimized if we allot the shortest codewords to the most probable symbols. Thus, $l_1 = l_2 = 1, l_3 = l_4 = l_5 = l_6 = 2$, etc. Show that in general $l_i = \left\lceil \log\left(\frac{i}{2} + 1\right) \right\rceil$, and therefore $L^*_{1:1} = \sum_{i=1}^{m} p_i \left\lceil \log\left(\frac{i}{2} + 1\right) \right\rceil$.

(e) Part (d) shows that it is easy to find the optimal nonsingular code for a distribution. However, it is a little more tricky to deal with the average length of this code. We now bound this average length. It follows from part (d) that $L^*_{1:1} \geq$

$\tilde{L} \stackrel{\triangle}{=} \sum_{i=1}^{m} p_i \log\left(\frac{i}{2}+1\right)$. Consider the difference

$$F(\mathbf{p}) = H(X) - \tilde{L} = -\sum_{i=1}^{m} p_i \log p_i - \sum_{i=1}^{m} p_i \log\left(\frac{i}{2}+1\right). \tag{5.151}$$

Prove by the method of Lagrange multipliers that the maximum of $F(\mathbf{p})$ occurs when $p_i = c/(i+2)$, where $c = 1/(H_{m+2} - H_2)$ and H_k is the sum of the harmonic series:

$$H_k \stackrel{\triangle}{=} \sum_{i=1}^{k} \frac{1}{i}. \tag{5.152}$$

(This can also be done using the nonnegativity of relative entropy.)

(f) Complete the arguments for

$$H(X) - L^*_{1:1} \le H(X) - \tilde{L} \tag{5.153}$$

$$\le \log(2(H_{m+2} - H_2)). \tag{5.154}$$

Now it is well known (see, e.g., Knuth [315]) that $H_k \approx \ln k$ (more precisely, $H_k = \ln k + \gamma + \frac{1}{2k} - \frac{1}{12k^2} + \frac{1}{120k^4} - \epsilon$, where $0 < \epsilon < 1/252n^6$, and $\gamma =$ Euler's constant $= 0.577\ldots$). Using either this or a simple approximation that $H_k \le \ln k + 1$, which can be proved by integration of $\frac{1}{x}$, it can be shown that $H(X) - L^*_{1:1} < \log\log m + 2$. Thus, we have

$$H(X) - \log\log|\mathcal{X}| - 2 \le L^*_{1:1} \le H(X) + 1. \tag{5.155}$$

A nonsingular code cannot do much better than an instantaneous code!

5.32 *Bad wine*. One is given six bottles of wine. It is known that precisely one bottle has gone bad (tastes terrible). From inspection of the bottles it is determined that the probability p_i that the ith bottle is bad is given by $(p_1, p_2, \ldots, p_6) = (\frac{8}{23}, \frac{6}{23}, \frac{4}{23}, \frac{2}{23}, \frac{2}{23}, \frac{1}{23})$. Tasting will determine the bad wine. Suppose that you taste the wines one at a time. Choose the order of tasting to minimize the

expected number of tastings required to determine the bad bottle. Remember, if the first five wines pass the test, you don't have to taste the last.

(a) What is the expected number of tastings required?

(b) Which bottle should be tasted first?

Now you get smart. For the first sample, you mix some of the wines in a fresh glass and sample the mixture. You proceed, mixing and tasting, stopping when the bad bottle has been determined.

(a) What is the minimum expected number of tastings required to determine the bad wine?

(b) What mixture should be tasted first?

5.33 *Huffman vs. Shannon*. A random variable X takes on three values with probabilities 0.6, 0.3, and 0.1.

(a) What are the lengths of the binary Huffman codewords for X? What are the lengths of the binary Shannon codewords $\left(l(x) = \left\lceil \log\left(\frac{1}{p(x)}\right)\right\rceil\right)$ for X?

(b) What is the smallest integer D such that the expected Shannon codeword length with a D-ary alphabet equals the expected Huffman codeword length with a D-ary alphabet?

5.34 *Huffman algorithm for tree construction*. Consider the following problem: m binary signals $S_1, S_2, \ldots, S_m$ are available at times $T_1 \leq T_2 \leq \cdots \leq T_m$, and we would like to find their sum $S_1 \oplus S_2 \oplus \cdots \oplus S_m$ using two-input gates, each gate with one time unit delay, so that the final result is available as quickly as possible. A simple greedy algorithm is to combine the earliest two results, forming the partial result at time $\max(T_1, T_2) + 1$. We now have a new problem with $S_1 \oplus S_2, S_3, \ldots, S_m$, available at times $\max(T_1, T_2) + 1, T_3, \ldots, T_m$. We can now sort this list of T's and apply the same merging step again, repeating this until we have the final result.

(a) Argue that the foregoing procedure is optimal, in that it constructs a circuit for which the final result is available as quickly as possible.

(b) Show that this procedure finds the tree that minimizes

$$C(T) = \max_i (T_i + l_i), \tag{5.156}$$

where T_i is the time at which the result allotted to the ith leaf is available and l_i is the length of the path from the ith leaf to the root.

(c) Show that

$$C(T) \geq \log_2 \left(\sum_i 2^{T_i} \right) \tag{5.157}$$

for any tree T.

(d) Show that there exists a tree such that

$$C(T) \leq \log_2 \left(\sum_i 2^{T_i} \right) + 1. \tag{5.158}$$

Thus, $\log_2 \left(\sum_i 2^{T_i} \right)$ is the analog of entropy for this problem.

5.35 *Generating random variables*. One wishes to generate a random variable X

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p. \end{cases} \tag{5.159}$$

You are given fair coin flips $Z_1, Z_2, \ldots$. Let N be the (random) number of flips needed to generate X. Find a good way to use $Z_1, Z_2, \ldots$ to generate X. Show that $EN \leq 2$.

5.36 *Optimal word lengths*.

(a) Can $l = (1, 2, 2)$ be the word lengths of a binary Huffman code. What about (2,2,3,3)?

(b) What word lengths $l = (l_1, l_2, \ldots)$ can arise from binary Huffman codes?

5.37 *Codes*. Which of the following codes are

(a) Uniquely decodable?

(b) Instantaneous?

$$C_1 = \{00, 01, 0\}$$
$$C_2 = \{00, 01, 100, 101, 11\}$$
$$C_3 = \{0, 10, 110, 1110, \ldots\}$$
$$C_4 = \{0, 00, 000, 0000\}$$

5.38 *Huffman*. Find the Huffman D-ary code for $(p_1, p_2, p_3, p_4, p_5, p_6) = (\frac{6}{25}, \frac{6}{25}, \frac{4}{25}, \frac{4}{25}, \frac{3}{25}, \frac{2}{25})$ and the expected word length

(a) For $D = 2$.

(b) For $D = 4$.

5.39 *Entropy of encoded bits*. Let $C : X \longrightarrow \{0, 1\}^*$ be a nonsingular but nonuniquely decodable code. Let X have entropy $H(X)$.

(a) Compare $H(C(X))$ to $H(X)$.

(b) Compare $H(C(X^n))$ to $H(X^n)$.

5.40 *Code rate*. Let X be a random variable with alphabet $\{1, 2, 3\}$ and distribution

$$X = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ 2 & \text{with probability } \frac{1}{4} \\ 3 & \text{with probability } \frac{1}{4}. \end{cases}$$

The data compression code for X assigns codewords

$$C(x) = \begin{cases} 0 & \text{if } x = 1 \\ 10 & \text{if } x = 2 \\ 11 & \text{if } x = 3. \end{cases}$$

Let $X_1, X_2, \ldots$ be independent, identically distributed according to this distribution and let $Z_1 Z_2 Z_3 \cdots = C(X_1)C(X_2) \cdots$ be the string of binary symbols resulting from concatenating the corresponding codewords. For example, 122 becomes 01010.

(a) Find the entropy rate $H(\mathcal{X})$ and the entropy rate $H(\mathcal{Z})$ in bits per symbol. Note that Z is not compressible further.

(b) Now let the code be

$$C(x) = \begin{cases} 00 & \text{if } x = 1 \\ 10 & \text{if } x = 2 \\ 01 & \text{if } x = 3 \end{cases}$$

and find the entropy rate $H(\mathcal{Z})$.

(c) Finally, let the code be

$$C(x) = \begin{cases} 00 & \text{if } x = 1 \\ 1 & \text{if } x = 2 \\ 01 & \text{if } x = 3 \end{cases}$$

and find the entropy rate $H(\mathcal{Z})$.

5.41 *Optimal codes*. Let $l_1, l_2, \ldots, l_{10}$ be the binary Huffman codeword lengths for the probabilities $p_1 \geq p_2 \geq \cdots \geq p_{10}$. Suppose that we get a new distribution by splitting the last probability

mass. What can you say about the optimal binary codeword lengths $\tilde{l}_1, \tilde{l}_2, \ldots, \tilde{l}_{11}$ for the probabilities $p_1, p_2, \ldots, p_9, \alpha p_{10}, (1-\alpha)p_{10}$, where $0 \le \alpha \le 1$.

5.42 *Ternary codes*. Which of the following codeword lengths can be the word lengths of a 3-ary Huffman code, and which cannot?

(a) $(1, 2, 2, 2, 2)$

(b) $(2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3)$

5.43 *Piecewise Huffman*. Suppose the codeword that we use to describe a random variable $X \sim p(x)$ always starts with a symbol chosen from the set $\{A, B, C\}$, followed by binary digits $\{0, 1\}$. Thus, we have a ternary code for the first symbol and binary thereafter. Give the optimal uniquely decodable code (minimum expected number of symbols) for the probability distribution

$$p = \left(\frac{16}{69}, \frac{15}{69}, \frac{12}{69}, \frac{10}{69}, \frac{8}{69}, \frac{8}{69}\right). \tag{5.160}$$

5.44 *Huffman*. Find the word lengths of the optimal binary encoding of $p = \left(\frac{1}{100}, \frac{1}{100}, \ldots, \frac{1}{100}\right)$.

5.45 *Random 20 questions*. Let X be uniformly distributed over $\{1, 2, \ldots, m\}$. Assume that $m = 2^n$. We ask random questions: Is $X \in S_1$? Is $X \in S_2$?... until only one integer remains. All 2^m subsets S of $\{1, 2, \ldots, m\}$ are equally likely to be asked.

(a) Without loss of generality, suppose that $X = 1$ is the random object. What is the probability that object 2 yields the same answers for k questions as does object 1?

(b) What is the expected number of objects in $\{2, 3, \ldots, m\}$ that have the same answers to the questions as does the correct object 1?

(c) Suppose that we ask $n + \sqrt{n}$ random questions. What is the expected number of wrong objects agreeing with the answers?

(d) Use Markov's inequality $\Pr\{X \ge t\mu\} \le \frac{1}{t}$, to show that the probability of error (one or more wrong object remaining) goes to zero as $n \longrightarrow \infty$.

HISTORICAL NOTES

The foundations for the material in this chapter can be found in Shannon's original paper [469], in which Shannon stated the source coding

Conservation law. For uniform fair odds,

$$H(\mathbf{p}) + W^*(\mathbf{p}) = \log m. \tag{6.52}$$

Side information. In a horse race X, the increase ΔW in doubling rate due to side information Y is

$$\Delta W = I(X; Y). \tag{6.53}$$

PROBLEMS

6.1 *Horse race.* Three horses run a race. A gambler offers 3-for-1 odds on each horse. 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

$$\mathbf{p} = (p_1, p_2, p_3) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right). \tag{6.54}$$

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

$$W(\mathbf{b}) = \sum_{i=1}^{3} p_i \log 3b_i. \tag{6.55}$$

(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 1.

(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 1.

6.2 *Horse race with subfair odds.* If the odds are bad (due to a track take), the gambler may wish to keep money in his pocket. Let $b(0)$ be the amount in his pocket and let $b(1), b(2), \ldots, b(m)$ be the amount bet on horses $1, 2, \ldots, m$, with odds $o(1), o(2), \ldots, o(m)$, and win probabilities $p(1), p(2), \ldots, p(m)$. Thus, the resulting wealth is $S(x) = b(0) + b(x)o(x)$, with probability $p(x)$, $x = 1, 2, \ldots, m$.

(a) Find $\mathbf{b}^*$ maximizing $E \log S$ if $\sum 1/o(i) < 1$.

(b) Discuss $\mathbf{b}^*$ if $\sum 1/o(i) > 1$. (There isn't an easy closed-form solution in this case, but a "water-filling" solution results from the application of the Kuhn–Tucker conditions.)

6.3 *Cards*. An ordinary deck of cards containing 26 red cards and 26 black cards is shuffled and dealt out one card 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 $H(X_1, X_2, \ldots, X_{52})$.

6.4 *Gambling*. Suppose that one gambles sequentially on the card outcomes in Problem 6.6.3. Even odds of 2-for-1 are paid. Thus, the wealth S_n at time n is $S_n = 2^n b(x_1, x_2, \ldots, x_n)$, where $b(x_1, x_2, \ldots, x_n)$ is the proportion of wealth bet on $x_1, x_2, \ldots, x_n$. Find $\max_{b(\cdot)} E \log S_{52}$.

6.5 *Beating the public odds*. Consider a three-horse race with win probabilities

$$(p_1, p_2, p_3) = \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right)$$

and fair odds with respect to the (false) distribution

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

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.

6.6 *Horse race*. A three-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 exponentially to zero.

(a) Find the exponent.

(b) Find the optimal gambling scheme $\mathbf{b}$ (i.e., the bet $\mathbf{b}^*$ that maximizes the exponent).

(c) Assuming that $\mathbf{b}$ is chosen as in part (b), what distribution $\mathbf{p}$ causes S_n to go to zero at the fastest rate?

6.7 *Horse race*. Consider a horse race with four horses. Assume that each horse pays 4-for-1 if it wins. Let the probabilities of winning of the horses be $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\}$. If you started with \$100 and bet optimally to maximize your long-term growth rate, what are your optimal bets on each horse? Approximately how much money would you have after 20 races with this strategy?

6.8 *Lotto*. The following analysis is a crude approximation to the games of Lotto conducted by various states. Assume that the player of the game is required to pay \$1 to play and is asked to choose one number from a range 1 to 8. At the end of every day, the state lottery commission picks a number uniformly over the same range. The jackpot (i.e., all the money collected that day) is split among all the people who chose the same number as the one chosen by the state. For example, if 100 people played today, 10 of them chose the number 2, and the drawing at the end of the day picked 2, the \$100 collected is split among the 10 people (i.e., each person who picked 2 will receive \$10, and the others will receive nothing).

The general population does not choose numbers uniformly—numbers such as 3 and 7 are supposedly lucky and are more popular than 4 or 8. Assume that the fraction of people choosing the various numbers $1, 2, \ldots, 8$ is $(f_1, f_2, \ldots, f_8)$, and assume that n people play every day. Also assume that n is very large, so that any single person's choice does not change the proportion of people betting on any number.

(a) What is the optimal strategy to divide your money among the various possible tickets so as to maximize your long-term growth rate? (Ignore the fact that you cannot buy fractional tickets.)

(b) What is the optimal growth rate that you can achieve in this game?

(c) If $(f_1, f_2, \ldots, f_8) = (\frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{4}, \frac{1}{16})$, and you start with \$1, how long will it be before you become a millionaire?

6.9 *Horse race*. Suppose that one is interested in maximizing the doubling rate for a horse race. Let $p_1, p_2, \ldots, p_m$ denote the win probabilities of the m horses. When do the odds $(o_1, o_2, \ldots, o_m)$ yield a higher doubling rate than the odds $(o'_1, o'_2, \ldots, o'_m)$?

6.10 *Horse race with probability estimates*.

(a) Three horses race. Their probabilities of winning are $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. The odds are 4-for-1, 3-for-1, and 3-for-1. Let W^* be the optimal doubling rate. Suppose you believe that the probabilities are $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. If you try to maximize the doubling rate, what doubling rate W will you achieve? By how much has your doubling rate decrease due to your poor estimate of the probabilities (i.e., what is $\Delta W = W^* - W$)?

(b) Now let the horse race be among m horses, with probabilities $p = (p_1, p_2, \ldots, p_m)$ and odds $o = (o_1, o_2, \ldots, o_m)$. If you believe the true probabilities to be $q = (q_1, q_2, \ldots, q_m)$, and try to maximize the doubling rate W, what is $W^* - W$?

6.11 *Two-envelope problem*. One envelope contains b dollars, the other $2b$ dollars. The amount b is unknown. An envelope is selected at random. Let X be the amount observed in this envelope, and let Y be the amount in the other envelope. Adopt the strategy of switching to the other envelope with probability $p(x)$, where $p(x) = \frac{e^{-x}}{(e^{-x}+e^{x})}$. Let Z be the amount that the player receives. Thus,

$$(X, Y) = \begin{cases} (b, 2b) & \text{with probability } \frac{1}{2} \\ (2b, b) & \text{with probability } \frac{1}{2} \end{cases} \tag{6.56}$$

$$Z = \begin{cases} X & \text{with probability } 1 - p(x) \\ Y & \text{with probability } p(x). \end{cases} \tag{6.57}$$

(a) Show that $E(X) = E(Y) = \frac{3b}{2}$.

(b) Show that $E(Y/X) = \frac{5}{4}$. Since the expected ratio of the amount in the other envelope is $\frac{5}{4}$, it seems that one should always switch. (This is the origin of the switching paradox.) However, observe that $E(Y) \neq E(X)E(Y/X)$. Thus, although $E(Y/X) > 1$, it does not follow that $E(Y) > E(X)$.

(c) Let J be the index of the envelope containing the maximum amount of money, and let J' be the index of the envelope chosen by the algorithm. Show that for any b, $I(J; J') > 0$. Thus, the amount in the first envelope always contains some information about which envelope to choose.

(d) Show that $E(Z) > E(X)$. Thus, you can do better than always staying or always switching. In fact, this is true for any monotonic decreasing switching function $p(x)$. By randomly switching according to $p(x)$, you are more likely to trade up than to trade down.

6.12 *Gambling*. Find the horse win probabilities $p_1, p_2, \ldots, p_m$:

(a) Maximizing the doubling rate W^* for given *fixed* known odds $o_1, o_2, \ldots, o_m$.

(b) Minimizing the doubling rate for given fixed odds $o_1, o_2, \ldots, o_m$.

6.13 *Dutch book*. Consider a horse race with $m = 2$ horses,

$$\begin{aligned} X &= 1,\ 2 \\ p &= \tfrac{1}{2},\ \tfrac{1}{2} \\ \text{odds (for one)} &= 10,\ 30 \\ \text{bets} &= b,\ 1-b. \end{aligned}$$

The odds are superfair.

(a) There is a bet b that guarantees the same payoff regardless of which horse wins. Such a bet is called a *Dutch book*. Find this b and the associated wealth factor $S(X)$.

(b) What is the maximum growth rate of the wealth for the optimal choice of b? Compare it to the growth rate for the Dutch book.

6.14 *Horse race*. Suppose that one is interested in maximizing the doubling rate for a horse race. Let $p_1, p_2, \ldots, p_m$ denote the win probabilities of the m horses. When do the odds $(o_1, o_2, \ldots, o_m)$ yield a higher doubling rate than the odds $(o_1', o_2', \ldots, o_m')$?

6.15 *Entropy of a fair horse race*. Let $X \sim p(x)$, $x = 1, 2, \ldots, m$, denote the winner of a horse race. Suppose that the odds $o(x)$ are fair with respect to $p(x)$ [i.e., $o(x) = \frac{1}{p(x)}$]. Let $b(x)$ be the amount bet on horse x, $b(x) \geq 0$, $\sum_1^m b(x) = 1$. Then the resulting wealth factor is $S(x) = b(x)o(x)$, with probability $p(x)$.

(a) Find the expected wealth $ES(X)$.

(b) Find W^*, the optimal growth rate of wealth.

(c) Suppose that

$$Y = \begin{cases} 1, & X = 1 \text{ or } 2 \\ 0, & \text{otherwise.} \end{cases}$$

If this side information is available before the bet, how much does it increase the growth rate W^*?

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

6.16 *Negative horse race*. Consider a horse race with m horses with win probabilities $p_1, p_2, \ldots, p_m$. Here the gambler hopes that a given horse will lose. He places bets $(b_1, b_2, \ldots, b_m)$, $\sum_{i=1}^{m} b_i = 1$, on the horses, loses his bet b_i if horse i wins, and retains the rest of his bets. (No odds.) Thus, $S = \sum_{j \neq i} b_j$, with probability p_i, and one wishes to maximize $\sum p_i \ln(1 - b_i)$ subject to the constraint $\sum b_i = 1$.

(a) Find the growth rate optimal investment strategy b^*. Do *not* constrain the bets to be positive, but do constrain the bets to sum to 1. (This effectively allows short selling and margin.)

(b) What is the optimal growth rate?

6.17 *St. Petersburg paradox*. Many years ago in ancient St. Petersburg the following gambling proposition caused great consternation. For an entry fee of c units, a gambler receives a payoff of 2^k units with probability 2^{-k}, $k = 1, 2, \ldots$.

(a) Show that the expected payoff for this game is infinite. For this reason, it was argued that $c = \infty$ was a "fair" price to pay to play this game. Most people find this answer absurd.

(b) Suppose that the gambler can buy a share of the game. For example, if he invests $c/2$ units in the game, he receives $\frac{1}{2}$ a share and a return $X/2$, where $\Pr(X = 2^k) = 2^{-k}$, $k = 1, 2, \ldots$. Suppose that $X_1, X_2, \ldots$ are i.i.d. according to this distribution and that the gambler reinvests all his wealth each time. Thus, his wealth S_n at time n is given by

$$S_n = \prod_{i=1}^{n} \frac{X_i}{c}. \tag{6.58}$$

Show that this limit is ∞ or 0, with probability 1, accordingly as $c < c^*$ or $c > c^*$. Identify the "fair" entry fee c^*.

More realistically, the gambler should be allowed to keep a proportion $\overline{b} = 1 - b$ of his money in his pocket and invest the rest in the St. Petersburg game. His wealth at time n is then

$$S_n = \prod_{i=1}^{n} \left(\overline{b} + \frac{bX_i}{c} \right). \tag{6.59}$$

Let

$$W(b, c) = \sum_{k=1}^{\infty} 2^{-k} \log \left(1 - b + \frac{b2^k}{c} \right). \tag{6.60}$$

We have

$$S_n \doteq 2^{nW(b,c)}. \tag{6.61}$$

Let

$$W^*(c) = \max_{0 \le b \le 1} W(b, c). \tag{6.62}$$

Here are some questions about $W^*(c)$.

(a) For what value of the entry fee c does the optimizing value b^* drop below 1?

(b) How does b^* vary with c?

(c) 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.

6.18 *Super St. Petersburg*. Finally, we have the super St. Petersburg paradox, where $\Pr(X = 2^{2^k}) = 2^{-k}, k = 1, 2, \ldots$. Here the expected log wealth is infinite for all $b > 0$, for all c, and the gambler's wealth grows to infinity faster than exponentially for any $b > 0$. But that doesn't mean that all investment ratios b are equally good. To see this, we wish to maximize the relative growth rate with respect to some other portfolio, say, $\mathbf{b} = (\frac{1}{2}, \frac{1}{2})$. Show that there exists a unique b maximizing

$$E \ln \frac{\overline{b} + bX/c}{\frac{1}{2} + \frac{1}{2}X/c}$$

and interpret the answer.

HISTORICAL NOTES

The original treatment of gambling on a horse race is due to Kelly [308], who found that $\Delta W = I$. Log-optimal portfolios go back to the work of Bernoulli, Kelly [308], Latané [346], and Latané and Tuttle [347]. Proportional gambling is sometimes referred to as the *Kelly gambling scheme*. The improvement in the probability of winning by switching envelopes in Problem 6.11 is based on Cover [130].

Shannon studied stochastic models for English in his original paper [472]. His guessing game for estimating the entropy rate of English is described in [482]. Cover and King [131] provide a gambling estimate for the entropy of English. The analysis of the St. Petersburg paradox is from Bell and Cover [39]. An alternative analysis can be found in Feller [208].

$$\left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \tag{7.153}$$

$$\left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon \Bigg\}, \tag{7.154}$$

where $p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$.

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:

1. $\Pr((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$.
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.
3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then $\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \leq 2^{-n(I(X;Y)-3\epsilon)}$.

Channel coding theorem. All rates below capacity C are achievable, and all rates above capacity are not; that is, for all rates $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with probability of error $\lambda^{(n)} \to 0$. Conversely, for rates $R > C$, $\lambda^{(n)}$ is bounded away from 0.

Feedback capacity. Feedback does not increase capacity for discrete memoryless channels (i.e., $C_{FB} = C$).

Source–channel theorem. A stochastic process with entropy rate H cannot be sent reliably over a discrete memoryless channel if $H > C$. Conversely, if the process satisfies the AEP, the source can be transmitted reliably if $H < C$.

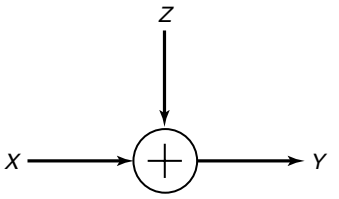
PROBLEMS

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

(a) Show that he is wrong.

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

7.2 *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.

7.3 *Channels with memory have higher capacity.* Consider a binary symmetric channel with $Y_i = X_i \oplus Z_i$, where $\oplus$ is mod 2 addition, and $X_i, Y_i \in \{0, 1\}$. Suppose that $\{Z_i\}$ has constant marginal probabilities $\Pr\{Z_i = 1\} = p = 1 - \Pr\{Z_i = 0\}$, but that $Z_1, Z_2, \ldots, Z_n$ are not necessarily independent. Assume that Z^n is independent of the input X^n. Let $C = 1 - H(p, 1 - p)$. Show that $\max_{p(x_1, x_2, \ldots, x_n)} I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) \geq nC$.

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

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

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

(a) Find the capacity.

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

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

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

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

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

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

7.7 *Cascade of binary symmetric channels.* Show that a cascade of n identical independent binary symmetric channels,

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

each with raw error probability p, is equivalent to a single BSC with error probability $\frac{1}{2}\bigl(1 - (1-2p)^n\bigr)$ 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.

7.8 *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.

7.9 *Suboptimal codes.* For the Z-channel of Problem 7.8, 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.

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

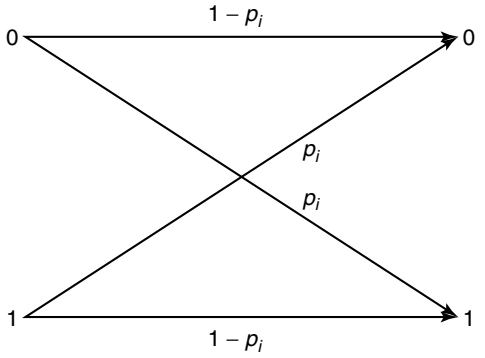
$$p(y|x) = \begin{cases} 1/2 & \text{if } y = x \pm 1 \bmod 5 \\ 0 & \text{otherwise.} \end{cases}$$

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

(b) The zero-error capacity of a channel is the number of bits per channel use that can be transmitted with zero probability of error. Clearly, the zero-error capacity of this pentagonal channel is at least 1 bit (transmit 0 or 1 with probability 1/2). Find a block code that shows that the zero-error capacity is greater than 1 bit. Can you estimate the exact value of the zero-error capacity? (*Hint*: Consider codes of length 2 for this channel.) The zero-error capacity of this channel was finally found by Lovasz [365].

7.11 *Time-varying channels.* Consider a time-varying discrete *memoryless* channel.

Let $Y_1, Y_2, \ldots, Y_n$ be conditionally independent given $X_1, X_2, \ldots, X_n$, with conditional distribution given by $p(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{n} p_i(y_i \mid x_i)$. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. Find $\max_{p(\mathbf{x})} I(\mathbf{X}; \mathbf{Y})$.



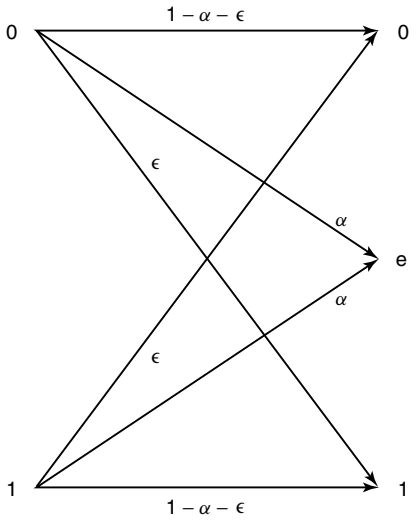
7.12 *Unused symbols.* Show that the capacity of the channel with probability transition matrix

$$P_{y|x} = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix} \tag{7.155}$$

is achieved by a distribution that places zero probability on one of input symbols. What is the capacity of this channel? Give an intuitive reason why that letter is not used.

7.13 *Erasures and errors in a binary channel.* Consider a channel with binary inputs that has both erasures and errors. Let the probability

of error be ϵ and the probability of erasure be α, so the channel is follows:



(a) Find the capacity of this channel.

(b) Specialize to the case of the binary symmetric channel ($\alpha = 0$).

(c) Specialize to the case of the binary erasure channel ($\epsilon = 0$).

7.14 *Channels with dependence between the letters.* Consider the following channel over a binary alphabet that takes in 2-bit symbols and produces a 2-bit output, as determined by the following mapping: $00 \rightarrow 01$, $01 \rightarrow 10$, $10 \rightarrow 11$, and $11 \rightarrow 00$. Thus, if the 2-bit sequence 01 is the input to the channel, the output is 10 with probability 1. Let X_1, X_2 denote the two input symbols and Y_1, Y_2 denote the corresponding output symbols.

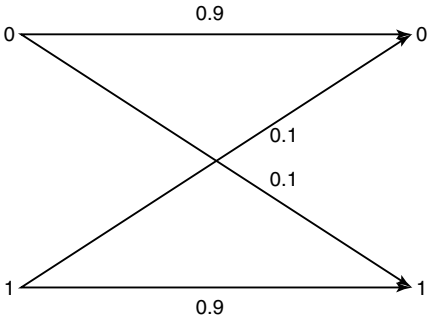
(a) Calculate the mutual information $I(X_1, X_2; Y_1, Y_2)$ as a function of the input distribution on the four possible pairs of inputs.

(b) Show that the capacity of a pair of transmissions on this channel is 2 bits.

(c) Show that under the maximizing input distribution, $I(X_1; Y_1) = 0$. Thus, the distribution on the input sequences that achieves capacity does not necessarily maximize the mutual information between individual symbols and their corresponding outputs.

7.15 *Jointly typical sequences.* As we did in Problem 3.13 for the typical set for a single random variable, we will calculate the jointly typical set for a pair of random variables connected by a binary symmetric

channel, and the probability of error for jointly typical decoding for such a channel.



We consider a binary symmetric channel with crossover probability 0.1. The input distribution that achieves capacity is the uniform distribution [i.e., $p(x) = (\frac{1}{2}, \frac{1}{2})$], which yields the joint distribution $p(x, y)$ for this channel is given by

X \ Y	0	1
0	0.45	0.05
1	0.05	0.45

The marginal distribution of Y is also $(\frac{1}{2}, \frac{1}{2})$.

(a) Calculate $H(X)$, $H(Y)$, $H(X, Y)$, and $I(X; Y)$ for the joint distribution above.

(b) Let $X_1, X_2, \ldots, X_n$ be drawn i.i.d. according the Bernoulli($\frac{1}{2}$) distribution. Of the 2^n possible input sequences of length n, which of them are typical [i.e., member of $A_\epsilon^{(n)}(X)$ for $\epsilon = 0.2$]? Which are the typical sequences in $A_\epsilon^{(n)}(Y)$?

(c) The jointly typical set $A_\epsilon^{(n)}(X, Y)$ is defined as the set of sequences that satisfy equations (7.35-7.37). The first two equations correspond to the conditions that x^n and y^n are in $A_\epsilon^{(n)}(X)$ and $A_\epsilon^{(n)}(Y)$, respectively. Consider the last condition, which can be rewritten to state that $-\frac{1}{n} \log p(x^n, y^n) \in (H(X, Y) - \epsilon, H(X, Y) + \epsilon)$. Let k be the number of places in which the sequence x^n differs from y^n (k is a function of the two sequences). Then we can write

$$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i) \tag{7.156}$$

$$= (0.45)^{n-k}(0.05)^k \tag{7.157}$$

$$= \left(\frac{1}{2}\right)^n (1-p)^{n-k} p^k. \tag{7.158}$$

An alternative way at looking at this probability is to look at the binary symmetric channel as in additive channel $Y = X \oplus Z$, where Z is a binary random variable that is equal to 1 with probability p, and is independent of X. In this case,

$$p(x^n, y^n) = p(x^n)p(y^n|x^n) \tag{7.159}$$

$$= p(x^n)p(z^n|x^n) \tag{7.160}$$

$$= p(x^n)p(z^n) \tag{7.161}$$

$$= \left(\frac{1}{2}\right)^n (1-p)^{n-k} p^k. \tag{7.162}$$

Show that the condition that (x^n, y^n) being jointly typical is equivalent to the condition that x^n is typical and $z^n = y^n - x^n$ is typical.

(d) We now calculate the size of $A_\epsilon^{(n)}(Z)$ for $n = 25$ and $\epsilon = 0.2$. As in Problem 3.13, here is a table of the probabilities and numbers of sequences with k ones:

k	$\binom{n}{k}$	$\binom{n}{k} p^k (1-p)^{n-k}$	$-\frac{1}{n}\log p(x^n)$
0	1	0.071790	0.152003
1	25	0.199416	0.278800
2	300	0.265888	0.405597
3	2300	0.226497	0.532394
4	12650	0.138415	0.659191
5	53130	0.064594	0.785988
6	177100	0.023924	0.912785
7	480700	0.007215	1.039582
8	1081575	0.001804	1.166379
9	2042975	0.000379	1.293176
10	3268760	0.000067	1.419973
11	4457400	0.000010	1.546770
12	5200300	0.000001	1.673567

[Sequences with more than 12 ones are omitted since their total probability is negligible (and they are not in the typical set).] What is the size of the set $A_\epsilon^{(n)}(Z)$?

(e) Now consider random coding for the channel, as in the proof of the channel coding theorem. Assume that 2^{nR} codewords $X^n(1)$, $X^n(2)$,..., $X^n(2^{nR})$ are chosen uniformly over the 2^n possible binary sequences of length n. One of these codewords is chosen and sent over the channel. The receiver looks at the received sequence and tries to find a codeword in the code that is jointly typical with the received sequence. As argued above, this corresponds to finding a codeword $X^n(i)$ such that $Y^n - X^n(i) \in A_\epsilon^{(n)}(Z)$. For a fixed codeword $x^n(i)$, what is the probability that the received sequence Y^n is such that $(x^n(i), Y^n)$ is jointly typical?

(f) Now consider a particular received sequence $y^n = 000000\ldots0$, say. Assume that we choose a sequence X^n at random, uniformly distributed among all the 2^n possible binary n-sequences. What is the probability that the chosen sequence is jointly typical with this y^n? [*Hint*: This is the probability of all sequences x^n such that $y^n - x^n \in A_\epsilon^{(n)}(Z)$.]

(g) Now consider a code with $2^9 = 512$ codewords of length 12 chosen at random, uniformly distributed among all the 2^n sequences of length $n = 25$. One of these codewords, say the one corresponding to $i = 1$, is chosen and sent over the channel. As calculated in part (e), the received sequence, with high probability, is jointly typical with the codeword that was sent. What is the probability that one or more of the other codewords (which were chosen at random, independent of the sent codeword) is jointly typical with the received sequence? [*Hint*: You could use the union bound, but you could also calculate this probability exactly, using the result of part (f) and the independence of the codewords.]

(h) Given that a particular codeword was sent, the probability of error (averaged over the probability distribution of the channel and over the random choice of other codewords) can be written as

$$\Pr(\text{Error}|x^n(1) \text{ sent}) = \sum_{y^n : y^n \text{causes error}} p(y^n|x^n(1)). \quad (7.163)$$

There are two kinds of error: the first occurs if the received sequence y^n is not jointly typical with the transmitted codeword, and the second occurs if there is another codeword jointly typical with the received sequence. Using the result of the preceding parts, calculate this probability of error. By

the symmetry of the random coding argument, this does not depend on which codeword was sent.

The calculations above show that average probability of error for a random code with 512 codewords of length 25 over the binary symmetric channel of crossover probability 0.1 is about 0.34. This seems quite high, but the reason for this is that the value of ϵ that we have chosen is too large. By choosing a smaller ϵ and a larger n in the definitions of $A_\epsilon^{(n)}$, we can get the probability of error to be as small as we want as long as the rate of the code is less than $I(X; Y) - 3\epsilon$.

Also note that the decoding procedure described in the problem is not optimal. The optimal decoding procedure is maximum likelihood (i.e., to choose the codeword that is closest to the received sequence). It is possible to calculate the average probability of error for a random code for which the decoding is based on an approximation to maximum likelihood decoding, where we decode a received sequence to the unique codeword that differs from the received sequence in ≤ 4 bits, and declare an error otherwise. The only difference with the jointly typical decoding described above is that in the case when the codeword is equal to the received sequence! The average probability of error for this decoding scheme can be shown to be about 0.285.

7.16 *Encoder and decoder as part of the channel.* Consider a binary symmetric channel with crossover probability 0.1. A possible coding scheme for this channel with two codewords of length 3 is to encode message a_1 as 000 and a_2 as 111. With this coding scheme, we can consider the combination of encoder, channel, and decoder as forming a new BSC, with two inputs a_1 and a_2 and two outputs a_1 and a_2.

(a) Calculate the crossover probability of this channel.

(b) What is the capacity of this channel in bits per transmission of the original channel?

(c) What is the capacity of the original BSC with crossover probability 0.1?

(d) Prove a general result that for any channel, considering the encoder, channel, and decoder together as a new channel from messages to estimated messages will not increase the capacity in bits per transmission of the original channel.

7.17 *Codes of length 3 for a BSC and BEC.* In Problem 7.16, the probability of error was calculated for a code with two codewords of

length 3 (000 and 111) sent over a binary symmetric channel with crossover probability ϵ. For this problem, take $\epsilon = 0.1$.

(a) Find the best code of length 3 with four codewords for this channel. What is the probability of error for this code? (Note that all possible received sequences should be mapped onto possible codewords.)

(b) What is the probability of error if we used all eight possible sequences of length 3 as codewords?

(c) Now consider a binary erasure channel with erasure probability 0.1. Again, if we used the two-codeword code 000 and 111, received sequences 00E, 0E0, E00, 0EE, E0E, EE0 would all be decoded as 0, and similarly, we would decode 11E, 1E1, E11, 1EE, E1E, EE1 as 1. If we received the sequence EEE, we would not know if it was a 000 or a 111 that was sent—so we choose one of these two at random, and are wrong half the time. What is the probability of error for this code over the erasure channel?

(d) What is the probability of error for the codes of parts (a) and (b) when used over the binary erasure channel?

7.18 *Channel capacity.* Calculate the capacity of the following channels with probability transition matrices:

(a) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \tag{7.164}$$

(b) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$

$$p(y|x) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \tag{7.165}$$

(c) $\mathcal{X} = \mathcal{Y} = \{0, 1, 2, 3\}$

$$p(y|x) = \begin{bmatrix} p & 1-p & 0 & 0 \\ 1-p & p & 0 & 0 \\ 0 & 0 & q & 1-q \\ 0 & 0 & 1-q & q \end{bmatrix} \tag{7.166}$$

7.19 *Capacity of the carrier pigeon channel.* Consider a commander of an army besieged in a fort for whom the only means of communication to his allies is a set of carrier pigeons. Assume that each carrier pigeon can carry one letter (8 bits), that pigeons are released once every 5 minutes, and that each pigeon takes exactly 3 minutes to reach its destination.

(a) Assuming that all the pigeons reach safely, what is the capacity of this link in bits/hour?

(b) Now assume that the enemies try to shoot down the pigeons and that they manage to hit a fraction α of them. Since the pigeons are sent at a constant rate, the receiver knows when the pigeons are missing. What is the capacity of this link?

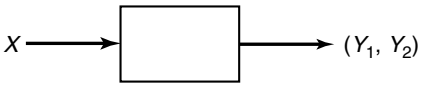
(c) Now assume that the enemy is more cunning and that every time they shoot down a pigeon, they send out a dummy pigeon carrying a random letter (chosen uniformly from all 8-bit letters). What is the capacity of this link in bits/hour?

Set up an appropriate model for the channel in each of the above cases, and indicate how to go about finding the capacity.

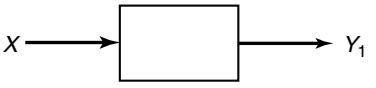
7.20 *Channel with two independent looks at Y.* Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X.

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

(b) Conclude that the capacity of the channel



is less than twice the capacity of the channel



7.21 *Tall, fat people*. Suppose that the average height of people in a room is 5 feet. Suppose that the average weight is 100 lb.

(a) Argue that no more than one-third of the population is 15 feet tall.

(b) Find an upper bound on the fraction of 300-lb 10-footers in the room.

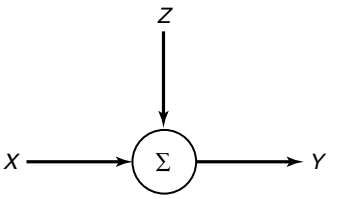
7.22 *Can signal alternatives lower capacity?* Show that adding a row to a channel transition matrix does not decrease capacity.

7.23 *Binary multiplier channel*

(a) Consider the channel $Y = XZ$, where X and Z are independent binary random variables that take on values 0 and 1. Z is Bernoulli(α) [i.e., $P(Z = 1) = \alpha$]. Find the capacity of this channel and the maximizing distribution on X.

(b) Now suppose that the receiver can observe Z as well as Y. What is the capacity?

7.24 *Noise alphabets.* Consider the channel

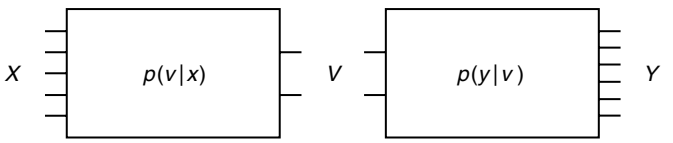


$\mathcal{X} = \{0, 1, 2, 3\}$, where $Y = X + Z$, and Z is uniformly distributed over three distinct integer values $\mathcal{Z} = \{z_1, z_2, z_3\}$.

(a) What is the maximum capacity over all choices of the $\mathcal{Z}$ alphabet? Give distinct integer values z_1, z_2, z_3 and a distribution on $\mathcal{X}$ achieving this.

(b) What is the minimum capacity over all choices for the $\mathcal{Z}$ alphabet? Give distinct integer values z_1, z_2, z_3 and a distribution on $\mathcal{X}$ achieving this.

7.25 *Bottleneck channel.* Suppose that a signal $X \in \mathcal{X} = \{1, 2, \ldots, m\}$ goes through an intervening transition $X \longrightarrow V \longrightarrow Y$:



where $x = \{1, 2, \ldots, m\}$, $y = \{1, 2, \ldots, m\}$, and $v = \{1, 2, \ldots, k\}$. Here $p(v|x)$ and $p(y|v)$ are arbitrary and the channel has transition probability $p(y|x) = \sum_v p(v|x)p(y|v)$. Show that $C \leq \log k$.

7.26 *Noisy typewriter.* Consider the channel with $x, y \in \{0, 1, 2, 3\}$ and transition probabilities $p(y|x)$ given by the following matrix:

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{bmatrix}$$

(a) Find the capacity of this channel.

(b) Define the random variable $z = g(y)$, where

$$g(y) = \begin{cases} A & \text{if } y \in \{0, 1\} \\ B & \text{if } y \in \{2, 3\}. \end{cases}$$

For the following two PMFs for x, compute $I(X; Z)$:

(i)

$$p(x) = \begin{cases} \frac{1}{2} & \text{if } x \in \{1, 3\} \\ 0 & \text{if } x \in \{0, 2\}. \end{cases}$$

(ii)

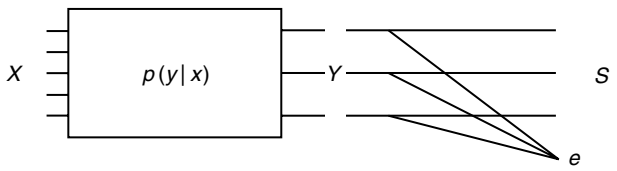
$$p(x) = \begin{cases} 0 & \text{if } x \in \{1, 3\} \\ \frac{1}{2} & \text{if } x \in \{0, 2\}. \end{cases}$$

(c) Find the capacity of the channel between x and z, specifically where $x \in \{0, 1, 2, 3\}$, $z \in \{A, B\}$, and the transition probabilities $P(z|x)$ are given by

$$p(Z = z|X = x) = \sum_{g(y_0)=z} P(Y = y_0|X = x).$$

(d) For the X distribution of part (i) of (b), does $X \to Z \to Y$ form a Markov chain?

7.27 *Erasure channel.* Let $\{\mathcal{X}, p(y|x), \mathcal{Y}\}$ be a discrete memoryless channel with capacity C. Suppose that this channel is cascaded immediately with an erasure channel $\{\mathcal{Y}, p(s|y), \mathcal{S}\}$ that erases α of its symbols.



Specifically, $\mathcal{S} = \{y_1, y_2, \ldots, y_m, e\}$, and

$$\Pr\{S = y|X = x\} = \overline{\alpha}p(y|x), \qquad y \in \mathcal{Y},$$
$$\Pr\{S = e|X = x\} = \alpha.$$

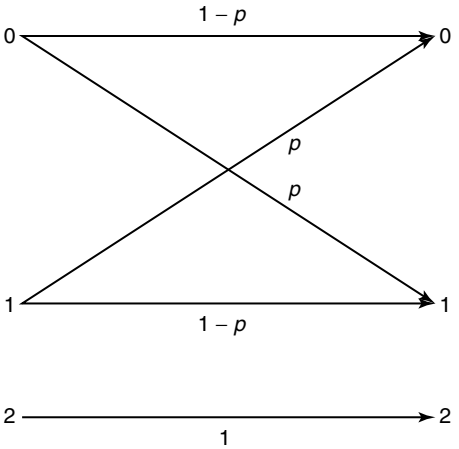
Determine the capacity of this channel.

7.28 *Choice of channels.* Find the capacity C of the union of two channels $(\mathcal{X}_1, p_1(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p_2(y_2|x_2), \mathcal{Y}_2)$, where at each time, one can send a symbol over channel 1 or channel 2 but not both. Assume that the output alphabets are distinct and do not intersect.

(a) Show that $2^C = 2^{C_1} + 2^{C_2}$. Thus, 2^C is the effective alphabet size of a channel with capacity C.

(b) Compare with Problem 2.10 where $2^H = 2^{H_1} + 2^{H_2}$, and interpret part (a) in terms of the effective number of noise-free symbols.

(c) Use the above result to calculate the capacity of the following channel.

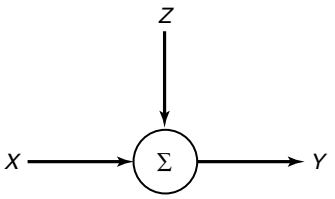


7.29 *Binary multiplier channel*

(a) Consider the discrete memoryless channel $Y = XZ$, where X and Z are independent binary random variables that take on values 0 and 1. Let $P(Z = 1) = \alpha$. Find the capacity of this channel and the maximizing distribution on X.

(b) Now suppose that the receiver can observe Z as well as Y. What is the capacity?

7.30 *Noise alphabets.* Consider the channel

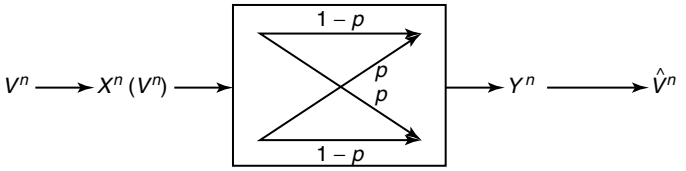


$\mathcal{X} = \{0, 1, 2, 3\}$, where $Y = X + Z$, and Z is uniformly distributed over three distinct integer values $\mathcal{Z} = \{z_1, z_2, z_3\}$.

(a) What is the maximum capacity over all choices of the $\mathcal{Z}$ alphabet? Give distinct integer values z_1, z_2, z_3 and a distribution on $\mathcal{X}$ achieving this.

(b) What is the minimum capacity over all choices for the $\mathcal{Z}$ alphabet? Give distinct integer values z_1, z_2, z_3 and a distribution on $\mathcal{X}$ achieving this.

7.31 *Source and channel.* We wish to encode a Bernoulli(α) process $V_1, V_2, \ldots$ for transmission over a binary symmetric channel with crossover probability p.



Find conditions on α and p so that the probability of error $P(\hat{V}^n \neq V^n)$ can be made to go to zero as $n \longrightarrow \infty$.

7.32 *Random 20 questions.* Let X be uniformly distributed over $\{1, 2, \ldots, m\}$. Assume that $m = 2^n$. We ask random questions: Is $X \in S_1$? Is $X \in S_2$? ... until only one integer remains. All 2^m subsets S of $\{1, 2, \ldots, m\}$ are equally likely.

(a) How many deterministic questions are needed to determine X?

(b) Without loss of generality, suppose that $X = 1$ is the random object. What is the probability that object 2 yields the same answers as object 1 for k questions?

(c) What is the expected number of objects in $\{2, 3, \ldots, m\}$ that have the same answers to the questions as those of the correct object 1?

(d) Suppose that we ask $n + \sqrt{n}$ *random* questions. What is the expected number of wrong objects agreeing with the answers?

(e) Use Markov's inequality $\Pr\{X \geq t\mu\} \leq \frac{1}{t}$, to show that the probability of error (one or more wrong object remaining) goes to zero as $n \longrightarrow \infty$.

7.33 *BSC with feedback.* Suppose that feedback is used on a binary symmetric channel with parameter p. Each time a Y is received, it becomes the next transmission. Thus, X_1 is Bern($\frac{1}{2}$), $X_2 = Y_1$, $X_3 = Y_2, \ldots, X_n = Y_{n-1}$.

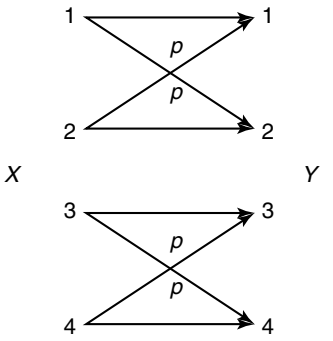
(a) Find $\lim_{n\to\infty} \frac{1}{n} I(X^n; Y^n)$.

(b) Show that for some values of p, this can be higher than capacity.

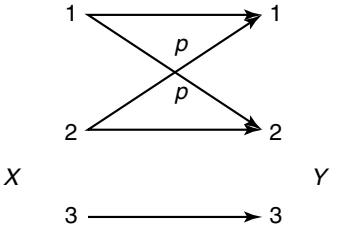
(c) Using this feedback transmission scheme, $X^n(W, Y^n) = (X_1(W), Y_1, Y_2, \ldots, Y_{m-1})$, what is the asymptotic communication rate achieved; that is, what is $\lim_{n\to\infty} \frac{1}{n} I(W; Y^n)$?

7.34 *Capacity.* Find the capacity of

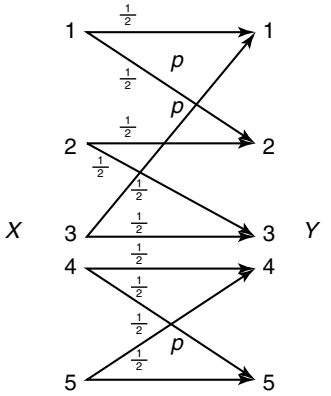
(a) Two parallel BSCs:



(b) BSC and a single symbol:



(c) BSC and a ternary channel:



(d) Ternary channel:

$$p(y|x) = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix}. \tag{7.167}$$

7.35 *Capacity.* Suppose that channel $\mathcal{P}$ has capacity C, where $\mathcal{P}$ is an $m \times n$ channel matrix.

(a) What is the capacity of

$$\tilde{\mathcal{P}} = \begin{bmatrix} \mathcal{P} & 0 \\ 0 & 1 \end{bmatrix}?$$

(b) What about the capacity of

$$\hat{\mathcal{P}} = \begin{bmatrix} \mathcal{P} & 0 \\ 0 & I_k \end{bmatrix}?$$

where I_k if the $k \times k$ identity matrix.

7.36 *Channel with memory.* Consider the discrete memoryless channel $Y_i = Z_i X_i$ with input alphabet $X_i \in \{-1, 1\}$.

(a) What is the capacity of this channel when $\{Z_i\}$ is i.i.d. with

$$Z_i = \begin{cases} 1, & p = 0.5 \\ -1, & p = 0.5? \end{cases} \tag{7.168}$$

Now consider the channel with memory. Before transmission begins, Z is randomly chosen and fixed for all time. Thus, $Y_i = ZX_i$.

(b) What is the capacity if

$$Z = \begin{cases} 1, & p = 0.5 \\ -1, & p = 0.5? \end{cases} \tag{7.169}$$

7.37 *Joint typicality.* Let (X_i, Y_i, Z_i) be i.i.d. according to $p(x, y, z)$. We will say that (x^n, y^n, z^n) is jointly typical [written $(x^n, y^n, z^n) \in A_\epsilon^{(n)}$] if

- $p(x^n) \in 2^{-n(H(X)\pm\epsilon)}$.
- $p(y^n) \in 2^{-n(H(Y)\pm\epsilon)}$.
- $p(z^n) \in 2^{-n(H(Z)\pm\epsilon)}$.
- $p(x^n, y^n) \in 2^{-n(H(X,Y)\pm\epsilon)}$.
- $p(x^n, z^n) \in 2^{-n(H(X,Z)\pm\epsilon)}$.
- $p(y^n, z^n) \in 2^{-n(H(Y,Z)\pm\epsilon)}$.
- $p(x^n, y^n, z^n) \in 2^{-n(H(X,Y,Z)\pm\epsilon)}$.

Now suppose that $(\tilde{X}^n, \tilde{Y}^n, \tilde{Z}^n)$ is drawn according to $p(x^n)p(y^n)p(z^n)$. Thus, $\tilde{X}^n, \tilde{Y}^n, \tilde{Z}^n$ have the same marginals as $p(x^n, y^n, z^n)$ but are independent. Find (bounds on) $\Pr\{(\tilde{X}^n,\tilde{Y}^n,\tilde{Z}^n) \in A_\epsilon^{(n)}\}$ in terms of the entropies $H(X)$,$H(Y)$,$H(Z)$,$H(X, Y)$,$H(X, Z)$, $H(Y, Z)$, and $H(X, Y, Z)$.

HISTORICAL NOTES

The idea of mutual information and its relationship to channel capacity was developed by Shannon in his original paper [472]. In this paper, he stated the channel capacity theorem and outlined the proof using typical sequences in an argument similar to the one described here. The first rigorous proof was due to Feinstein [205], who used a painstaking "cookie-cutting" argument to find the number of codewords that can be sent with a low probability of error. A simpler proof using a random coding exponent was developed by Gallager [224]. Our proof is based on Cover [121] and on Forney's unpublished course notes [216].

The converse was proved by Fano [201], who used the inequality bearing his name. The strong converse was first proved by Wolfowitz [565], using techniques that are closely related to typical sequences. An iterative algorithm to calculate the channel capacity was developed independently by Arimoto [25] and Blahut [65].

The idea of the zero-error capacity was developed by Shannon [474]; in the same paper, he also proved that feedback does not increase the capacity of a discrete memoryless channel. The problem of finding the zero-error capacity is essentially combinatorial; the first important result in this area is due to Lovasz [365]. The general problem of finding the zero error capacity is still open; see a survey of related results in Körner and Orlitsky [327].

Quantum information theory, the quantum mechanical counterpart to the classical theory in this chapter, is emerging as a large research area in its own right and is well surveyed in an article by Bennett and Shor [49] and in the text by Nielsen and Chuang [395].

CHAPTER 8

DIFFERENTIAL ENTROPY

We now introduce the concept of *differential entropy*, which is the entropy of a continuous random variable. Differential entropy is also related to the shortest description length and is similar in many ways to the entropy of a discrete random variable. But there are some important differences, and there is need for some care in using the concept.

8.1 DEFINITIONS

Definition Let X be a random variable with cumulative distribution function $F(x) = \Pr(X \leq x)$. If $F(x)$ is continuous, the random variable is said to be continuous. Let $f(x) = F'(x)$ when the derivative is defined. If $\int_{-\infty}^{\infty} f(x) = 1$, $f(x)$ is called the *probability density function* for X. The set where $f(x) > 0$ is called the *support set* of X.

Definition The *differential entropy* $h(X)$ of a continuous random variable X with density $f(x)$ is defined as

$$h(X) = -\int_S f(x) \log f(x)\, dx, \tag{8.1}$$

where S is the support set of the random variable.

As in the discrete case, the differential entropy depends only on the probability density of the random variable, and therefore the differential entropy is sometimes written as $h(f)$ rather than $h(X)$.

Remark As in every example involving an integral, or even a density, we should include the statement *if it exists*. It is easy to construct examples

Elements of Information Theory, Second Edition, By Thomas M. Cover and Joy A. Thomas

of random variables for which a density function does not exist or for which the above integral does not exist.

Example 8.1.1 (*Uniform distribution*) Consider a random variable distributed uniformly from 0 to a so that its density is $1/a$ from 0 to a and 0 elsewhere. Then its differential entropy is

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a}\, dx = \log a. \tag{8.2}$$

Note: For $a < 1$, $\log a < 0$, and the differential entropy is negative. Hence, unlike discrete entropy, differential entropy can be negative. However, $2^{h(X)} = 2^{\log a} = a$ is the volume of the support set, which is always nonnegative, as we expect.

Example 8.1.2 (*Normal distribution*) Let $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}$. Then calculating the differential entropy in nats, we obtain

$$h(\phi) = -\int \phi \ln \phi \tag{8.3}$$

$$= -\int \phi(x) \left[-\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right] \tag{8.4}$$

$$= \frac{EX^2}{2\sigma^2} + \frac{1}{2} \ln 2\pi\sigma^2 \tag{8.5}$$

$$= \frac{1}{2} + \frac{1}{2} \ln 2\pi\sigma^2 \tag{8.6}$$

$$= \frac{1}{2} \ln e + \frac{1}{2} \ln 2\pi\sigma^2 \tag{8.7}$$

$$= \frac{1}{2} \ln 2\pi e\sigma^2 \qquad \text{nats.} \tag{8.8}$$

Changing the base of the logarithm, we have

$$h(\phi) = \frac{1}{2} \log 2\pi e\sigma^2 \qquad \text{bits.} \tag{8.9}$$

8.2 AEP FOR CONTINUOUS RANDOM VARIABLES

One of the important roles of the entropy for discrete random variables is in the AEP, which states that for a sequence of i.i.d. random variables, $p(X_1, X_2, \ldots, X_n)$ is close to $2^{-nH(X)}$ with high probability. This enables us to define the typical set and characterize the behavior of typical sequences.

We can do the same for a continuous random variable.

Theorem 8.2.1 *Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables drawn i.i.d. according to the density $f(x)$. Then*

$$-\frac{1}{n}\log f(X_1, X_2, \ldots, X_n) \to E[-\log f(X)] = h(X) \quad \textit{in probability}. \tag{8.10}$$

Proof: The proof follows directly from the weak law of large numbers. □

This leads to the following definition of the typical set.

Definition For $\epsilon > 0$ and any n, we define the *typical set* $A_\epsilon^{(n)}$ with respect to $f(x)$ as follows:

$$A_\epsilon^{(n)} = \left\{(x_1, x_2, \ldots, x_n) \in S^n : \left|-\frac{1}{n}\log f(x_1, x_2, \ldots, x_n) - h(X)\right| \le \epsilon\right\}, \tag{8.11}$$

where $f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n f(x_i)$.

The properties of the typical set for continuous random variables parallel those for discrete random variables. The analog of the cardinality of the typical set for the discrete case is the volume of the typical set for continuous random variables.

Definition The *volume* $\text{Vol}(A)$ of a set $A \subset \mathcal{R}^n$ is defined as

$$\text{Vol}(A) = \int_A dx_1\, dx_2 \cdots dx_n. \tag{8.12}$$

Theorem 8.2.2 *The typical set $A_\epsilon^{(n)}$ has the following properties:*

1. $\Pr\left(A_\epsilon^{(n)}\right) > 1 - \epsilon$ *for n sufficiently large.*
2. $\text{Vol}\left(A_\epsilon^{(n)}\right) \le 2^{n(h(X)+\epsilon)}$ *for all n.*
3. $\text{Vol}\left(A_\epsilon^{(n)}\right) \ge (1-\epsilon)2^{n(h(X)-\epsilon)}$ *for n sufficiently large.*

Proof: By Theorem 8.2.1, $-\frac{1}{n}\log f(X^n) = -\frac{1}{n}\sum \log f(X_i) \to h(X)$ in probability, establishing property 1. Also,

$$1 = \int_{S^n} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \tag{8.13}$$

$$\geq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \tag{8.14}$$

$$\geq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)+\epsilon)}\, dx_1\, dx_2 \cdots dx_n \tag{8.15}$$

$$= 2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} dx_1\, dx_2 \cdots dx_n \tag{8.16}$$

$$= 2^{-n(h(X)+\epsilon)}\, \mathrm{Vol}\left(A_\epsilon^{(n)}\right). \tag{8.17}$$

Hence we have property 2. We argue further that the volume of the typical set is at least this large. If n is sufficiently large so that property 1 is satisfied, then

$$1 - \epsilon \leq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \tag{8.18}$$

$$\leq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)-\epsilon)}\, dx_1\, dx_2 \cdots dx_n \tag{8.19}$$

$$= 2^{-n(h(X)-\epsilon)} \int_{A_\epsilon^{(n)}} dx_1\, dx_2 \cdots dx_n \tag{8.20}$$

$$= 2^{-n(h(X)-\epsilon)}\, \mathrm{Vol}\left(A_\epsilon^{(n)}\right), \tag{8.21}$$

establishing property 3. Thus for n sufficiently large, we have

$$(1-\epsilon)2^{n(h(X)-\epsilon)} \leq \mathrm{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}. \qquad \square \tag{8.22}$$

Theorem 8.2.3 *The set $A_\epsilon^{(n)}$ is the smallest volume set with probability $\geq 1-\epsilon$, to first order in the exponent.*

Proof: Same as in the discrete case. $\square$

This theorem indicates that the volume of the smallest set that contains most of the probability is approximately 2^{nh}. This is an n-dimensional volume, so the corresponding side length is $(2^{nh})^{\frac{1}{n}} = 2^h$. This provides

an interpretation of the differential entropy: It is the logarithm of the equivalent side length of the smallest set that contains most of the probability. Hence low entropy implies that the random variable is confined to a small effective volume and high entropy indicates that the random variable is widely dispersed.

Note. Just as the entropy is related to the volume of the typical set, there is a quantity called Fisher information which is related to the surface area of the typical set. We discuss Fisher information in more detail in Sections 11.10 and 17.8.

8.3 RELATION OF DIFFERENTIAL ENTROPY TO DISCRETE ENTROPY

Consider a random variable X with density $f(x)$ illustrated in Figure 8.1. Suppose that we divide the range of X into bins of length Δ. Let us assume that the density is continuous within the bins. Then, by the mean value theorem, there exists a value x_i within each bin such that

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx. \tag{8.23}$$

Consider the quantized random variable X^Δ, which is defined by

$$X^\Delta = x_i \qquad \text{if } i\Delta \le X < (i+1)\Delta. \tag{8.24}$$

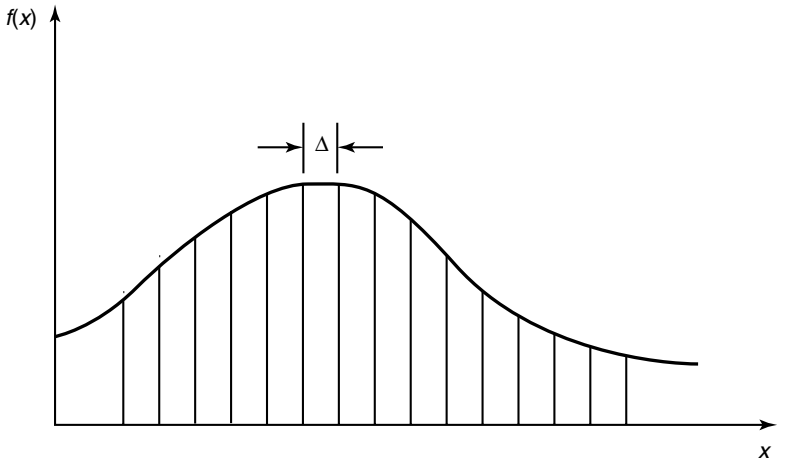


FIGURE 8.1. Quantization of a continuous random variable.

Then the probability that $X^\Delta = x_i$ is

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx = f(x_i)\Delta. \tag{8.25}$$

The entropy of the quantized version is

$$H(X^\Delta) = -\sum_{-\infty}^{\infty} p_i \log p_i \tag{8.26}$$

$$= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta) \tag{8.27}$$

$$= -\sum \Delta f(x_i) \log f(x_i) - \sum f(x_i)\Delta \log \Delta \tag{8.28}$$

$$= -\sum \Delta f(x_i) \log f(x_i) - \log \Delta, \tag{8.29}$$

since $\sum f(x_i)\Delta = \int f(x) = 1$. If $f(x)\log f(x)$ is Riemann integrable (a condition to ensure that the limit is well defined [556]), the first term in (8.29) approaches the integral of $-f(x)\log f(x)$ as $\Delta \to 0$ by definition of Riemann integrability. This proves the following.

Theorem 8.3.1 *If the density $f(x)$ of the random variable X is Riemann integrable, then*

$$H(X^\Delta) + \log \Delta \to h(f) = h(X), \quad \text{as } \Delta \to 0. \tag{8.30}$$

Thus, the entropy of an n-bit quantization of a continuous random variable X is approximately $h(X) + n$.

Example 8.3.1

1. If X has a uniform distribution on $[0, 1]$ and we let $\Delta = 2^{-n}$, then $h = 0$, $H(X^\Delta) = n$, and n bits suffice to describe X to n bit accuracy.
2. If X is uniformly distributed on $[0, \frac{1}{8}]$, the first 3 bits to the right of the decimal point must be 0. To describe X to n-bit accuracy requires only $n - 3$ bits, which agrees with $h(X) = -3$.
3. If $X \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 100$, describing X to n bit accuracy would require on the average $n + \frac{1}{2}\log(2\pi e\sigma^2) = n + 5.37$ bits.

In general, $h(X) + n$ is the number of bits *on the average* required to describe X to n-bit accuracy.

The differential entropy of a discrete random variable can be considered to be $-\infty$. Note that $2^{-\infty} = 0$, agreeing with the idea that the volume of the support set of a discrete random variable is zero.

8.4 JOINT AND CONDITIONAL DIFFERENTIAL ENTROPY

As in the discrete case, we can extend the definition of differential entropy of a single random variable to several random variables.

Definition The *differential entropy* of a set $X_1, X_2, \ldots, X_n$ of random variables with density $f(x_1, x_2, \ldots, x_n)$ is defined as

$$h(X_1, X_2, \ldots, X_n) = -\int f(x^n) \log f(x^n)\, dx^n. \tag{8.31}$$

Definition If X, Y have a joint density function $f(x, y)$, we can define the conditional differential entropy $h(X|Y)$ as

$$h(X|Y) = -\int f(x, y) \log f(x|y)\, dx\, dy. \tag{8.32}$$

Since in general $f(x|y) = f(x, y)/f(y)$, we can also write

$$h(X|Y) = h(X, Y) - h(Y). \tag{8.33}$$

But we must be careful if any of the differential entropies are infinite.

The next entropy evaluation is used frequently in the text.

Theorem 8.4.1 (*Entropy of a multivariate normal distribution*) *Let $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with mean μ and covariance matrix K. Then*

$$h(X_1, X_2, \ldots, X_n) = h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log(2\pi e)^n |K| \quad \textit{bits}, \tag{8.34}$$

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

Proof: The probability density function of $X_1, X_2, \ldots, X_n$ is

$$f(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n |K|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)}. \tag{8.35}$$

Then

$$h(f) = -\int f(\mathbf{x}) \left[-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu) - \ln\left(\sqrt{2\pi}\right)^n |K|^{\frac{1}{2}} \right] d\mathbf{x} \tag{8.36}$$

$$= \frac{1}{2} E\left[\sum_{i,j} (X_i - \mu_i) \left(K^{-1}\right)_{ij} (X_j - \mu_j) \right] + \frac{1}{2} \ln(2\pi)^n |K| \tag{8.37}$$

$$= \frac{1}{2} E\left[\sum_{i,j} (X_i - \mu_i)(X_j - \mu_j) \left(K^{-1}\right)_{ij} \right] + \frac{1}{2} \ln(2\pi)^n |K| \tag{8.38}$$

$$= \frac{1}{2} \sum_{i,j} E[(X_j - \mu_j)(X_i - \mu_i)] \left(K^{-1}\right)_{ij} + \frac{1}{2} \ln(2\pi)^n |K| \tag{8.39}$$

$$= \frac{1}{2} \sum_{j} \sum_{i} K_{ji} \left(K^{-1}\right)_{ij} + \frac{1}{2} \ln(2\pi)^n |K| \tag{8.40}$$

$$= \frac{1}{2} \sum_{j} (K K^{-1})_{jj} + \frac{1}{2} \ln(2\pi)^n |K| \tag{8.41}$$

$$= \frac{1}{2} \sum_{j} I_{jj} + \frac{1}{2} \ln(2\pi)^n |K| \tag{8.42}$$

$$= \frac{n}{2} + \frac{1}{2} \ln(2\pi)^n |K| \tag{8.43}$$

$$= \frac{1}{2} \ln(2\pi e)^n |K| \qquad \text{nats} \tag{8.44}$$

$$= \frac{1}{2} \log(2\pi e)^n |K| \qquad \text{bits.} \qquad \square \tag{8.45}$$

8.5 RELATIVE ENTROPY AND MUTUAL INFORMATION

We now extend the definition of two familiar quantities, $D(f||g)$ and $I(X;Y)$, to probability densities.

Definition The *relative entropy* (or *Kullback–Leibler distance*) $D(f||g)$ between two densities f and g is defined by

$$D(f||g) = \int f \log \frac{f}{g}. \tag{8.46}$$

Note that $D(f||g)$ is finite only if the support set of f is contained in the support set of g. [Motivated by continuity, we set $0 \log \frac{0}{0} = 0$.]

Definition The *mutual information* $I(X; Y)$ between two random variables with joint density $f(x, y)$ is defined as

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x) f(y)} \, dx \, dy. \tag{8.47}$$

From the definition it is clear that

$$I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X, Y) \tag{8.48}$$

and

$$I(X; Y) = D(f(x, y)||f(x) f(y)). \tag{8.49}$$

The properties of $D(f||g)$ and $I(X; Y)$ are the same as in the discrete case. In particular, the mutual information between two random variables is the limit of the mutual information between their quantized versions, since

$$I(X^\Delta; Y^\Delta) = H(X^\Delta) - H(X^\Delta|Y^\Delta) \tag{8.50}$$

$$\approx h(X) - \log \Delta - (h(X|Y) - \log \Delta) \tag{8.51}$$

$$= I(X; Y). \tag{8.52}$$

More generally, we can define mutual information in terms of finite partitions of the range of the random variable. Let $\mathcal{X}$ be the range of a random variable X. A partition $\mathcal{P}$ of $\mathcal{X}$ is a finite collection of disjoint sets P_i such that $\cup_i P_i = \mathcal{X}$. The quantization of X by $\mathcal{P}$ (denoted $[X]_\mathcal{P}$) is the discrete random variable defined by

$$\Pr([X]_\mathcal{P} = i) = \Pr(X \in P_i) = \int_{P_i} dF(x). \tag{8.53}$$

For two random variables X and Y with partitions $\mathcal{P}$ and $\mathcal{Q}$, we can calculate the mutual information between the quantized versions of X and Y using (2.28). Mutual information can now be defined for arbitrary pairs of random variables as follows:

Definition The *mutual information* between two random variables X and Y is given by

$$I(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}), \tag{8.54}$$

where the supremum is over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$.

This is the master definition of mutual information that always applies, even to joint distributions with atoms, densities, and singular parts. Moreover, by continuing to refine the partitions P and Q, one finds a monotonically increasing sequence $I([X]_P;[Y]_Q) \nearrow I$.

By arguments similar to (8.52), we can show that this definition of mutual information is equivalent to (8.47) for random variables that have a density. For discrete random variables, this definition is equivalent to the definition of mutual information in (2.28).

Example 8.5.1 (*Mutual information between correlated Gaussian random variables with correlation* ρ) Let $(X,Y) \sim \mathcal{N}(0,K)$, where

$$K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}. \tag{8.55}$$

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

$$I(X;Y) = h(X) + h(Y) - h(X,Y) = -\frac{1}{2}\log(1-\rho^2). \tag{8.56}$$

If $\rho = 0$, X and Y are independent and the mutual information is 0. If $\rho = \pm 1$, X and Y are perfectly correlated and the mutual information is infinite.

8.6 PROPERTIES OF DIFFERENTIAL ENTROPY, RELATIVE ENTROPY, AND MUTUAL INFORMATION

Theorem 8.6.1

$$D(f||g) \geq 0 \tag{8.57}$$

with equality iff $f = g$ *almost everywhere (a.e.).*

Proof: Let S be the support set of f. Then

$$-D(f||g) = \int_S f \log \frac{g}{f} \tag{8.58}$$

$$\leq \log \int_S f \frac{g}{f} \qquad \text{(by Jensen's inequality)} \tag{8.59}$$

$$= \log \int_S g \tag{8.60}$$

$$\leq \log 1 = 0. \tag{8.61}$$

We have equality iff we have equality in Jensen's inequality, which occurs iff $f = g$ a.e. □

Corollary *$I(X; Y) \geq 0$ with equality iff X and Y are independent.*

Corollary *$h(X|Y) \leq h(X)$ with equality iff X and Y are independent.*

Theorem 8.6.2 (*Chain rule for differential entropy*)

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, X_2, \ldots, X_{i-1}). \tag{8.62}$$

Proof: Follows directly from the definitions. □

Corollary

$$h(X_1, X_2, \ldots, X_n) \leq \sum h(X_i), \tag{8.63}$$

with equality iff $X_1, X_2, \ldots, X_n$ are independent.

Proof: Follows directly from Theorem 8.6.2 and the corollary to Theorem 8.6.1. □

Application (*Hadamard's inequality*) *If we let $\mathbf{X} \sim \mathcal{N}(0, K)$ be a multivariate normal random variable, calculating the entropy in the above inequality gives us*

$$|K| \leq \prod_{i=1}^{n} K_{ii}, \tag{8.64}$$

which is Hadamard's inequality. A number of determinant inequalities can be derived in this fashion from information-theoretic inequalities (Chapter 17).

Theorem 8.6.3

$$h(X + c) = h(X). \tag{8.65}$$

Translation does not change the differential entropy.

Proof: Follows directly from the definition of differential entropy. □

Theorem 8.6.4

$$h(aX) = h(X) + \log |a|. \tag{8.66}$$

Proof: Let $Y = aX$. Then $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$, and

$$h(aX) = -\int f_Y(y) \log f_Y(y)\, dy \tag{8.67}$$

$$= -\int \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right) dy \tag{8.68}$$

$$= -\int f_X(x) \log f_X(x)\, dx + \log |a| \tag{8.69}$$

$$= h(X) + \log |a|, \tag{8.70}$$

after a change of variables in the integral. □

Similarly, we can prove the following corollary for vector-valued random variables.

Corollary

$$h(A\mathbf{X}) = h(\mathbf{X}) + \log |\det(A)|. \tag{8.71}$$

We now show that the multivariate normal distribution maximizes the entropy over all distributions with the same covariance.

Theorem 8.6.5 *Let the random vector $\mathbf{X} \in \mathbf{R}^n$ have zero mean and covariance $K = E\mathbf{X}\mathbf{X}^t$ (i.e., $K_{ij} = EX_iX_j$, $1 \le i, j \le n$). Then $h(\mathbf{X}) \le \frac{1}{2} \log(2\pi e)^n |K|$, with equality iff $\mathbf{X} \sim \mathcal{N}(0, K)$.*

Proof: Let $g(\mathbf{x})$ be any density satisfying $\int g(\mathbf{x}) x_i x_j d\mathbf{x} = K_{ij}$ for all i, j. Let ϕ_K be the density of a $\mathcal{N}(0, K)$ vector as given in (8.35), where we set $\mu = 0$. Note that $\log \phi_K(\mathbf{x})$ is a quadratic form and $\int x_i x_j \phi_K(\mathbf{x})\, d\mathbf{x} = K_{ij}$. Then

$$0 \le D(g||\phi_K) \tag{8.72}$$

$$= \int g \log(g/\phi_K) \tag{8.73}$$

$$= -h(g) - \int g \log \phi_K \tag{8.74}$$

$$= -h(g) - \int \phi_K \log \phi_K \tag{8.75}$$

$$= -h(g) + h(\phi_K), \tag{8.76}$$

where the substitution $\int g \log \phi_K = \int \phi_K \log \phi_K$ follows from the fact that g and ϕ_K yield the same moments of the quadratic form $\log \phi_K(\mathbf{x})$. □

In particular, the Gaussian distribution maximizes the entropy over all distributions with the same variance. This leads to the estimation counterpart to Fano's inequality. Let X be a random variable with differential entropy $h(X)$. Let $\hat{X}$ be an estimate of X, and let $E(X - \hat{X})^2$ be the expected prediction error. Let $h(X)$ be in nats.

Theorem 8.6.6 (*Estimation error and differential entropy*) *For any random variable X and estimator $\hat{X}$,*

$$E(X - \hat{X})^2 \geq \frac{1}{2\pi e} e^{2h(X)},$$

with equality if and only if X is Gaussian and $\hat{X}$ is the mean of X.

Proof: Let $\hat{X}$ be any estimator of X; then

$$E(X - \hat{X})^2 \geq \min_{\hat{X}} E(X - \hat{X})^2 \tag{8.77}$$

$$= E\,(X - E(X))^2 \tag{8.78}$$

$$= \mathrm{var}(X) \tag{8.79}$$

$$\geq \frac{1}{2\pi e} e^{2h(X)}, \tag{8.80}$$

where (8.78) follows from the fact that the mean of X is the best estimator for X and the last inequality follows from the fact that the Gaussian distribution has the maximum entropy for a given variance. We have equality only in (8.78) only if $\hat{X}$ is the best estimator (i.e., $\hat{X}$ is the mean of X and equality in (8.80) only if X is Gaussian). □

Corollary *Given side information Y and estimator $\hat{X}(Y)$, it follows that*

$$E(X - \hat{X}(Y))^2 \geq \frac{1}{2\pi e} e^{2h(X|Y)}.$$

SUMMARY

$$h(X) = h(f) = -\int_S f(x) \log f(x)\, dx \tag{8.81}$$

$$f(X^n) \doteq 2^{-nh(X)} \tag{8.82}$$

$$\text{Vol}(A_\epsilon^{(n)}) \doteq 2^{nh(X)}. \tag{8.83}$$

$$H([X]_{2^{-n}}) \approx h(X) + n. \tag{8.84}$$

$$h(\mathcal{N}(0, \sigma^2)) = \frac{1}{2} \log 2\pi e \sigma^2. \tag{8.85}$$

$$h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log (2\pi e)^n |K|. \tag{8.86}$$

$$D(f||g) = \int f \log \frac{f}{g} \geq 0. \tag{8.87}$$

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, X_2, \ldots, X_{i-1}). \tag{8.88}$$

$$h(X|Y) \leq h(X). \tag{8.89}$$

$$h(aX) = h(X) + \log |a|. \tag{8.90}$$

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x) f(y)} \geq 0. \tag{8.91}$$

$$\max_{E\mathbf{X}\mathbf{X}^t = K} h(\mathbf{X}) = \frac{1}{2} \log (2\pi e)^n |K|. \tag{8.92}$$

$$E(X - \hat{X}(Y))^2 \geq \frac{1}{2\pi e} e^{2h(X|Y)}.$$

$2^{nH(X)}$ is the effective alphabet size for a discrete random variable.
$2^{nh(X)}$ is the effective support set size for a continuous random variable.
2^C is the effective alphabet size of a channel of capacity C.

PROBLEMS

8.1 *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 σ_i^2, $i = 1, 2$.

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 [199]:

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

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 $h(\mathbf{Z} \mid \theta) \leq h(\mathbf{Z})$.]

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

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

(b) For $a = 1$ find the capacity of the channel when the input X is peak-limited; that is, the range of X is limited to $-\frac{1}{2} \leq x \leq +\frac{1}{2}$. 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 $-\frac{1}{2} \leq x \leq +\frac{1}{2}$.

8.4 *Quantized random variables*. Roughly how many bits are required on the average to describe to three-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.

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

8.6 *Variational inequality*. Verify for positive random variables X that

$$\log E_P(X) = \sup_Q \left[E_Q(\log X) - D(Q||P) \right], \tag{8.93}$$

where $E_P(X) = \sum_x x P(x)$ and $D(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$, and the supremum is over all $Q(x) \geq 0$, $\sum Q(x) = 1$. It is enough to extremize $J(Q) = E_Q \ln X - D(Q||P) + \lambda(\sum Q(x) - 1)$.

8.7 *Differential entropy bound on discrete entropy*. Let X be a discrete random variable on the set $\mathcal{X} = \{a_1, a_2, \ldots\}$ with $\Pr(X = a_i) = p_i$. Show that

$$H(p_1, p_2, \ldots) \le \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{\infty} p_i i^2 - \left(\sum_{i=1}^{\infty} i p_i\right)^2 + \frac{1}{12}\right). \tag{8.94}$$

Moreover, for every permutation σ,

$$H(p_1, p_2, \ldots) \le \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{\infty} p_{\sigma(i)} i^2 - \left(\sum_{i=1}^{\infty} i p_{\sigma(i)}\right)^2 + \frac{1}{12}\right). \tag{8.95}$$

[*Hint:* Construct a random variable X' such that $\Pr(X' = i) = p_i$. Let U be a uniform (0,1] random variable and let $Y = X' + U$, where X' and U are independent. Use the maximum entropy bound on Y to obtain the bounds in the problem. This bound is due to Massey (unpublished) and Willems (unpublished).]

8.8 *Channel with uniformly distributed noise*. Consider a additive channel whose input alphabet $\mathcal{X} = \{0, \pm 1, \pm 2\}$ and whose output $Y = X + Z$, where Z is distributed uniformly over the interval $[-1, 1]$. Thus, the input of the channel is a discrete random variable, whereas the output is continuous. Calculate the capacity $C = \max_{p(x)} I(X; Y)$ of this channel.

8.9 *Gaussian mutual information*. Suppose that (X, Y, Z) are jointly Gaussian and that $X \to Y \to Z$ forms a Markov chain. Let X and Y have correlation coefficient ρ_1 and let Y and Z have correlation coefficient ρ_2. Find $I(X; Z)$.

8.10 *Shape of the typical set*. Let X_i be i.i.d. $\sim f(x)$, where

$$f(x) = ce^{-x^4}.$$

Let $h = -\int f \ln f$. Describe the shape (or form) or the typical set $A_\epsilon^{(n)} = \{x^n \in \mathcal{R}^n : f(x^n) \in 2^{-n(h \pm \epsilon)}\}$.

8.11 *Nonergodic Gaussian process*. Consider a constant signal V in the presence of iid observational noise $\{Z_i\}$. Thus, $X_i = V + Z_i$, where $V \sim N(0, S)$ and Z_i are iid $\sim N(0, N)$. Assume that V and $\{Z_i\}$ are independent.

(a) Is $\{X_i\}$ stationary?

(b) Find $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i$. Is the limit random?

(c) What is the entropy rate h of $\{X_i\}$?

(d) Find the least-mean-squared error predictor $\hat{X}_{n+1}(X^n)$, and find $\sigma_\infty^2 = \lim_{n\to\infty} E(\hat{X}_n - X_n)^2$.

(e) Does $\{X_i\}$ have an AEP? That is, does $-\frac{1}{n}\log f(X^n) \longrightarrow h$?

HISTORICAL NOTES

Differential entropy and discrete entropy were introduced in Shannon's original paper [472]. The general rigorous definition of relative entropy and mutual information for arbitrary random variables was developed by Kolmogorov [319] and Pinsker [425], who defined mutual information as $\sup_{\mathcal{P},\mathcal{Q}} I([X]_{\mathcal{P}}; [Y]_{\mathcal{Q}})$, where the supremum is over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$.

Capacity with feedback

$$C_{n,\mathrm{FB}} = \max_{\mathrm{tr}(K_X) \le nP} \frac{1}{2n} \log \frac{|K_{X+Z}|}{|K_Z|}. \tag{9.168}$$

Feedback bounds

$$C_{n,\mathrm{FB}} \le C_n + \frac{1}{2}. \tag{9.169}$$

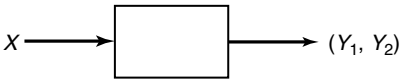
$$C_{n,\mathrm{FB}} \le 2C_n. \tag{9.170}$$

PROBLEMS

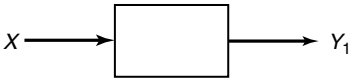
9.1 *Channel with two independent looks at Y.* Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X.

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

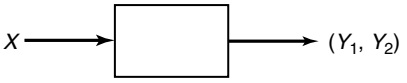
(b) Conclude that the capacity of the channel



is less than twice the capacity of the channel



9.2 *Two-look Gaussian channel*



Consider the ordinary Gaussian channel with two correlated looks at X, that is, $Y = (Y_1, Y_2)$, where

$$Y_1 = X + Z_1 \tag{9.171}$$

$$Y_2 = X + Z_2 \tag{9.172}$$

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

$$K = \begin{bmatrix} N & N\rho \\ N\rho & N \end{bmatrix}. \tag{9.173}$$

Find the capacity C for

(a) $\rho = 1$

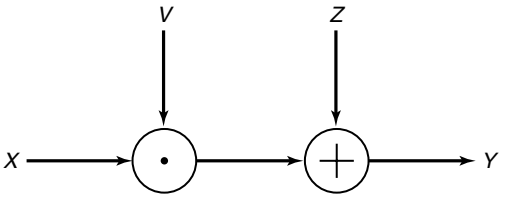
(b) $\rho = 0$

(c) $\rho = -1$

9.3 *Output power constraint*. Consider an additive white Gaussian noise channel with an expected *output* power constraint P. Thus, $Y = X + Z$, $Z \sim N(0, \sigma^2)$, Z is independent of X, and $EY^2 \leq P$. Find the channel capacity.

9.4 *Exponential noise channels*. $Y_i = X_i + Z_i$, where Z_i is i.i.d. exponentially distributed noise with mean μ. Assume that we have a mean constraint on the signal (i.e., $EX_i \leq \lambda$). Show that the capacity of such a channel is $C = \log(1 + \frac{\lambda}{\mu})$.

9.5 *Fading channel*. Consider an additive noise fading channel



$$Y = XV + Z,$$

where Z is additive noise, V is a random variable representing fading, and Z and V are independent of each other and of X. Argue that knowledge of the fading factor V improves capacity by showing that

$$I(X; Y|V) \geq I(X; Y).$$

9.6 *Parallel channels and water-filling*. Consider a pair of parallel Gaussian channels:

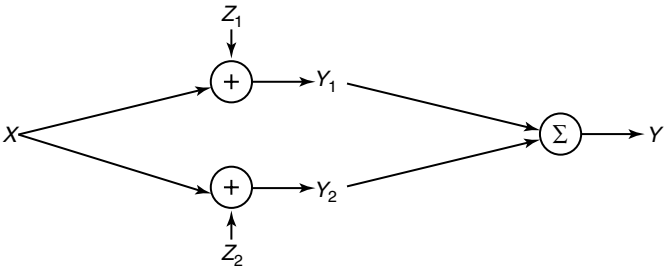
$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}, \tag{9.174}$$

where

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}\right), \tag{9.175}$$

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?

9.7 *Multipath Gaussian channel*. Consider a Gaussian noise channel with power constraint P, where the signal takes two different paths and the received noisy signals are added together at the antenna.

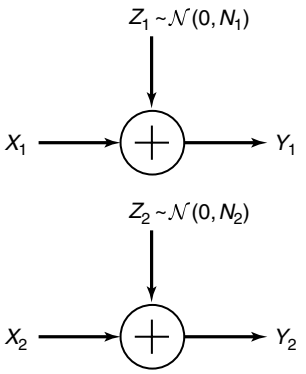


(a) Find the capacity of this channel if Z_1 and Z_2 are jointly normal with covariance matrix

$$K_Z = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}.$$

(b) What is the capacity for $\rho = 0$, $\rho = 1$, $\rho = -1$?

9.8 *Parallel Gaussian channels*. Consider the following parallel Gaussian channel:



where $Z_1 \sim \mathcal{N}(0,N_1)$ and $Z_2 \sim \mathcal{N}(0,N_2)$ are independent Gaussian random variables and $Y_i = X_i + Z_i$. We wish to allocate power to the two parallel channels. Let β_1 and β_2 be fixed. Consider a total cost constraint $\beta_1 P_1 + \beta_2 P_2 \leq \beta$, where P_i is the power allocated to the ith channel and β_i is the cost per unit power in that channel. Thus, $P_1 \geq 0$ and $P_2 \geq 0$ can be chosen subject to the cost constraint β.

(a) For what value of β does the channel stop acting like a single channel and start acting like a pair of channels?

(b) Evaluate the capacity and find P_1 and P_2 that achieve capacity for $\beta_1 = 1$, $\beta_2 = 2$, $N_1 = 3$, $N_2 = 2$, and $\beta = 10$.

9.9 *Vector Gaussian channel*. Consider the vector Gaussian noise channel

$$Y = X + Z,$$

where $X = (X_1, X_2, X_3)$, $Z = (Z_1, Z_2, Z_3)$, $Y = (Y_1, Y_2, Y_3)$, $E\|X\|^2 \leq P$, and

$$Z \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}\right).$$

Find the capacity. The answer may be surprising.

9.10 *Capacity of photographic film*. Here is a problem with a nice answer that takes a little time. We're interested in the capacity of photographic film. The film consists of silver iodide crystals, Poisson distributed, with a density of λ particles per square inch. The film is illuminated without knowledge of the position of the silver iodide particles. It is then developed and the receiver sees only the silver iodide particles that have been illuminated. It is assumed that light incident on a cell exposes the grain if it is there and otherwise results in a blank response. Silver iodide particles that are not illuminated and vacant portions of the film remain blank. The question is: What is the capacity of this film?

We make the following assumptions. We grid the film very finely into cells of area dA. It is assumed that there is at most one silver iodide particle per cell and that no silver iodide particle is intersected by the cell boundaries. Thus, the film can be considered to be a large number of parallel binary asymmetric channels with crossover probability $1 - \lambda dA$. By calculating the capacity of this binary asymmetric channel to first order in dA (making the

necessary approximations), one can calculate the capacity of the film in bits per square inch. It is, of course, proportional to λ. The question is: What is the multiplicative constant?
The answer would be λ bits per unit area if both illuminator and receiver knew the positions of the crystals.

9.11 *Gaussian mutual information*. Suppose that (X, Y, Z) are jointly Gaussian and that $X \to Y \to Z$ forms a Markov chain. Let X and Y have correlation coefficient ρ_1 and let Y and Z have correlation coefficient ρ_2. Find $I(X; Z)$.

9.12 *Time-varying channel*. A train pulls out of the station at constant velocity. The received signal energy thus falls off with time as $1/i^2$. The total received signal at time i is

$$Y_i = \frac{1}{i} X_i + Z_i,$$

where $Z_1, Z_2, \ldots$ are i.i.d. $\sim N(0, N)$. The transmitter constraint for block length n is

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2(w) \le P, \quad w \in \{1, 2, \ldots, 2^{nR}\}.$$

Using Fano's inequality, show that the capacity C is equal to zero for this channel.

9.13 *Feedback capacity*. Let $(Z_1, Z_2) \sim N(0, K)$, $K = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. Find the maximum of $\frac{1}{2} \log \frac{|K_{X+Z}|}{|K_Z|}$ with and without feedback given a trace (power) constraint $\text{tr}(K_X) \le 2P$.

9.14 *Additive noise channel*. Consider the channel $Y = X + Z$, where X is the transmitted signal with power constraint P, Z is independent additive noise, and Y is the received signal. Let

$$Z = \begin{cases} 0 & \text{with probability } \frac{1}{10} \\ Z^* & \text{with probability } \frac{9}{10}, \end{cases}$$

where $Z^* \sim N(0, N)$. Thus, Z has a mixture distribution that is the mixture of a Gaussian distribution and a degenerate distribution with mass 1 at 0.

(a) What is the capacity of this channel? This should be a pleasant surprise.

(b) How would you signal to achieve capacity?

9.15 *Discrete input, continuous output channel*. Let $\Pr\{X = 1\} = p$, $\Pr\{X = 0\} = 1 - p$, and let $Y = X + Z$, where Z is uniform over the interval $[0, a]$, $a > 1$, and Z is independent of X.

(a) Calculate

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

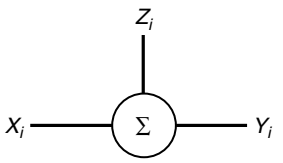
(b) Now calculate $I(X; Y)$ the other way by

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

(c) Calculate the capacity of this channel by maximizing over p.

9.16 *Gaussian mutual information*. Suppose that (X, Y, Z) are jointly Gaussian and that $X \to Y \to Z$ forms a Markov chain. Let X and Y have correlation coefficient ρ_1 and let Y and Z have correlation coefficient ρ_2. Find $I(X; Z)$.

9.17 *Impulse power*. Consider the additive white Gaussian channel



where $Z_i \sim N(0, N)$, and the input signal has average power constraint P.

(a) Suppose that we use all our power at time 1 (i.e., $EX_1^2 = nP$ and $EX_i^2 = 0$ for $i = 2, 3, \ldots, n$). Find

$$\max_{f(x^n)} \frac{I(X^n; Y^n)}{n},$$

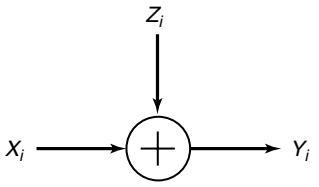
where the maximization is over all distributions $f(x^n)$ subject to the constraint $EX_1^2 = nP$ and $EX_i^2 = 0$ for $i = 2, 3, \ldots, n$.

(b) Find

$$\max_{f(x^n):\ E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)\le P} \frac{1}{n} I(X^n; Y^n)$$

and compare to part (a).

9.18 *Gaussian channel with time-varying mean*. Find the capacity of the following Gaussian channel:



Let $Z_1, Z_2, \ldots$ be independent and let there be a power constraint P on $x^n(W)$. Find the capacity when:

(a) $\mu_i = 0$, for all i.

(b) $\mu_i = e^i, \quad i = 1, 2, \ldots$. Assume that μ_i is known to the transmitter and receiver.

(c) μ_i unknown, but μ_i i.i.d. $\sim N(0, N_1)$ for all i.

9.19 *Parametric form for channel capacity*. Consider m parallel Gaussian channels, $Y_i = X_i + Z_i$, where $Z_i \sim N(0, \lambda_i)$ and the noises X_i are independent random variables. Thus, $C = \sum_{i=1}^{m} \frac{1}{2} \log(1 + \frac{(\lambda - \lambda_i)^+}{\lambda_i})$, where λ is chosen to satisfy $\sum_{i=1}^{m} (\lambda - \lambda_i)^+ = P$. Show that this can be rewritten in the form

$$\begin{aligned} P(\lambda) &= \textstyle\sum_{i:\lambda_i \le \lambda} (\lambda - \lambda_i) \\ C(\lambda) &= \textstyle\sum_{i:\lambda_i \le \lambda} \frac{1}{2} \log \frac{\lambda}{\lambda_i}. \end{aligned}$$

Here $P(\lambda)$ is piecewise linear and $C(\lambda)$ is piecewise logarithmic in λ.

9.20 *Robust decoding*. Consider an additive noise channel whose output Y is given by

$$Y = X + Z,$$

where the channel input X is average power limited,

$$EX^2 \le P,$$

and the noise process $\{Z_k\}_{k=-\infty}^{\infty}$ is i.i.d. with marginal distribution $p_Z(z)$ (not necessarily Gaussian) of power N,

$$EZ^2 = N.$$

(a) Show that the channel capacity, $C = \max_{EX^2 \le P} I(X;Y)$, is lower bounded by C_G, where

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

(i.e., the capacity C_G corresponding to white Gaussian noise).

(b) Decoding the received vector to the codeword that is closest to it in Euclidean distance is in general suboptimal if the noise is non-Gaussian. Show, however, that the rate C_G is achievable even if one insists on performing nearest-neighbor decoding (minimum Euclidean distance decoding) rather than the optimal maximum-likelihood or joint typicality decoding (with respect to the true noise distribution).

(c) Extend the result to the case where the noise is not i.i.d. but is stationary and ergodic with power N.

(*Hint for b and c:* Consider a size 2^{nR} random codebook whose codewords are drawn independently of each other according to a uniform distribution over the n-dimensional sphere of radius $\sqrt{nP}$.)

(a) Using a symmetry argument, show that conditioned on the noise vector, the ensemble average probability of error depends on the noise vector only via its Euclidean norm $\|\mathbf{z}\|$.

(b) Use a geometric argument to show that this dependence is monotonic.

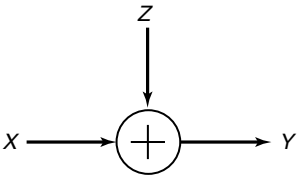
(c) Given a rate $R < C_G$, choose some $N' > N$ such that

$$R < \frac{1}{2}\log\left(1 + \frac{P}{N'}\right).$$

Compare the case where the noise is i.i.d. $\mathcal{N}(0, N')$ to the case at hand.

(d) Conclude the proof using the fact that the above ensemble of codebooks can achieve the capacity of the Gaussian channel (no need to prove that).

9.21 *Mutual information game*. Consider the following channel:



Throughout this problem we shall constrain the signal power

$$EX = 0, \qquad EX^2 = P, \tag{9.176}$$

and the noise power

$$EZ = 0, \qquad EZ^2 = N, \tag{9.177}$$

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

Now for the game. The noise player chooses a distribution on Z to minimize $I(X; X + Z)$, while the signal player chooses a distribution on X to maximize $I(X; X + Z)$. Letting $X^* \sim \mathcal{N}(0, P)$, $Z^* \sim \mathcal{N}(0, N)$, show that Gaussian X^* and Z^* satisfy the saddlepoint conditions

$$I(X; X + Z^*) \le I(X^*; X^* + Z^*) \le I(X^*; X^* + Z). \tag{9.178}$$

Thus,

$$\min_Z \max_X I(X; X + Z) = \max_X \min_Z I(X; X + Z) \tag{9.179}$$

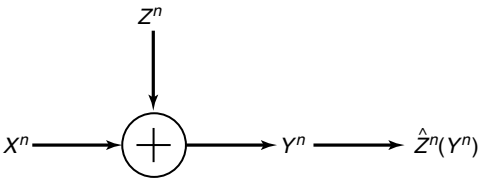
$$= \frac{1}{2} \log \left(1 + \frac{P}{N} \right), \tag{9.180}$$

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 17.8, which states that if $\mathbf{X}$ and $\mathbf{Y}$ are independent random n-vectors with densities, then

$$2^{\frac{2}{n}h(\mathbf{X}+\mathbf{Y})} \ge 2^{\frac{2}{n}h(\mathbf{X})} + 2^{\frac{2}{n}h(\mathbf{Y})}. \tag{9.181}$$

9.22 *Recovering the noise*. Consider a standard Gaussian channel $Y^n = X^n + Z^n$, where Z_i is i.i.d. $\sim \mathcal{N}(0, N)$, $i = 1, 2, \ldots, n$, and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \leq P$. Here we are interested in recovering the noise Z^n and we don't care about the signal X^n. By sending $X^n = (0, 0, \ldots, 0)$, the receiver gets $Y^n = Z^n$ and can fully determine the value of Z^n. We wonder how much variability there can be in X^n and still recover the Gaussian noise Z^n. Use of the channel looks like



Argue that for some $R > 0$, the transmitter can arbitrarily send one of 2^{nR} different sequences of x^n without affecting the recovery of the noise in the sense that

$$\Pr\{\hat{Z}^n \neq Z^n\} \to 0 \qquad \text{as } n \to \infty.$$

For what R is this possible?

HISTORICAL NOTES

The Gaussian channel was first analyzed by Shannon in his original paper [472]. The water-filling solution to the capacity of the colored noise Gaussian channel was developed by Shannon [480] and treated in detail by Pinsker [425]. The time-continuous Gaussian channel is treated in Wyner [576], Gallager [233], and Landau, Pollak, and Slepian [340, 341, 500].

Pinsker [421] and Ebert [178] argued that feedback at most doubles the capacity of a nonwhite Gaussian channel; the proof in the text is from Cover and Pombra [136], who also show that feedback increases the capacity of the nonwhite Gaussian channel by at most half a bit. The most recent feedback capacity results for nonwhite Gaussian noise channels are due to Kim [314].

Rate distortion theorem. If $R > R(D)$, there exists a sequence of codes $\hat{X}^n(X^n)$ with the number of codewords $|\hat{X}^n(\cdot)| \leq 2^{nR}$ with $Ed(X^n, \hat{X}^n(X^n)) \to D$. If $R < R(D)$, no such codes exist.

Bernoulli source. For a Bernoulli source with Hamming distortion,

$$R(D) = H(p) - H(D). \tag{10.149}$$

Gaussian source. For a Gaussian source with squared-error distortion,

$$R(D) = \frac{1}{2} \log \frac{\sigma^2}{D}. \tag{10.150}$$

Source–channel separation. A source with rate distortion $R(D)$ can be sent over a channel of capacity C and recovered with distortion D if and only if $R(D) < C$.

Multivariate Gaussian source. The rate distortion function for a multivariate normal vector with Euclidean mean-squared-error distortion is given by reverse water-filling on the eigenvalues.

PROBLEMS

10.1 *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 $R = 1$.

10.2 *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

$$d(x, \hat{x}) = \begin{cases} 0, & x = \hat{x} \\ 1, & x = 1, \hat{x} = 0 \\ \infty, & x = 0, \hat{x} = 1. \end{cases}$$

10.3 *Rate distortion for binary source with asymmetric distortion* . Fix $p(\hat{x}|x)$ and evaluate $I(X;\hat{X})$ and D for

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

$$d(x,\hat{x}) = \begin{bmatrix} 0 & a \\ b & 0 \end{bmatrix}.$$

(The rate distortion function cannot be expressed in closed form.)

10.4 *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 $R'(D) = R(D + \overline{w})$, where $\overline{w} = \sum p_i w_i$, 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 $\hat{x}$ that reproduces the source with zero distortion). This result is due to Pinkston [420].

10.5 *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; that is,

$$d(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x}, \\ 1 & \text{if } x \neq \hat{x}. \end{cases}$$

10.6 *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

$$\phi(D) = \max_{\mathbf{p}:\sum_{i=1}^{m} p_i d_i \leq D} H(\mathbf{p}). \tag{10.151}$$

The Shannon lower bound on the rate distortion function [485] 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 $Ed(X, \hat{X}) \leq D$,

$$I(X;\hat{X}) = H(X) - H(X|\hat{X}) \tag{10.152}$$

$$= H(X) - \sum_{\hat{x}} p(\hat{x}) H(X|\hat{X} = \hat{x}) \tag{10.153}$$

$$\geq H(X) - \sum_{\hat{x}} p(\hat{x}) \phi(D_{\hat{x}}) \tag{10.154}$$

$$\geq H(X) - \phi\left(\sum_{\hat{x}} p(\hat{x}) D_{\hat{x}} \right) \tag{10.155}$$

$$\geq H(X) - \phi(D), \tag{10.156}$$

where $D_{\hat{x}} = \sum_x p(x|\hat{x}) d(x, \hat{x})$.

(c) Argue that

$$R(D) \geq H(X) - \phi(D), \tag{10.157}$$

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).

10.7 *Erasure distortion*. Consider $X \sim$ Bernoulli $(\frac{1}{2})$, and let the distortion measure be given by the matrix

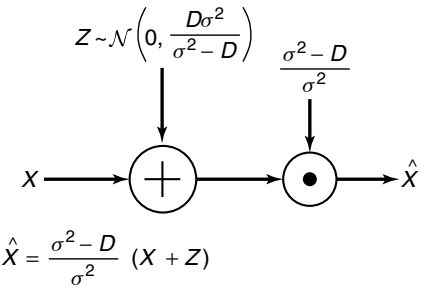
$$d(x, \hat{x}) = \begin{bmatrix} 0 & 1 & \infty \\ \infty & 1 & 0 \end{bmatrix}. \tag{10.158}$$

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?

10.8 *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

$$h(X) - \frac{1}{2} \log(2\pi e D) \leq R(D) \leq \frac{1}{2} \log \frac{\sigma^2}{D}. \tag{10.159}$$

For the upper bound, consider the following joint distribution:



Are Gaussian random variables harder or easier to describe than other random variables with the same variance?

10.9 *Properties of optimal rate distortion code.* A good (R, D) rate distortion code with $R \approx R(D)$ puts severe constraints on the relationship of the source X^n and the representations $\hat{X}^n$. Examine the chain of inequalities (10.58–10.71) considering the conditions for equality and interpret as properties of a good code. For example, equality in (10.59) implies that $\hat{X}^n$ is a deterministic function of X^n.

10.10 *Rate distortion.* Find and verify the rate distortion function $R(D)$ for X uniform on $\mathcal{X} = \{1, 2, \ldots, 2m\}$ and

$$d(x, \hat{x}) = \begin{cases} 1 & \text{for } x - \hat{x} \text{ odd}, \\ 0 & \text{for } x - \hat{x} \text{ even}, \end{cases}$$

where $\hat{X}$ is defined on $\hat{\mathcal{X}} = \{1, 2, \ldots, 2m\}$. (You may wish to use the Shannon lower bound in your argument.)

10.11 *Lower bound.* Let

$$X \sim \frac{e^{-x^4}}{\int_{-\infty}^{\infty} e^{-x^4} dx}$$

and

$$\frac{\int x^4 e^{-x^4} dx}{\int e^{-x^4} dx} = c.$$

Define $g(a) = \max h(X)$ over all densities such that $EX^4 \leq a$. Let $R(D)$ be the rate distortion function for X with the density above and with distortion criterion $d(x, \hat{x}) = (x - \hat{x})^4$. Show that $R(D) \geq g(c) - g(D)$.

10.12 *Adding a column to the distortion matrix*. Let $R(D)$ be the rate distortion function for an i.i.d. process with probability mass function $p(x)$ and distortion function $d(x, \hat{x})$, $x \in \mathcal{X}$, $\hat{x} \in \hat{\mathcal{X}}$. Now suppose that we add a new reproduction symbol $\hat{x}_0$ to $\hat{\mathcal{X}}$ with associated distortion $d(x, \hat{x}_0)$, $x \in \mathcal{X}$. Does this increase or decrease $R(D)$, and why?

10.13 *Simplification*. Suppose that $\mathcal{X} = \{1, 2, 3, 4\}$, $\hat{\mathcal{X}} = \{1, 2, 3, 4\}$, $p(i) = \frac{1}{4}$, $i = 1, 2, 3, 4$, and $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$. The distortion matrix $d(x, \hat{x})$ is given by

	1	2	3	4
1	0	0	1	1
2	0	0	1	1
3	1	1	0	0
4	1	1	0	0

(a) Find $R(0)$, the rate necessary to describe the process with zero distortion.

(b) Find the rate distortion function $R(D)$. There are some irrelevant distinctions in alphabets $\mathcal{X}$ and $\hat{\mathcal{X}}$, which allow the problem to be collapsed.

(c) Suppose that we have a nonuniform distribution $p(i) = p_i$, $i = 1, 2, 3, 4$. What is $R(D)$?

10.14 *Rate distortion for two independent sources*. Can one compress two independent sources simultaneously better than by compressing the sources individually? The following problem addresses this question. Let $\{X_i\}$ be i.i.d. $\sim p(x)$ with distortion $d(x, \hat{x})$ and rate distortion function $R_X(D)$. Similarly, let $\{Y_i\}$ be i.i.d. $\sim p(y)$ with distortion $d(y, \hat{y})$ and rate distortion function $R_Y(D)$. Suppose we now wish to describe the process $\{(X_i, Y_i)\}$ subject to distortions $Ed(X, \hat{X}) \leq D_1$ and $Ed(Y, \hat{Y}) \leq D_2$. Thus, a rate $R_{X,Y}(D_1, D_2)$ is sufficient, where

$$R_{X,Y}(D_1, D_2) = \min_{p(\hat{x}, \hat{y}|x, y): Ed(X, \hat{X}) \leq D_1, Ed(Y, \hat{Y}) \leq D_2} I(X, Y; \hat{X}, \hat{Y}).$$

Now suppose that the $\{X_i\}$ process and the $\{Y_i\}$ process are independent of each other.

(a) Show that

$$R_{X,Y}(D_1, D_2) \geq R_X(D_1) + R_Y(D_2).$$

(b) Does equality hold?

Now answer the question.

10.15 *Distortion rate function*. Let

$$D(R) = \min_{p(\hat{x}|x): I(X;\hat{X}) \le R} Ed(X, \hat{X}) \tag{10.160}$$

be the distortion rate function.

(a) Is $D(R)$ increasing or decreasing in R?

(b) Is $D(R)$ convex or concave in R?

(c) Converse for distortion rate functions: We now wish to prove the converse by focusing on $D(R)$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. Suppose that one is given a $(2^{nR}, n)$ rate distortion code $X^n \to i(X^n) \to \hat{X}^n(i(X^n))$, with $i(X^n) \in 2^{nR}$, and suppose that the resulting distortion is $D = Ed(X^n, \hat{X}^n(i(X^n)))$. We must show that $D \ge D(R)$. Give reasons for the following steps in the proof:

$$D = Ed(X^n, \hat{X}^n(i(X^n))) \tag{10.161}$$

$$\stackrel{(a)}{=} E\frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i) \tag{10.162}$$

$$\stackrel{(b)}{=} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Ed}(X_i, \hat{X}_i) \tag{10.163}$$

$$\stackrel{(c)}{\ge} \frac{1}{n}\sum_{i=1}^{n} D\left(I(X_i; \hat{X}_i)\right) \tag{10.164}$$

$$\stackrel{(d)}{\ge} D\left(\frac{1}{n}\sum_{i=1}^{n} I(X_i; \hat{X}_i)\right) \tag{10.165}$$

$$\stackrel{(e)}{\ge} D\left(\frac{1}{n} I(X^n; \hat{X}^n)\right) \tag{10.166}$$

$$\stackrel{(f)}{\ge} D(R). \tag{10.167}$$

10.16 *Probability of conditionally typical sequences*. In Chapter 7 we calculated the probability that two independently drawn sequences X^n and Y^n are weakly jointly typical. To prove the rate distortion theorem, however, we need to calculate this probability when

one of the sequences is fixed and the other is random. The techniques of weak typicality allow us only to calculate the average set size of the conditionally typical set. Using the ideas of strong typicality, on the other hand, provides us with stronger bounds that work for all typical x^n sequences. We 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 [53] and is fully developed in the book by Csiszár and Körner [149].

Let (X_i, Y_i) be drawn i.i.d. $\sim p(x, y)$. Let the marginals of X and Y be $p(x)$ and $p(y)$, respectively.

(a) Let $A_\epsilon^{*(n)}$ be the strongly typical set for X. Show that

$$|A_\epsilon^{*(n)}| \doteq 2^{nH(X)}. \tag{10.168}$$

(*Hint:* Theorems 11.1.1 and 11.1.3.)

(b) The *joint type* of a pair of sequences (x^n, y^n) is the proportion of times $(x_i, y_i) = (a, b)$ in the pair of sequences:

$$p_{x^n, y^n}(a, b) = \frac{1}{n} N(a, b|x^n, y^n) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = a, y_i = b). \tag{10.169}$$

The *conditional type* of a sequence y^n given x^n is a stochastic matrix that gives the proportion of times a particular element of $\mathcal{Y}$ occurred with each element of $\mathcal{X}$ in the pair of sequences. Specifically, the conditional type $V_{y^n|x^n}(b|a)$ is defined as

$$V_{y^n|x^n}(b|a) = \frac{N(a, b|x^n, y^n)}{N(a|x^n)}. \tag{10.170}$$

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

$$\frac{1}{(n + 1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{nH(Y|X)} \leq |T_V(x^n)| \leq 2^{nH(Y|X)}. \tag{10.171}$$

(d) The sequence $y^n \in \mathcal{Y}^n$ is said to be *ϵ-strongly conditionally typical* with the sequence x^n with respect to the conditional distribution $V(\cdot|\cdot)$ if the conditional type is close to V. The conditional type should satisfy the following two conditions:

(i) For all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $V(b|a) > 0$,

$$\frac{1}{n}\left|N(a, b|x^n, y^n) - V(b|a)N(a|x^n)\right| \leq \frac{\epsilon}{|\mathcal{Y}| + 1}. \tag{10.172}$$

(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

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(Y|X)-\epsilon_1)} \leq |A_\epsilon^{*(n)}(Y|x^n)| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(Y|X)+\epsilon_1)}, \tag{10.173}$$

where $\epsilon_1 \to 0$ as $\epsilon \to 0$.

(e) For a pair of random variables (X, Y) with joint distribution $p(x, y)$, the *ϵ-strongly typical* set $A_\epsilon^{*(n)}$ is the set of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ satisfying

(i)

$$\left|\frac{1}{n}N(a, b|x^n, y^n) - p(a, b)\right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|} \tag{10.174}$$

for every pair $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a, b) > 0$.

(ii) $N(a, b|x^n, y^n) = 0$ for all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ with $p(a, b) = 0$.

The set of ϵ-strongly jointly typical sequences is called the *ϵ-strongly jointly typical set* and is denoted $A_\epsilon^{*(n)}(X, Y)$. Let (X, Y) be drawn i.i.d. $\sim p(x, y)$. For any x^n such that there exists at least one pair $(x^n, y^n) \in A_\epsilon^{*(n)}(X, Y)$, the set of sequences y^n such that $(x^n, y^n) \in A_\epsilon^{*(n)}$ satisfies

$$\frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} 2^{n(H(Y|X)-\delta(\epsilon))} \leq |\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} 2^{n(H(Y|X)+\delta(\epsilon))}, \tag{10.175}$$

where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. In particular, we can write

$$2^{n(H(Y|X)-\epsilon_2)} \leq |\{y^n : (x^n, y^n) \in A_\epsilon^{*(n)}\}| \leq 2^{n(H(Y|X)+\epsilon_2)}, \tag{10.176}$$

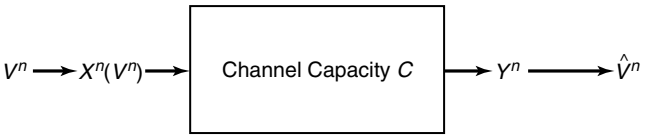
where we can make ϵ_2 arbitrarily small with an appropriate choice of ϵ and n.

(f) Let $Y_1, Y_2, \ldots, Y_n$ be drawn i.i.d. $\sim \prod p(y_i)$. For $x^n \in A_\epsilon^{*(n)}$, the probability that $(x^n, Y^n) \in A_\epsilon^{*(n)}$ is bounded by

$$2^{-n(I(X;Y)+\epsilon_3)} \leq \Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon_3)}, \tag{10.177}$$

where ϵ_3 goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

10.17 *Source–channel separation theorem with distortion.* Let $V_1, V_2, \ldots, V_n$ be a finite alphabet i.i.d. source which is encoded as a sequence of n input symbols X^n of a discrete memoryless channel. The output of the channel Y^n is mapped onto the reconstruction alphabet $\hat{V}^n = g(Y^n)$. Let $D = Ed(V^n, \hat{V}^n) = \frac{1}{n}\sum_{i=1}^n Ed(V_i, \hat{V}_i)$ be the average distortion achieved by this combined source and channel coding scheme.



(a) Show that if $C > R(D)$, where $R(D)$ is the rate distortion function for V, it is possible to find encoders and decoders that achieve a average distortion arbitrarily close to D.

(b) (Converse) Show that if the average distortion is equal to D, the capacity of the channel C must be greater than $R(D)$.

10.18 *Rate distortion.* Let $d(x, \hat{x})$ be a distortion function. We have a source $X \sim p(x)$. Let $R(D)$ be the associated rate distortion function.

(a) Find $\tilde{R}(D)$ in terms of $R(D)$, where $\tilde{R}(D)$ is the rate distortion function associated with the distortion $\tilde{d}(x, \hat{x}) = d(x, \hat{x}) + a$ for some constant $a > 0$. (They are not equal.)

(b) Now suppose that $d(x, \hat{x}) \geq 0$ for all $x, \hat{x}$ and define a new distortion function $d^*(x, \hat{x}) = bd(x, \hat{x})$, where b is some number ≥ 0. Find the associated rate distortion function $R^*(D)$ in terms of $R(D)$.

(c) Let $X \sim N(0, \sigma^2)$ and $d(x, \hat{x}) = 5(x - \hat{x})^2 + 3$. What is $R(D)$?

10.19 *Rate distortion with two constraints.* Let X_i be iid $\sim p(x)$. We are given two distortion functions, $d_1(x, \hat{x})$ and $d_2(x, \hat{x})$. We wish to describe X^n at rate R and reconstruct it with distortions $Ed_1(X^n, \hat{X}_1^n) \leq D_1$, and $Ed_2(X^n, \hat{X}_2^n) \leq D_2$, as shown here:

$$X^n \longrightarrow i(X^n) \longrightarrow (\hat{X}_1^n(i), \hat{X}_2^n(i))$$

$$D_1 = Ed_1(X_1^n, \hat{X}_1^n)$$
$$D_2 = Ed_2(X_1^n, \hat{X}_2^n).$$

Here $i(\cdot)$ takes on 2^{nR} values. What is the rate distortion function $R(D_1, D_2)$?

10.20 *Rate distortion.* Consider the standard rate distortion problem, X_i i.i.d. $\sim p(x)$, $X^n \to i(X^n) \to \hat{X}^n$, $|i(\cdot)| = 2^{nR}$. Consider two distortion criteria $d_1(x, \hat{x})$ and $d_2(x, \hat{x})$. Suppose that $d_1(x, \hat{x}) \leq d_2(x, \hat{x})$ for all $x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}$. Let $R_1(D)$ and $R_2(D)$ be the corresponding rate distortion functions.

(a) Find the inequality relationship between $R_1(D)$ and $R_2(D)$.

(b) Suppose that we must describe the source $\{X_i\}$ at the minimum rate R achieving $d_1(X^n, \hat{X}_1^n) \leq D$ and $d_2(X^n, \hat{X}_2^n) \leq D$. Thus,

$$X^n \to i(X^n) \to \begin{cases} \hat{X}_1^n(i(X^n)) \\ \hat{X}_2^n(i(X^n)) \end{cases}$$

and $|i(\cdot)| = 2^{nR}$.
Find the minimum rate R.

HISTORICAL NOTES

The idea of rate distortion was introduced by Shannon in his original paper [472]. He returned to it and dealt with it exhaustively in his 1959 paper [485], which proved the first rate distortion theorem. Meanwhile, Kolmogorov and his school in the Soviet Union began to develop rate distortion theory in 1956. Stronger versions of the rate distortion theorem have been proved for more general sources in the comprehensive book by Berger [52].

The inverse water-filling solution for the rate distortion function for parallel Gaussian sources was established by McDonald and Schultheiss