3. Assuming that s_i is decoded correctly at the receiver, the receiver constructs a list $Ł(\mathbf{y}(i-1))$ of indices that the receiver considers to be jointly typical with $\mathbf{y}(i-1)$ in the $(i-1)$th block. The receiver then declares $\hat{w}_{i-1} = w$ as the index sent in block $i-1$ if there is a unique w in $S_{s_i} \cap Ł(\mathbf{y}(i-1))$. If n is sufficiently large and if

$$R < I(X; Y|X_1) + R_0, \tag{15.255}$$

then $\hat{w}_{i-1} = w_{i-1}$ with arbitrarily small probability of error. Combining the two constraints (15.254) and (15.255), R_0 drops out, leaving

$$R < I(X; Y|X_1) + I(X_1; Y) = I(X, X_1; Y). \tag{15.256}$$

For a detailed analysis of the probability of error, the reader is referred to Cover and El Gamal [127]. □

Theorem 15.7.2 can also shown to be the capacity for the following classes of relay channels:

1. Reversely degraded relay channel, that is,

$$p(y, y_1|x, x_1) = p(y|x, x_1)p(y_1|y, x_1). \tag{15.257}$$

2. Relay channel with feedback
3. Deterministic relay channel,

$$y_1 = f(x, x_1), \qquad y = g(x, x_1). \tag{15.258}$$

15.8 SOURCE CODING WITH SIDE INFORMATION

We now consider the distributed source coding problem where two random variables X and Y are encoded separately but only X is to be recovered.We now ask how many bits R_1 are required to describe X if we are allowed R_2 bits to describe Y. If $R_2 > H(Y)$, then Y can be described perfectly, and by the results of Slepian–Wolf coding, $R_1 = H(X|Y)$ bits suffice to describe X. At the other extreme, if $R_2 = 0$, we must describe X without any help, and $R_1 = H(X)$ bits are then necessary to describe X. In general, we use $R_2 = I(Y; \hat{Y})$ bits to describe an approximate version of Y. This will allow us to describe X using $H(X|\hat{Y})$ bits in the presence of side information $\hat{Y}$. The following theorem is consistent with this intuition.

Theorem 15.8.1 *Let $(X, Y) \sim p(x, y)$. If Y is encoded at rate R_2 and X is encoded at rate R_1, we can recover X with an arbitrarily small probability of error if and only if*

$$R_1 \geq H(X|U), \tag{15.259}$$

$$R_2 \geq I(Y; U) \tag{15.260}$$

for some joint probability mass function $p(x, y)p(u|y)$, where $|\mathcal{U}| \leq |\mathcal{Y}| + 2$.

We prove this theorem in two parts. We begin with the converse, in which we show that for any encoding scheme that has a small probability of error, we can find a random variable U with a joint probability mass function as in the theorem.

Proof: (*Converse*). Consider any source code for Figure 15.32. The source code consists of mappings $f_n(X^n)$ and $g_n(Y^n)$ such that the rates of f_n and g_n are less than R_1 and R_2, respectively, and a decoding mapping h_n such that

$$P_e^{(n)} = \Pr\{h_n(f_n(X^n), g_n(Y^n)) \neq X^n\} < \epsilon. \tag{15.261}$$

Define new random variables $S = f_n(X^n)$ and $T = g_n(Y^n)$. Then since we can recover X^n from S and T with low probability of error, we have, by Fano's inequality,

$$H(X^n|S, T) \leq n\epsilon_n. \tag{15.262}$$

Then

$$nR_2 \overset{(a)}{\geq} H(T) \tag{15.263}$$

$$\overset{(b)}{\geq} I(Y^n; T) \tag{15.264}$$

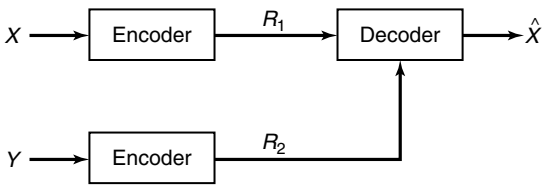


FIGURE 15.32. Encoding with side information.

$$= \sum_{i=1}^{n} I(Y_i; T|Y_1, \ldots, Y_{i-1}) \tag{15.265}$$

$$\overset{(c)}{=} \sum_{i=1}^{n} I(Y_i; T, Y_1, \ldots, Y_{i-1}) \tag{15.266}$$

$$\overset{(d)}{=} \sum_{i=1}^{n} I(Y_i; U_i) \tag{15.267}$$

where

(a) follows from the fact that the range of g_n is $\{1, 2, \ldots, 2^{nR_2}\}$
(b) follows from the properties of mutual information
(c) follows from the chain rule and the fact that Y_i is independent of $Y_1, \ldots, Y_{i-1}$ and hence $I(Y_i; Y_1, \ldots, Y_{i-1}) = 0$
(d) follows if we define $U_i = (T, Y_1, \ldots, Y_{i-1})$

We also have another chain for R_1,

$$nR_1 \overset{(a)}{\geq} H(S) \tag{15.268}$$

$$\overset{(b)}{\geq} H(S|T) \tag{15.269}$$

$$= H(S|T) + H(X^n|S, T) - H(X^n|S, T) \tag{15.270}$$

$$\overset{(c)}{\geq} H(X^n, S|T) - n\epsilon_n \tag{15.271}$$

$$\overset{(d)}{=} H(X^n|T) - n\epsilon_n \tag{15.272}$$

$$\overset{(e)}{=} \sum_{i=1}^{n} H(X_i|T, X_1, \ldots, X_{i-1}) - n\epsilon_n \tag{15.273}$$

$$\overset{(f)}{\geq} \sum_{i=1}^{n} H(X_i|T, X^{i-1}, Y^{i-1}) - n\epsilon_n \tag{15.274}$$

$$\overset{(g)}{=} \sum_{i=1}^{n} H(X_i|T, Y^{i-1}) - n\epsilon_n \tag{15.275}$$

$$\overset{(h)}{=} \sum_{i=1}^{n} H(X_i|U_i) - n\epsilon_n, \tag{15.276}$$

where
(a) follows from the fact that the range of S is $\{1, 2, \ldots, 2^{nR_1}\}$
(b) follows from the fact that conditioning reduces entropy
(c) follows from Fano's inequality
(d) follows from the chain rule and the fact that S is a function of X^n
(e) follows from the chain rule for entropy
(f) follows from the fact that conditioning reduces entropy
(g) follows from the (subtle) fact that $X_i \to (T, Y^{i-1}) \to X^{i-1}$ forms a Markov chain since X_i does not contain any information about X^{i-1} that is not there in Y^{i-1} and T
(h) follows from the definition of U

Also, since X_i contains no more information about U_i than is present in Y_i, it follows that $X_i \to Y_i \to U_i$ forms a Markov chain. Thus we have the following inequalities:

$$R_1 \geq \frac{1}{n} \sum_{i=1}^{n} H(X_i|U_i), \tag{15.277}$$

$$R_2 \geq \frac{1}{n} \sum_{i=1}^{n} I(Y_i; U_i). \tag{15.278}$$

We now introduce a timesharing random variable Q so that we can rewrite these equations as

$$R_1 \geq \frac{1}{n} \sum_{i=1}^{n} H(X_i|U_i, Q = i) = H(X_Q|U_Q, Q), \tag{15.279}$$

$$R_2 \geq \frac{1}{n} \sum_{i=1}^{n} I(Y_i; U_i|Q = i) = I(Y_Q; U_Q|Q). \tag{15.280}$$

Now since Q is independent of Y_Q (the distribution of Y_i does not depend on i), we have

$$I(Y_Q; U_Q|Q) = I(Y_Q; U_Q, Q) - I(Y_Q; Q) = I(Y_Q; U_Q, Q). \tag{15.281}$$

Now X_Q and Y_Q have the joint distribution $p(x, y)$ in the theorem. Defining $U = (U_Q, Q)$, $X = X_Q$, and $Y = Y_Q$, we have shown the existence of a random variable U such that

$$R_1 \geq H(X|U), \tag{15.282}$$

$$R_2 \geq I(Y; U) \tag{15.283}$$

for any encoding scheme that has a low probability of error. Thus, the converse is proved. □

Before we proceed to the proof of the achievability of this pair of rates, we will need a new lemma about strong typicality and Markov chains. Recall the definition of strong typicality for a triple of random variables X, Y, and Z. A triplet of sequences x^n, y^n, z^n is said to be ϵ-strongly typical if

$$\left| \frac{1}{n} N(a, b, c|x^n, y^n, z^n) - p(a, b, c) \right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|}. \tag{15.284}$$

In particular, this implies that (x^n, y^n) are jointly strongly typical and that (y^n, z^n) are also jointly strongly typical. But the converse is not true: The fact that $(x^n, y^n) \in A_\epsilon^{*(n)}(X, Y)$ and $(y^n, z^n) \in A_\epsilon^{*(n)}(Y, Z)$ does not in general imply that $(x^n, y^n, z^n) \in A_\epsilon^{*(n)}(X, Y, Z)$. But if $X \to Y \to Z$ forms a Markov chain, this implication is true. We state this as a lemma without proof [53, 149].

Lemma 15.8.1 *Let (X, Y, Z) form a Markov chain $X \to Y \to Z$ [i.e., $p(x, y, z) = p(x, y)p(z|y)$]. If for a given $(y^n, z^n) \in A_\epsilon^{*(n)}(Y, Z)$, X^n is drawn $\sim \prod_{i=1}^n p(x_i|y_i)$, then $\Pr\{(X^n, y^n, z^n) \in A_\epsilon^{*(n)}(X, Y, Z)\} > 1 - \epsilon$ for n sufficiently large.*

Remark The theorem is true from the strong law of large numbers if $X^n \sim \prod_{i=1}^n p(x_i|y_i, z_i)$. The Markovity of $X \to Y \to Z$ is used to show that $X^n \sim p(x_i|y_i)$ is sufficient for the same conclusion.

We now outline the proof of achievability in Theorem 15.8.1.

Proof: (*Achievability in Theorem 15.8.1*). Fix $p(u|y)$. Calculate $p(u) = \sum_y p(y)p(u|y)$.

Generation of codebooks: Generate 2^{nR_2} independent codewords of length n, $\mathbf{U}(w_2)$, $w_2 \in \{1, 2, \ldots, 2^{nR_2}\}$ according to $\prod_{i=1}^n p(u_i)$. Randomly bin all the X^n sequences into 2^{nR_1} bins by independently generating an index b distributed uniformly on $\{1, 2, \ldots, 2^{nR_1}\}$ for each X^n. Let $B(i)$ denote the set of X^n sequences allotted to bin i.

Encoding: The X sender sends the index i of the bin in which X^n falls.

The Y sender looks for an index s such that $(Y^n, U^n(s)) \in A_\epsilon^{*(n)}(Y, U)$. If there is more than one such s, it sends the least. If there is no such $U^n(s)$ in the codebook, it sends $s = 1$.

Decoding: The receiver looks for a unique $X^n \in B(i)$ such that $(X^n, U^n(s)) \in A_\epsilon^{*(n)}(X, U)$. If there is none or more than one, it declares an error.

Analysis of the probability of error: The various sources of error are as follows:

1. The pair (X^n, Y^n) generated by the source is not typical. The probability of this is small if n is large. Hence, without loss of generality, we can condition on the event that the source produces a particular typical sequence $(x^n, y^n) \in A_\epsilon^{*(n)}$.
2. The sequence Y^n is typical, but there does not exist a $U^n(s)$ in the codebook that is jointly typical with it. The probability of this is small from the arguments of Section 10.6, where we showed that if there are enough codewords; that is, if

$$R_2 > I(Y; U), \tag{15.285}$$

   we are very likely to find a codeword that is jointly strongly typical with the given source sequence.
3. The codeword $U^n(s)$ is jointly typical with y^n but not with x^n. But by Lemma 15.8.1, the probability of this is small since $X \to Y \to U$ forms a Markov chain.
4. We also have an error if there exists another typical $X^n \in B(i)$ which is jointly typical with $U^n(s)$. The probability that any other X^n is jointly typical with $U^n(s)$ is less than $2^{-n(I(U;X)-3\epsilon)}$, and therefore the probability of this kind of error is bounded above by

$$|B(i) \cap A_\epsilon^{*(n)}(X)|2^{-n(I(X;U)-3\epsilon)} \leq 2^{n(H(X)+\epsilon)}2^{-nR_1}2^{-n(I(X;U)-3\epsilon)}, \tag{15.286}$$

   which goes to 0 if $R_1 > H(X|U)$.

Hence, it is likely that the actual source sequence X^n is jointly typical with $U^n(s)$ and that no other typical source sequence in the same bin is also jointly typical with $U^n(s)$. We can achieve an arbitrarily low probability of error with an appropriate choice of n and ϵ, and this completes the proof of achievability. □

15.9 RATE DISTORTION WITH SIDE INFORMATION

We know that $R(D)$ bits are sufficient to describe X within distortion D. We now ask how many bits are required given side information Y.

We begin with a few definitions. Let (X_i, Y_i) be i.i.d. $\sim p(x, y)$ and encoded as shown in Figure 15.33.

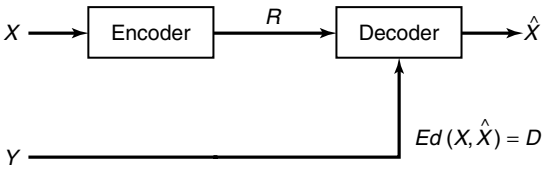


FIGURE 15.33. Rate distortion with side information.

Definition The *rate distortion function with side information* $R_Y(D)$ is defined as the minimum rate required to achieve distortion D if the side information Y is available to the decoder. Precisely, $R_Y(D)$ is the infimum of rates R such that there exist maps $i_n : \mathcal{X}^n \to \{1, \ldots, 2^{nR}\}$, $g_n : \mathcal{Y}^n \times \{1, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n$ such that

$$\limsup_{n\to\infty} Ed(X^n, g_n(Y^n, i_n(X^n))) \le D. \tag{15.287}$$

Clearly, since the side information can only help, we have $R_Y(D) \le R(D)$. For the case of zero distortion, this is the Slepian–Wolf problem and we will need $H(X|Y)$ bits. Hence, $R_Y(0) = H(X|Y)$. We wish to determine the entire curve $R_Y(D)$. The result can be expressed in the following theorem.

Theorem 15.9.1 (*Rate distortion with side information (Wyner and Ziv)*) *Let* (X, Y) *be drawn i.i.d.* $\sim p(x, y)$ *and let* $d(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^n d(x_i, \hat{x}_i)$ *be given. The rate distortion function with side information is*

$$R_Y(D) = \min_{p(w|x)} \min_f (I(X; W) - I(Y; W)) \tag{15.288}$$

where the minimization is over all functions $f : \mathcal{Y} \times \mathcal{W} \to \hat{\mathcal{X}}$ *and conditional probability mass functions* $p(w|x)$, $|\mathcal{W}| \le |\mathcal{X}| + 1$, *such that*

$$\sum_x \sum_w \sum_y p(x, y)p(w|x)d(x, f(y, w)) \le D. \tag{15.289}$$

The function f in the theorem corresponds to the decoding map that maps the encoded version of the X symbols and the side information Y to the output alphabet. We minimize over all conditional distributions on W and functions f such that the expected distortion for the joint distribution is less than D.

We first prove the converse after considering some of the properties of the function $R_Y(D)$ defined in (15.288).

Lemma 15.9.1 *The rate distortion function with side information $R_Y(D)$ defined in (15.288) is a nonincreasing convex function of D.*

Proof: The monotonicity of $R_Y(D)$ follows immediately from the fact that the domain of minimization in the definition of $R_Y(D)$ increases with D. As in the case of rate distortion without side information, we expect $R_Y(D)$ to be convex. However, the proof of convexity is more involved because of the double rather than single minimization in the definition of $R_Y(D)$ in (15.288). We outline the proof here.

Let D_1 and D_2 be two values of the distortion and let W_1, f_1 and W_2, f_2 be the corresponding random variables and functions that achieve the minima in the definitions of $R_Y(D_1)$ and $R_Y(D_2)$, respectively. Let Q be a random variable independent of X, Y, W_1, and W_2 which takes on the value 1 with probability λ and the value 2 with probability $1-\lambda$.

Define $W = (Q, W_Q)$ and let $f(W, Y) = f_Q(W_Q, Y)$. Specifically, $f(W, Y) = f_1(W_1, Y)$ with probability λ and $f(W, Y) = f_2(W_2, Y)$ with probability $1-\lambda$. Then the distortion becomes

$$D = Ed(X, \hat{X}) \tag{15.290}$$

$$= \lambda Ed(X, f_1(W_1, Y)) + (1-\lambda)Ed(X, f_2(W_2, Y)) \tag{15.291}$$

$$= \lambda D_1 + (1-\lambda)D_2, \tag{15.292}$$

and (15.288) becomes

$$I(W; X) - I(W; Y) = H(X) - H(X|W) - H(Y) + H(Y|W) \tag{15.293}$$

$$= H(X) - H(X|W_Q, Q) - H(Y) + H(Y|W_Q, Q) \tag{15.294}$$

$$= H(X) - \lambda H(X|W_1) - (1-\lambda)H(X|W_2) - H(Y) + \lambda H(Y|W_1) + (1-\lambda)H(Y|W_2) \tag{15.295}$$

$$= \lambda\,(I(W_1, X) - I(W_1; Y)) + (1-\lambda)\,(I(W_2, X) - I(W_2; Y))\,, \tag{15.296}$$

and hence

$$R_Y(D) = \min_{U: Ed \le D} (I(U; X) - I(U; Y)) \tag{15.297}$$

$$\le I(W; X) - I(W; Y) \tag{15.298}$$

$$= \lambda (I(W_1, X) - I(W_1; Y)) + (1 - \lambda) (I(W_2, X) - I(W_2; Y))$$
$$= \lambda R_Y(D_1) + (1 - \lambda) R_Y(D_2), \tag{15.299}$$

proving the convexity of $R_Y(D)$. □

We are now in a position to prove the converse to the conditional rate distortion theorem.

Proof: (*Converse to Theorem 15.9.1*). Consider any rate distortion code with side information. Let the encoding function be $f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$. Let the decoding function be $g_n : \mathcal{Y}^n \times \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n$, and let $g_{ni} : \mathcal{Y}^n \times \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}$ denote the ith symbol produced by the decoding function. Let $T = f_n(X^n)$ denote the encoded version of X^n. We must show that if $Ed(X^n, g_n(Y^n, f_n(X^n))) \leq D$, then $R \geq R_Y(D)$. We have the following chain of inequalities:

$$nR \overset{(a)}{\geq} H(T) \tag{15.300}$$
$$\overset{(b)}{\geq} H(T|Y^n) \tag{15.301}$$
$$\geq I(X^n; T|Y^n) \tag{15.302}$$
$$\overset{(c)}{=} \sum_{i=1}^{n} I(X_i; T|Y^n, X^{i-1}) \tag{15.303}$$
$$= \sum_{i=1}^{n} H(X_i|Y^n, X^{i-1}) - H(X_i|T, Y^n, X^{i-1}) \tag{15.304}$$
$$\overset{(d)}{=} \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|T, Y^{i-1}, Y_i, Y_{i+1}^n, X^{i-1}) \tag{15.305}$$
$$\overset{(e)}{\geq} \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|T, Y^{i-1}, Y_i, Y_{i+1}^n) \tag{15.306}$$
$$\overset{(f)}{=} \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|W_i, Y_i) \tag{15.307}$$
$$\overset{(g)}{=} \sum_{i=1}^{n} I(X_i; W_i|Y_i) \tag{15.308}$$

$$= \sum_{i=1}^{n} H(W_i|Y_i) - H(W_i|X_i, Y_i) \tag{15.309}$$

$$\stackrel{\text{(h)}}{=} \sum_{i=1}^{n} H(W_i|Y_i) - H(W_i|X_i) \tag{15.310}$$

$$= \sum_{i=1}^{n} H(W_i) - H(W_i|X_i) - H(W_i) + H(W_i|Y_i) \tag{15.311}$$

$$= \sum_{i=1}^{n} I(W_i; X_i) - I(W_i; Y_i) \tag{15.312}$$

$$\stackrel{\text{(i)}}{\geq} \sum_{i=1}^{n} R_Y(Ed(X_i, g'_{ni}(W_i, Y_i))) \tag{15.313}$$

$$= n\frac{1}{n}\sum_{i=1}^{n} R_Y(Ed(X_i, g'_{ni}(W_i, Y_i))) \tag{15.314}$$

$$\stackrel{\text{(j)}}{\geq} nR_Y\left(\frac{1}{n}\sum_{i=1}^{n} Ed(X_i, g'_{ni}(W_i, Y_i))\right) \tag{15.315}$$

$$\stackrel{\text{(k)}}{\geq} nR_Y(D), \tag{15.316}$$

where

(a) follows from the fact that the range of T is $\{1, 2, \ldots, 2^{nR}\}$
(b) follows from the fact that conditioning reduces entropy
(c) follows from the chain rule for mutual information
(d) follows from the fact that X_i is independent of the past and future Y's and X's given Y_i
(e) follows from the fact that conditioning reduces entropy
(f) follows by defining $W_i = (T, Y^{i-1}, Y_{i+1}^n)$
(g) follows from the definition of mutual information
(h) follows from the fact that since Y_i depends only on X_i and is conditionally independent of T and the past and future Y's, $W_i \to X_i \to Y_i$ forms a Markov chain
(i) follows from the definition of the (information) conditional rate distortion function since $\hat{X}_i = g_{ni}(T, Y^n) \stackrel{\triangle}{=} g'_{ni}(W_i, Y_i)$, and hence $I(W_i; X_i) - I(W_i; Y_i) \geq \min_{W:Ed(X,\hat{X})\leq D_i} I(W; X) - I(W; Y) = R_Y(D_i)$

(j) follows from Jensen's inequality and the convexity of the conditional rate distortion function (Lemma 15.9.1)
(k) follows from the definition of $D = E[\frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i)]$ □

It is easy to see the parallels between this converse and the converse for rate distortion without side information (Section 10.4). The proof of achievability is also parallel to the proof of the rate distortion theorem using strong typicality. However, instead of sending the index of the codeword that is jointly typical with the source, we divide these codewords into bins and send the bin index instead. If the number of codewords in each bin is small enough, the side information can be used to isolate the particular codeword in the bin at the receiver. Hence again we are combining random binning with rate distortion encoding to find a jointly typical reproduction codeword. We outline the details of the proof below.

Proof: (*Achievability of Theorem 15.9.1*). Fix $p(w|x)$ and the function $f(w, y)$. Calculate $p(w) = \sum_x p(x)p(w|x)$.

Generation of codebook: Let $R_1 = I(X; W) + \epsilon$. Generate 2^{nR} i.i.d. codewords $W^n(s) \sim \prod_{i=1}^{n} p(w_i)$, and index them by $s \in \{1, 2, \ldots, 2^{nR_1}\}$. Let $R_2 = I(X; W) - I(Y; W) + 5\epsilon$. Randomly assign the indices $s \in \{1, 2, \ldots, 2^{nR_1}\}$ to one of 2^{nR_2} bins using a uniform distribution over the bins. Let $B(i)$ denote the indices assigned to bin i. There are approximately $2^{n(R_1-R_2)}$ indices in each bin.

Encoding: Given a source sequence X^n, the encoder looks for a codeword $W^n(s)$ such that $(X^n, W^n(s)) \in A_\epsilon^{*(n)}$. If there is no such W^n, the encoder sets $s = 1$. If there is more than one such s, the encoder uses the lowest s. The encoder sends the index of the bin in which s belongs.

Decoding: The decoder looks for a $W^n(s)$ such that $s \in B(i)$ and $(W^n(s), Y^n) \in A_\epsilon^{*(n)}$. If he finds a unique s, he then calculates $\hat{X}^n$, where $\hat{X}_i = f(W_i, Y_i)$. If he does not find any such s or more than one such s, he sets $\hat{X}^n = \hat{x}^n$, where $\hat{x}^n$ is an arbitrary sequence in $\hat{\mathcal{X}}^n$. It does not matter which default sequence is used; we will show that the probability of this event is small.

Analysis of the probability of error: As usual, we have various error events:

1. The pair $(X^n, Y^n) \notin A_\epsilon^{*(n)}$. The probability of this event is small for large enough n by the weak law of large numbers.
2. The sequence X^n is typical, but there does not exist an s such that $(X^n, W^n(s)) \in A_\epsilon^{*(n)}$. As in the proof of the rate distortion theorem,

the probability of this event is small if

$$R_1 > I(W; X). \tag{15.317}$$

3. The pair of sequences $(X^n, W^n(s)) \in A_\epsilon^{*(n)}$ but $(W^n(s), Y^n) \notin A_\epsilon^{*(n)}$ (i.e., the codeword is not jointly typical with the Y^n sequence). By the Markov lemma (Lemma 15.8.1), the probability of this event is small if n is large enough.
4. There exists another s' with the same bin index such that $(W^n(s'), Y^n) \in A_\epsilon^{*(n)}$. Since the probability that a randomly chosen W^n is jointly typical with Y^n is $\approx 2^{-nI(Y;W)}$, the probability that there is another W^n in the same bin that is typical with Y^n is bounded by the number of codewords in the bin times the probability of joint typicality, that is,

$$\Pr(\exists s' \in B(i) : (W^n(s'), Y^n) \in A_\epsilon^{*(n)}) \leq 2^{n(R_1 - R_2)} 2^{-n(I(W;Y) - 3\epsilon)}, \tag{15.318}$$

which goes to zero since $R_1 - R_2 < I(Y; W) - 3\epsilon$.
5. If the index s is decoded correctly, $(X^n, W^n(s)) \in A_\epsilon^{*(n)}$. By item 1 we can assume that $(X^n, Y^n) \in A_\epsilon^{*(n)}$. Thus, by the Markov lemma, we have $(X^n, Y^n, W^n) \in A_\epsilon^{*(n)}$ and therefore the empirical joint distribution is close to the original distribution $p(x, y)p(w|x)$ that we started with, and hence $(X^n, \hat{X}^n)$ will have a joint distribution that is close to the distribution that achieves distortion D.

Hence with high probability, the decoder will produce $\hat{X}^n$ such that the distortion between X^n and $\hat{X}^n$ is close to nD. This completes the proof of the theorem. □

The reader is referred to Wyner and Ziv [574] for details of the proof. After the discussion of the various situations of compressing distributed data, it might be expected that the problem is almost completely solved, but unfortunately, this is not true. An immediate generalization of all the above problems is the rate distortion problem for correlated sources, illustrated in Figure 15.34. This is essentially the Slepian–Wolf problem with distortion in both X and Y. It is easy to see that the three distributed source coding problems considered above are all special cases of this setup. Unlike the earlier problems, though, this problem has not yet been solved and the general rate distortion region remains unknown.

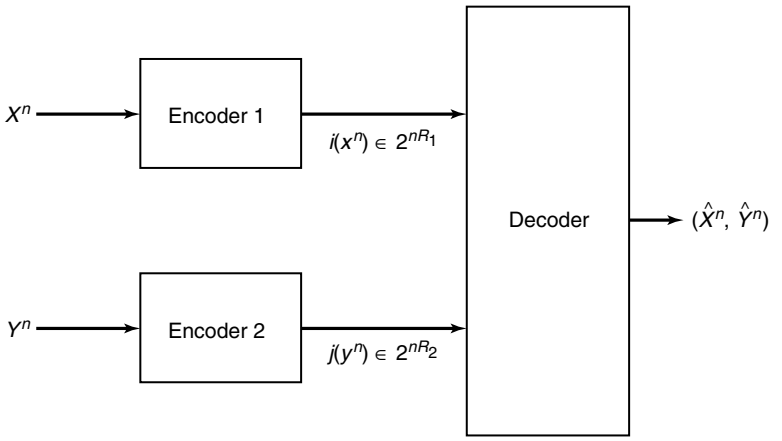


FIGURE 15.34. Rate distortion for two correlated sources.

15.10 GENERAL MULTITERMINAL NETWORKS

We conclude this chapter by considering a general multiterminal network of senders and receivers and deriving some bounds on the rates achievable for communication in such a network. A general multiterminal network is illustrated in Figure 15.35. In this section, superscripts denote node indices and subscripts denote time indices. There are m nodes, and node i has an associated transmitted variable $X^{(i)}$ and a received variable $Y^{(i)}$.

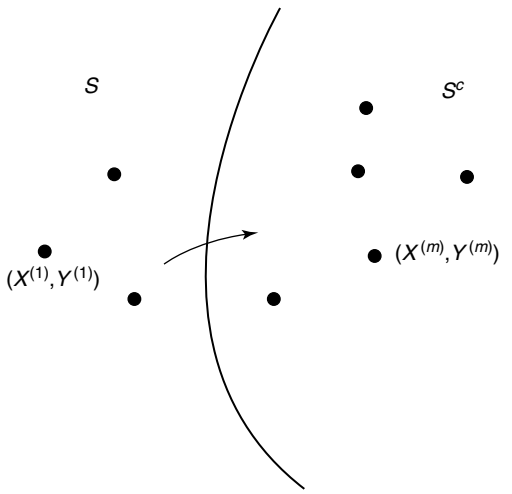


FIGURE 15.35. General multiterminal network.

The node i sends information at rate $R^{(ij)}$ to node j. We assume that all the messages $W^{(ij)}$ being sent from node i to node j are independent and uniformly distributed over their respective ranges $\{1, 2, \ldots, 2^{nR^{(ij)}}\}$.

The channel is represented by the channel transition function $p(y^{(1)}, \ldots, y^{(m)}|x^{(1)}, \ldots, x^{(m)})$, which is the conditional probability mass function of the outputs given the inputs. This probability transition function captures the effects of the noise and the interference in the network. The channel is assumed to be memoryless (i.e., the outputs at any time instant depend only the current inputs and are conditionally independent of the past inputs).

Corresponding to each transmitter–receiver node pair is a message $W^{(ij)} \in \{1, 2, \ldots, 2^{nR^{(ij)}}\}$. The input symbol $X^{(i)}$ at node i depends on $W^{(ij)}$, $j \in \{1, \ldots, m\}$ and also on the past values of the received symbol $Y^{(i)}$ at node i. Hence, an encoding scheme of block length n consists of a set of encoding and decoding functions, one for each node:

- *Encoders:* $X_k^{(i)}(W^{(i1)}, W^{(i2)}, \ldots, W^{(im)}, Y_1^{(i)}, Y_2^{(i)}, \ldots, Y_{k-1}^{(i)})$, $k = 1, \ldots, n$. The encoder maps the messages and past received symbols into the symbol $X_k^{(i)}$ transmitted at time k.
- *Decoders:* $\hat{W}^{(ji)}\left(Y_1^{(i)}, \ldots, Y_n^{(i)}, W^{(i1)}, \ldots, W^{(im)}\right)$, $j = 1, 2, \ldots, m$. The decoder j at node i maps the received symbols in each block and his own transmitted information to form estimates of the messages intended for him from node j, $j = 1, 2, \ldots, m$.

Associated with every pair of nodes is a rate and a corresponding probability of error that the message will not be decoded correctly,

$$P_e^{(n)(ij)} = \Pr\left(\hat{W}^{(ij)}\left(\mathbf{Y}^{(j)}, W^{(j1)}, \ldots, W^{(jm)}\right) \neq W^{(ij)}\right), \tag{15.319}$$

where $P_e^{(n)(ij)}$ is defined under the assumption that all the messages are independent and distributed uniformly over their respective ranges.

A set of rates $\{R^{(ij)}\}$ is said to be *achievable* if there exist encoders and decoders with block length n with $P_e^{(n)(ij)} \to 0$ as $n \to \infty$ for all $i, j \in \{1, 2, \ldots, m\}$. We use this formulation to derive an upper bound on the flow of information in any multiterminal network. We divide the nodes into two sets, S and the complement S^c. We now bound the rate of flow of information from nodes in S to nodes in S^c. See [514]

Theorem 15.10.1 *If the information rates* $\{R^{(ij)}\}$ *are achievable, there exists some joint probability distribution* $p(x^{(1)}, x^{(2)}, \ldots, x^{(m)})$ *such that*

$$\sum_{i \in S, j \in S^c} R^{(ij)} \leq I(X^{(S)}; Y^{(S^c)} | X^{(S^c)}) \tag{15.320}$$

for all $S \subset \{1, 2, \ldots, m\}$. *Thus, the total rate of flow of information across cut sets is bounded by the conditional mutual information.*

Proof: The proof follows the same lines as the proof of the converse for the multiple access channel. Let $T = \{(i, j) : i \in S, j \in S^c\}$ be the set of links that cross from S to S^c, and let T^c be all the other links in the network. Then

$$n \sum_{i \in S, j \in S^c} R^{(ij)} \tag{15.321}$$

$$\stackrel{(a)}{=} \sum_{i \in S, j \in S^c} H\left(W^{(ij)}\right) \tag{15.322}$$

$$\stackrel{(b)}{=} H\left(W^{(T)}\right) \tag{15.323}$$

$$\stackrel{(c)}{=} H\left(W^{(T)} | W^{(T^c)}\right) \tag{15.324}$$

$$= I\left(W^{(T)}; Y_1^{(S^c)}, \ldots, Y_n^{(S^c)} | W^{(T^c)}\right) \tag{15.325}$$

$$+ H\left(W^{(T)} | Y_1^{(S^c)}, \ldots, Y_n^{(S^c)}, W^{(T^c)}\right) \tag{15.326}$$

$$\stackrel{(d)}{\leq} I\left(W^{(T)}; Y_1^{(S^c)}, \ldots, Y_n^{(S^c)} | W^{(T^c)}\right) + n\epsilon_n \tag{15.327}$$

$$\stackrel{(e)}{=} \sum_{k=1}^{n} I\left(W^{(T)}; Y_k^{(S^c)} | Y_1^{(S^c)}, \ldots, Y_{k-1}^{(S^c)}, W^{(T^c)}\right) + n\epsilon_n \tag{15.328}$$

$$\stackrel{(f)}{=} \sum_{k=1}^{n} H\left(Y_k^{(S^c)} | Y_1^{(S^c)}, \ldots, Y_{k-1}^{(S^c)}, W^{(T^c)}\right)$$

$$- H\left(Y_k^{(S^c)} | Y_1^{(S^c)}, \ldots, Y_{k-1}^{(S^c)}, W^{(T^c)}, W^{(T)}\right) + n\epsilon_n \tag{15.329}$$

$$\stackrel{(g)}{\leq} \sum_{k=1}^{n} H\left(Y_k^{(S^c)}|Y_1^{(S^c)}, \ldots, Y_{k-1}^{(S^c)}, W^{(T^c)}, X_k^{(S^c)}\right) - H\left(Y_k^{(S^c)}|Y_1^{(S^c)}, \ldots, Y_{k-1}^{(S^c)}, W^{(T^c)}, W^{(T)}, X_k^{(S)}, X_k^{(S^c)}\right) + n\epsilon_n \tag{15.330}$$

$$\stackrel{(h)}{\leq} \sum_{k=1}^{n} H\left(Y_k^{(S^c)}|X_k^{(S^c)}\right) - H\left(Y_k^{(S^c)}|X_k^{(S^c)}, X_k^{(S)}\right) + n\epsilon_n \tag{15.331}$$

$$= \sum_{k=1}^{n} I\left(X_k^{(S)}; Y_k^{(S^c)}|X_k^{(S^c)}\right) + n\epsilon_n \tag{15.332}$$

$$\stackrel{(i)}{=} n\frac{1}{n}\sum_{k=1}^{n} I\left(X_Q^{(S)}; Y_Q^{(S^c)}|X_Q^{(S^c)}, Q = k\right) + n\epsilon_n \tag{15.333}$$

$$\stackrel{(j)}{=} nI\left(X_Q^{(S)}; Y_Q^{(S^c)}|X_Q^{(S^c)}, Q\right) + n\epsilon_n \tag{15.334}$$

$$= n\left(H\left(Y_Q^{(S^c)}|X_Q^{(S^c)}, Q\right) - H\left(Y_Q^{(S^c)}|X_Q^{(S)}, X_Q^{(S^c)}, Q\right)\right) + n\epsilon_n \tag{15.335}$$

$$\stackrel{(k)}{\leq} n\left(H\left(Y_Q^{(S^c)}|X_Q^{(S^c)}\right) - H\left(Y_Q^{(S^c)}|X_Q^{(S)}, X_Q^{(S^c)}, Q\right)\right) + n\epsilon_n \tag{15.336}$$

$$\stackrel{(l)}{=} n\left(H\left(Y_Q^{(S^c)}|X_Q^{(S^c)}\right) - H\left(Y_Q^{(S^c)}|X_Q^{(S)}, X_Q^{(S^c)}\right)\right) + n\epsilon_n \tag{15.337}$$

$$= nI\left(X_Q^{(S)}; Y_Q^{(S^c)}|X_Q^{(S^c)}\right) + n\epsilon_n, \tag{15.338}$$

where

(a) follows from the fact that the messages $W^{(ij)}$ are uniformly distributed over their respective ranges $\{1, 2, \ldots, 2^{nR^{(ij)}}\}$
(b) follows from the definition of $W^{(T)} = \{W^{(ij)} : i \in S, j \in S^c\}$ and the fact that the messages are independent
(c) follows from the independence of the messages for T and T^c
(d) follows from Fano's inequality since the messages $W^{(T)}$ can be decoded from $Y^{(S)}$ and $W^{(T^c)}$
(e) is the chain rule for mutual information
(f) follows from the definition of mutual information
(g) follows from the fact that $X_k^{(S^c)}$ is a function of the past received symbols $Y^{(S^c)}$ and the messages $W^{(T^c)}$ and the fact that adding conditioning reduces the second term

(h) follows from the fact that $Y_k^{(S^c)}$ depends only on the current input symbols $X_k^{(S)}$ and $X_k^{(S^c)}$
(i) follows after we introduce a new timesharing random variable Q distributed uniformly on $\{1, 2, \ldots, n\}$
(j) follows from the definition of mutual information
(k) follows from the fact that conditioning reduces entropy
(l) follows from the fact that $Y_Q^{(S^c)}$ depends only on the inputs $X_Q^{(S)}$ and $X_Q^{(S^c)}$ and is conditionally independent of Q

Thus, there exist random variables $X^{(S)}$ and $X^{(S^c)}$ with some arbitrary joint distribution that satisfy the inequalities of the theorem. □

The theorem has a simple max-flow min-cut interpretation. The rate of flow of information across any boundary is less than the mutual information between the inputs on one side of the boundary and the outputs on the other side, conditioned on the inputs on the other side.

The problem of information flow in networks would be solved if the bounds of the theorem were achievable. But unfortunately, these bounds are not achievable even for some simple channels. We now apply these bounds to a few of the channels that we considered earlier.

- *Multiple-access channel.* The multiple access channel is a network with many input nodes and one output node. For the case of a two-user multiple-access channel, the bounds of Theorem 15.10.1 reduce to

$$R_1 \le I(X_1; Y|X_2), \tag{15.339}$$

$$R_2 \le I(X_2; Y|X_1), \tag{15.340}$$

$$R_1 + R_2 \le I(X_1, X_2; Y) \tag{15.341}$$

  for some joint distribution $p(x_1, x_2)p(y|x_1, x_2)$. These bounds coincide with the capacity region if we restrict the input distribution to be a product distribution and take the convex hull (Theorem 15.3.1).
- *Relay channel.* For the relay channel, these bounds give the upper bound of Theorem 15.7.1 with different choices of subsets as shown in Figure 15.36. Thus,

$$C \le \sup_{p(x,x_1)} \min\{I(X, X_1; Y), I(X; Y, Y_1|X_1)\}. \tag{15.342}$$

  This upper bound is the capacity of a physically degraded relay channel and for the relay channel with feedback [127].

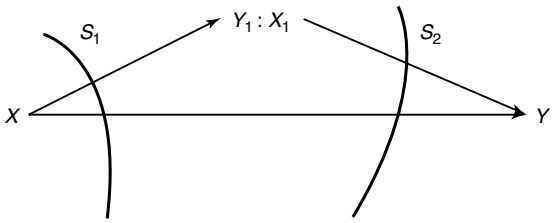


FIGURE 15.36. Relay channel.

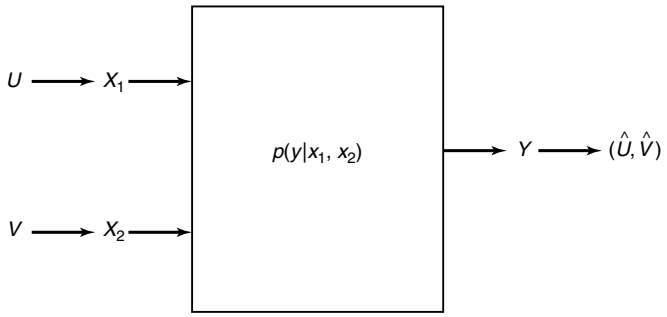


FIGURE 15.37. Transmission of correlated sources over a multiple-access channel.

To complement our discussion of a general network, we should mention two features of single-user channels that do not apply to a multiuser network.

- *Source–channel separation theorem.* In Section 7.13 we discussed the source–channel separation theorem, which proves that we can transmit the source noiselessly over the channel if and only if the entropy rate is less than the channel capacity. This allows us to characterize a source by a single number (the entropy rate) and the channel by a single number (the capacity). What about the multiuser case? We would expect that a distributed source could be transmitted over a channel if and only if the rate region for the noiseless coding of the source lay within the capacity region of the channel. To be specific, consider the transmission of a distributed source over a multiple-access channel, as shown in Figure 15.37. Combining the results of Slepian–Wolf encoding with the capacity results for the multiple-access channel, we can show that we can transmit the source over the channel and recover it with a low probability of error if

$$H(U|V) \leq I(X_1; Y|X_2, Q), \tag{15.343}$$

$$H(V|U) \leq I(X_2; Y|X_1, Q), \tag{15.344}$$

$$H(U, V) \leq I(X_1, X_2; Y|Q) \tag{15.345}$$

for some distribution $p(q)p(x_1|q)p(x_2|q)p(y|x_1, x_2)$. This condition is equivalent to saying that the Slepian–Wolf rate region of the source has a nonempty intersection with the capacity region of the multiple-access channel.

But is this condition also necessary? No, as a simple example illustrates. Consider the transmission of the source of Example 15.4.2 over the binary erasure multiple-access channel (Example 15.3.3). The Slepian–Wolf region does not intersect the capacity region, yet it is simple to devise a scheme that allows the source to be transmitted over the channel. We just let $X_1 = U$ and $X_2 = V$, and the value of Y will tell us the pair (U, V) with no error. Thus, the conditions (15.345) are not necessary.

The reason for the failure of the source–channel separation theorem lies in the fact that the capacity of the multiple-access channel increases with the correlation between the inputs of the channel. Therefore, to maximize the capacity, one should preserve the correlation between the inputs of the channel. Slepian–Wolf encoding, on the other hand, gets rid of the correlation. Cover et al. [129] proposed an achievable region for transmission of a correlated source over a multiple access channel based on the idea of preserving the correlation. Han and Costa [273] have proposed a similar region for the transmission of a correlated source over a broadcast channel.

- *Capacity regions with feedback.* Theorem 7.12.1 shows that feedback does not increase the capacity of a single-user discrete memoryless channel. For channels with memory, on the other hand, feedback enables the sender to predict something about the noise and to combat it more effectively, thus increasing capacity.

  What about multiuser channels? Rather surprisingly, feedback does increase the capacity region of multiuser channels, even when the channels are memoryless. This was first shown by Gaarder and Wolf [220], who showed how feedback helps increase the capacity of the binary erasure multiple-access channel. In essence, feedback from the receiver to the two senders acts as a separate channel between the two senders. The senders can decode each other's transmissions before the receiver does. They then cooperate to resolve the uncertainty at the receiver, sending information at the higher cooperative capacity rather than the noncooperative capacity. Using this scheme, Cover and Leung [133] established an achievable region for a multiple-access

channel with feedback. Willems [557] showed that this region was the capacity for a class of multiple-access channels that included the binary erasure multiple-access channel. Ozarow [410] established the capacity region for a two-user Gaussian multiple-access channel. The problem of finding the capacity region for a multiple-access channel with feedback is closely related to the capacity of a two-way channel with a common output.

There is as yet no unified theory of network information flow. But there can be no doubt that a complete theory of communication networks would have wide implications for the theory of communication and computation.

SUMMARY

Multiple-access channel. The capacity of a multiple-access channel $(\mathcal{X}_1 \times \mathcal{X}_2, p(y|x_1, x_2), \mathcal{Y})$ is the closure of the convex hull of all (R_1, R_2) satisfying

$$R_1 < I(X_1; Y|X_2), \tag{15.346}$$

$$R_2 < I(X_2; Y|X_1), \tag{15.347}$$

$$R_1 + R_2 < I(X_1, X_2; Y) \tag{15.348}$$

for some distribution $p_1(x_1)p_2(x_2)$ on $\mathcal{X}_1 \times \mathcal{X}_2$.
The capacity region of the m-user multiple-access channel is the closure of the convex hull of the rate vectors satisfying

$$R(S) \leq I(X(S); Y|X(S^c)) \quad \text{for all } S \subseteq \{1, 2, \ldots, m\} \tag{15.349}$$

for some product distribution $p_1(x_1)p_2(x_2)\cdots p_m(x_m)$.

Gaussian multiple-access channel. The capacity region of a two-user Gaussian multiple-access channel is

$$R_1 \leq C\left(\frac{P_1}{N}\right), \tag{15.350}$$

$$R_2 \leq C\left(\frac{P_2}{N}\right), \tag{15.351}$$

$$R_1 + R_2 \le C\left(\frac{P_1 + P_2}{N}\right), \tag{15.352}$$

where

$$C(x) = \frac{1}{2}\log(1 + x). \tag{15.353}$$

Slepian–Wolf coding. Correlated sources X and Y can be described separately at rates R_1 and R_2 and recovered with arbitrarily low probability of error by a common decoder if and only if

$$R_1 \ge H(X|Y), \tag{15.354}$$

$$R_2 \ge H(Y|X), \tag{15.355}$$

$$R_1 + R_2 \ge H(X, Y). \tag{15.356}$$

Broadcast channels. The capacity region of the degraded broadcast channel $X \to Y_1 \to Y_2$ is the convex hull of the closure of all (R_1, R_2) satisfying

$$R_2 \le I(U; Y_2), \tag{15.357}$$

$$R_1 \le I(X; Y_1|U) \tag{15.358}$$

for some joint distribution $p(u)p(x|u)p(y_1, y_2|x)$.

Relay channel. The capacity C of the physically degraded relay channel $p(y, y_1|x, x_1)$ is given by

$$C = \sup_{p(x,x_1)} \min\{I(X, X_1; Y), I(X; Y_1|X_1)\}, \tag{15.359}$$

where the supremum is over all joint distributions on $\mathcal{X} \times \mathcal{X}_1$.

Source coding with side information. Let $(X, Y) \sim p(x, y)$. If Y is encoded at rate R_2 and X is encoded at rate R_1, we can recover X with an arbitrarily small probability of error iff

$$R_1 \ge H(X|U), \tag{15.360}$$

$$R_2 \ge I(Y; U) \tag{15.361}$$

for some distribution $p(y, u)$ such that $X \to Y \to U$.

Rate distortion with side information. Let $(X, Y) \sim p(x, y)$. The rate distortion function with side information is given by

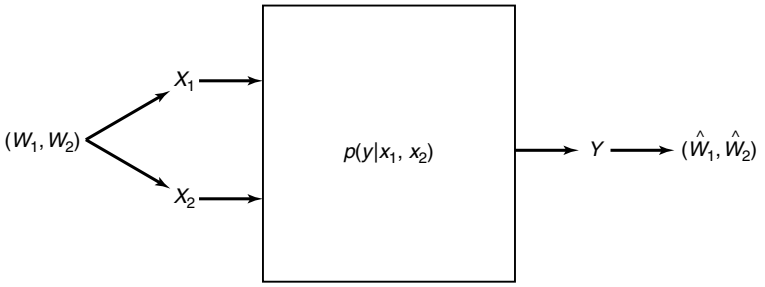
$$R_Y(D) = \min_{p(w|x)} \min_{f:\mathcal{Y}\times\mathcal{W}\to\hat{\mathcal{X}}} I(X; W) - I(Y; W), \tag{15.362}$$

where the minimization is over all functions f and conditional distributions $p(w|x)$, $|\mathcal{W}| \le |\mathcal{X}| + 1$, such that

$$\sum_x \sum_w \sum_y p(x, y)p(w|x)d(x, f(y, w)) \le D. \tag{15.363}$$

PROBLEMS

15.1 *Cooperative capacity of a multiple-access channel*



(a) Suppose that X_1 and X_2 have access to *both* indices $W_1 \in \{1, 2^{nR}\}$, $W_2 \in \{1, 2^{nR_2}\}$. Thus, the codewords $\mathbf{X}_1(W_1, W_2)$, $\mathbf{X}_2(W_1, W_2)$ depend on both indices. Find the capacity region.

(b) Evaluate this region for the binary erasure multiple access channel $Y = X_1 + X_2$, $X_i \in \{0, 1\}$. Compare to the noncooperative region.

15.2 *Capacity of multiple-access channels*. Find the capacity region for each of the following multiple-access channels:

(a) Additive modulo 2 multiple-access channel. $X_1 \in \{0, 1\}$, $X_2 \in \{0, 1\}$, $Y = X_1 \oplus X_2$.

(b) Multiplicative multiple-access channel. $X_1 \in \{-1, 1\}$, $X_2 \in \{-1, 1\}$, $Y = X_1 \cdot X_2$.

15.3 *Cut-set interpretation of capacity region of multiple-access channel*. For the multiple-access channel we know that (R_1, R_2) is achievable if

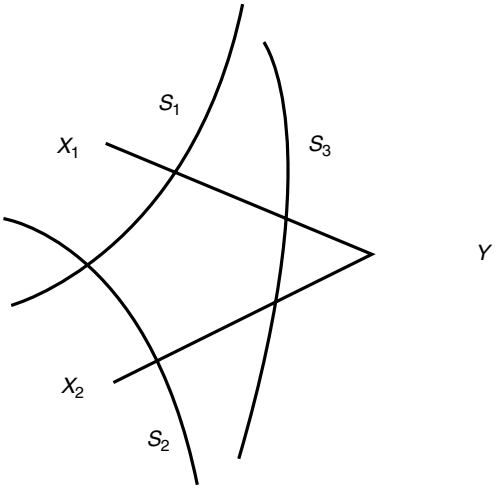
$$R_1 < I(X_1; Y \mid X_2), \tag{15.364}$$

$$R_2 < I(X_2; Y \mid X_1), \tag{15.365}$$

$$R_1 + R_2 < I(X_1, X_2; Y) \tag{15.366}$$

for X_1, X_2 independent. Show, for X_1, X_2 independent that

$$I(X_1; Y \mid X_2) = I(X_1; Y, X_2).$$



Interpret the information bounds as bounds on the rate of flow across cut sets S_1, S_2, and S_3.

15.4 *Gaussian multiple-access channel capacity*. For the AWGN multiple-access channel, prove, using typical sequences, the achievability of any rate pairs (R_1, R_2) satisfying

$$R_1 < \frac{1}{2}\log\left(1 + \frac{P_1}{N}\right), \tag{15.367}$$

$$R_2 < \frac{1}{2}\log\left(1 + \frac{P_2}{N}\right), \tag{15.368}$$

$$R_1 + R_2 < \frac{1}{2}\log\left(1 + \frac{P_1 + P_2}{N}\right). \tag{15.369}$$

The proof extends the proof for the discrete multiple-access channel in the same way as the proof for the single-user Gaussian channel extends the proof for the discrete single-user channel.

15.5 *Converse for the Gaussian multiple-access channel*. Prove the converse for the Gaussian multiple-access channel by extending the converse in the discrete case to take into account the power constraint on the codewords.

15.6 *Unusual multiple-access channel*. Consider the following multiple-access channel: $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y} = \{0, 1\}$. If $(X_1, X_2) = (0, 0)$, then $Y = 0$. If $(X_1, X_2) = (0, 1)$, then $Y = 1$. If $(X_1, X_2) = (1, 0)$, then $Y = 1$. If $(X_1, X_2) = (1, 1)$, then $Y = 0$ with probability $\frac{1}{2}$ and $Y = 1$ with probability $\frac{1}{2}$.

(a) Show that the rate pairs (1,0) and (0,1) are achievable.

(b) Show that for any nondegenerate distribution $p(x_1)p(x_2)$, we have $I(X_1, X_2; Y) < 1$.

(c) Argue that there are points in the capacity region of this multiple-access channel that can only be achieved by time-sharing; that is, there exist achievable rate pairs (R_1, R_2) that lie in the capacity region for the channel but not in the region defined by

$$R_1 \leq I(X_1; Y|X_2), \tag{15.370}$$

$$R_2 \leq I(X_2; Y|X_1), \tag{15.371}$$

$$R_1 + R_2 \leq I(X_1, X_2; Y) \tag{15.372}$$

for any product distribution $p(x_1)p(x_2)$. Hence the operation of convexification strictly enlarges the capacity region. This channel was introduced independently by Csiszár and Körner [149] and Bierbaum and Wallmeier [59].

15.7 *Convexity of capacity region of broadcast channel*. Let $\mathbf{C} \subseteq \mathbf{R}^2$ be the capacity region of all achievable rate pairs $\mathbf{R} = (R_1, R_2)$ for the broadcast channel. Show that $\mathbf{C}$ is a convex set by using a time-sharing argument. Specifically, show that if $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ are achievable, $\lambda\mathbf{R}^{(1)} + (1 - \lambda)\mathbf{R}^{(2)}$ is achievable for $0 \leq \lambda \leq 1$.

15.8 *Slepian–Wolf for deterministically related sources*. Find and sketch the Slepian–Wolf rate region for the simultaneous data compression of (X, Y), where $y = f(x)$ is some deterministic function of x.

15.9 *Slepian–Wolf*. Let X_i be i.i.d. Bernoulli(p). Let Z_i be i.i.d. $\sim$ Bernoulli(r), and let $\mathbf{Z}$ be independent of $\mathbf{X}$. Finally, let $\mathbf{Y} = \mathbf{X} \oplus \mathbf{Z}$ (mod 2 addition). Let $\mathbf{X}$ be described at rate R_1 and $\mathbf{Y}$ be described at rate R_2. What region of rates allows recovery of $\mathbf{X}, \mathbf{Y}$ with probability of error tending to zero?

15.10 *Broadcast capacity depends only on the conditional marginals*. Consider the general broadcast channel $(X, Y_1 \times Y_2, p(y_1, y_2 \mid x))$. Show that the capacity region depends only on $p(y_1 \mid x)$ and $p(y_2 \mid x)$. To do this, for any given $((2^{nR_1}, 2^{nR_2}), n)$ code, let

$$P_1^{(n)} = P\{\hat{W}_1(\mathbf{Y}_1) \neq W_1\}, \tag{15.373}$$

$$P_2^{(n)} = P\{\hat{W}_2(\mathbf{Y}_2) \neq W_2\}, \tag{15.374}$$

$$P^{(n)} = P\{(\hat{W}_1, \hat{W}_2) \neq (W_1, W_2)\}. \tag{15.375}$$

Then show that

$$\max\{P_1^{(n)}, P_2^{(n)}\} \leq P^{(n)} \leq P_1^{(n)} + P_2^{(n)}.$$

The result now follows by a simple argument. (*Remark:* The probability of error $P^{(n)}$ *does* depend on the conditional joint distribution $p(y_1, y_2 \mid x)$. But whether or not $P^{(n)}$ can be driven to zero [at rates (R_1, R_2)] *does not* [except through the conditional marginals $p(y_1 \mid x), p(y_2 \mid x)$] .)

15.11 *Converse for the degraded broadcast channel*. The following chain of inequalities proves the converse for the degraded discrete memoryless broadcast channel. Provide reasons for each of the labeled inequalities.
Setup for converse for degraded broadcast channel capacity:

$$(W_1, W_2)_{\text{indep.}} \to X^n(W_1, W_2) \to Y_1^n \to Y_2^n.$$

- Encoding $f_n : 2^{nR_1} \times 2^{nR_2} \to \mathcal{X}^n$
- Decoding: $g_n : \mathcal{Y}_1^n \to 2^{nR_1}, h_n : \mathcal{Y}_2^n \to 2^{nR_2}$. Let $U_i = (W_2, Y_1^{i-1})$. Then

$$nR_2 \overset{\cdot}{\leq}_{\text{Fano}} I(W_2; Y_2^n) \tag{15.376}$$

$$\overset{\text{(a)}}{=} \sum_{i=1}^{n} I(W_2; Y_{2i} \mid Y_2^{i-1}) \tag{15.377}$$

$$\stackrel{(b)}{=} \sum_i (H(Y_{2i} \mid Y_2^{i-1}) - H(Y_{2i} \mid W_2, Y_2^{i-1})) \quad (15.378)$$

$$\stackrel{(c)}{\leq} \sum_i (H(Y_{2i}) - H(Y_{2i} \mid W_2, Y_2^{i-1}, Y_1^{i-1})) \quad (15.379)$$

$$\stackrel{(d)}{=} \sum_i (H(Y_{2i}) - H(Y_{2i} \mid W_2, Y_1^{i-1})) \quad (15.380)$$

$$\stackrel{(e)}{=} \sum_{i=1}^{n} I(U_i; Y_{2i}). \quad (15.381)$$

Continuation of converse: Give reasons for the labeled inequalities:

$$nR_1 \stackrel{.}{\leq}_{\text{Fano}} I(W_1; Y_1^n) \quad (15.382)$$

$$\stackrel{(f)}{\leq} I(W_1; Y_1^n, W_2) \quad (15.383)$$

$$\stackrel{(g)}{\leq} I(W_1; Y_1^n \mid W_2) \quad (15.384)$$

$$\stackrel{(h)}{=} \sum_{i-1}^{n} I(W_1; Y_{1i} \mid Y_1^{i-1}, W_2) \quad (15.385)$$

$$\stackrel{(i)}{\leq} \sum_{i=1}^{n} I(X_i; Y_{1i} \mid U_i). \quad (15.386)$$

Now let Q be a time-sharing random variable with $\Pr(Q = i) = 1/n$, $i = 1, 2, \ldots, n$. Justify the following:

$$R_1 \leq I(X_Q; Y_{1Q} | U_Q, Q), \quad (15.387)$$

$$R_2 \leq I(U_Q; Y_{2Q} | Q) \quad (15.388)$$

for some distribution $p(q)p(u|q)p(x|u, q)p(y_1, y_2|x)$. By appropriately redefining U, argue that this region is equal to the convex closure of regions of the form

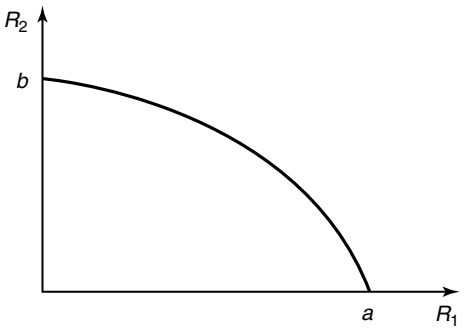
$$R_1 \leq I(X; Y_1|U), \quad (15.389)$$

$$R_2 \leq I(U; Y_2) \quad (15.390)$$

for some joint distribution $p(u)p(x|u)p(y_1, y_2|x)$.

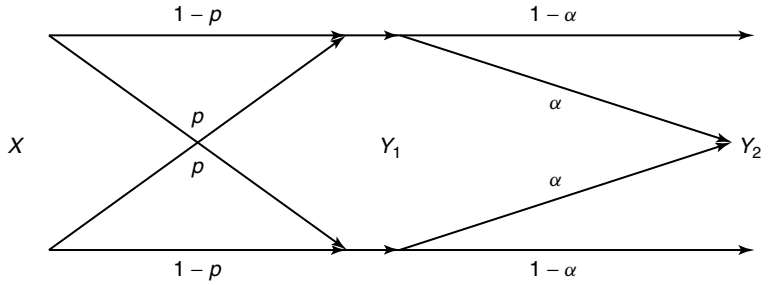
15.12 *Capacity points*.

(a) For the degraded broadcast channel $X \rightarrow Y_1 \rightarrow Y_2$, find the points a and b where the capacity region hits the R_1 and R_2 axes.



(b) Show that $b \leq a$.

15.13 *Degraded broadcast channel*. Find the capacity region for the degraded broadcast channel shown below.



15.14 *Channels with unknown parameters*. We are given a binary symmetric channel with parameter p. The capacity is $C = 1 - H(p)$. Now we change the problem slightly. The receiver knows only that $p \in \{p_1, p_2\}$ (i.e., $p = p_1$ or $p = p_2$, where p_1 and p_2 are given real numbers). The transmitter knows the actual value of p. Devise two codes for use by the transmitter, one to be used if $p = p_1$, the other to be used if $p = p_2$, such that transmission to the receiver can take place at rate $\approx C(p_1)$ if $p = p_1$ and at rate $\approx C(p_2)$ if $p = p_2$. (*Hint:* Devise a method for revealing p to the receiver without affecting the asymptotic rate. Prefixing the codeword by a sequence of 1's of appropriate length should work.)

15.15 *Two-way channel*. Consider the two-way channel shown in Figure 15.6. The outputs Y_1 and Y_2 depend only on the current inputs X_1 and X_2.

(a) By using independently generated codes for the two senders, show that the following rate region is achievable:

$$R_1 < I(X_1; Y_2|X_2), \tag{15.391}$$

$$R_2 < I(X_2; Y_1|X_1) \tag{15.392}$$

for some product distribution $p(x_1)p(x_2)p(y_1, y_2|x_1, x_2)$.

(b) Show that the rates for any code for a two-way channel with arbitrarily small probability of error must satisfy

$$R_1 \leq I(X_1; Y_2|X_2), \tag{15.393}$$

$$R_2 \leq I(X_2; Y_1|X_1) \tag{15.394}$$

for some joint distribution $p(x_1, x_2)p(y_1, y_2|x_1, x_2)$.

The inner and outer bounds on the capacity of the two-way channel are due to Shannon [486]. He also showed that the inner bound and the outer bound do not coincide in the case of the binary multiplying channel $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y}_1 = \mathcal{Y}_2 = \{0, 1\}$, $Y_1 = Y_2 = X_1X_2$. The capacity of the two-way channel is still an open problem.

15.16 *Multiple-access channel*. Let the output Y of a multiple-access channel be given by

$$Y = X_1 + \text{sgn}(X_2),$$

where X_1, X_2 are both real and power limited,

$$\begin{aligned} E(X_1^2) &\leq P_1, \\ E(X_2^2) &\leq P_2, \end{aligned}$$

and $\text{sgn}(x) = \begin{cases} 1, & x > 0, \\ -1, & x \leq 0. \end{cases}$

Note that there is interference but no noise in this channel.

(a) Find the capacity region.

(b) Describe a coding scheme that achieves the capacity region.

15.17 *Slepian–Wolf*. Let (X, Y) have the joint probability mass function $p(x, y)$:

$p(x, y)$	1	2	3
1	α	β	β
2	β	α	β
3	β	β	α

where $\beta = \frac{1}{6} - \frac{\alpha}{2}$. (*Note*: This is a joint, not a conditional, probability mass function.)

(a) Find the Slepian–Wolf rate region for this source.

(b) What is $\Pr\{X = Y\}$ in terms of α?

(c) What is the rate region if $\alpha = \frac{1}{3}$?

(d) What is the rate region if $\alpha = \frac{1}{9}$?

15.18 *Square channel*. What is the capacity of the following multiple-access channel?

$$X_1 \in \{-1, 0, 1\},$$

$$X_2 \in \{-1, 0, 1\},$$

$$Y = X_1^2 + X_2^2.$$

(a) Find the capacity region.

(b) Describe $p^*(x_1)$, $p^*(x_2)$ achieving a point on the boundary of the capacity region.

15.19 *Slepian–Wolf*. Two senders know random variables U_1 and U_2, respectively. Let the random variables (U_1, U_2) have the following joint distribution:

$U_1 \backslash U_2$	0	1	2	$\cdots$	$m-1$
0	α	$\frac{\beta}{m-1}$	$\frac{\beta}{m-1}$	$\cdots$	$\frac{\beta}{m-1}$
1	$\frac{\gamma}{m-1}$	0	0	$\cdots$	0
2	$\frac{\gamma}{m-1}$	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$m-1$	$\frac{\gamma}{m-1}$	0	0	$\cdots$	0

where $\alpha + \beta + \gamma = 1$. Find the region of rates (R_1, R_2) that would allow a common receiver to decode both random variables reliably.

15.20 *Multiple access*

(a) Find the capacity region for the multiple-access channel

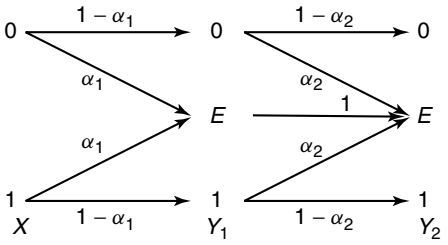
$$Y = X_1^{X_2},$$

where

$$X_1 \epsilon \{2, 4\}\ , \quad X_2 \epsilon \{1, 2\}.$$

(b) Suppose that the range of X_1 is $\{1, 2\}$. Is the capacity region decreased? Why or why not?

15.21 *Broadcast channel*. Consider the following degraded broadcast channel.



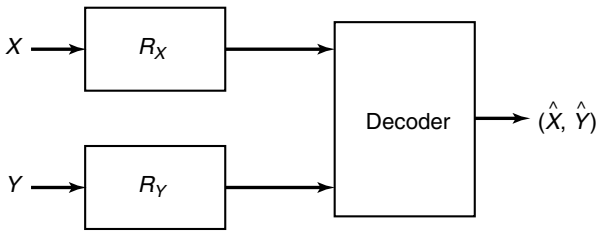
(a) What is the capacity of the channel from X to Y_1?

(b) What is the channel capacity from X to Y_2?

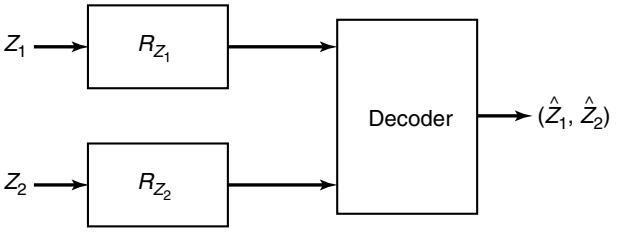
(c) What is the capacity region of all (R_1, R_2) achievable for this broadcast channel? Simplify and sketch.

15.22 *Stereo*. The sum and the difference of the right and left ear signals are to be individually compressed for a common receiver. Let Z_1 be Bernoulli (p_1) and Z_2 be Bernoulli (p_2) and suppose that Z_1 and Z_2 are independent. Let $X = Z_1 + Z_2$, and $Y = Z_1 - Z_2$.

(a) What is the Slepian–Wolf rate region of achievable (R_X, R_Y)?

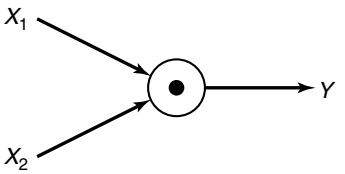


(b) Is this larger or smaller than the rate region of (R_{Z_1}, R_{Z_2})? Why?



There is a simple way to do this part.

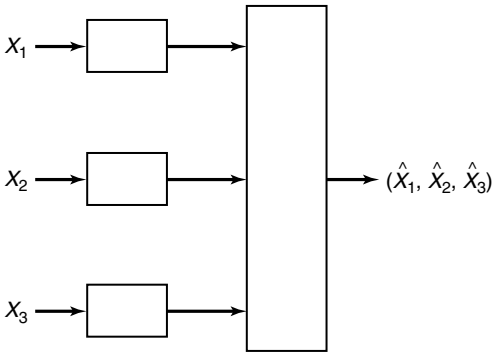
15.23 *Multiplicative multiple-access channel*. Find and sketch the capacity region of the following multiplicative multiple-access channel:



with $X_1 \in \{0, 1\}$, $X_2 \in \{1, 2, 3\}$, and $Y = X_1 X_2$.

15.24 *Distributed data compression*. Let Z_1, Z_2, Z_3 be independent Bernoulli(p). Find the Slepian–Wolf rate region for the description of (X_1, X_2, X_3), where

$$X_1 = Z_1$$
$$X_2 = Z_1 + Z_2$$
$$X_3 = Z_1 + Z_2 + Z_3.$$



15.25 *Noiseless multiple-access channel*. Consider the following multiple-access channel with two binary inputs $X_1, X_2 \in \{0, 1\}$ and output $Y = (X_1, X_2)$.

(a) Find the capacity region. Note that each sender can send at capacity.

(b) Now consider the cooperative capacity region, $R_1 \geq 0$, $R_2 \geq 0$, $R_1 + R_2 \leq \max_{p(x_1,x_2)} I(X_1, X_2; Y)$. Argue that the throughput $R_1 + R_2$ does not increase but the capacity region increases.

15.26 *Infinite bandwidth multiple-access channel*. Find the capacity region for the Gaussian multiple-access channel with infinite bandwidth. Argue that all senders can send at their individual capacities (i.e., infinite bandwidth eliminates interference).

15.27 *Multiple-access identity*. Let $C(x) = \frac{1}{2}\log(1 + x)$ denote the channel capacity of a Gaussian channel with signal-to-noise ratio x. Show that

$$C\left(\frac{P_1}{N}\right) + C\left(\frac{P_2}{P_1 + N}\right) = C\left(\frac{P_1 + P_2}{N}\right).$$

This suggests that two independent users can send information as well as if they had pooled their power.

15.28 *Frequency-division multiple access* (*FDMA*). Maximize the throughput $R_1 + R_2 = W_1 \log(1 + \frac{P_1}{NW_1}) + (W - W_1)\log(1 + \frac{P_2}{N(W-W_1)})$ over W_1 to show that bandwidth should be proportional to transmitted power for FDMA.

15.29 *Trilingual-speaker broadcast channel*. A speaker of Dutch, Spanish, and French wishes to communicate simultaneously to three people: D, S, and F. D knows only Dutch but can distinguish when a Spanish word is being spoken as distinguished from a French word; similarly for the other two, who know only Spanish and French, respectively, but can distinguish when a foreign word is spoken and which language is being spoken. Suppose that each language, Dutch, Spanish, and French, has M words: M words of Dutch, M words of French, and M words of Spanish.

(a) What is the maximum rate at which the trilingual speaker can speak to D?

(b) If he speaks to D at the maximum rate, what is the maximum rate at which he can speak simultaneously to S?

(c) If he is speaking to D and S at the joint rate in part (b), can he also speak to F at some positive rate? If so, what is it? If not, why not?

15.30 *Parallel Gaussian channels from a mobile telephone.* Assume that a sender X is sending to two fixed base stations. Assume that the sender sends a signal X that is constrained to have average power P. Assume that the two base stations receive signals Y_1 and Y_2, where

$$Y_1 = \alpha_1 X + Z_1$$

$$Y_2 = \alpha_2 X + Z_2,$$

where $Z_i \sim \mathcal{N}(0, N_1)$, $Z_2 \sim \mathcal{N}(0, N_2)$, and Z_1 and Z_2 are independent. We will assume the α's are constant over a transmitted block.

(a) Assuming that both signals Y_1 and Y_2 are available at a common decoder $Y = (Y_1, Y_2)$, what is the capacity of the channel from the sender to the common receiver?

(b) If, instead, the two receivers Y_1 and Y_2 each decode their signals independently, this becomes a broadcast channel. Let R_1 be the rate to base station 1 and R_2 be the rate to base station 2. Find the capacity region of this channel.

15.31 *Gaussian multiple access.* A group of m users, each with power P, is using a Gaussian multiple-access channel at capacity, so that

$$\sum_{i=1}^{m} R_i = C\left(\frac{mP}{N}\right), \tag{15.395}$$

where $C(x) = \frac{1}{2}\log(1 + x)$ and N is the receiver noise power. A new user of power P_0 wishes to join in.

(a) At what rate can he send without disturbing the other users?

(b) What should his power P_0 be so that the new users' rate is equal to the combined communication rate $C(mP/N)$ of all the other users?

15.32 *Converse for deterministic broadcast channel.* A deterministic broadcast channel is defined by an input X and two outputs, Y_1 and Y_2, which are functions of the input X. Thus, $Y_1 = f_1(X)$ and $Y_2 = f_2(X)$. Let R_1 and R_2 be the rates at which information can be sent to the two receivers. Prove that

$$R_1 \leq H(Y_1) \tag{15.396}$$

$$R_2 \le H(Y_2) \tag{15.397}$$

$$R_1 + R_2 \le H(Y_1, Y_2). \tag{15.398}$$

15.33 *Multiple-access channel*. Consider the multiple-access channel $Y = X_1 + X_2 \pmod 4$, where $X_1 \in \{0, 1, 2, 3\}$, $X_2 \in \{0, 1\}$.

(a) Find the capacity region (R_1, R_2).

(b) What is the maximum throughput $R_1 + R_2$?

15.34 *Distributed source compression*. Let

$$Z_1 = \begin{cases} 1, & p \\ 0, & q, \end{cases}$$

$$Z_2 = \begin{cases} 1, & p \\ 0, & q, \end{cases}$$

and let $U = Z_1 Z_2$, $V = Z_1 + Z_2$. Assume that Z_1 and Z_2 are independent. This induces a joint distribution on (U, V). Let (U_i, V_i) be i.i.d. according to this distribution. Sender 1 describes U^n at rate R_1, and sender 2 describes V^n at rate R_2.

(a) Find the Slepian–Wolf rate region for recovering (U^n, V^n) at the receiver.

(b) What is the residual uncertainty (conditional entropy) that the receiver has about (X^n, Y^n).

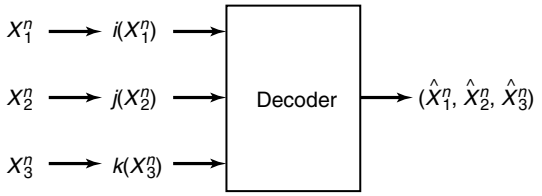
15.35 *Multiple-access channel capacity with costs*. The cost of using symbol x is $r(x)$. The cost of a codeword x^n is $r(x^n) = \frac{1}{n}\sum_{i=1}^n r(x_i)$. A $(2^{nR}, n)$ codebook satisfies cost constraint r if $\frac{1}{n}\sum_{i=1}^n r(x_i(w)) \le r$ for all $w \in 2^{nR}$.

(a) Find an expression for the capacity $C(r)$ of a discrete memoryless channel with cost constraint r.

(b) Find an expression for the multiple-access channel capacity region for $(\mathcal{X}_1 \times \mathcal{X}_2, p(y|x_1, x_2), \mathcal{Y})$ if sender X_1 has cost constraint r_1 and sender X_2 has cost constraint r_2.

(c) Prove the converse for part (b).

15.36 *Slepian–Wolf*. Three cards from a three-card deck are dealt, one to sender X_1, one to sender X_2, and one to sender X_3. At what rates do X_1, X_2, and X_3 need to communicate to some receiver so that their card information can be recovered?



Assume that (X_{1i}, X_{2i}, X_{3i}) are drawn i.i.d. from a uniform distribution over the permutations of $\{1, 2, 3\}$.

HISTORICAL NOTES

This chapter is based on the review in El Gamal and Cover [186]. The two-way channel was studied by Shannon [486] in 1961. He derived inner and outer bounds on the capacity region. Dueck [175] and Schalkwijk [464, 465] suggested coding schemes for two-way channels that achieve rates exceeding Shannon's inner bound; outer bounds for this channel were derived by Zhang et al. [596] and Willems and Hekstra [558].

The multiple-access channel capacity region was found by Ahlswede [7] and Liao [355] and was extended to the case of the multiple-access channel with common information by Slepian and Wolf [501]. Gaarder and Wolf [220] were the first to show that feedback increases the capacity of a discrete memoryless multiple-access channel. Cover and Leung [133] proposed an achievable region for the multiple-access channel with feedback, which was shown to be optimal for a class of multiple-access channels by Willems [557]. Ozarow [410] has determined the capacity region for a two-user Gaussian multiple-access channel with feedback. Cover et al. [129] and Ahlswede and Han [12] have considered the problem of transmission of a correlated source over a multiple-access channel. The Slepian–Wolf theorem was proved by Slepian and Wolf [502] and was extended to jointly ergodic sources by a binning argument in Cover [122].

Superposition coding for broadcast channels was suggested by Cover in 1972 [119]. The capacity region for the degraded broadcast channel was determined by Bergmans [55] and Gallager [225]. The superposition codes for the degraded broadcast channel are also optimal for the less noisy broadcast channel (Körner and Marton [324]), the more capable broadcast channel (El Gamal [185]), and the broadcast channel with degraded message sets (Körner and Marton [325]). Van der Meulen [526] and Cover [121] proposed achievable regions for the general broadcast channel. The capacity of a deterministic broadcast channel was found by Gelfand and Pinsker [242, 243, 423] and Marton [377]. The best known

achievable region for the broadcast channel is due to Marton [377]; a simpler proof of Marton's region was given by El Gamal and Van der Meulen [188]. El Gamal [184] showed that feedback does not increase the capacity of a physically degraded broadcast channel. Dueck [176] introduced an example to illustrate that feedback can increase the capacity of a memoryless broadcast channel; Ozarow and Leung [411] described a coding procedure for the Gaussian broadcast channel with feedback that increased the capacity region.

The relay channel was introduced by Van der Meulen [528]; the capacity region for the degraded relay channel was determined by Cover and El Gamal [127]. Carleial [85] introduced the Gaussian interference channel with power constraints and showed that very strong interference is equivalent to no interference at all. Sato and Tanabe [459] extended the work of Carleial to discrete interference channels with strong interference. Sato [457] and Benzel [51] dealt with degraded interference channels. The best known achievable region for the general interference channel is due to Han and Kobayashi [274]. This region gives the capacity for Gaussian interference channels with interference parameters greater than 1, as was shown in Han and Kobayashi [274] and Sato [458]. Carleial [84] proved new bounds on the capacity region for interference channels.

The problem of coding with side information was introduced by Wyner and Ziv [573] and Wyner [570]; the achievable region for this problem was described in Ahlswede and Körner [13], Gray and Wyner [261], and Wyner [571],[572]. The problem of finding the rate distortion function with side information was solved by Wyner and Ziv [574]. The channel capacity counterpart of rate distortion with side information was solved by Gelfand and Pinsker [243]; the duality between the two results is explored in Cover and Chiang [113]. The problem of multiple descriptions is treated in El Gamal and Cover [187].

The special problem of encoding a function of two random variables was discussed by Körner and Marton [326], who described a simple method to encode the modulo 2 sum of two binary random variables. A general framework for the description of source networks may be found in Csiszár and Körner [148],[149]. A common model that includes Slepian–Wolf encoding, coding with side information, and rate distortion with side information as special cases was described by Berger and Yeung [54].

In 1989, Ahlswede and Dueck [17] introduced the problem of identification via communication channels, which can be viewed as a problem where the sender sends information to the receivers but each receiver only needs to know whether or not a single message was sent. In this case, the set of possible messages that can be sent reliably is doubly exponential in

the block length, and the key result of this paper was to show that $2^{2^{nC}}$ messages could be identified for any noisy channel with capacity C. This problem spawned a set of papers [16, 18, 269, 434], including extensions to channels with feedback and multiuser channels.

Another active area of work has been the analysis of MIMO (multiple-input multiple-output) systems or space-time coding, which use multiple antennas at the transmitter and receiver to take advantage of the diversity gains from multipath for wireless systems. The analysis of these multiple antenna systems by Foschini [217], Teletar [512], and Rayleigh and Cioffi [246] show that the capacity gains from the diversity obtained using multiple antennas in fading environments can be substantial relative to the single-user capacity achieved by traditional equalization and interleaving techniques. A special issue of the *IEEE Transactions in Information Theory* [70] has a number of papers covering different aspects of this technology.

Comprehensive surveys of network information theory may be found in El Gamal and Cover [186], Van der Meulen [526–528], Berger [53], Csiszár and Körner [149], Verdu [538], Cover [111], and Ephremides and Hajek [197].

CHAPTER 16

INFORMATION THEORY AND PORTFOLIO THEORY

The duality between the growth rate of wealth in the stock market and the entropy rate of the market is striking. In particular, we shall find the competitively optimal and growth rate optimal portfolio strategies. They are the same, just as the Shannon code is optimal both competitively and in the expected description rate. We also find the asymptotic growth rate of wealth for an ergodic stock market process. We end with a discussion of universal portfolios that enable one to achieve the same asymptotic growth rate as the best constant rebalanced portfolio in hindsight.

In Section 16.8 we provide a "sandwich" proof of the asymptotic equipartition property for general ergodic processes that is motivated by the notion of optimal portfolios for stationary ergodic stock markets.

16.1 THE STOCK MARKET: SOME DEFINITIONS

A stock market is represented as a vector of stocks $\mathbf{X} = (X_1, X_2, \ldots, X_m)$, $X_i \geq 0$, $i = 1, 2, \ldots, m$, where m is the number of stocks and the *price relative* X_i is the ratio of the price at the end of the day to the price at the beginning of the day. So typically, X_i is near 1. For example, $X_i = 1.03$ means that the ith stock went up 3 percent that day.

Let $\mathbf{X} \sim F(\mathbf{x})$, where $F(\mathbf{x})$ is the joint distribution of the vector of price relatives. A *portfolio* $\mathbf{b} = (b_1, b_2, \ldots, b_m)$, $b_i \geq 0$, $\sum b_i = 1$, is an allocation of wealth across the stocks. Here b_i is the fraction of one's wealth invested in stock i. If one uses a portfolio $\mathbf{b}$ and the stock vector is $\mathbf{X}$, the wealth relative (ratio of the wealth at the end of the day to the wealth at the beginning of the day) is $S = \mathbf{b}^t\mathbf{X} = \sum_{i=1}^{m} b_i X_i$.

We wish to maximize S in some sense. But S is a random variable, the distribution of which depends on portfolio $\mathbf{b}$, so there is controversy

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

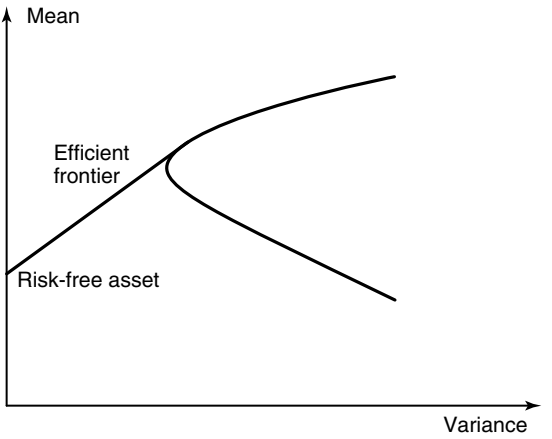


FIGURE 16.1. Sharpe–Markowitz theory: set of achievable mean–variance pairs.

over the choice of the best distribution for S. The standard theory of stock market investment is based on consideration of the first and second moments of S. The objective is to maximize the expected value of S subject to a constraint on the variance. Since it is easy to calculate these moments, the theory is simpler than the theory that deals with the entire distribution of S.

The mean–variance approach is the basis of the Sharpe–Markowitz theory of investment in the stock market and is used by business analysts and others. It is illustrated in Figure 16.1. The figure illustrates the set of achievable mean–variance pairs using various portfolios. The set of portfolios on the boundary of this region corresponds to the undominated portfolios: These are the portfolios that have the highest mean for a given variance. This boundary is called the *efficient frontier*, and if one is interested only in mean and variance, one should operate along this boundary.

Normally, the theory is simplified with the introduction of a *risk-free* asset (e.g., cash or Treasury bonds, which provide a fixed interest rate with zero variance). This stock corresponds to a point on the Y axis in the figure. By combining the risk-free asset with various stocks, one obtains all points below the tangent from the risk-free asset to the efficient frontier. This line now becomes part of the efficient frontier.

The concept of the efficient frontier also implies that there is a true price for a stock corresponding to its risk. This theory of stock prices, called the *capital asset pricing model* (CAPM), is used to decide whether the market price for a stock is too high or too low. Looking at the mean of a random variable gives information about the long-term behavior of

the sum of i.i.d. versions of the random variable. But in the stock market, one normally reinvests every day, so that the wealth at the end of n days is the product of factors, one for each day of the market. The behavior of the product is determined not by the expected value but by the expected logarithm. This leads us to define the growth rate as follows:

Definition The *growth rate* of a stock market portfolio $\mathbf{b}$ with respect to a stock distribution $F(\mathbf{x})$ is defined as

$$W(\mathbf{b}, F) = \int \log \mathbf{b}^t \mathbf{x} \, dF(\mathbf{x}) = E\left(\log \mathbf{b}^t \mathbf{X}\right). \tag{16.1}$$

If the logarithm is to base 2, the growth rate is also called the *doubling rate*.

Definition The *optimal growth rate* $W^*(F)$ is defined as

$$W^*(F) = \max_{\mathbf{b}} W(\mathbf{b}, F), \tag{16.2}$$

where the maximum is over all possible portfolios $b_i \geq 0$, $\sum_i b_i = 1$.

Definition A portfolio $\mathbf{b}^*$ that achieves the maximum of $W(\mathbf{b}, F)$ is called a *log-optimal portfolio* or *growth optimal portfolio*.

The definition of growth rate is justified by the following theorem, which shows that wealth grows as 2^{nW^*}.

Theorem 16.1.1 *Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ *be i.i.d. according to* $F(\mathbf{x})$. *Let*

$$S_n^* = \prod_{i=1}^{n} \mathbf{b}^{*t} \mathbf{X}_i \tag{16.3}$$

be the wealth after n days using the constant rebalanced portfolio $\mathbf{b}^*$. *Then*

$$\frac{1}{n} \log S_n^* \to W^* \quad \textit{with probability } 1. \tag{16.4}$$

Proof: By the strong law of large numbers,

$$\frac{1}{n} \log S_n^* = \frac{1}{n} \sum_{i=1}^{n} \log \mathbf{b}^{*t} \mathbf{X}_i \tag{16.5}$$

$$\to W^* \qquad \text{with probability } 1. \tag{16.6}$$

Hence, $S_n^* \doteq 2^{nW^*}$. □

We now consider some of the properties of the growth rate.

Lemma 16.1.1 *$W(\mathbf{b}, F)$ is concave in $\mathbf{b}$ and linear in F. $W^*(F)$ is convex in F.*

Proof: The growth rate is

$$W(\mathbf{b}, F) = \int \log \mathbf{b}^t \mathbf{x}\, dF(\mathbf{x}). \tag{16.7}$$

Since the integral is linear in F, so is $W(\mathbf{b}, F)$. Since

$$\log(\lambda \mathbf{b}_1 + (1-\lambda)\mathbf{b}_2)^t \mathbf{X} \geq \lambda \log \mathbf{b}_1^t \mathbf{X} + (1-\lambda) \log \mathbf{b}_2^t \mathbf{X}, \tag{16.8}$$

by the concavity of the logarithm, it follows, by taking expectations, that $W(\mathbf{b}, F)$ is concave in $\mathbf{b}$. Finally, to prove the convexity of $W^*(F)$ as a function of F, let F_1 and F_2 be two distributions on the stock market and let the corresponding optimal portfolios be $\mathbf{b}^*(F_1)$ and $\mathbf{b}^*(F_2)$, respectively. Let the log-optimal portfolio corresponding to $\lambda F_1 + (1-\lambda)F_2$ be $\mathbf{b}^*(\lambda F_1 + (1-\lambda)F_2)$. Then by linearity of $W(\mathbf{b}, F)$ with respect to F, we have

$$\begin{aligned} & W^*(\lambda F_1 + (1-\lambda)F_2) \\ & \quad = W(\mathbf{b}^*(\lambda F_1 + (1-\lambda)F_2), \lambda F_1 + (1-\lambda)F_2) \qquad (16.9) \\ & \quad = \lambda W(\mathbf{b}^*(\lambda F_1 + (1-\lambda)F_2), F_1) \\ & \qquad + (1-\lambda) W(\mathbf{b}^*(\lambda F_1 + (1-\lambda)F_2), F_2) \\ & \quad \leq \lambda W(\mathbf{b}^*(F_1), F_1) + (1-\lambda) W^*(\mathbf{b}^*(F_2), F_2), \qquad (16.10) \end{aligned}$$

since $\mathbf{b}^*(F_1)$ maximizes $W(\mathbf{b}, F_1)$ and $\mathbf{b}^*(F_2)$ maximizes $W(\mathbf{b}, F_2)$. □

Lemma 16.1.2 *The set of log-optimal portfolios with respect to a given distribution is convex.*

Proof: Suppose that $\mathbf{b}_1$ and $\mathbf{b}_2$ are log-optimal (i.e., $W(\mathbf{b}_1, F) = W(\mathbf{b}_2, F) = W^*(F)$). By the concavity of $W(\mathbf{b}, F)$ in $\mathbf{b}$, we have

$$W(\lambda \mathbf{b}_1 + (1-\lambda)\mathbf{b}_2, F) \geq \lambda W(\mathbf{b}_1, F) + (1-\lambda) W(\mathbf{b}_2, F) = W^*(F). \tag{16.11}$$

Thus, $\lambda \mathbf{b}_1 + (1-\lambda)\mathbf{b}_2$ is also log-optimal. □

In the next section we use these properties to characterize the log-optimal portfolio.

16.2 KUHN–TUCKER CHARACTERIZATION OF THE LOG-OPTIMAL PORTFOLIO

Let $\mathcal{B} = \{\mathbf{b} \in \mathcal{R}^m : \mathbf{b}_i \geq 0, \sum_{i=1}^m \mathbf{b}_i = 1\}$ denote the *set of allowed portfolios*. The determination of $\mathbf{b}^*$ that achieves $W^*(F)$ is a problem of maximization of a concave function $W(\mathbf{b}, F)$ over a convex set $\mathcal{B}$. The maximum may lie on the boundary. We can use the standard Kuhn–Tucker conditions to characterize the maximum. Instead, we derive these conditions from first principles.

Theorem 16.2.1 *The log-optimal portfolio $\mathbf{b}^*$ for a stock market $\mathbf{X} \sim F$ (i.e., the portfolio that maximizes the growth rate $W(\mathbf{b}, F)$) satisfies the following necessary and sufficient conditions:*

$$E\left(\frac{X_i}{\mathbf{b}^{*t}\mathbf{X}}\right) = 1 \qquad \text{if } b_i^* > 0,$$
$$\leq 1 \qquad \text{if } b_i^* = 0. \tag{16.12}$$

Proof: The growth rate $W(\mathbf{b}) = E(\ln \mathbf{b}^t\mathbf{X})$ is concave in $\mathbf{b}$, where $\mathbf{b}$ ranges over the simplex of portfolios. It follows that $\mathbf{b}^*$ is log-optimum iff the directional derivative of $W(\cdot)$ in the direction from $\mathbf{b}^*$ to any alternative portfolio $\mathbf{b}$ is nonpositive. Thus, letting $\mathbf{b}_\lambda = (1-\lambda)\mathbf{b}^* + \lambda\mathbf{b}$ for $0 \leq \lambda \leq 1$, we have

$$\frac{d}{d\lambda} W(\mathbf{b}_\lambda)\Big|_{\lambda=0+} \leq 0, \qquad \mathbf{b} \in \mathcal{B}. \tag{16.13}$$

These conditions reduce to (16.12) since the one-sided derivative at $\lambda = 0+$ of $W(\mathbf{b}_\lambda)$ is

$$\frac{d}{d\lambda} E(\ln(\mathbf{b}_\lambda^t\mathbf{X}))\Big|_{\lambda=0+}$$
$$= \lim_{\lambda \downarrow 0} \frac{1}{\lambda} E\left(\ln\left(\frac{(1-\lambda)\mathbf{b}^{*t}\mathbf{X} + \lambda\mathbf{b}^t\mathbf{X}}{\mathbf{b}^{*t}\mathbf{X}}\right)\right) \tag{16.14}$$
$$= E\left(\lim_{\lambda \downarrow 0} \frac{1}{\lambda} \ln\left(1 + \lambda\left(\frac{\mathbf{b}^t\mathbf{X}}{\mathbf{b}^{*t}\mathbf{X}} - 1\right)\right)\right) \tag{16.15}$$
$$= E\left(\frac{\mathbf{b}^t\mathbf{X}}{\mathbf{b}^{*t}\mathbf{X}}\right) - 1, \tag{16.16}$$

where the interchange of limit and expectation can be justified using the dominated convergence theorem [39]. Thus, (16.13) reduces to

$$E\left(\frac{\mathbf{b}^t\mathbf{X}}{\mathbf{b}^{*t}\mathbf{X}}\right) - 1 \leq 0 \tag{16.17}$$

for all $\mathbf{b} \in \mathcal{B}$. If the line segment from $\mathbf{b}$ to $\mathbf{b}^*$ can be extended beyond $\mathbf{b}^*$ in the simplex, the two-sided derivative at $\lambda = 0$ of $W(\mathbf{b}_\lambda)$ vanishes and (16.17) holds with equality. If the line segment from $\mathbf{b}$ to $\mathbf{b}^*$ cannot be extended because of the inequality constraint on $\mathbf{b}$, we have an inequality in (16.17).

The Kuhn–Tucker conditions will hold for all portfolios $\mathbf{b} \in \mathcal{B}$ if they hold for all extreme points of the simplex $\mathcal{B}$ since $E(\mathbf{b}^t\mathbf{X}/\mathbf{b}^{*t}\mathbf{X})$ is linear in $\mathbf{b}$. Furthermore, the line segment from the jth extreme point ($\mathbf{b} : b_j = 1, b_i = 0, i \neq j$) to $\mathbf{b}^*$ can be extended beyond $\mathbf{b}^*$ in the simplex iff $b_j^* > 0$. Thus, the Kuhn–Tucker conditions that characterize the log-optimum $\mathbf{b}^*$ are equivalent to the following necessary and sufficient conditions:

$$E\left(\frac{X_i}{\mathbf{b}^{*t}\mathbf{X}}\right) = 1 \qquad \text{if } b_i^* > 0,$$
$$\leq 1 \qquad \text{if } b_i^* = 0. \quad \square \tag{16.18}$$

This theorem has a few immediate consequences. One useful equivalence is expressed in the following theorem.

Theorem 16.2.2 *Let $S^* = \mathbf{b}^{*t}\mathbf{X}$ be the random wealth resulting from the log-optimal portfolio $\mathbf{b}^*$. Let $S = \mathbf{b}^t\mathbf{X}$ be the wealth resulting from any other portfolio $\mathbf{b}$. Then*

$$E \ln \frac{S}{S^*} \leq 0 \quad \textit{for all } S \quad \Leftrightarrow \quad E\frac{S}{S^*} \leq 1 \quad \textit{for all } S. \tag{16.19}$$

Proof: From Theorem 16.2.1 it follows that for a log-optimal portfolio $\mathbf{b}^*$,

$$E\left(\frac{X_i}{\mathbf{b}^{*t}\mathbf{X}}\right) \leq 1 \tag{16.20}$$

for all i. Multiplying this equation by b_i and summing over i, we have

$$\sum_{i=1}^{m} b_i E\left(\frac{X_i}{\mathbf{b}^{*t}\mathbf{X}}\right) \leq \sum_{i=1}^{m} b_i = 1, \tag{16.21}$$

which is equivalent to

$$E\frac{\mathbf{b}^t\mathbf{X}}{\mathbf{b}^{*t}\mathbf{X}} = E\frac{S}{S^*} \leq 1. \tag{16.22}$$

The converse follows from Jensen's inequality, since

$$E \log \frac{S}{S^*} \leq \log E\frac{S}{S^*} \leq \log 1 = 0. \quad \square \tag{16.23}$$

Maximizing the expected logarithm was motivated by the asymptotic growth rate. But we have just shown that the log-optimal portfolio, in addition to maximizing the asymptotic growth rate, also "maximizes" the expected wealth relative $E(S/S^*)$ for one day. We shall say more about the short-term optimality of the log-optimal portfolio when we consider the game-theoretic optimality of this portfolio.

Another consequence of the Kuhn–Tucker characterization of the log-optimal portfolio is the fact that the expected proportion of wealth in each stock under the log-optimal portfolio is unchanged from day to day. Consider the stocks at the end of the first day. The initial allocation of wealth is $\mathbf{b}^*$. The proportion of the wealth in stock i at the end of the day is $\frac{b_i^* X_i}{\mathbf{b}^{*t}\mathbf{X}}$, and the expected value of this proportion is

$$E\frac{b_i^* X_i}{\mathbf{b}^{*t}\mathbf{X}} = b_i^* E\frac{X_i}{\mathbf{b}^{*t}\mathbf{X}} = b_i^*. \tag{16.24}$$

Hence, the proportion of wealth in stock i expected at the end of the day is the same as the proportion invested in stock i at the beginning of the day. This is a counterpart to Kelly proportional gambling, where one invests in proportions that remain unchanged in expected value after the investment period.

16.3 ASYMPTOTIC OPTIMALITY OF THE LOG-OPTIMAL PORTFOLIO

In Section 16.2 we introduced the log-optimal portfolio and explained its motivation in terms of the long-term behavior of a sequence of investments in a repeated independent versions of the stock market. In this section we expand on this idea and prove that with probability 1, the conditionally log-optimal investor will not do any worse than any other investor who uses a causal investment strategy.

We first consider an i.i.d. stock market (i.e., $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are i.i.d. according to $F(\mathbf{x})$). Let

$$S_n = \prod_{i=1}^{n} \mathbf{b}_i^t \mathbf{X}_i \tag{16.25}$$

be the wealth after n days for an investor who uses portfolio $\mathbf{b}_i$ on day i. Let

$$W^* = \max_{\mathbf{b}} W(\mathbf{b}, F) = \max_{\mathbf{b}} E \log \mathbf{b}^t \mathbf{X} \tag{16.26}$$

be the maximal growth rate, and let $\mathbf{b}^*$ be a portfolio that achieves the maximum growth rate. We only allow alternative portfolios $\mathbf{b}_i$ that depend causally on the past and are independent of the future values of the stock market.

Definition A *nonanticipating* or *causal* portfolio strategy is a sequence of mappings $b_i : \mathcal{R}^{m(i-1)} \to \mathcal{B}$, with the interpretation that portfolio $b_i(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1})$ is used on day i.

From the definition of W^*, it follows immediately that the log-optimal portfolio maximizes the expected log of the final wealth. This is stated in the following lemma.

Lemma 16.3.1 *Let S_n^* be the wealth after n days using the log-optimal strategy $\mathbf{b}^*$ on i.i.d. stocks, and let S_n be the wealth using a causal portfolio strategy $\mathbf{b}_i$. Then*

$$E \log S_n^* = nW^* \geq E \log S_n. \tag{16.27}$$

Proof

$$\max_{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n} E \log S_n = \max_{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n} E \sum_{i=1}^{n} \log \mathbf{b}_i^t \mathbf{X}_i \tag{16.28}$$

$$= \sum_{i=1}^{n} \max_{\mathbf{b}_i(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1})} E \log \mathbf{b}_i^t(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1})\mathbf{X}_i \tag{16.29}$$

$$= \sum_{i=1}^{n} E \log \mathbf{b}^{*t} \mathbf{X}_i \tag{16.30}$$

$$= nW^*, \tag{16.31}$$

and the maximum is achieved by a constant portfolio strategy $\mathbf{b}^*$. □

So far, we have proved two simple consequences of the definition of log-optimal portfolios: that $\mathbf{b}^*$ (satisfying (16.12)) maximizes the expected log wealth, and that the resulting wealth S_n^* is equal to 2^{nW^*} to first order in the exponent, with high probability.

Now we prove a much stronger result, which shows that S_n^* exceeds the wealth (to first order in the exponent) of any other investor for almost every sequence of outcomes from the stock market.

Theorem 16.3.1 (*Asymptotic optimality of the log-optimal portfolio*) *Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a sequence of i.i.d. stock vectors drawn according*

to $F(\mathbf{x})$. Let $S_n^* = \prod_{i=1}^n \mathbf{b}^{*t}\mathbf{X}_i$, where $\mathbf{b}^*$ is the log-optimal portfolio, and let $S_n = \prod_{i=1}^n \mathbf{b}_i^t\mathbf{X}_i$ be the wealth resulting from any other causal portfolio. Then

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{S_n}{S_n^*} \le 0 \qquad \textit{with probability 1.} \tag{16.32}$$

Proof: From the Kuhn–Tucker conditions and the log optimality of S_n^*, we have

$$E\frac{S_n}{S_n^*} \le 1. \tag{16.33}$$

Hence by Markov's inequality, we have

$$\Pr\left(S_n > t_n S_n^*\right) = \Pr\left(\frac{S_n}{S_n^*} > t_n\right) < \frac{1}{t_n}. \tag{16.34}$$

Hence,

$$\Pr\left(\frac{1}{n}\log\frac{S_n}{S_n^*} > \frac{1}{n}\log t_n\right) \le \frac{1}{t_n}. \tag{16.35}$$

Setting $t_n = n^2$ and summing over n, we have

$$\sum_{n=1}^{\infty} \Pr\left(\frac{1}{n}\log\frac{S_n}{S_n^*} > \frac{2\log n}{n}\right) \le \sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}. \tag{16.36}$$

Then, by the Borel–Cantelli lemma,

$$\Pr\left(\frac{1}{n}\log\frac{S_n}{S_n^*} > \frac{2\log n}{n}, \text{ infinitely often}\right) = 0. \tag{16.37}$$

This implies that for almost every sequence from the stock market, there exists an N such that for all $n > N$, $\frac{1}{n}\log\frac{S_n}{S_n^*} < \frac{2\log n}{n}$. Thus,

$$\limsup \frac{1}{n}\log\frac{S_n}{S_n^*} \le 0 \qquad \text{with probability 1.} \quad \square \tag{16.38}$$

The theorem proves that the log-optimal portfolio will perform as well as or better than any other portfolio to first order in the exponent.

16.4 SIDE INFORMATION AND THE GROWTH RATE

We showed in Chapter 6 that side information Y for the horse race X can be used to increase the growth rate by the mutual information $I(X;Y)$.

We now extend this result to the stock market. Here, $I(X; Y)$ is an upper bound on the increase in the growth rate, with equality if X is a horse race. We first consider the decrease in growth rate incurred by believing in the wrong distribution.

Theorem 16.4.1 *Let $\mathbf{X} \sim f(\mathbf{x})$. Let $\mathbf{b}_f$ be a log-optimal portfolio corresponding to $f(\mathbf{x})$, and let $\mathbf{b}_g$ be a log-optimal portfolio corresponding to some other density $g(\mathbf{x})$. Then the increase in growth rate ΔW by using $\mathbf{b}_f$ instead of $\mathbf{b}_g$ is bounded by*

$$\Delta W = W(\mathbf{b}_f, F) - W(\mathbf{b}_g, F) \le D(f||g). \tag{16.39}$$

Proof: We have

$$\Delta W = \int f(\mathbf{x}) \log \mathbf{b}_f^t \mathbf{x} - \int f(\mathbf{x}) \log \mathbf{b}_g^t \mathbf{x} \tag{16.40}$$

$$= \int f(\mathbf{x}) \log \frac{\mathbf{b}_f^t \mathbf{x}}{\mathbf{b}_g^t \mathbf{x}} \tag{16.41}$$

$$= \int f(\mathbf{x}) \log \frac{\mathbf{b}_f^t \mathbf{x}}{\mathbf{b}_g^t \mathbf{x}} \frac{g(\mathbf{x})}{f(\mathbf{x})} \frac{f(\mathbf{x})}{g(\mathbf{x})} \tag{16.42}$$

$$= \int f(\mathbf{x}) \log \frac{\mathbf{b}_f^t \mathbf{x}}{\mathbf{b}_g^t \mathbf{x}} \frac{g(\mathbf{x})}{f(\mathbf{x})} + D(f||g) \tag{16.43}$$

$$\overset{(a)}{\le} \log \int f(\mathbf{x}) \frac{\mathbf{b}_f^t \mathbf{x}}{\mathbf{b}_g^t \mathbf{x}} \frac{g(\mathbf{x})}{f(\mathbf{x})} + D(f||g) \tag{16.44}$$

$$= \log \int g(\mathbf{x}) \frac{\mathbf{b}_f^t \mathbf{x}}{\mathbf{b}_g^t \mathbf{x}} + D(f||g) \tag{16.45}$$

$$\overset{(b)}{\le} \log 1 + D(f||g) \tag{16.46}$$

$$= D(f||g), \tag{16.47}$$

where (a) follows from Jensen's inequality and (b) follows from the Kuhn–Tucker conditions and the fact that $\mathbf{b}_g$ is log-optimal for g. □

Theorem 16.4.2 *The increase ΔW in growth rate due to side information Y is bounded by*

$$\Delta W \le I(\mathbf{X}; Y). \tag{16.48}$$

Proof: Let $(\mathbf{X}, Y) \sim f(\mathbf{x}, y)$, where $\mathbf{X}$ is the market vector and Y is the related side information. Given side information $Y = y$, the log-optimal investor uses the conditional log-optimal portfolio for the conditional distribution $f(\mathbf{x}|Y = y)$. Hence, conditional on $Y = y$, we have, from Theorem 16.4.1,

$$\Delta W_{Y=y} \le D(f(\mathbf{x}|Y = y)||f(\mathbf{x})) = \int_{\mathbf{x}} f(\mathbf{x}|Y = y) \log \frac{f(\mathbf{x}|Y = y)}{f(\mathbf{x})}\, d\mathbf{x}. \tag{16.49}$$

Averaging this over possible values of Y, we have

$$\Delta W \le \int_y f(y) \int_{\mathbf{x}} f(\mathbf{x}|Y = y) \log \frac{f(\mathbf{x}|Y = y)}{f(\mathbf{x})}\, d\mathbf{x}\, dy \tag{16.50}$$

$$= \int_y \int_{\mathbf{x}} f(y) f(\mathbf{x}|Y = y) \log \frac{f(\mathbf{x}|Y = y)}{f(\mathbf{x})} \frac{f(y)}{f(y)}\, d\mathbf{x}\, dy \tag{16.51}$$

$$= \int_y \int_{\mathbf{x}} f(\mathbf{x}, y) \log \frac{f(\mathbf{x}, y)}{f(\mathbf{x}) f(y)}\, d\mathbf{x}\, dy \tag{16.52}$$

$$= I(\mathbf{X}; Y). \tag{16.53}$$

Hence, the increase in growth rate is bounded above by the mutual information between the side information Y and the stock market $\mathbf{X}$. □

16.5 INVESTMENT IN STATIONARY MARKETS

We now extend some of the results of Section 16.4 from i.i.d. markets to time-dependent market processes. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n, \ldots$ be a vector-valued stochastic process with $\mathbf{X}_i \ge 0$. We consider investment strategies that depend on the past values of the market in a causal fashion (i.e., $\mathbf{b}_i$ may depend on $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1}$). Let

$$S_n = \prod_{i=1}^{n} \mathbf{b}_i^t(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1})\mathbf{X}_i. \tag{16.54}$$

Our objective is to maximize $E \log S_n$ over all such causal portfolio strategies $\{\mathbf{b}_i(\cdot)\}$. Now

$$\max_{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n} E \log S_n = \sum_{i=1}^{n} \max_{\mathbf{b}_i(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1})} E \log \mathbf{b}_i^t \mathbf{X}_i \tag{16.55}$$

$$= \sum_{i=1}^{n} E \log \mathbf{b}_i^{*t} \mathbf{X}_i, \tag{16.56}$$

where $\mathbf{b}_i^*$ is the log-optimal portfolio for the conditional distribution of $\mathbf{X}_i$ given the past values of the stock market; that is, $\mathbf{b}_i^*(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1})$ is the portfolio that achieves the conditional maximum, which is denoted by

$$\max_{\mathbf{b}} E[\log \mathbf{b}^t \mathbf{X}_i | (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1}) = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1})] = W^*(\mathbf{X}_i | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1}). \quad (16.57)$$

Taking the expectation over the past, we write

$$W^*(\mathbf{X}_i | \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1}) = E \max_{\mathbf{b}} E\left[\log \mathbf{b}^t \mathbf{X}_i | \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1}\right] \quad (16.58)$$

as the conditional optimal growth rate, where the maximum is over all portfolio-valued functions $\mathbf{b}$ defined on $\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}$. Thus, the highest expected log return is achieved by using the conditional log-optimal portfolio at each stage. Let

$$W^*(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) = \max_{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n} E \log S_n, \quad (16.59)$$

where the maximum is over all causal portfolio strategies. Then since $\log S_n^* = \sum_{i=1}^m \log \mathbf{b}_i^{*t} \mathbf{X}_i$, we have the following chain rule for W^*:

$$W^*(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) = \sum_{i=1}^n W^*(\mathbf{X}_i | \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1}). \quad (16.60)$$

This chain rule is formally the same as the chain rule for H. In some ways, W is the dual of H. In particular, conditioning reduces H but increases W. We now define the counterpart of the entropy rate for time-dependent stochastic processes.

Definition The *growth rate* W_∞^* is defined as

$$W_\infty^* = \lim_{n \to \infty} \frac{W^*(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)}{n} \quad (16.61)$$

if the limit exists.

Theorem 16.5.1 *For a stationary market, the growth rate exists and is equal to*

$$W_\infty^* = \lim_{n \to \infty} W^*(\mathbf{X}_n | \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n-1}). \quad (16.62)$$

Proof: By stationarity, $W^*(\mathbf{X}_n|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n-1})$ is nondecreasing in n. Hence, it must have a limit, possibly infinity. Since

$$\frac{W^*(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} W^*(\mathbf{X}_i|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1}), \tag{16.63}$$

it follows by the theorem of the Cesáro mean (Theorem 4.2.3) that the left-hand side has the same limit as the limit of the terms on the right-hand side. Hence, W_∞^* exists and

$$W_\infty^* = \lim_{n\to\infty} \frac{W^*(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)}{n} = \lim_{n\to\infty} W^*(\mathbf{X}_n|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{n-1}). \quad \square \tag{16.64}$$

We can now extend the asymptotic optimality property to stationary markets. We have the following theorem.

Theorem 16.5.2 *Consider an arbitrary stochastic process $\{X_i\}$, $X_i \in \mathcal{R}_+^m$, conditionally log-optimal portfolios, $\mathbf{b}_i^*(X^{i-1})$ and wealth S_n^*. Let S_n be the wealth generated by any other causal portfolio strategy $\mathbf{b}_i(X^{i-1})$. Then S_n/S_n^* is a positive supermartingale with respect to the sequence of σ-fields generated by the past $X_1, X_2, \ldots, X_n$. Consequently, there exists a random variable V such that*

$$\frac{S_n}{S_n^*} \to V \qquad \textit{with probability 1} \tag{16.65}$$

$$EV \le 1 \tag{16.66}$$

and

$$\Pr\left\{\sup_n \frac{S_n}{S_n^*} \ge t\right\} \le \frac{1}{t}. \tag{16.67}$$

Proof: S_n/S_n^* is a positive supermartingale because

$$E\left[\frac{S_{n+1}(X^{n+1})}{S_{n+1}^*(X^{n+1})}\,\middle|\, X^n\right] = E\left[\frac{(\mathbf{b}_{n+1}^t\mathbf{X}_{n+1})S_n(X^n)}{(\mathbf{b}_{n+1}^{*t}\mathbf{X}_{n+1})S_n^*(X^n)}\,\middle|\, X^n\right] \tag{16.68}$$

$$= \frac{S_n(X^n)}{S_n^*(X^n)} E\left[\frac{\mathbf{b}_{n+1}^t\mathbf{X}_{n+1}}{\mathbf{b}_{n+1}^{*t}\mathbf{X}_{n+1}}\,\middle|\, X^n\right] \tag{16.69}$$

$$\le \frac{S_n(X^n)}{S_n^*(X^n)}, \tag{16.70}$$

by the Kuhn–Tucker condition on the conditionally log-optimal portfolio. Thus, by the martingale convergence theorem, S_n/S_n^* has a limit, call it V, and $EV \le E(S_0/S_0^*) = 1$. Finally, the result for $\sup(S_n/S_n^*)$ follows from Kolmogorov's inequality for positive martingales. □

We remark that (16.70) shows how strong the competitive optimality of S_n^* is. Apparently, the probability is less than 1/10 that $S_n(X^n)$ will ever be 10 times as large as $S_n^*(X^n)$. For a stationary ergodic market, we can extend the asymptotic equipartition property to prove the following theorem.

Theorem 16.5.3 (*AEP for the stock market*) *Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ *be a stationary ergodic vector-valued stochastic process. Let* S_n^* *be the wealth at time n for the conditionally log-optimal strategy, where*

$$S_n^* = \prod_{i=1}^{n} \mathbf{b}_i^{*t}(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1})\mathbf{X}_i. \tag{16.71}$$

Then

$$\frac{1}{n} \log S_n^* \to W_\infty^* \qquad \textit{with probability } 1. \tag{16.72}$$

Proof: The proof involves a generalization of the sandwich argument [20] used to prove the AEP in Section 16.8. The details of the proof (in Algoet and Cover [21]) are omitted. □

Finally, we consider the example of the horse race once again. The horse race is a special case of the stock market in which there are m stocks corresponding to the m horses in the race. At the end of the race, the value of the stock for horse i is either 0 or o_i, the value of the odds for horse i. Thus, $\mathbf{X}$ is nonzero only in the component corresponding to the winning horse.

In this case, the log-optimal portfolio is proportional betting, known as *Kelly gambling* (i.e., $b_i^* = p_i$), and in the case of uniform fair odds (i.e., $o_i = m$, for all i),

$$W^* = \log m - H(X). \tag{16.73}$$

When we have a sequence of correlated horse races, the optimal portfolio is conditional proportional betting and the asymptotic growth rate is

$$W_\infty^* = \log m - H(\mathcal{X}), \tag{16.74}$$

where $H(\mathcal{X}) = \lim \frac{1}{n} H(X_1, X_2, \ldots, X_n)$ if the limit exists. Then Theorem 16.5.3 asserts that

$$S_n^* \doteq 2^{nW^*}, \tag{16.75}$$

in agreement with the results in chapter 6.

16.6 COMPETITIVE OPTIMALITY OF THE LOG-OPTIMAL PORTFOLIO

We now ask whether the log-optimal portfolio outperforms alternative portfolios at a given finite time n. As a direct consequence of the Kuhn–Tucker conditions, we have

$$E\frac{S_n}{S_n^*} \leq 1, \tag{16.76}$$

and hence by Markov's inequality,

$$\Pr(S_n > tS_n^*) \leq \frac{1}{t}. \tag{16.77}$$

This result is similar to the result derived in Chapter 5 for the competitive optimality of Shannon codes.

By considering examples, it can be seen that it is not possible to get a better bound on the probability that $S_n > S_n^*$. Consider a stock market with two stocks and two possible outcomes,

$$(X_1, X_2) = \begin{cases} \left(1, \dfrac{1}{1-\epsilon}\right) & \text{with probability } 1-\epsilon, \\ (1, 0) & \text{with probability } \epsilon. \end{cases} \tag{16.78}$$

In this market the log-optimal portfolio invests all the wealth in the first stock. [It is easy to verify that $\mathbf{b} = (1, 0)$ satisfies the Kuhn–Tucker conditions.] However, an investor who puts all his wealth in the second stock earns more money with probability $1 - \epsilon$. Hence, it is not true that with high probability the log-optimal investor will do better than any other investor.

The problem with trying to prove that the log-optimal investor does best with a probability of at least $\frac{1}{2}$ is that there exist examples like the one above, where it is possible to beat the log-optimal investor by a small amount most of the time. We can get around this by allowing each investor an additional fair randomization, which has the effect of reducing the effect of small differences in the wealth.

Theorem 16.6.1 (*Competitive optimality*) *Let S^* be the wealth at the end of one period of investment in a stock market* $\mathbf{X}$ *with the log-optimal portfolio, and let S be the wealth induced by any other portfolio. Let U^* be a random variable independent of* $\mathbf{X}$ *uniformly distributed on* $[0, 2]$, *and let V be any other random variable independent of* $\mathbf{X}$ *and U^* with $V \geq 0$ and $EV = 1$. Then*

$$\Pr(VS \geq U^*S^*) \leq \frac{1}{2}. \tag{16.79}$$

Remark Here U^* and V correspond to initial "fair" randomizations of the initial wealth. This exchange of initial wealth $S_0 = 1$ for "fair" wealth U^* can be achieved in practice by placing a fair bet. The effect of the fair randomization is to randomize small differences, so that only the significant deviations of the ratio S/S^* affect the probability of winning.

Proof: We have

$$\Pr(VS \geq U^*S^*) = \Pr\left(\frac{VS}{S^*} \geq U^*\right) \tag{16.80}$$

$$= \Pr(W \geq U^*), \tag{16.81}$$

where $W = \frac{VS}{S^*}$ is a non-negative-valued random variable with mean

$$EW = E(V)E\left(\frac{S_n}{S_n^*}\right) \leq 1 \tag{16.82}$$

by the independence of V from $\mathbf{X}$ and the Kuhn–Tucker conditions. Let F be the distribution function of W. Then since U^* is uniform on $[0, 2]$,

$$\Pr(W \geq U^*) = \int_0^2 \Pr(W > w) f_{U^*}(w)\, dw \tag{16.83}$$

$$= \int_0^2 \Pr(W > w) \frac{1}{2}\, dw \tag{16.84}$$

$$= \int_0^2 \frac{1 - F(w)}{2}\, dw \tag{16.85}$$

$$\leq \int_0^\infty \frac{1 - F(w)}{2}\, dw \tag{16.86}$$

$$= \frac{1}{2} EW \tag{16.87}$$

$$\leq \frac{1}{2}, \tag{16.88}$$

using the easily proved fact (by integrating by parts) that

$$EW = \int_0^\infty (1 - F(w))\, dw \tag{16.89}$$

for a positive random variable W. Hence, we have

$$\Pr(VS \geq U^*S^*) = \Pr(W \geq U^*) \leq \frac{1}{2}. \quad \square \tag{16.90}$$

Theorem 16.6.1 provides a short-term justification for the use of the log-optimal portfolio. If the investor's only objective is to be ahead of his opponent at the end of the day in the stock market, and if fair randomization is allowed, Theorem 16.6.1 says that the investor should exchange his wealth for a uniform $[0, 2]$ wealth and then invest using the log-optimal portfolio. This is the game-theoretic solution to the problem of gambling competitively in the stock market.

16.7 UNIVERSAL PORTFOLIOS

The development of the log-optimal portfolio strategy in Section 16.1 relies on the assumption that we know the distribution of the stock vectors and can therefore calculate the optimal portfolio $\mathbf{b}^*$. In practice, though, we often do not know the distribution. In this section we describe a causal portfolio that performs well on individual sequences. Thus, we make no statistical assumptions about the market sequence. We assume that the stock market can be represented by a sequence of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots \in \mathcal{R}_+^m$, where x_{ij} is the price relative for stock j on day i and $\mathbf{x}_i$ is the vector of price relatives for all stocks on day i. We begin with a finite-horizon problem, where we have n vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. We later extend the results to the infinite-horizon case.

Given this sequence of stock market outcomes, what is the best we can do? A realistic target is the growth achieved by the best constant rebalanced portfolio strategy in hindsight (i.e., the best constant rebalanced portfolio on the known sequence of stock market vectors). Note that constant rebalanced portfolios are optimal against i.i.d. stock market sequences with known distribution, so that this set of portfolios is reasonably natural.

Let us assume that we have a number of mutual funds, each of which follows a constant rebalanced portfolio strategy chosen in advance. Our objective is to perform as well as the best of these funds. In this section we show that we can do almost as well as the best constant rebalanced

portfolio without advance knowledge of the distribution of the stock market vectors.

One approach is to distribute the wealth among a continuum of fund managers, each of which follows a different constantly rebalanced portfolio strategy. Since one of the managers will do exponentially better than the others, the total wealth after n days will be dominated by the largest term. We will show that we can achieve a performance of the best fund manager within a factor of $n^{\frac{m-1}{2}}$. This is the essence of the argument for the infinite-horizon universal portfolio strategy.

A second approach to this problem is as a game against a malicious opponent or nature who is allowed to choose the sequence of stock market vectors. We define a causal (nonanticipating) portfolio strategy $\hat{\mathbf{b}}_i(\mathbf{x}_{i-1}, \ldots, \mathbf{x}_1)$ that depends only on the past values of the stock market sequence. Then nature, with knowledge of the strategy $\hat{\mathbf{b}}_i(\mathbf{x}^{i-1})$, chooses a sequence of vectors $\mathbf{x}_i$ to make the strategy perform as poorly as possible relative to the best constantly rebalanced portfolio for that stock sequence. Let $\mathbf{b}^*(\mathbf{x}^n)$ be the best constantly rebalanced portfolio for a stock market sequence $\mathbf{x}^n$. Note that $\mathbf{b}^*(\mathbf{x}^n)$ depends only on the empirical distribution of the sequence, not on the order in which the vectors occur. At the end of n days, a constantly rebalanced portfolio $\mathbf{b}$ achieves wealth:

$$S_n(\mathbf{b}, \mathbf{x}^n) = \prod_{i=1}^{n} \mathbf{b}^t \mathbf{x}_i, \tag{16.91}$$

and the best constant portfolio $\mathbf{b}^*(\mathbf{x}^n)$ achieves a wealth

$$S_n^*(\mathbf{x}^n) = \max_{\mathbf{b}} \prod_{i=1}^{n} \mathbf{b}^t \mathbf{x}_i, \tag{16.92}$$

whereas the nonanticipating portfolio $\hat{\mathbf{b}}_i(\mathbf{x}^{i-1})$ strategy achieves

$$\hat{S}_n(\mathbf{x}^n) = \prod_{i=1}^{n} \hat{\mathbf{b}}_i^t(\mathbf{x}^{i-1})\mathbf{x}_i. \tag{16.93}$$

Our objective is to find a nonanticipating portfolio strategy $\hat{\mathbf{b}}(\cdot) = (\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_2(\mathbf{x}_1), \ldots, \hat{\mathbf{b}}_i(\mathbf{x}^{i-1}))$ that does well in the worst case in terms of the ratio of $\hat{S}_n$ to S_n^*. We will find the optimal universal strategy and show that this strategy for each stock sequence achieves wealth $\hat{S}_n$ that is within a factor $V_n \approx n^{-\frac{m-1}{2}}$ of the wealth S_n^* achieved by the best constantly rebalanced portfolio on that sequence. This strategy depends on n, the horizon of

the game. Later we describe some horizon-free results that have the same worst-case asymptotic performance as that of the finite-horizon game.

16.7.1 Finite-Horizon Universal Portfolios

We begin by analyzing a stock market of n periods, where n is known in advance, and attempt to find a portfolio strategy that does well against all possible sequences of n stock market vectors. The main result can be stated in the following theorem.

Theorem 16.7.1 *For a stock market sequence $\mathbf{x}^n = \mathbf{x}_1, \ldots, \mathbf{x}_n$, $\mathbf{x}_i \in \mathcal{R}_+^m$ of length n with m assets, let $S_n^*(\mathbf{x}^n)$ be the wealth achieved by the optimal constantly rebalanced portfolio on $\mathbf{x}^n$, and let $\hat{S}_n(\mathbf{x}^n)$ be the wealth achieved by any causal portfolio strategy $\hat{\mathbf{b}}_i(\cdot)$ on $\mathbf{x}^n$; then*

$$\max_{\hat{\mathbf{b}}_i(\cdot)} \min_{\mathbf{x}_1,\ldots,\mathbf{x}_n} \frac{\hat{S}_n(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} = V_n, \tag{16.94}$$

where

$$V_n = \left[\sum_{n_1+\cdots+n_m=n} \binom{n}{n_1, n_2, \ldots, n_m} 2^{-nH(\frac{n_1}{n},\ldots,\frac{n_m}{n})} \right]^{-1}. \tag{16.95}$$

Using Stirling's approximation, we can show that V_n is on the order of $n^{-\frac{m-1}{2}}$, and therefore the growth rate for the universal portfolio on the worst sequence differs from the growth rate of the best constantly rebalanced portfolio on that sequence by at most a polynomial factor. The logarithm of the ratio of growth of wealth of the universal portfolio $\hat{\mathbf{b}}$ to the growth of wealth of the best constant portfolio behaves like the redundancy of a universal source code. (See Shtarkov [496], where $\log V_n$ appears as the minimax individual sequence redundancy in data compression.)

We first illustrate the main results by means of an example for $n = 1$. Consider the case of two stocks and a single day. Let the stock vector for the day be $\mathbf{x} = (x_1, x_2)$. If $x_1 > x_2$, the best portfolio is one that puts all its money on stock 1, and if $x_2 > x_1$, the best portfolio puts all its money on stock 2. (If $x_1 = x_2$, all portfolios are equivalent.)

Now assume that we must choose a portfolio in advance and our opponent can choose the stock market sequence after we have chosen our portfolio to make us do as badly as possible relative to the best portfolio. Given our portfolio, the opponent can ensure that we do as badly as possible by making the stock on which we have put more weight equal to 0 and the other stock equal to 1. Our best strategy is therefore to put equal

weight on both stocks, and with this, we will achieve a growth factor at least equal to half the growth factor of the best stock, and hence we will achieve at least half the gain of the best constantly rebalanced portfolio. It is not hard to calculate that $V_n = 2$ when $n = 1$ and $m = 2$ in equation (16.94).

However, this result seems misleading, since it appears to suggest that for n days, we would use a constant uniform portfolio, putting half our money on each stock every day. If our opponent then chose the stock sequence so that only the first stock was 1 (and the other was 0) every day, this uniform strategy would achieve a wealth of $1/2^n$, and we would achieve a wealth only within a factor of 2^n of the best constant portfolio, which puts all the money on the first stock for all time.

The result of the theorem shows that we can do significantly better. The main part of the argument is to reduce a sequence of stock vectors to the extreme cases where only one of the stocks is nonzero for each day. If we can ensure that we do well on such sequences, we can guarantee that we do well on any sequence of stock vectors, and achieve the bounds of the theorem.

Before we prove the theorem, we need the following lemma.

Lemma 16.7.1 *For $p_1, p_2, \ldots, p_m \geq 0$ and $q_1, q_2, \ldots, q_m \geq 0$,*

$$\frac{\sum_{i=1}^m p_i}{\sum_{i=1}^m q_i} \geq \min_i \frac{p_i}{q_i}. \tag{16.96}$$

Proof: Let I denote the index i that minimizes the right-hand side in (16.96). Assume that $p_I > 0$ (if $p_I = 0$, the lemma is trivially true). Also, if $q_I = 0$, both sides of (16.96) are infinite (all the other q_i's must also be zero), and again the inequality holds. Therefore, we can also assume that $q_I > 0$. Then

$$\frac{\sum_{i=1}^m p_i}{\sum_{i=1}^m q_i} = \frac{p_I}{q_I} \frac{1 + \sum_{i \neq I} (p_i/p_I)}{1 + \sum_{i \neq I} (q_i/q_I)} \geq \frac{p_I}{q_I} \tag{16.97}$$

because

$$\frac{p_i}{q_i} \geq \frac{p_I}{q_I} \longrightarrow \frac{p_i}{p_I} \geq \frac{q_i}{q_I} \tag{16.98}$$

for all i. □

First consider the case when $n = 1$. The wealth at the end of the first day is

$$\hat{S}_1(\mathbf{x}) = \hat{\mathbf{b}}^t \mathbf{x}, \tag{16.99}$$

$$S_1(\mathbf{x}) = \mathbf{b}^t \mathbf{x} \tag{16.100}$$

and

$$\frac{\hat{S}_1(\mathbf{x})}{S_1(\mathbf{x})} = \frac{\sum \hat{b}_i x_i}{\sum b_i x_i} \geq \min\left\{\frac{\hat{b}_i}{b_i}\right\}. \tag{16.101}$$

We wish to find $\max_{\hat{\mathbf{b}}} \min_{\mathbf{b},\mathbf{x}} \frac{\hat{\mathbf{b}}^t\mathbf{x}}{\mathbf{b}^t\mathbf{x}}$. Nature should choose $\mathbf{x} = \mathbf{e}_i$, where $\mathbf{e}_i$ is the ith basis vector with 1 in the component i that minimizes $\frac{\hat{b}_i}{b_i^*}$, and the investor should choose $\hat{\mathbf{b}}$ to maximize this minimum. This is achieved by choosing $\hat{\mathbf{b}} = (\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m})$.

The important point to realize is that

$$\frac{\hat{S}_n(\mathbf{x}^n)}{S_n(\mathbf{x}^n)} = \frac{\prod_{i=1}^n \hat{\mathbf{b}}_i^t\mathbf{x}_i}{\prod_{i=1}^n \mathbf{b}_i^t\mathbf{x}_i} \tag{16.102}$$

can also be rewritten in the form of a ratio of terms

$$\frac{\hat{S}_n(\mathbf{x}^n)}{S_n(\mathbf{x}^n)} = \frac{\hat{\mathbf{b}}^t\mathbf{x}'}{\mathbf{b}^t\mathbf{x}'}, \tag{16.103}$$

where $\hat{\mathbf{b}}$, $\mathbf{b}$, $\mathbf{x}' \in \mathcal{R}_+^{m^n}$. Here the m^n components of the constantly rebalanced portfolios $\mathbf{b}$ are all of the product form $b_1^{n_1} b_2^{n_2} \cdots b_m^{n_m}$. One wishes to find a universal $\hat{\mathbf{b}}$ that is uniformly close to the $\mathbf{b}$'s corresponding to constantly rebalanced portfolios.

We can now prove the main theorem (Theorem 16.7.1).

Proof of Theorem 16.7.1: We will prove the theorem for $m = 2$. The proof extends in a straightforward fashion to the case $m > 2$. Denote the stocks by 1 and 2. The key idea is to express the wealth at time n,

$$S_n(\mathbf{x}^n) = \prod_{i=1}^n \mathbf{b}_i^t\mathbf{x}_i, \tag{16.104}$$

which is a product of sums, into a sum of products. Each term in the sum corresponds to a sequence of stock price relatives for stock 1 or stock 2 times the proportion b_{i1} or b_{i2} that the strategy places on stock 1 or stock 2 at time i. We can therefore view the wealth S_n as a sum over all 2^n possible n-sequences of 1's and 2's of the product of the portfolio proportions times the stock price relatives:

$$S_n(\mathbf{x}^n) = \sum_{j^n \in \{1,2\}^n} \prod_{i=1}^n b_{ij_i} x_{ij_i} = \sum_{j^n \in \{1,2\}^n} \prod_{i=1}^n b_{ij_i} \prod_{i=1}^n x_{ij_i}. \tag{16.105}$$

If we let $w(j^n)$ denote the product $\prod_{i=1}^{n} b_{i j_i}$, the total fraction of wealth invested in the sequence j^n, and let

$$x(j^n) = \prod_{i=1}^{n} x_{i j_i} \tag{16.106}$$

be the corresponding return for this sequence, we can write

$$S_n(\mathbf{x}^n) = \sum_{j^n \in \{1,2\}^n} w(j^n) x(j^n). \tag{16.107}$$

Similar expressions apply to both the best constantly rebalanced portfolio and the universal portfolio strategy. Thus, we have

$$\frac{\hat{S}_n(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} = \frac{\sum_{j^n \in \{1,2\}^n} \hat{w}(j^n) x(j^n)}{\sum_{j^n \in \{1,2\}^n} w^*(j^n) x(j^n)}, \tag{16.108}$$

where $\hat{w}^n$ is the amount of wealth placed on the sequence j^n by the universal nonanticipating strategy, and $w^*(j^n)$ is the amount placed by the best constant rebalanced portfolio strategy. Now applying Lemma 16.7.1, we have

$$\frac{\hat{S}_n(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} \geq \min_{j^n} \frac{\hat{w}(j^n) x(j^n)}{w^*(j^n) x(j^n)} = \min_{j^n} \frac{\hat{w}(j^n)}{w^*(j^n)}. \tag{16.109}$$

Thus, the problem of maximizing the performance ratio $\hat{S}_n / S_n^*$ is reduced to ensuring that the proportion of money bet on a sequence of stocks by the universal portfolio is uniformly close to the proportion bet by $\mathbf{b}^*$. As might be obvious by now, this formulation of S_n reduces the n-period stock market to a special case of a single-period stock market—there are 2^n stocks, one invests $w(j^n)$ in stock j^n and receives a return $x(j^n)$ for stock j^n, and the total wealth S_n is $\sum_{j^n} w(j^n) x(j^n)$.

We first calculate the weight $w^*(j^n)$ associated with the best constant rebalanced portfolio $\mathbf{b}^*$. We observe that a constantly rebalanced portfolio $\mathbf{b}$ results in

$$w(j^n) = \prod_{i=1}^{n} b_{i j_i} = b^k (1-b)^{n-k}, \tag{16.110}$$

where k is the number of times 1 appears in the sequence j^n. Thus, $w(j^n)$ depends only on k, the number of 1's in j^n. Fixing attention on j^n, we

find by differentiating with respect to b that the maximum value

$$w^*(j^n) = \max_{0 \le b \le 1} b^k (1-b)^{n-k} \tag{16.111}$$

$$= \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}, \tag{16.112}$$

which is achieved by

$$\mathbf{b}^* = \left(\frac{k}{n}, \frac{n-k}{n}\right). \tag{16.113}$$

Note that $\sum w^*(j^n) > 1$, reflecting the fact that the amount "bet" on j^n is chosen in hindsight, thus relieving the hindsight investor of the responsibility of allocating his investments $w^*(j^n)$ to sum to 1. The causal investor has no such luxury. How can the causal investor choose initial investments $\hat{w}(j^n)$, $\sum \hat{w}(j^n) = 1$, to protect himself from all possible j^n and hindsight-determined $w^*(j^n)$? The answer will be to choose $\hat{w}(j^n)$ proportional to $w^*(j^n)$. Then the worst-case ratio of $\hat{w}(j^n)/w^*(j^n)$ will be maximized. To proceed, we define V_n by

$$\frac{1}{V_n} = \sum_{j^n} \left(\frac{k(j^n)}{n}\right)^{k(j^n)} \left(\frac{n-k(j^n)}{n}\right)^{n-k(j^n)} \tag{16.114}$$

$$= \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \tag{16.115}$$

and let

$$\hat{w}(j^n) = V_n \left(\frac{k(j^n)}{n}\right)^{k(j^n)} \left(\frac{n-k(j^n)}{n}\right)^{n-k(j^n)}. \tag{16.116}$$

It is clear that $\hat{w}(j^n)$ is a legitimate distribution of wealth over the 2^n stock sequences (i.e., $\hat{w}(j^n) \ge 0$ and $\sum_{j^n} \hat{w}(j^n) = 1$). Here V_n is the normalization factor that makes $\hat{w}(j^n)$ a probability mass function. Also, from (16.109) and (16.113), for all sequences $\mathbf{x}^n$,

$$\frac{\hat{S}_n(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} \ge \min_{j^n} \frac{\hat{w}(j^n)}{w^*(j^n)} \tag{16.117}$$

$$= \min_k \frac{V_n (\frac{k}{n})^k (\frac{n-k}{n})^{n-k}}{b^{*k}(1-b^*)^{n-k}} \tag{16.118}$$

$$\ge V_n, \tag{16.119}$$

where (16.117) follows from (16.109) and (16.119) follows from (16.112). Consequently, we have

$$\max_{\hat{\mathbf{b}}} \min_{\mathbf{x}^n} \frac{\hat{S}_n(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} \geq V_n. \tag{16.120}$$

We have thus demonstrated a portfolio on the 2^n possible sequences of length n that achieves wealth $\hat{S}_n(\mathbf{x}^n)$ within a factor V_n of the wealth $S_n^*(\mathbf{x}^n)$ achieved by the best constant rebalanced portfolio in hindsight. To complete the proof of the theorem, we show that this is the best possible, that is, that any nonanticipating portfolio $\mathbf{b}_i(\mathbf{x}^{i-1})$ cannot do better than a factor V_n in the worst case (i.e., for the worst choice of $\mathbf{x}^n$). To prove this, we construct a set of extremal stock market sequences and show that the performance of any nonanticipating portfolio strategy is bounded by V_n for at least one of these sequences, proving the worst-case bound.

For each $j^n \in \{1, 2\}^n$, we define the corresponding extremal stock market vector $\mathbf{x}^n(j^n)$ as

$$\mathbf{x}_i(j_i) = \begin{cases} (1, 0)^t & \text{if } j_i = 1, \\ (0, 1)^t & \text{if } j_i = 2, \end{cases} \tag{16.121}$$

Let $\mathbf{e}_1 = (1, 0)^t$, $\mathbf{e}_2 = (0, 1)^t$ be standard basis vectors. Let

$$\mathcal{K} = \{\mathbf{x}(j^n) : j^n \in \{1, 2\}^n, \mathbf{x}_{ij_i} = \mathbf{e}_{j_i}\} \tag{16.122}$$

be the set of extremal sequences. There are 2^n such extremal sequences, and for each sequence at each time, there is only one stock that yields a nonzero return. The wealth invested in the other stock is lost. Therefore, the wealth at the end of n periods for extremal sequence $\mathbf{x}^n(j^n)$ is the product of the amounts invested in the stocks $j_1, j_2, \ldots, j_n$, [i.e., $S_n(\mathbf{x}^n(j^n)) = \prod_i b_{j_i} = w(j^n)$]. Again, we can view this as an investment on sequences of length n, and given the 0–1 nature of the return, it is easy to see for $\mathbf{x}^n \in \mathcal{K}$ that

$$\sum_{j^n} S_n(\mathbf{x}^n(j^n)) = 1. \tag{16.123}$$

For any extremal sequence $\mathbf{x}^n(j^n) \in \mathcal{K}$, the best constant rebalanced portfolio is

$$\mathbf{b}^*(\mathbf{x}^n(j^n)) = \left(\frac{n_1(j^n)}{n}, \quad \frac{n_2(j^n)}{n}\right)^t, \tag{16.124}$$

where $n_1(j^n)$ is the number of occurrences of 1 in the sequence j^n. The corresponding wealth at the end of n periods is

$$S_n^*(\mathbf{x}^n(j^n)) = \left(\frac{n_1(j^n)}{n}\right)^{n_1(j^n)} \left(\frac{n_2(j^n)}{n}\right)^{n_2(j^n)} = \frac{\hat{w}(j^n)}{V_n}, \tag{16.125}$$

from (16.116) and it therefore follows that

$$\sum_{\mathbf{x}^n \in \mathcal{K}} S_n^*(\mathbf{x}^n) = \frac{1}{V_n} \sum_{j^n} \hat{w}(j^n) = \frac{1}{V_n}. \tag{16.126}$$

We then have the following inequality for any portfolio sequence $\{\mathbf{b}_i\}_{i=1}^n$, with $S_n(\mathbf{x}^n)$ defined as in (16.104):

$$\min_{\mathbf{x}^n \in \mathcal{K}} \frac{S_n(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} \le \sum_{\tilde{\mathbf{x}}^n \in \mathcal{K}} \frac{S_n^*(\tilde{\mathbf{x}}^n)}{\sum_{\mathbf{x}^n \in \mathcal{K}} S_n^*(\mathbf{x}^n)} \frac{S_n(\tilde{\mathbf{x}}^n)}{S_n^*(\tilde{\mathbf{x}}^n)} \tag{16.127}$$

$$= \sum_{\tilde{\mathbf{x}}^n \in \mathcal{K}} \frac{S_n(\tilde{\mathbf{x}}^n)}{\sum_{\mathbf{x}^n \in \mathcal{K}} S_n^*(\mathbf{x}^n)} \tag{16.128}$$

$$= \frac{1}{\sum_{\mathbf{x}^n \in \mathcal{K}} S_n^*(\mathbf{x}^n)} \tag{16.129}$$

$$= V_n, \tag{16.130}$$

where the inequality follows from the fact that the minimum is less than the average. Thus,

$$\max_{\mathbf{b}} \min_{\mathbf{x}^n \in \mathcal{K}} \frac{S_n(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} \le V_n. \quad \square \tag{16.131}$$

The strategy described in the theorem puts mass on all sequences of length n and is clearly dependent on n. We can recast the strategy in incremental terms (i.e., in terms of the amount bet on stock 1 and stock 2 at time 1), then, conditional on the outcome at time 1, the amount bet on each of the two stocks at time 2, and so on. Consider the weight $\hat{b}_{i,1}$ assigned by the algorithm to stock 1 at time i given the previous sequence of stock vectors $\mathbf{x}^{i-1}$. We can calculate this by summing over all sequences j^n that have a 1 in position i, giving

$$\hat{\mathbf{b}}_{i,1}(\mathbf{x}^{i-1}) = \frac{\sum_{j^{i-1} \in M^{i-1}} \hat{w}(j^{i-1}1)x(j^{i-1})}{\sum_{j^i \in M^i} \hat{w}(j^i)x(j^{i-1})}, \tag{16.132}$$

where

$$\hat{w}(j^i) = \sum_{j^n : j^i \subseteq j^n} w(j^n) \tag{16.133}$$

is the weight put on all sequences j^n that start with j^i, and

$$x(j^{i-1}) = \prod_{k=1}^{i-1} x_{kj_k} \tag{16.134}$$

is the return on those sequences as defined in (16.106).

Investigation of the asymptotics of V_n reveals [401, 496] that

$$V_n \sim \left(\sqrt{\frac{2}{n}}\right)^{m-1} \Gamma(m/2)/\sqrt{\pi} \tag{16.135}$$

for m assets. In particular, for $m = 2$ assets,

$$V_n \sim \sqrt{\frac{2}{\pi n}} \tag{16.136}$$

and

$$\frac{1}{2\sqrt{n+1}} \leq V_n \leq \frac{2}{\sqrt{n+1}} \tag{16.137}$$

for all n [400]. Consequently, for $m = 2$ stocks, the causal portfolio strategy $\hat{\mathbf{b}}_i(\mathbf{x}^{i-1})$ given in (16.132) achieves wealth $\hat{S}_n(x^n)$ such that

$$\frac{\hat{S}_n(x^n)}{S_n^*(x^n)} \geq V_n \geq \frac{1}{2\sqrt{n+1}} \tag{16.138}$$

for all market sequences x^n.

16.7.2 Horizon-Free Universal Portfolios

We describe the horizon-free strategy in terms of a weighting of different portfolio strategies. As described earlier, each constantly rebalanced portfolio $\mathbf{b}$ can be viewed as corresponding to a mutual fund that rebalances the m assets according to $\mathbf{b}$. Initially, we distribute the wealth among these funds according to a distribution $\mu(\mathbf{b})$, where $d\mu(\mathbf{b})$ is the amount of wealth invested in portfolios in the neighborhood $d\mathbf{b}$ of the constantly rebalanced portfolio $\mathbf{b}$.

Let

$$S_n(\mathbf{b}, \mathbf{x}^n) = \prod_{i=1}^{n} \mathbf{b}^t \mathbf{x}_i \tag{16.139}$$

be the wealth generated by a constant rebalanced portfolio $\mathbf{b}$ on the stock sequence $\mathbf{x}^n$. Recall that

$$S_n^*(\mathbf{x}^n) = \max_{b \in \mathcal{B}} S_n(\mathbf{b}, \mathbf{x}^n) \tag{16.140}$$

is the wealth of the best constant rebalanced portfolio in hindsight.

We investigate the causal portfolio defined by

$$\hat{\mathbf{b}}_{i+1}(\mathbf{x}^i) = \frac{\int_{\mathcal{B}} \mathbf{b} S_i(\mathbf{b}, \mathbf{x}^i)\, d\mu(\mathbf{b})}{\int_{\mathcal{B}} S_i(\mathbf{b}, \mathbf{x}^i)\, d\mu(\mathbf{b})}. \tag{16.141}$$

We note that

$$\hat{\mathbf{b}}_{i+1}^t(\mathbf{x}^i)\mathbf{x}_{i+1} = \frac{\int_{\mathcal{B}} \mathbf{b}^t \mathbf{x}_{i+1} S_i(\mathbf{b}, \mathbf{x}^i)\, d\mu(\mathbf{b})}{\int_{\mathcal{B}} S_i(\mathbf{b}, \mathbf{x}^i)\, d\mu(\mathbf{b})} \tag{16.142}$$

$$= \frac{\int_{\mathcal{B}} S_{i+1}(\mathbf{b}, \mathbf{x}^{i+1})\, d\mu(\mathbf{b})}{\int_{\mathcal{B}} S_i(\mathbf{b}, \mathbf{x}^i)\, d\mu(\mathbf{b})}. \tag{16.143}$$

Thus, the product $\prod \hat{\mathbf{b}}_i^t \mathbf{x}_i$ telescopes and we see that the wealth $\hat{S}_n(\mathbf{x}^n)$ resulting from this portfolio is given by

$$\hat{S}_n(\mathbf{x}^n) = \prod_{i=1}^{n} \hat{\mathbf{b}}_i^t(\mathbf{x}^{i-1})\mathbf{x}_i \tag{16.144}$$

$$= \int_{b \in \mathcal{B}} S_n(\mathbf{b}, \mathbf{x}^n)\, d\mu(\mathbf{b}). \tag{16.145}$$

There is another way to interpret (16.145). The amount given to portfolio manager $\mathbf{b}$ is $d\mu(\mathbf{b})$, the resulting growth factor for the manager rebalancing to $\mathbf{b}$ is $S(\mathbf{b}, \mathbf{x}^n)$, and the total wealth of this batch of investments is

$$\hat{S}_n(\mathbf{x}^n) = \int_{\mathcal{B}} S_n(\mathbf{b}, \mathbf{x}^n)\, d\mu(\mathbf{b}). \tag{16.146}$$

Then $\hat{\mathbf{b}}_{i+1}$, defined in (16.141), is the performance-weighted total "buy order" of the individual portfolio manager $\mathbf{b}$.

So far, we have not specified what distribution $\mu(\mathbf{b})$ we use to apportion the initial wealth. We now use a distribution μ that puts mass on all possible portfolios, so that we approximate the performance of the best portfolio for the actual distribution of stock price vectors.

In the next lemma, we bound $\hat{S}_n/S_n^*$ as a function of the initial wealth distribution $\mu(\mathbf{b})$.

Lemma 16.7.2 *Let $S_n^*(\mathbf{x}^n)$ in 16.140 be the wealth achieved by the best constant rebalanced portfolio and let $\hat{S}_n(\mathbf{x}^n)$ in (16.144) be the wealth achieved by the universal mixed portfolio $\hat{\mathbf{b}}(\cdot)$, given by*

$$\hat{\mathbf{b}}_{i+1}(\mathbf{x}^i) = \frac{\int \mathbf{b} S_i(\mathbf{b}, \mathbf{x}^i)\, d\mu(\mathbf{b})}{\int S_i(\mathbf{b}, \mathbf{x}^i)\, d\mu(\mathbf{b})}. \tag{16.147}$$

Then

$$\frac{\hat{S}_n(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} \geq \min_{j^n} \frac{\int_{\mathcal{B}} \prod_{i=1}^n b_{j_i}\, d\mu(\mathbf{b})}{\prod_{i=1}^n b_{j_i}^*}. \tag{16.148}$$

Proof: As before, we can write

$$S_n^*(\mathbf{x}^n) = \sum_{j^n} w^*(j^n) x(j^n), \tag{16.149}$$

where $w^*(j^n) = \prod_{i=1}^n b_{j_i}^*$ is the amount invested on the sequence j^n and $x(j^n) = \prod_{i=1}^n x_{ij_i}$ is the corresponding return. Similarly, we can write

$$\hat{S}_n(\mathbf{x}^n) = \int \prod_{i=1}^n \mathbf{b}^t \mathbf{x}_i\, d\mu(\mathbf{b}) \tag{16.150}$$

$$= \sum_{j^n} \int \prod_{i=1}^n b_{j_i} x_{ij_i}\, d\mu(\mathbf{b}) \tag{16.151}$$

$$= \sum_{j^n} \hat{w}(j^n) x(j^n), \tag{16.152}$$

where $\hat{w}(j^n) = \int \prod_{i=1}^n b_{j_i}\, d\mu(\mathbf{b})$. Now applying Lemma 16.7.1, we have

$$\frac{\hat{S}_n(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} = \frac{\sum_{j^n} \hat{w}(j^n) x(j^n)}{\sum_{j^n} w^*(j^n) x(j^n)} \tag{16.153}$$

$$\geq \min_{j^n} \frac{\hat{w}(j^n) x(j^n)}{w^*(j^n) x(j^n)} \tag{16.154}$$

$$= \min_{j^n} \frac{\int_{\mathcal{B}} \prod_{i=1}^n b_{j_i}\, d\mu(\mathbf{b})}{\prod_{i=1}^n b_{j_i}^*}. \quad \square \tag{16.155}$$

We now apply this lemma when $\mu(\mathbf{b})$ is the Dirichlet($\frac{1}{2}$) distribution.

Theorem 16.7.2 *For the causal universal portfolio $\hat{b}_i(\,)$, $i = 1, 2, \ldots$, given in (16.141), with $m = 2$ stocks and $d\mu(\mathbf{b})$ the Dirichlet($\frac{1}{2}, \frac{1}{2}$) distribution, we have*

$$\frac{\hat{S}_n(x^n)}{S_n^*(x^n)} \geq \frac{1}{2\sqrt{n+1}},$$

for all n and all stock sequences x^n.

Proof: As in the discussion preceding (16.112), we can show that the weight put by the best constant portfolio b^* on the sequence j^n is

$$\prod_{i=1}^{n} b_{j_i}^* = \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} = 2^{-nH(k/n)}, \tag{16.156}$$

where k is the number of indices where $j_i = 1$. We can also explicitly calculate the integral in the numerator of (16.148) in Lemma 16.7.2 for the Dirichlet($\frac{1}{2}$) density, defined for m variables as

$$d\mu(\mathbf{b}) = \frac{\Gamma(\frac{m}{2})}{\left[\Gamma\left(\frac{1}{2}\right)\right]^m} \prod_{j=1}^{m} b_j^{-\frac{1}{2}}\, d\mathbf{b}, \tag{16.157}$$

where $\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt$ denotes the gamma function. For simplicity, we consider the case of two stocks, in which case

$$d\mu(b) = \frac{1}{\pi} \frac{1}{\sqrt{b(1-b)}}\, db, \qquad 0 \leq b \leq 1, \tag{16.158}$$

where b is the fraction of wealth invested in stock 1. Now consider any sequence $j^n \in \{1, 2\}^n$, and consider the amount invested in that sequence,

$$b(j^n) = \prod_{i=1}^{n} b_{j_i} = b^l (1-b)^{n-l}, \tag{16.159}$$

where l is the number of indices where $j_i = 1$. Then

$$\int b(j^n)\, d\mu(\mathbf{b}) = \int b^l (1-b)^{n-l} \frac{1}{\pi} \frac{1}{\sqrt{b(1-b)}}\, db \tag{16.160}$$

$$= \frac{1}{\pi} \int b^{l-\frac{1}{2}} (1-b)^{n-l-\frac{1}{2}} \, db \tag{16.161}$$

$$\stackrel{\triangle}{=} \frac{1}{\pi} B\left(l + \frac{1}{2}, n - l + \frac{1}{2}\right), \tag{16.162}$$

where $B(\lambda_1, \lambda_2)$ is the beta function, defined as

$$B(\lambda_1, \lambda_2) = \int_0^1 x^{\lambda_1 - 1} (1-x)^{\lambda_2 - 1} \, dx \tag{16.163}$$

$$= \frac{\Gamma(\lambda_1)\Gamma(\lambda_2)}{\Gamma(\lambda_1 + \lambda_2)} \tag{16.164}$$

and

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda - 1} e^{-x} \, dx. \tag{16.165}$$

Note that for any integer n, $\Gamma(n+1) = n!$ and $\Gamma(n + \frac{1}{2}) = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2^n} \sqrt{\pi}$.

We can calculate $B(l + \frac{1}{2}, n - l + \frac{1}{2})$ by means of simple recursion using integration by parts. Alternatively, using (16.164), we obtain

$$B\left(l + \frac{1}{2}, n - l + \frac{1}{2}\right) = \frac{\pi}{2^{2n}} \frac{\binom{2n}{n}\binom{n}{l}}{\binom{2n}{2l}}. \tag{16.166}$$

Combining all the results with Lemma 16.7.2, we have

$$\frac{\hat{S}_n(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} \geq \min_{j^n} \frac{\int_{\mathcal{B}} \prod_{i=1}^n b_{j_i} \, d\mu(\mathbf{b})}{\prod_{i=1}^n b_{j_i}^*} \tag{16.167}$$

$$\geq \min_l \frac{\frac{1}{\pi} B(l + \frac{1}{2}, n - l + \frac{1}{2})}{2^{-nH(l/n)}} \tag{16.168}$$

$$\geq \frac{1}{2\sqrt{n+1}}, \tag{16.169}$$

using the results in [135, Theorem 2]. □

It follows for $m = 2$ stocks that

$$\frac{\hat{S}_n}{S_n^*} \geq \frac{1}{\sqrt{2\pi}} V_n \tag{16.170}$$

for all n and all market sequences $x_1, x_2, \ldots, x_n$. Thus, good minimax performance for all n costs at most an extra factor $\sqrt{2\pi}$ over the fixed horizon minimax portfolio. The cost of universality is V_n, which is asymptotically negligible in the growth rate in the sense that

$$\frac{1}{n} \ln \hat{S}_n(\mathbf{x}^n) - \frac{1}{n} \ln S_n^*(\mathbf{x}^n) \geq \frac{1}{n} \ln \frac{V_n}{\sqrt{2\pi}} \to 0. \tag{16.171}$$

Thus, the universal causal portfolio achieves the same asymptotic growth rate of wealth as the best hindsight portfolio.

Let's now consider how this portfolio algorithm performs on two real stocks. We consider a 14-year period (ending in 2004) and two stocks, Hewlett-Packard and Altria (formerly, Phillip Morris), which are both components of the Dow Jones Index. Over these 14 years, HP went up by a factor of 11.8, while Altria went up by a factor of 11.5. The performance of the different constantly rebalanced portfolios that contain HP and Altria are shown in Figure 16.2. The best constantly rebalanced portfolio (which can be computed only in hindsight) achieves a growth of a factor of 18.7 using a mixture of about 51% HP and 49% Altria. The universal portfolio strategy described in this section achieves a growth factor of 15.7 without foreknowledge.

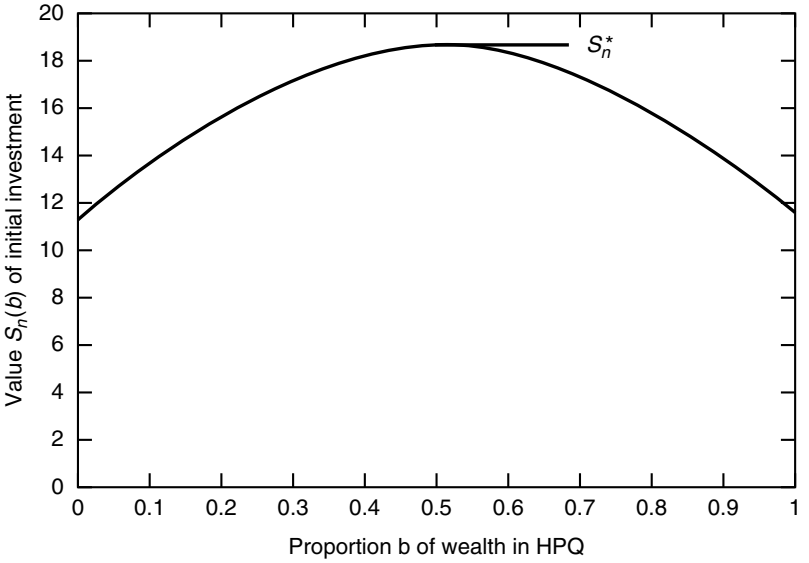


FIGURE 16.2. Performance of different constant rebalanced portfolios **b** for HP and Altria.

16.8 SHANNON–MCMILLAN–BREIMAN THEOREM (GENERAL AEP)

The AEP for ergodic processes has come to be known as the *Shannon–McMillan–Breiman theorem*. In Chapter 3 we proved the AEP for i.i.d. processes. In this section we offer a proof of the theorem for a general ergodic process. We prove the convergence of $\frac{1}{n}\log p(X^n)$ by sandwiching it between two ergodic sequences.

In a sense, an ergodic process is the most general dependent process for which the strong law of large numbers holds. For finite alphabet processes, ergodicity is equivalent to the convergence of the kth-order empirical distributions to their marginals for all k.

The technical definition requires some ideas from probability theory. To be precise, an ergodic source is defined on a probability space $(\Omega, \mathcal{B}, P)$, where $\mathcal{B}$ is a σ-algebra of subsets of Ω and P is a probability measure. A random variable X is defined as a function $X(\omega)$, $\omega \in \Omega$, on the probability space. We also have a transformation $T : \Omega \to \Omega$, which plays the role of a time shift. We will say that the transformation is *stationary* if $P(TA) = P(A)$ for all $A \in \mathcal{B}$. The transformation is called *ergodic* if every set A such that $TA = A$, a.e., satisfies $P(A) = 0$ or 1. If T is stationary and ergodic, we say that the process defined by $X_n(\omega) = X(T^n\omega)$ is stationary and ergodic. For a stationary ergodic source, Birkhoff's ergodic theorem states that

$$\frac{1}{n}\sum_{i=1}^{n} X_i(\omega) \to EX = \int X\,dP \quad \text{with probability 1.} \tag{16.172}$$

Thus, the law of large numbers holds for ergodic processes.

We wish to use the ergodic theorem to conclude that

$$\begin{aligned} -\frac{1}{n}\log p(X_0, X_1, \ldots, X_{n-1}) &= -\frac{1}{n}\sum_{i=0}^{n-1} \log p(X_i|X_0^{i-1}) \\ &\to \lim_{n\to\infty} E[-\log p(X_n|X_0^{n-1})]. \end{aligned} \tag{16.173}$$

But the stochastic sequence $p(X_i|X_0^{i-1})$ is not ergodic. However, the closely related quantities $p(X_i|X_{i-k}^{i-1})$ and $p(X_i|X_{-\infty}^{i-1})$ are ergodic and have expectations easily identified as entropy rates. We plan to sandwich $p(X_i|X_0^{i-1})$ between these two more tractable processes.

We define the kth-order entropy H^k as

$$H^k = E\{-\log p(X_k|X_{k-1}, X_{k-2}, \ldots, X_0)\} \tag{16.174}$$

$$= E\{-\log p(X_0|X_{-1}, X_{-2}, \ldots, X_{-k})\}, \tag{16.175}$$

where the last equation follows from stationarity. Recall that the entropy rate is given by

$$H = \lim_{k\to\infty} H^k \tag{16.176}$$

$$= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} H^k. \tag{16.177}$$

Of course, $H^k \searrow H$ by stationarity and the fact that conditioning does not increase entropy. It will be crucial that $H^k \searrow H = H^\infty$, where

$$H^\infty = E\{-\log p(X_0|X_{-1}, X_{-2}, \ldots)\}. \tag{16.178}$$

The proof that $H^\infty = H$ involves exchanging expectation and limit.

The main idea in the proof goes back to the idea of (conditional) proportional gambling. A gambler receiving uniform odds with the knowledge of the k past will have a growth rate of wealth $\log|\mathcal{X}| - H^k$, while a gambler with a knowledge of the infinite past will have a growth rate of wealth of $\log|\mathcal{X}| - H^\infty$. We don't know the wealth growth rate of a gambler with growing knowledge of the past X_0^n, but it is certainly sandwiched between $\log|\mathcal{X}| - H^k$ and $\log|\mathcal{X}| - H^\infty$. But $H^k \searrow H = H^\infty$. Thus, the sandwich closes and the growth rate must be $\log|\mathcal{X}| - H$.

We will prove the theorem based on lemmas that will follow the proof.

Theorem 16.8.1 (*AEP: Shannon–McMillan–Breiman Theorem*) *If H is the entropy rate of a finite-valued stationary ergodic process $\{X_n\}$, then*

$$-\frac{1}{n}\log p(X_0, \ldots, X_{n-1}) \to H \quad \textit{with probability 1.} \tag{16.179}$$

Proof: We prove this for finite alphabet $\mathcal{X}$; this proof and the proof for countable alphabets and densities is given in Algoet and Cover [20]. We argue that the sequence of random variables $-\frac{1}{n}\log p(X_0^{n-1})$ is asymptotically sandwiched between the upper bound H^k and the lower bound H^∞ for all $k \geq 0$. The AEP will follow since $H^k \to H^\infty$ and $H^\infty = H$. The kth-order Markov approximation to the probability is defined for $n \geq k$ as

$$p^k(X_0^{n-1}) = p(X_0^{k-1}) \prod_{i=k}^{n-1} p(X_i|X_{i-k}^{i-1}). \tag{16.180}$$

From Lemma 16.8.3 we have

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{p^k(X_0^{n-1})}{p(X_0^{n-1})} \le 0, \tag{16.181}$$

which we rewrite, taking the existence of the limit $\frac{1}{n} \log p^k(X_0^n)$ into account (Lemma 16.8.1), as

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{p(X_0^{n-1})} \le \lim_{n\to\infty} \frac{1}{n} \log \frac{1}{p^k(X_0^{n-1})} = H^k \tag{16.182}$$

for $k = 1, 2, \ldots$. Also, from Lemma 16.8.3, we have

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{p(X_0^{n-1})}{p(X_0^{n-1}|X_{-\infty}^{-1})} \le 0, \tag{16.183}$$

which we rewrite as

$$\liminf \frac{1}{n} \log \frac{1}{p(X_0^{n-1})} \ge \lim \frac{1}{n} \log \frac{1}{p(X_0^{n-1}|X_{-\infty}^{-1})} = H^\infty \tag{16.184}$$

from the definition of H^∞ in Lemma 16.8.1.

Putting together (16.182) and (16.184), we have

$$\begin{aligned} H^\infty &\le \liminf -\frac{1}{n} \log p(X_0^{n-1}) \le \limsup -\frac{1}{n} \log p(X_0^{n-1}) \\ &\le H^k \quad \text{for all } k. \end{aligned} \tag{16.185}$$

But by Lemma 16.8.2, $H^k \to H^\infty = H$. Consequently,

$$\lim -\frac{1}{n} \log p(X_0^n) = H. \quad \square \tag{16.186}$$

We now prove the lemmas that were used in the main proof. The first lemma uses the ergodic theorem.

Lemma 16.8.1 (*Markov approximations*) *For a stationary ergodic stochastic process* $\{X_n\}$,

$$-\frac{1}{n} \log p^k(X_0^{n-1}) \to H^k \quad \textit{with probability 1}, \tag{16.187}$$

$$-\frac{1}{n} \log p(X_0^{n-1}|X_{-\infty}^{-1}) \to H^\infty \quad \textit{with probability 1}. \tag{16.188}$$

Proof: Functions $Y_n = f(X_{-\infty}^n)$ of ergodic processes $\{X_i\}$ are ergodic processes. Thus, $\log p(X_n|X_{n-k}^{n-1})$ and $\log p(X_n|X_{n-1}, X_{n-2}, \ldots,)$ are also ergodic processes, and

$$-\frac{1}{n}\log p^k(X_0^{n-1}) = -\frac{1}{n}\log p(X_0^{k-1}) - \frac{1}{n}\sum_{i=k}^{n-1}\log p(X_i|X_{i-k}^{i-1}) \quad (16.189)$$

$$\to 0 + H^k \quad \text{with probability 1,} \quad (16.190)$$

by the ergodic theorem. Similarly, by the ergodic theorem,

$$-\frac{1}{n}\log p(X_0^{n-1}|X_{-1}, X_{-2}, \ldots) = -\frac{1}{n}\sum_{i=0}^{n-1}\log p(X_i|X_{i-1}, X_{i-2}, \ldots) \quad (16.191)$$

$$\to H^\infty \quad \text{with probability 1.} \quad \square \quad (16.192)$$

Lemma 16.8.2 (*No gap*) $H^k \searrow H^\infty$ *and* $H = H^\infty$.

Proof: We know that for stationary processes, $H^k \searrow H$, so it remains to show that $H^k \searrow H^\infty$, thus yielding $H = H^\infty$. Levy's martingale convergence theorem for conditional probabilities asserts that

$$p(x_0|X_{-k}^{-1}) \to p(x_0|X_{-\infty}^{-1}) \quad \text{with probability 1} \quad (16.193)$$

for all $x_0 \in \mathcal{X}$. Since $\mathcal{X}$ is finite and $p \log p$ is bounded and continuous in p for all $0 \le p \le 1$, the bounded convergence theorem allows interchange of expectation and limit, yielding

$$\lim_{k\to\infty} H^k = \lim_{k\to\infty} E\left\{ -\sum_{x_0\in\mathcal{X}} p(x_0|X_{-k}^{-1})\log p(x_0|X_{-k}^{-1})\right\} \quad (16.194)$$

$$= E\left\{ -\sum_{x_0\in\mathcal{X}} p(x_0|X_{-\infty}^{-1})\log p(x_0|X_{-\infty}^{-1})\right\} \quad (16.195)$$

$$= H^\infty. \quad (16.196)$$

Thus, $H^k \searrow H = H^\infty$. $\square$

Lemma 16.8.3 (*Sandwich*)

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{p^k(X_0^{n-1})}{p(X_0^{n-1})} \le 0, \tag{16.197}$$

$$\limsup \frac{1}{n} \log \frac{p(X_0^{n-1})}{p(X_0^{n-1}|X_{-\infty}^{-1})} \le 0. \tag{16.198}$$

Proof: Let A be the support set of $p(X_0^{n-1})$. Then

$$E\left\{\frac{p^k(X_0^{n-1})}{p(X_0^{n-1})}\right\} = \sum_{x_0^{n-1}\in A} p(x_0^{n-1}) \frac{p^k(x_0^{n-1})}{p(x_0^{n-1})} \tag{16.199}$$

$$= \sum_{x_0^{n-1}\in A} p^k(x_0^{n-1}) \tag{16.200}$$

$$= p^k(A) \tag{16.201}$$

$$\le 1. \tag{16.202}$$

Similarly, let $B(X_{-\infty}^{-1})$ denote the support set of $p(\cdot|X_{-\infty}^{-1})$. Then we have

$$E\left\{\frac{p(X_0^{n-1})}{p(X_0^{n-1}|X_{-\infty}^{-1})}\right\} = E\left[E\left\{\frac{p(X_0^{n-1})}{p(X_0^{n-1}|X_{-\infty}^{-1})}\middle| X_{-\infty}^{-1}\right\}\right] \tag{16.203}$$

$$= E\left[\sum_{x^n\in B(X_{-\infty}^{-1})} \frac{p(x^n)}{p(x^n|X_{-\infty}^{-1})} p(x^n|X_{-\infty}^{-1})\right] \tag{16.204}$$

$$= E\left[\sum_{x^n\in B(X_{-\infty}^{-1})} p(x^n)\right] \tag{16.205}$$

$$\le 1. \tag{16.206}$$

By Markov's inequality and (16.202), we have

$$\Pr\left\{\frac{p^k(X_0^{n-1})}{p(X_0^{n-1})} \ge t_n\right\} \le \frac{1}{t_n} \tag{16.207}$$

or

$$\Pr\left\{\frac{1}{n}\log\frac{p^k(X_0^{n-1})}{p(X_0^{n-1})} \geq \frac{1}{n}\log t_n\right\} \leq \frac{1}{t_n}. \tag{16.208}$$

Letting $t_n = n^2$ and noting that $\sum_{n=1}^{\infty}\frac{1}{n^2} < \infty$, we see by the Borel–Cantelli lemma that the event

$$\left\{\frac{1}{n}\log\frac{p^k(X_0^{n-1})}{p(X_0^{n-1})} \geq \frac{1}{n}\log t_n\right\} \tag{16.209}$$

occurs only finitely often with probability 1. Thus,

$$\limsup\frac{1}{n}\log\frac{p^k(X_0^{n-1})}{p(X_0^{n-1})} \leq 0 \qquad \text{with probability 1.} \tag{16.210}$$

Applying the same arguments using Markov's inequality to (16.206), we obtain

$$\limsup\frac{1}{n}\log\frac{p(X_0^{n-1})}{p(X_0^{n-1}|X_{-\infty}^{-1})} \leq 0 \qquad \text{with probability 1,} \tag{16.211}$$

proving the lemma. □

The arguments used in the proof can be extended to prove the AEP for the stock market (Theorem 16.5.3).

SUMMARY

Growth rate. The *growth rate* of a stock market portfolio $\mathbf{b}$ with respect to a distribution $F(\mathbf{x})$ is defined as

$$W(\mathbf{b}, F) = \int \log \mathbf{b}^t\mathbf{x}\, dF(\mathbf{x}) = E\left(\log \mathbf{b}^t\mathbf{x}\right). \tag{16.212}$$

Log-optimal portfolio. The *optimal growth rate* with respect to a distribution $F(x)$ is

$$W^*(F) = \max_{\mathbf{b}} W(\mathbf{b}, F). \tag{16.213}$$

The portfolio $\mathbf{b}^*$ that achieves the maximum of $W(\mathbf{b}, F)$ is called the *log-optimal portfolio*.

Concavity. $W(\mathbf{b}, F)$ is concave in $\mathbf{b}$ and linear in F. $W^*(F)$ is convex in F.

Optimality conditions. The portfolio $\mathbf{b}^*$ is log-optimal if and only if

$$E\left(\frac{X_i}{\mathbf{b}^{*t}\mathbf{X}}\right) = 1 \quad \text{if } b_i^* > 0,$$
$$\leq 1 \quad \text{if } b_i^* = 0. \tag{16.214}$$

Expected ratio optimality. If $S_n^* = \prod_{i=1}^n \mathbf{b}^{*t}\mathbf{X}_i$, $S_n = \prod_{i=1}^n \mathbf{b}_i^t\mathbf{X}_i$, then

$$E\frac{S_n}{S_n^*} \leq 1 \quad \text{if and only if} \quad E\ln\frac{S_n}{S_n^*} \leq 0. \tag{16.215}$$

Growth rate (AEP)

$$\frac{1}{n}\log S_n^* \to W^*(F) \qquad \text{with probability 1.} \tag{16.216}$$

Asymptotic optimality

$$\limsup_{n\to\infty} \frac{1}{n}\log\frac{S_n}{S_n^*} \leq 0 \qquad \text{with probability 1.} \tag{16.217}$$

Wrong information. Believing g when f is true loses

$$\Delta W = W(\mathbf{b}_f^*, F) - W(\mathbf{b}_g^*, F) \leq D(f||g). \tag{16.218}$$

Side information Y

$$\Delta W \leq I(\mathbf{X}; Y). \tag{16.219}$$

Chain rule

$$W^*(\mathbf{X}_i|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1}) = \max_{\mathbf{b}_i(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{i-1})} E\log\mathbf{b}_i^t\mathbf{X}_i \tag{16.220}$$

$$W^*(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) = \sum_{i=1}^n W^*(\mathbf{X}_i|\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{i-1}). \tag{16.221}$$

Growth rate for a stationary market.

$$W_\infty^* = \lim \frac{W^*(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)}{n} \tag{16.222}$$

$$\frac{1}{n} \log S_n^* \to W_\infty^*. \tag{16.223}$$

Competitive optimality of log-optimal portfolios.

$$\Pr(VS \geq U^* S^*) \leq \frac{1}{2}. \tag{16.224}$$

Universal portfolio.

$$\max_{\hat{\mathbf{b}}_i(\cdot)} \min_{\mathbf{x}^n, \mathbf{b}} \frac{\prod_{i=1}^n \hat{\mathbf{b}}_i^t(\mathbf{x}^{i-1})\mathbf{x}_i}{\prod_{i=1}^n \mathbf{b}^t \mathbf{x}_i} = V_n, \tag{16.225}$$

where

$$V_n = \left[\sum_{n_1+\cdots+n_m=n} \binom{n}{n_1, n_2, \ldots, n_m} 2^{-nH(n_1/n,\ldots,n_m/n)} \right]^{-1}. \tag{16.226}$$

For $m = 2$,

$$V_n \sim \sqrt{2/\pi n} \tag{16.227}$$

The causal universal portfolio

$$\hat{\mathbf{b}}_{i+1}(\mathbf{x}^i) = \frac{\int \mathbf{b} S_i(\mathbf{b}, \mathbf{x}^i)\, d\mu(\mathbf{b})}{\int S_i(\mathbf{b}, \mathbf{x}^i)\, d\mu(\mathbf{b})} \tag{16.228}$$

achieves

$$\frac{\hat{S}_n(\mathbf{x}^n)}{S_n^*(\mathbf{x}^n)} \geq \frac{1}{2\sqrt{n+1}} \tag{16.229}$$

for all n and all $\mathbf{x}^n$.

AEP. If $\{X_i\}$ is stationary ergodic, then

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(\mathcal{X}) \qquad \text{with probability 1.} \tag{16.230}$$

PROBLEMS

16.1 *Growth rate*. Let

$$\mathbf{X} = \begin{cases} (1, a) & \text{with probability } \frac{1}{2} \\ (1, 1/a) & \text{with probability } \frac{1}{2} \end{cases}$$

where $a > 1$. This vector $\mathbf{X}$ represents a stock market vector of cash vs. a hot stock. Let

$$W(\mathbf{b}, F) = E \log \mathbf{b}^t \mathbf{X}$$

and

$$W^* = \max_{\mathbf{b}} W(\mathbf{b}, F)$$

be the growth rate.

(a) Find the log optimal portfolio $\mathbf{b}^*$.

(b) Find the growth rate W^*.

(c) Find the asymptotic behavior of

$$S_n = \prod_{i=1}^{n} \mathbf{b}^t \mathbf{X}_i$$

for all $\mathbf{b}$.

16.2 *Side information*. Suppose, in Problem 16.1, that

$$\mathbf{Y} = \begin{cases} 1 & \text{if } (X_1, X_2) \geq (1, 1), \\ 0 & \text{if } (X_1, X_2) \leq (1, 1). \end{cases}$$

Let the portfolio $\mathbf{b}$ depend on $\mathbf{Y}$. Find the new growth rate W^{**} and verify that $\Delta W = W^{**} - W^*$ satisfies

$$\Delta W \leq I(X; Y).$$

16.3 *Stock dominance*. Consider a stock market vector

$$\mathbf{X} = (X_1, X_2).$$

Suppose that $X_1 = 2$ with probability 1. Thus an investment in the first stock is doubled at the end of the day.

(a) Find necessary and sufficient conditions on the distribution of stock X_2 such that the log-optimal portfolio $\mathbf{b}^*$ invests all the wealth in stock X_2 [i.e., $\mathbf{b}^* = (0, 1)$].

(b) Argue for any distribution on X_2 that the growth rate satisfies $W^* \geq 1$.

16.4 *Including experts and mutual funds*. Let $\mathbf{X} \sim F(\mathbf{x}), \mathbf{x} \in \mathcal{R}_+^m$, be the vector of price relatives for a stock market. Suppose that an "expert" suggests a portfolio $\mathbf{b}$. This would result in a wealth factor $\mathbf{b}^t\mathbf{X}$. We add this to the stock alternatives to form $\tilde{\mathbf{X}} = (X_1, X_2, \ldots, X_m, \mathbf{b}^t\mathbf{X})$. Show that the new growth rate,

$$\tilde{W}^* = \max_{b_1,\ldots,b_m,b_{m+1}} \int \ln(\mathbf{b}^t\tilde{\mathbf{x}})\, dF(\tilde{\mathbf{x}}), \tag{16.231}$$

is equal to the old growth rate,

$$W^* = \max_{b_1,\ldots,b_m} \int \ln(\mathbf{b}^t\mathbf{x})\, dF(\mathbf{x}). \tag{16.232}$$

16.5 *Growth rate for symmetric distribution*. Consider a stock vector $\mathbf{X} \sim F(\mathbf{x}), \quad \mathbf{X} \in \mathcal{R}^m, \quad \mathbf{X} \geq 0$, where the component stocks are exchangeable. Thus, $F(x_1, x_2, \ldots, x_m) = F(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(m)})$ for all permutations σ.

(a) Find the portfolio $\mathbf{b}^*$ optimizing the growth rate and establish its optimality. Now assume that $\mathbf{X}$ has been normalized so that $\frac{1}{m}\sum_{i=1}^m X_i = 1$, and F is symmetric as before.

(b) Again assuming $\mathbf{X}$ to be normalized, show that all symmetric distributions F have the same growth rate against $\mathbf{b}^*$.

(c) Find this growth rate.

16.6 *Convexity*. We are interested in the set of stock market densities that yield the same optimal porfolio. Let $P_{\mathbf{b}_0}$ be the set of all probability densities on $\mathcal{R}_+^m$ for which $\mathbf{b}_0$ is optimal. Thus, $P_{\mathbf{b}_0} = \{p(x) : \int \ln(\mathbf{b}^t x)p(x)\, dx$ is maximized by $\mathbf{b} = \mathbf{b}_0\}$. Show that $P_{\mathbf{b}_0}$ is a convex set. It may be helpful to use Theorem 16.2.2.

16.7 *Short selling*. Let

$$X = \begin{cases} (1, 2), & p, \\ (1, \frac{1}{2}), & 1 - p. \end{cases}$$

Let $B = \{(b_1, b_2) : b_1 + b_2 = 1\}$. Thus, this set of portfolios B does not include the constraint $b_i \geq 0$. (This allows short selling.)

(a) Find the log optimal portfolio $\mathbf{b}^*(p)$.

(b) Relate the growth rate $W^*(p)$ to the entropy rate $H(p)$.

16.8 *Normalizing* $\mathbf{x}$. Suppose that we define the log-optimal portfolio $\mathbf{b}^*$ to be the portfolio maximizing the relative growth rate

$$\int \ln \frac{\mathbf{b}^t \mathbf{x}}{\frac{1}{m}\sum_{i=1}^{m} x_i} dF(x_1, \ldots, x_m).$$

The virtue of the normalization $\frac{1}{m}\sum X_i$, which can be viewed as the wealth associated with a uniform portfolio, is that the relative growth rate is finite even when the growth rate $\int \ln b^t x dF(x)$ is not. This matters, for example, if X has a St. Petersburg-like distribution. Thus, the log-optimal portfolio $\mathbf{b}^*$ is defined for all distributions F, even those with infinite growth rates $W^*(F)$.

(a) Show that if $\mathbf{b}$ maximizes $\int \ln(\mathbf{b}^t \mathbf{x})\, dF(x)$, it also maximizes $\int \ln \frac{\mathbf{b}^t \mathbf{x}}{u^t x}\, dF(x)$, where $u = (\frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m})$.

(b) Find the log optimal portfolio $\mathbf{b}^*$ for

$$\mathbf{X} = \begin{cases} (2^{2^k+1}, 2^{2^k}), & 2^{-(k+1)}, \\ (2^{2^k}, 2^{2^k+1}), & 2^{-(k+1)}, \end{cases}$$

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

(c) Find EX and W^*.

(d) Argue that $\mathbf{b}^*$ is competitively better than any portfolio $\mathbf{b}$ in the sense that $\Pr\{\mathbf{b}^t \mathbf{X} > c\mathbf{b}^{*t} \mathbf{X}\} \leq \frac{1}{c}$.

16.9 *Universal portfolio*. We examine the first $n = 2$ steps of the implementation of the universal portfolio in (16.7.2) for $\mu(b)$ uniform for $m = 2$ stocks. Let the stock vectors for days 1 and 2 be $\mathbf{x}_1 = (1, \frac{1}{2})$, and $\mathbf{x}_2 = (1, 2)$. Let $\mathbf{b} = (b, 1-b)$ denote a portfolio.

(a) Graph $S_2(\mathbf{b}) = \prod_{i=1}^{2} \mathbf{b}^t \mathbf{x}_i, \quad 0 \leq b \leq 1$.

(b) Calculate $S_2^* = \max_{\mathbf{b}} S_2(\mathbf{b})$.

(c) Argue that $\log S_2(\mathbf{b})$ is concave in $\mathbf{b}$.

(d) Calculate the (universal) wealth $\hat{S}_2 = \int_0^1 S_2(\mathbf{b}) d\mathbf{b}$.

(e) Calculate the universal portfolio at times $n = 1$ and $n = 2$:

$$\hat{\mathbf{b}}_1 = \int_0^1 \mathbf{b}\, d\mathbf{b}$$

$$\hat{\mathbf{b}}_2(\mathbf{x}_1) = \frac{\int_0^1 \mathbf{b} S_1(\mathbf{b})\, d\mathbf{b}}{\int_0^1 S_1(\mathbf{b})\, d\mathbf{b}}.$$

(f) Which of $S_2(\mathbf{b})$, S_2^*, $\hat{S}_2$, $\hat{\mathbf{b}}_2$ are unchanged if we permute the order of appearance of the stock vector outcomes [i.e., if the sequence is now $(1, 2), (1, \frac{1}{2})$]?

16.10 *Growth optimal*. Let $X_1, X_2 \geq 0$, be price relatives of two independent stocks. Suppose that $EX_1 > EX_2$. Do you always want some of X_1 in a growth rate optimal portfolio $S(\mathbf{b}) = bX_1 + \overline{b}X_2$? Prove or provide a counterexample.

16.11 *Cost of universality*. In the discussion of finite-horizon universal portfolios, it was shown that the loss factor due to universality is

$$\frac{1}{V_n} = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}. \tag{16.233}$$

Evaluate V_n for $n = 1, 2, 3$.

16.12 *Convex families*. This problem generalizes Theorem 16.2.2. We say that $\mathcal{S}$ is a convex family of random variables if $S_1, S_2 \in \mathcal{S}$ implies that $\lambda S_1 + (1 - \lambda) S_2 \in \mathcal{S}$. Let $\mathcal{S}$ be a closed convex family of random variables. Show that there is a random variable $S^* \in \mathcal{S}$ such that

$$E \ln \left(\frac{S}{S^*}\right) \leq 0 \tag{16.234}$$

for all $S \in \mathcal{S}$ if and only if

$$E \left(\frac{S}{S^*}\right) \leq 1 \tag{16.235}$$

for all $S \in \mathcal{S}$.

HISTORICAL NOTES

There is an extensive literature on the mean–variance approach to investment in the stock market. A good introduction is the book by Sharpe [491]. Log-optimal portfolios were introduced by Kelly [308] and Latané [346], and generalized by Breiman [75]. The bound on the increase in the

growth rate in terms of the mutual information is due to Barron and Cover [31]. See Samuelson [453, 454] for a criticism of log-optimal investment.

The proof of the competitive optimality of the log-optimal portfolio is due to Bell and Cover [39, 40]. Breiman [75] investigated asymptotic optimality for random market processes.

The AEP was introduced by Shannon. The AEP for the stock market and the asymptotic optimality of log-optimal investment are given in Algoet and Cover [21]. The relatively simple sandwich proof for the AEP is due to Algoet and Cover [20]. The AEP for real-valued ergodic processes was proved in full generality by Barron [34] and Orey [402].

The universal portfolio was defined in Cover [110] and the proof of universality was given in Cover [110] and more exactly in Cover and Ordentlich [135]. The fixed-horizon exact calculation of the cost of universality V_n is given in Ordentlich and Cover [401]. The quantity V_n also appears in data compression in the work of Shtarkov [496].

CHAPTER 17

INEQUALITIES IN INFORMATION THEORY

This chapter summarizes and reorganizes the inequalities found throughout this book. A number of new inequalities on the entropy rates of subsets and the relationship of entropy and $\mathbb{L}_p$ norms are also developed. The intimate relationship between Fisher information and entropy is explored, culminating in a common proof of the entropy power inequality and the Brunn–Minkowski inequality. We also explore the parallels between the inequalities in information theory and inequalities in other branches of mathematics, such as matrix theory and probability theory.

17.1 BASIC INEQUALITIES OF INFORMATION THEORY

Many of the basic inequalities of information theory follow directly from convexity.

Definition A function f is said to be *convex* if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \tag{17.1}$$

for all $0 \leq \lambda \leq 1$ and all x_1 and x_2.

Theorem 17.1.1 (*Theorem 2.6.2: Jensen's inequality*) *If f is convex, then*

$$f(EX) \leq Ef(X). \tag{17.2}$$

Lemma 17.1.1 *The function* $\log x$ *is concave and* $x \log x$ *is convex, for* $0 < x < \infty$.

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

Theorem 17.1.2 (*Theorem 2.7.1: Log sum inequality*) *For positive numbers* $a_1, a_2, \ldots, a_n$ *and* $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{17.3}$$

with equality iff $\frac{a_i}{b_i} = \text{constant}$.

We recall the following properties of entropy from Section 2.1.

Definition The *entropy* $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x). \tag{17.4}$$

Theorem 17.1.3 (*Lemma 2.1.1, Theorem 2.6.4: Entropy bound*)

$$0 \leq H(X) \leq \log |\mathcal{X}|. \tag{17.5}$$

Theorem 17.1.4 (*Theorem 2.6.5: Conditioning reduces entropy*) *For any two random variables* X *and* Y,

$$H(X|Y) \leq H(X), \tag{17.6}$$

with equality iff X *and* Y *are independent.*

Theorem 17.1.5 (*Theorem 2.5.1 with Theorem 2.6.6: Chain rule*)

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \leq \sum_{i=1}^{n} H(X_i), \tag{17.7}$$

with equality iff $X_1, X_2, \ldots, X_n$ *are independent.*

Theorem 17.1.6 (*Theorem 2.7.3*) $H(p)$ *is a concave function of* p.

We now state some properties of relative entropy and mutual information (Section 2.3).

Definition The *relative entropy* or *Kullback–Leibler distance* between two probability mass functions $p(x)$ and $q(x)$ is defined by

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}. \tag{17.8}$$

Definition The mutual information between two random variables X and Y is defined by

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y)). \tag{17.9}$$

The following basic information inequality can be used to prove many of the other inequalities in this chapter.

Theorem 17.1.7 (*Theorem 2.6.3: Information inequality*) *For any two probability mass functions p and q,*

$$D(p||q) \geq 0 \tag{17.10}$$

with equality iff $p(x) = q(x)$ for all $x \in \mathcal{X}$.

Corollary *For any two random variables X and Y,*

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0 \tag{17.11}$$

with equality iff $p(x,y) = p(x)p(y)$ (i.e., X and Y are independent).

Theorem 17.1.8 (*Theorem 2.7.2: Convexity of relative entropy*) *$D(p||q)$ is convex in the pair (p,q).*

Theorem 17.1.9 (*Theorem 2.4.1*)

$$I(X;Y) = H(X) - H(X|Y). \tag{17.12}$$

$$I(X;Y) = H(Y) - H(Y|X). \tag{17.13}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{17.14}$$

$$I(X;X) = H(X). \tag{17.15}$$

Theorem 17.1.10 (*Section 4.4*) *For a Markov chain:*

1. *Relative entropy $D(\mu_n||\mu'_n)$ decreases with time.*
2. *Relative entropy $D(\mu_n||\mu)$ between a distribution and the stationary distribution decreases with time.*
3. *Entropy $H(X_n)$ increases if the stationary distribution is uniform.*
4. *The conditional entropy $H(X_n|X_1)$ increases with time for a stationary Markov chain.*

Theorem 17.1.11 *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. Let $\hat{p}_n$ be the empirical probability mass function of $X_1, X_2, \ldots, X_n$. Then*

$$ED(\hat{p}_n||p) \leq ED(\hat{p}_{n-1}||p). \tag{17.16}$$

17.2 DIFFERENTIAL ENTROPY

We now review some of the basic properties of differential entropy (Section 8.1).

Definition The *differential entropy* $h(X_1, X_2, \ldots, X_n)$, sometimes written $h(f)$, is defined by

$$h(X_1, X_2, \ldots, X_n) = -\int f(\mathbf{x}) \log f(\mathbf{x})\, d\mathbf{x}. \tag{17.17}$$

The differential entropy for many common densities is given in Table 17.1.

Definition The *relative entropy* between probability densities f and g is

$$D(f||g) = \int f(\mathbf{x}) \log \left(f(\mathbf{x})/g(\mathbf{x})\right)\, d\mathbf{x}. \tag{17.18}$$

The properties of the continuous version of relative entropy are identical to the discrete version. Differential entropy, on the other hand, has some properties that differ from those of discrete entropy. For example, differential entropy may be negative.

We now restate some of the theorems that continue to hold for differential entropy.

Theorem 17.2.1 (*Theorem 8.6.1: Conditioning reduces entropy*) *$h(X|Y) \leq h(X)$, with equality iff X and Y are independent.*

Theorem 17.2.2 (*Theorem 8.6.2: Chain rule*)

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_{i-1}, X_{i-2}, \ldots, X_1) \leq \sum_{i=1}^{n} h(X_i) \tag{17.19}$$

with equality iff $X_1, X_2, \ldots, X_n$ are independent.

Lemma 17.2.1 *If X and Y are independent, then $h(X + Y) \geq h(X)$.*

Proof: $h(X + Y) \geq h(X + Y|Y) = h(X|Y) = h(X)$. □

TABLE 17.1 Differential Entropies[a]

Distribution		
Name	Density	Entropy (nats)
Beta	$f(x) = \dfrac{x^{p-1}(1-x)^{q-1}}{B(p,q)}$, $0 \le x \le 1,\ p,q > 0$	$\ln B(p,q) - (p-1)$ $\times[\psi(p) - \psi(p+q)]$ $-(q-1)[\psi(q) - \psi(p+q)]$
Cauchy	$f(x) = \dfrac{\lambda}{\pi}\dfrac{1}{\lambda^2 + x^2}$, $-\infty < x < \infty,\ \lambda > 0$	$\ln(4\pi\lambda)$
Chi	$f(x) = \dfrac{2}{2^{n/2}\sigma^n\Gamma(n/2)} x^{n-1} e^{-\frac{x^2}{2\sigma^2}}$, $x > 0, n > 0$	$\ln\dfrac{\sigma\Gamma(n/2)}{\sqrt{2}} - \dfrac{n-1}{2}\psi\left(\dfrac{n}{2}\right) + \dfrac{n}{2}$
Chi-squared	$f(x) = \dfrac{1}{2^{n/2}\sigma^n\Gamma(n/2)} x^{\frac{n}{2}-1} e^{-\frac{x}{2\sigma^2}}$, $x > 0, n > 0$	$\ln 2\sigma^2\Gamma\left(\dfrac{n}{2}\right)$ $-\left(1-\dfrac{n}{2}\right)\psi\left(\dfrac{n}{2}\right) + \dfrac{n}{2}$
Erlang	$f(x) = \dfrac{\beta^n}{(n-1)!} x^{n-1} e^{-\beta x}$, $x, \beta > 0, n > 0$	$(1-n)\psi(n) + \ln\dfrac{\Gamma(n)}{\beta} + n$
Exponential	$f(x) = \dfrac{1}{\lambda} e^{-\frac{x}{\lambda}},\ x, \lambda > 0$	$1 + \ln\lambda$
F	$f(x) = \dfrac{n_1^{\frac{n_1}{2}} n_2^{\frac{n_2}{2}}}{B(\frac{n_1}{2}, \frac{n_2}{2})} \times \dfrac{x^{(\frac{n_1}{2})-1}}{(n_2 + n_1 x)^{\frac{n_1+n_2}{2}}}$, $x > 0, n_1, n_2 > 0$	$\ln\dfrac{n_1}{n_2} B\left(\dfrac{n_1}{2}, \dfrac{n_2}{2}\right)$ $+\left(1-\frac{n_1}{2}\right)\psi\left(\dfrac{n_1}{2}\right)$ $-\left(1-\dfrac{n_2}{2}\right)\psi\left(\dfrac{n_2}{2}\right)$ $+\dfrac{n_1+n_2}{2}\psi\left(\dfrac{n_1+n_2}{2}\right)$
Gamma	$f(x) = \dfrac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^\alpha\Gamma(\alpha)}, \quad x, \alpha, \beta > 0$	$\ln(\beta\Gamma(\alpha)) + (1-\alpha)\psi(\alpha) + \alpha$
Laplace	$f(x) = \dfrac{1}{2\lambda} e^{-\frac{\lvert x-\theta\rvert}{\lambda}}$, $-\infty < x, \theta < \infty, \lambda > 0$	$1 + \ln 2\lambda$
Logistic	$f(x) = \dfrac{e^{-x}}{(1+e^{-x})^2}$, $-\infty < x < \infty$	2

TABLE 17.1 (*continued*)

Distribution		
Name	Density	Entropy (nats)
Lognormal	$f(x) = \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{\ln(x-m)^2}{2\sigma^2}}$, $x > 0, -\infty < m < \infty, \sigma > 0$	$m + \frac{1}{2}\ln(2\pi e\sigma^2)$
Maxwell–Boltzmann	$f(x) = 4\pi^{-\frac{1}{2}}\beta^{\frac{3}{2}}x^2 e^{-\beta x^2}$, $x, \beta > 0$	$\frac{1}{2}\ln\frac{\pi}{\beta} + \gamma - \frac{1}{2}$
Normal	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $-\infty < x, \mu < \infty, \sigma > 0$	$\frac{1}{2}\ln(2\pi e\sigma^2)$
Generalized normal	$f(x) = \frac{2\beta^{\frac{\alpha}{2}}}{\Gamma(\frac{\alpha}{2})} x^{\alpha-1} e^{-\beta x^2}$, $x, \alpha, \beta > 0$	$\ln\frac{\Gamma(\frac{\alpha}{2})}{2\beta^{\frac{1}{2}}} - \frac{\alpha-1}{2}\psi\left(\frac{\alpha}{2}\right) + \frac{\alpha}{2}$
Pareto	$f(x) = \frac{ak^a}{x^{a+1}}$, $x \geq k > 0, a > 0$	$\ln\frac{k}{a} + 1 + \frac{1}{a}$
Rayleigh	$f(x) = \frac{x}{b^2} e^{-\frac{x^2}{2b^2}}$, $x, b > 0$	$1 + \ln\frac{\beta}{\sqrt{2}} + \frac{\gamma}{2}$
Student's t	$f(x) = \frac{(1 + x^2/n)^{-(n+1)/2}}{\sqrt{n}B(\frac{1}{2}, \frac{n}{2})}$, $-\infty < x < \infty, n > 0$	$\frac{n+1}{2}\psi\left(\frac{n+1}{2}\right) - \psi\left(\frac{n}{2}\right) + \ln\sqrt{n}B\left(\frac{1}{2}, \frac{n}{2}\right)$
Triangular	$f(x) = \begin{cases} \frac{2x}{a}, & 0 \leq x \leq a \\ \frac{2(1-x)}{1-a}, & a \leq x \leq 1 \end{cases}$	$\frac{1}{2} - \ln 2$
Uniform	$f(x) = \frac{1}{\beta - \alpha}$, $\alpha \leq x \leq \beta$	$\ln(\beta - \alpha)$
Weibull	$f(x) = \frac{c}{\alpha} x^{c-1} e^{-\frac{x^c}{\alpha}}$, $x, c, \alpha > 0$	$\frac{(c-1)\gamma}{c} + \ln\frac{\alpha^{\frac{1}{c}}}{c} + 1$

[a] All entropies are in nats; $\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\,dt$; $\psi(z) = \frac{d}{dz}\ln\Gamma(z)$; γ = Euler's constant = 0.57721566....

Source: Lazo and Rathie [543].

Theorem 17.2.3 (*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*

$$h(\mathbf{X}) \le \frac{1}{2}\log(2\pi e)^n|K| \tag{17.20}$$

with equality iff $\mathbf{X} \sim \mathcal{N}(0, K)$.

17.3 BOUNDS ON ENTROPY AND RELATIVE ENTROPY

In this section we revisit some of the bounds on the entropy function. The most useful is Fano's inequality, which is used to bound away from zero the probability of error of the best decoder for a communication channel at rates above capacity.

Theorem 17.3.1 (*Theorem 2.10.1: Fano's inequality*) *Given two random variables* X *and* Y*, let* $\hat{X} = g(Y)$ *be any estimator of* X *given* Y *and let* $P_e = \Pr(X \ne \hat{X})$ *be the probability of error. Then*

$$H(P_e) + P_e \log|\mathcal{X}| \ge H(X|\hat{X}) \ge H(X|Y). \tag{17.21}$$

Consequently, if $H(X|Y) > 0$*, then* $P_e > 0$.

A similar result is given in the following lemma.

Lemma 17.3.1 (*Lemma 2.10.1*) *If* X *and* X' *are i.i.d. with entropy* $H(X)$

$$\Pr(X = X') \ge 2^{-H(X)} \tag{17.22}$$

with equality if and only if X *has a uniform distribution.*

The continuous analog of Fano's inequality bounds the mean-squared error of an estimator.

Theorem 17.3.2 (*Theorem 8.6.6*) *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. Then*

$$E(X - \hat{X})^2 \ge \frac{1}{2\pi e}e^{2h(X)}. \tag{17.23}$$

Given side information Y *and estimator* $\hat{X}(Y)$*,*

$$E(X - \hat{X}(Y))^2 \ge \frac{1}{2\pi e}e^{2h(X|Y)}. \tag{17.24}$$

Theorem 17.3.3 *($\mathcal{L}_1$ bound on entropy) Let p and q be two probability mass functions on $\mathcal{X}$ such that*

$$||p - q||_1 = \sum_{x \in \mathcal{X}} |p(x) - q(x)| \leq \frac{1}{2}. \tag{17.25}$$

Then

$$|H(p) - H(q)| \leq -||p - q||_1 \log \frac{||p - q||_1}{|\mathcal{X}|}. \tag{17.26}$$

Proof: Consider the function $f(t) = -t \log t$ shown in Figure 17.1. It can be verified by differentiation that the function $f(\cdot)$ is concave. Also, $f(0) = f(1) = 0$. Hence the function is positive between 0 and 1. Consider the chord of the function from t to $t + \nu$ (where $\nu \leq \frac{1}{2}$). The maximum absolute slope of the chord is at either end (when $t = 0$ or $1 - \nu$). Hence for $0 \leq t \leq 1 - \nu$, we have

$$|f(t) - f(t + \nu)| \leq \max\{f(\nu), f(1 - \nu)\} = -\nu \log \nu. \tag{17.27}$$

Let $r(x) = |p(x) - q(x)|$. Then

$$|H(p) - H(q)| = \left| \sum_{x \in \mathcal{X}} (-p(x) \log p(x) + q(x) \log q(x)) \right| \tag{17.28}$$

$$\leq \sum_{x \in \mathcal{X}} |(-p(x) \log p(x) + q(x) \log q(x))| \tag{17.29}$$

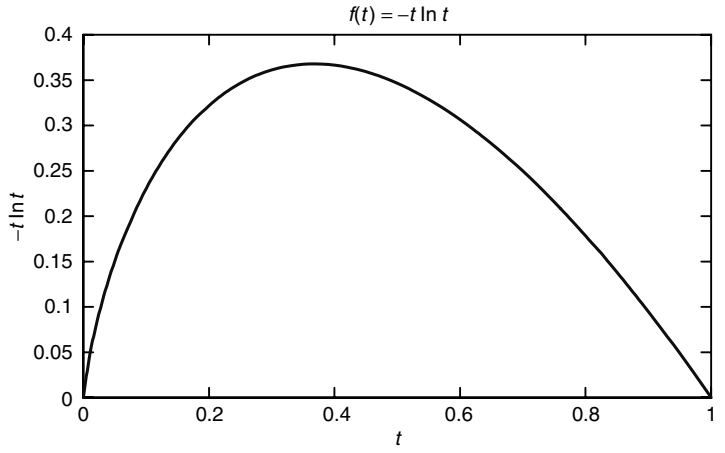


FIGURE 17.1. Function $f(t) = -t \ln t$.

$$\leq \sum_{x \in \mathcal{X}} -r(x) \log r(x) \tag{17.30}$$

$$= ||p - q||_1 \sum_{x \in \mathcal{X}} -\frac{r(x)}{||p - q||_1} \log \frac{r(x)}{||p - q||_1} ||p - q||_1 \tag{17.31}$$

$$= -||p - q||_1 \log ||p - q||_1 + ||p - q||_1 H\left(\frac{r(x)}{||p - q||_1}\right) \tag{17.32}$$

$$\leq -||p - q||_1 \log ||p - q||_1 + ||p - q||_1 \log |\mathcal{X}|, \tag{17.33}$$

where (17.30) follows from (17.27). □

Finally, relative entropy is stronger than the $\mathcal{L}_1$ norm in the following sense:

Lemma 17.3.2 (*Lemma 11.6.1*)

$$D(p_1||p_2) \geq \frac{1}{2 \ln 2} ||p_1 - p_2||_1^2. \tag{17.34}$$

The relative entropy between two probability mass functions $P(x)$ and $Q(x)$ is zero when $P = Q$. Around this point, the relative entropy has a quadratic behavior, and the first term in the Taylor series expansion of the relative entropy $D(P||Q)$ around the point $P = Q$ is the chi-squared distance between the distributions P and Q. Let

$$\chi^2(P, Q) = \sum_x \frac{(P(x) - Q(x))^2}{Q(x)}. \tag{17.35}$$

Lemma 17.3.3 *For P near Q,*

$$D(P \parallel Q) = \frac{1}{2} \chi^2 + \cdots. \tag{17.36}$$

Proof: See Problem 11.2. □

17.4 INEQUALITIES FOR TYPES

The method of types is a powerful tool for proving results in large deviation theory and error exponents. We repeat the basic theorems.

Theorem 17.4.1 (*Theorem 11.1.1*) *The number of types with denominator n is bounded by*

$$|\mathcal{P}_n| \leq (n + 1)^{|\mathcal{X}|}. \tag{17.37}$$

Theorem 17.4.2 (*Theorem 11.1.2*) *If $X_1, X_2, \ldots, X_n$ are drawn i.i.d. according to $Q(x)$, the probability of x^n depends only on its type and is given by*

$$Q^n(x^n) = 2^{-n(H(P_{x^n})+D(P_{x^n}||Q))}. \tag{17.38}$$

Theorem 17.4.3 (*Theorem 11.1.3: Size of a type class $T(P)$*) *For any type $P \in \mathcal{P}_n$,*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}. \tag{17.39}$$

Theorem 17.4.4 (*Theorem 11.1.4*) *For any $P \in \mathcal{P}_n$ and any distribution Q, the probability of the type class $T(P)$ under Q^n is $2^{-nD(P||Q)}$ to first order in the exponent. More precisely,*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}. \tag{17.40}$$

17.5 COMBINATORIAL BOUNDS ON ENTROPY

We give tight bounds on the size of $\binom{n}{k}$ when k is not 0 or n using the result of Wozencraft and Reiffen [568]:

Lemma 17.5.1 *For $0 < p < 1$, $q = 1 - p$, such that np is an integer,*

$$\frac{1}{\sqrt{8npq}} \leq \binom{n}{np} 2^{-nH(p)} \leq \frac{1}{\sqrt{\pi npq}}. \tag{17.41}$$

Proof: We begin with a strong form of Stirling's approximation [208], which states that

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}. \tag{17.42}$$

Applying this to find an upper bound, we obtain

$$\binom{n}{np} \leq \frac{\sqrt{2\pi n}(\frac{n}{e})^n e^{\frac{1}{12n}}}{\sqrt{2\pi np}(\frac{np}{e})^{np}\sqrt{2\pi nq}(\frac{nq}{e})^{nq}} \tag{17.43}$$

$$= \frac{1}{\sqrt{2\pi npq}} \frac{1}{p^{np}q^{nq}} e^{\frac{1}{12n}} \tag{17.44}$$

$$< \frac{1}{\sqrt{\pi npq}} 2^{nH(p)}, \tag{17.45}$$

since $e^{\frac{1}{12n}} < e^{\frac{1}{12}} = 1.087 < \sqrt{2}$, hence proving the upper bound.

The lower bound is obtained similarly. Using Stirling's formula, we obtain

$$\binom{n}{np} \geq \frac{\sqrt{2\pi n}(\frac{n}{e})^n e^{-\left(\frac{1}{12np} + \frac{1}{12nq}\right)}}{\sqrt{2\pi np}(\frac{np}{e})^{np}\sqrt{2\pi nq}(\frac{nq}{e})^{nq}} \tag{17.46}$$

$$= \frac{1}{\sqrt{2\pi npq}} \frac{1}{p^{np}q^{nq}} e^{-\left(\frac{1}{12np} + \frac{1}{12nq}\right)} \tag{17.47}$$

$$= \frac{1}{\sqrt{2\pi npq}} 2^{nH(p)} e^{-\left(\frac{1}{12np} + \frac{1}{12nq}\right)}. \tag{17.48}$$

If $np \geq 1$, and $nq \geq 3$, then

$$e^{-\left(\frac{1}{12np} + \frac{1}{12nq}\right)} \geq e^{-\frac{1}{9}} = 0.8948 > \frac{\sqrt{\pi}}{2} = 0.8862, \tag{17.49}$$

and the lower bound follows directly from substituting this into the equation. The exceptions to this condition are the cases where $np = 1, nq = 1$ or 2, and $np = 2, nq = 2$ (the case when $np \geq 3$, $nq = 1$ or 2 can be handled by flipping the roles of p and q). In each of these cases

$$np = 1, nq = 1 \rightarrow n = 2, p = \tfrac{1}{2}, \quad \binom{n}{np} = 2, \text{bound} = 2$$

$$np = 1, nq = 2 \rightarrow n = 3, p = \tfrac{1}{3}, \quad \binom{n}{np} = 3, \text{bound} = 2.92$$

$$np = 2, nq = 2 \rightarrow n = 4, p = \tfrac{1}{2}, \quad \binom{n}{np} = 6, \text{bound} = 5.66.$$

Thus, even in these special cases, the bound is valid, and hence the lower bound is valid for all $p \neq 0, 1$. Note that the lower bound blows up when $p = 0$ or $p = 1$, and is therefore not valid. □

17.6 ENTROPY RATES OF SUBSETS

We now generalize the chain rule for differential entropy. The chain rule provides a bound on the entropy rate of a collection of random variables in terms of the entropy of each random variable:

$$h(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} h(X_i). \tag{17.50}$$

We extend this to show that the entropy per element of a subset of a set of random variables decreases as the size of the subset increases. This is not true for each subset but is true on the average over subsets, as expressed in Theorem 17.6.1.

Definition Let $(X_1, X_2, \ldots, X_n)$ have a density, and for every $S \subseteq \{1, 2, \ldots, n\}$, denote by $X(S)$ the subset $\{X_i : i \in S\}$. Let

$$h_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:\,|S|=k} \frac{h(X(S))}{k}. \tag{17.51}$$

Here $h_k^{(n)}$ is the average entropy in bits per symbol of a randomly drawn k-element subset of $\{X_1, X_2, \ldots, X_n\}$.

The following theorem by Han [270] says that the average entropy decreases monotonically in the size of the subset.

Theorem 17.6.1

$$h_1^{(n)} \geq h_2^{(n)} \geq \cdots \geq h_n^{(n)}. \tag{17.52}$$

Proof: We first prove the last inequality, $h_n^{(n)} \leq h_{n-1}^{(n)}$. We write

$$\begin{aligned}
h(X_1, X_2, \ldots, X_n) &= h(X_1, X_2, \ldots, X_{n-1}) + h(X_n | X_1, X_2, \ldots, X_{n-1}),\\
h(X_1, X_2, \ldots, X_n) &= h(X_1, X_2, \ldots, X_{n-2}, X_n)\\
&\quad + h(X_{n-1} | X_1, X_2, \ldots, X_{n-2}, X_n),\\
&\leq h(X_1, X_2, \ldots, X_{n-2}, X_n)\\
&\quad + h(X_{n-1} | X_1, X_2, \ldots, X_{n-2}),\\
&\vdots\\
h(X_1, X_2, \ldots, X_n) &\leq h(X_2, X_3, \ldots, X_n) + h(X_1).
\end{aligned}$$

Adding these n inequalities and using the chain rule, we obtain

$$\begin{aligned}
n\, h(X_1, X_2, \ldots, X_n) \leq &\sum_{i=1}^{n} h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)\\
&+ h(X_1, X_2, \ldots, X_n)
\end{aligned} \tag{17.53}$$

or

$$\frac{1}{n} h(X_1, X_2, \ldots, X_n) \leq \frac{1}{n} \sum_{i=1}^{n} \frac{h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)}{n-1}, \tag{17.54}$$

which is the desired result $h_n^{(n)} \le h_{n-1}^{(n)}$. We now prove that $h_k^{(n)} \le h_{k-1}^{(n)}$ for all $k \le n$ by first conditioning on a k-element subset, and then taking a uniform choice over its $(k-1)$-element subsets. For each k-element subset, $h_k^{(k)} \le h_{k-1}^{(k)}$, and hence the inequality remains true after taking the expectation over all k-element subsets chosen uniformly from the n elements. □

Theorem 17.6.2 *Let $r > 0$, and define*

$$t_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:|S|=k} e^{\frac{r\,h(X(S))}{k}}. \tag{17.55}$$

Then

$$t_1^{(n)} \ge t_2^{(n)} \ge \cdots \ge t_n^{(n)}. \tag{17.56}$$

Proof: Starting from (17.54), we multiply both sides by r, exponentiate, and then apply the arithmetic mean geometric mean inequality, to obtain

$$\begin{aligned} & e^{\frac{1}{n} r h(X_1, X_2, \ldots, X_n)} \\ & \quad \le e^{\frac{1}{n}\sum_{i=1}^n \frac{r h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)}{(n-1)}} && (17.57) \\ & \quad \le \frac{1}{n} \sum_{i=1}^n e^{\frac{r h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)}{(n-1)}} \qquad \text{for all } r \ge 0, && (17.58) \end{aligned}$$

which is equivalent to $t_n^{(n)} \le t_{n-1}^{(n)}$. Now we use the same arguments as in Theorem 17.6.1, taking an average over all subsets to prove the result that for all $k \le n$, $t_k^{(n)} \le t_{k-1}^{(n)}$. □

Definition The average *conditional entropy rate per element* for all subsets of size k is the average of the above quantities for k-element subsets of $\{1, 2, \ldots, n\}$:

$$g_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:|S|=k} \frac{h(X(S)|X(S^c))}{k}. \tag{17.59}$$

Here $g_k(S)$ is the entropy per element of the set S conditional on the elements of the set S^c. When the size of the set S increases, one can expect a greater dependence among the elements of the set S, which explains Theorem 17.6.1.

In the case of the conditional entropy per element, as k increases, the size of the conditioning set S^c decreases and the entropy of the set S increases. The increase in entropy per element due to the decrease in conditioning dominates the decrease due to additional dependence among the elements, as can be seen from the following theorem due to Han [270]. Note that the conditional entropy ordering in the following theorem is the reverse of the unconditional entropy ordering in Theorem 17.6.1.

Theorem 17.6.3

$$g_1^{(n)} \leq g_2^{(n)} \leq \cdots \leq g_n^{(n)}. \tag{17.60}$$

Proof: The proof proceeds on lines very similar to the proof of the theorem for the unconditional entropy per element for a random subset. We first prove that $g_n^{(n)} \geq g_{n-1}^{(n)}$ and then use this to prove the rest of the inequalities. By the chain rule, the entropy of a collection of random variables is less than the sum of the entropies:

$$h(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} h(X_i). \tag{17.61}$$

Subtracting both sides of this inequality from $nh(X_1, X_2, \ldots, X_n)$, we have

$$(n-1)h(X_1, X_2, \ldots, X_n) \geq \sum_{i=1}^{n} (h(X_1, X_2, \ldots, X_n) - h(X_i)) \tag{17.62}$$

$$= \sum_{i=1}^{n} h(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n | X_i). \tag{17.63}$$

Dividing this by $n(n-1)$, we obtain

$$\frac{h(X_1, X_2, \ldots, X_n)}{n} \geq \frac{1}{n} \sum_{i=1}^{n} \frac{h(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n | X_i)}{n-1}, \tag{17.64}$$

which is equivalent to $g_n^{(n)} \geq g_{n-1}^{(n)}$. We now prove that $g_k^{(n)} \geq g_{k-1}^{(n)}$ for all $k \leq n$ by first conditioning on a k-element subset and then taking a uniform choice over its $(k-1)$-element subsets. For each k-element subset, $g_k^{(k)} \geq g_{k-1}^{(k)}$, and hence the inequality remains true after taking the expectation over all k-element subsets chosen uniformly from the n elements. □

Theorem 17.6.4 *Let*

$$f_k^{(n)} = \frac{1}{\binom{n}{k}} \sum_{S:|S|=k} \frac{I(X(S); X(S^c))}{k}. \tag{17.65}$$

Then

$$f_1^{(n)} \geq f_2^{(n)} \geq \cdots \geq f_n^{(n)}. \tag{17.66}$$

Proof: The theorem follows from the identity $I(X(S); X(S^c)) = h(X(S)) - h(X(S)|X(S^c))$ and Theorems 17.6.1 and 17.6.3. □

17.7 ENTROPY AND FISHER INFORMATION

The differential entropy of a random variable is a measure of its descriptive complexity. The Fisher information is a measure of the minimum error in estimating a parameter of a distribution. In this section we derive a relationship between these two fundamental quantities and use this to derive the entropy power inequality.

Let X be any random variable with density $f(x)$. We introduce a location parameter θ and write the density in a parametric form as $f(x - \theta)$. The Fisher information (Section 11.10) with respect to θ is given by

$$J(\theta) = \int_{-\infty}^{\infty} f(x-\theta) \left[\frac{\partial}{\partial \theta} \ln f(x-\theta) \right]^2 dx. \tag{17.67}$$

In this case, differentiation with respect to x is equivalent to differentiation with respect to θ. So we can write the Fisher information as

$$\begin{aligned} J(X) &= \int_{-\infty}^{\infty} f(x-\theta) \left[\frac{\partial}{\partial x} \ln f(x-\theta) \right]^2 dx \\ &= \int_{-\infty}^{\infty} f(x) \left[\frac{\partial}{\partial x} \ln f(x) \right]^2 dx, \end{aligned} \tag{17.68}$$

which we can rewrite as

$$J(X) = \int_{-\infty}^{\infty} f(x) \left[\frac{\frac{\partial}{\partial x} f(x)}{f(x)} \right]^2 dx. \tag{17.69}$$

We will call this the *Fisher information* of the distribution of X. Notice that like entropy, it is a function of the density.

The importance of Fisher information is illustrated in the following theorem.

Theorem 17.7.1 *(Theorem 11.10.1: Cramér–Rao inequality) The mean-squared error of any unbiased estimator $T(X)$ of the parameter θ is lower bounded by the reciprocal of the Fisher information:*

$$var(T) \geq \frac{1}{J(\theta)}. \tag{17.70}$$

We now prove a fundamental relationship between the differential entropy and the Fisher information:

Theorem 17.7.2 (*de Bruijn's identity: entropy and Fisher information*) *Let X be any random variable with a finite variance with a density $f(x)$. Let Z be an independent normally distributed random variable with zero mean and unit variance. Then*

$$\frac{\partial}{\partial t} h_e(X + \sqrt{t}Z) = \frac{1}{2} J(X + \sqrt{t}Z), \tag{17.71}$$

where h_e is the differential entropy to base e. In particular, if the limit exists as $t \to 0$,

$$\frac{\partial}{\partial t} h_e(X + \sqrt{t}Z)\bigg|_{t=0} = \frac{1}{2} J(X). \tag{17.72}$$

Proof: Let $Y_t = X + \sqrt{t}Z$. Then the density of Y_t is

$$g_t(y) = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}}\, dx. \tag{17.73}$$

Then

$$\frac{\partial}{\partial t} g_t(y) = \int_{-\infty}^{\infty} f(x) \frac{\partial}{\partial t} \left[\frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} \right] dx \tag{17.74}$$

$$= \int_{-\infty}^{\infty} f(x) \left[-\frac{1}{2t} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} + \frac{(y-x)^2}{2t^2} \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} \right] dx. \tag{17.75}$$

We also calculate

$$\frac{\partial}{\partial y} g_t(y) = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}} \frac{\partial}{\partial y} \left[e^{-\frac{(y-x)^2}{2t}} \right] dx \tag{17.76}$$

$$= \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}} \left[-\frac{y-x}{t} e^{-\frac{(y-x)^2}{2t}} \right] dx \tag{17.77}$$

and

$$\frac{\partial^2}{\partial y^2} g_t(y) = \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}} \frac{\partial}{\partial y} \left[-\frac{y-x}{t} e^{-\frac{(y-x)^2}{2t}} \right] dx \tag{17.78}$$

$$= \int_{-\infty}^{\infty} f(x) \frac{1}{\sqrt{2\pi t}} \left[-\frac{1}{t} e^{-\frac{(y-x)^2}{2t}} + \frac{(y-x)^2}{t^2} e^{-\frac{(y-x)^2}{2t}} \right] dx. \tag{17.79}$$

Thus,

$$\frac{\partial}{\partial t} g_t(y) = \frac{1}{2} \frac{\partial^2}{\partial y^2} g_t(y). \tag{17.80}$$

We will use this relationship to calculate the derivative of the entropy of Y_t, where the entropy is given by

$$h_e(Y_t) = -\int_{-\infty}^{\infty} g_t(y) \ln g_t(y)\, dy. \tag{17.81}$$

Differentiating, we obtain

$$\frac{\partial}{\partial t} h_e(Y_t) = -\int_{-\infty}^{\infty} \frac{\partial}{\partial t} g_t(y)\, dy - \int_{-\infty}^{\infty} \frac{\partial}{\partial t} g_t(y)\ \ln g_t(y)\, dy \tag{17.82}$$

$$= -\frac{\partial}{\partial t} \int_{-\infty}^{\infty} g_t(y)\, dy - \frac{1}{2} \int_{-\infty}^{\infty} \frac{\partial^2}{\partial y^2} g_t(y)\ \ln g_t(y)\, dy. \tag{17.83}$$

The first term is zero since $\int g_t(y)\, dy = 1$. The second term can be integrated by parts to obtain

$$\frac{\partial}{\partial t} h_e(Y_t) = -\frac{1}{2} \left[\frac{\partial g_t(y)}{\partial y} \ln g_t(y) \right]_{-\infty}^{\infty} + \frac{1}{2} \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial y} g_t(y) \right]^2 \frac{1}{g_t(y)}\ dy. \tag{17.84}$$

The second term in (17.84) is $\frac{1}{2} J(Y_t)$. So the proof will be complete if we show that the first term in (17.84) is zero. We can rewrite the first term as

$$\frac{\partial g_t(y)}{\partial y} \ln g_t(y) = \left[\frac{\frac{\partial g_t(y)}{\partial y}}{\sqrt{g_t(y)}} \right] \left[2\sqrt{g_t(y)} \ln \sqrt{g_t(y)} \right]. \tag{17.85}$$

The square of the first factor integrates to the Fisher information and hence must be bounded as $y \to \pm\infty$. The second factor goes to zero since $x \ln x \to 0$ as $x \to 0$ and $g_t(y) \to 0$ as $y \to \pm\infty$. Hence, the first term in

(17.84) goes to 0 at both limits and the theorem is proved. In the proof, we have exchanged integration and differentiation in (17.74), (17.76), (17.78), and (17.82). Strict justification of these exchanges requires the application of the bounded convergence and mean value theorems; the details may be found in Barron [30]. □

This theorem can be used to prove the entropy power inequality, which gives a lower bound on the entropy of a sum of independent random variables.

Theorem 17.7.3 *(Entropy power inequality) If* $\mathbf{X}$ *and* $\mathbf{Y}$ *are independent random* n*-vectors with densities, then*

$$2^{\frac{2}{n}h(\mathbf{X}+\mathbf{Y})} \geq 2^{\frac{2}{n}h(\mathbf{X})} + 2^{\frac{2}{n}h(\mathbf{Y})}. \tag{17.86}$$

We outline the basic steps in the proof due to Stam [505] and Blachman [61]. A different proof is given in Section 17.8.

Stam's proof of the entropy power inequality is based on a perturbation argument. Let $n = 1$. Let $X_t = X + \sqrt{f(t)}Z_1$, $Y_t = Y + \sqrt{g(t)}Z_2$, where Z_1 and Z_2 are independent $\mathcal{N}(0, 1)$ random variables. Then the entropy power inequality for $n = 1$ reduces to showing that $s(0) \leq 1$, where we define

$$s(t) = \frac{2^{2h(X_t)} + 2^{2h(Y_t)}}{2^{2h(X_t+Y_t)}}. \tag{17.87}$$

If $f(t) \to \infty$ and $g(t) \to \infty$ as $t \to \infty$, it is easy to show that $s(\infty) = 1$. If, in addition, $s'(t) \geq 0$ for $t \geq 0$, this implies that $s(0) \leq 1$. The proof of the fact that $s'(t) \geq 0$ involves a clever choice of the functions $f(t)$ and $g(t)$, an application of Theorem 17.7.2 and the use of a convolution inequality for Fisher information,

$$\frac{1}{J(X+Y)} \geq \frac{1}{J(X)} + \frac{1}{J(Y)}. \tag{17.88}$$

The entropy power inequality can be extended to the vector case by induction. The details may be found in the papers by Stam [505] and Blachman [61].

17.8 ENTROPY POWER INEQUALITY AND BRUNN–MINKOWSKI INEQUALITY

The entropy power inequality provides a lower bound on the differential entropy of a sum of two independent random vectors in terms of their individual differential entropies. In this section we restate and outline an