odes. He described a simple code
›blem 5.5.28), which he attributed
Shannon–Fano code construction

lecodable codes was first proved
ıere is due to Karush [306]. The
‹hibited and proved to be optimal

lerable interest in designing source
›lications, such as magnetic record-
ɔ design codes so that the output
›me of the results for this problem
er et al. [5] and Marcus [370].
as its roots in the Shannon–Fano
l), which was analyzed by Jelinek
ion of a prefix-free code described
oore [249]. The extension of the
ıences is based on the enumerative
ibed with finite-precision arithmetic
'he competitive optimality of Shan-
and extended to Huffman codes by
ration of discrete distributions from
ıuth and Yao[317].

CHAPTER 6

GAMBLING AND DATA COMPRESSION

At first sight, information theory and gambling seem to be unrelated. But as we shall see, there is strong duality between the growth rate of investment in a horse race and the entropy rate of the horse race. Indeed, the sum of the growth rate and the entropy rate is a constant. In the process of proving this, we shall argue that the financial value of side information is equal to the mutual information between the horse race and the side information. The horse race is a special case of investment in the stock market, studied in Chapter 16.

We also show how to use a pair of identical gamblers to compress a sequence of random variables by an amount equal to the growth rate of wealth on that sequence. Finally, we use these gambling techniques to estimate the entropy rate of English.

6.1 THE HORSE RACE

Assume that m horses run in a race. Let the ith horse win with probability p_i. If horse i wins, the payoff is o_i for 1 (i.e., an investment of 1 dollar on horse i results in o_i dollars if horse i wins and 0 dollars if horse i loses).

There are two ways of describing odds: a-for-1 and b-to-1. The first refers to an exchange that takes place before the race—the gambler puts down 1 dollar before the race and at a-for-1 odds will receive a dollars after the race if his horse wins, and will receive nothing otherwise. The second refers to an exchange after the race—at b-to-1 odds, the gambler will pay 1 dollar after the race if his horse loses and will pick up b dollars after the race if his horse wins. Thus, a bet at b-to-1 odds is equivalent to a bet at a-for-1 odds if $b = a - 1$. For example, fair odds on a coin flip would be 2-for-1 or 1-to-1, otherwise known as *even odds.*

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

We assume that the gambler distributes all of his wealth across the horses. Let b_i be the fraction of the gambler's wealth invested in horse i, where $b_i \geq 0$ and $\sum b_i = 1$. Then if horse i wins the race, the gambler will receive o_i times the amount of wealth bet on horse i. All the other bets are lost. Thus, at the end of the race, the gambler will have multiplied his wealth by a factor $b_i o_i$ if horse i wins, and this will happen with probability p_i. For notational convenience, we use $b(i)$ and b_i interchangeably throughout this chapter.

The wealth at the end of the race is a random variable, and the gambler wishes to "maximize" the value of this random variable. It is tempting to bet everything on the horse that has the maximum expected return (i.e., the one with the maximum $p_i o_i$). But this is clearly risky, since all the money could be lost.

Some clarity results from considering repeated gambles on this race. Now since the gambler can reinvest his money, his wealth is the product of the gains for each race. Let S_n be the gambler's wealth after n races. Then

$$S_n = \prod_{i=1}^{n} S(X_i), \tag{6.1}$$

where $S(X) = b(X)o(X)$ is the factor by which the gambler's wealth is multiplied when horse X wins.

Definition The *wealth relative* $S(X) = b(X)o(X)$ is the factor by which the gambler's wealth grows if horse X wins the race.

Definition The *doubling rate* of a horse race is

$$W(\mathbf{b}, \mathbf{p}) = E(\log S(X)) = \sum_{k=1}^{m} p_k \log b_k o_k. \tag{6.2}$$

The definition of doubling rate is justified by the following theorem.

Theorem 6.1.1 *Let the race outcomes $X_1, X_2, \ldots$ be i.i.d. $\sim p(x)$. Then the wealth of the gambler using betting strategy* $\mathbf{b}$ *grows exponentially at rate $W(\mathbf{b}, \mathbf{p})$; that is,*

$$S_n \doteq 2^{nW(\mathbf{b},\mathbf{p})}. \tag{6.3}$$

Proof: Functions of independent random variables are also independent, and hence $\log S(X_1), \log S(X_2), \ldots$ are i.i.d. Then, by the weak law of large numbers,

$$\frac{1}{n} \log S_n = \frac{1}{n} \sum_{i=1}^{n} \log S(X_i) \to E(\log S(X)) \quad \text{in probability.} \tag{6.4}$$

Thus,

$$S_n \doteq 2^{nW(\mathbf{b},\mathbf{p})}. \qquad \square \tag{6.5}$$

Now since the gambler's wealth grows as $2^{nW(\mathbf{b},\mathbf{p})}$, we seek to maximize the exponent $W(\mathbf{b}, \mathbf{p})$ over all choices of the portfolio $\mathbf{b}$.

Definition The *optimum doubling rate* $W^*(\mathbf{p})$ is the maximum doubling rate over all choices of the portfolio $\mathbf{b}$:

$$W^*(\mathbf{p}) = \max_{\mathbf{b}} W(\mathbf{b}, \mathbf{p}) = \max_{\mathbf{b}: b_i \geq 0, \sum_i b_i = 1} \sum_{i=1}^{m} p_i \log b_i o_i. \tag{6.6}$$

We maximize $W(\mathbf{b}, \mathbf{p})$ as a function of $\mathbf{b}$ subject to the constraint $\sum b_i = 1$. Writing the functional with a Lagrange multiplier and changing the base of the logarithm (which does not affect the maximizing $\mathbf{b}$), we have

$$J(\mathbf{b}) = \sum p_i \ln b_i o_i + \lambda \sum b_i. \tag{6.7}$$

Differentiating this with respect to b_i yields

$$\frac{\partial J}{\partial b_i} = \frac{p_i}{b_i} + \lambda, \quad i = 1, 2, \ldots, m. \tag{6.8}$$

Setting the partial derivative equal to 0 for a maximum, we have

$$b_i = -\frac{p_i}{\lambda}. \tag{6.9}$$

Substituting this in the constraint $\sum b_i = 1$ yields $\lambda = -1$ and $b_i = p_i$. Hence, we can conclude that $\mathbf{b} = \mathbf{p}$ is a stationary point of the function $J(\mathbf{b})$. To prove that this is actually a maximum is tedious if we take second derivatives. Instead, we use a method that works for many such problems: Guess and verify. We verify that proportional gambling $\mathbf{b} = \mathbf{p}$ is optimal in the following theorem. Proportional gambling is known as *Kelly gambling* [308].

Theorem 6.1.2 (*Proportional gambling is log-optimal*) *The optimum doubling rate is given by*

$$W^*(\mathbf{p}) = \sum p_i \log o_i - H(\mathbf{p}) \tag{6.10}$$

and is achieved by the proportional gambling scheme $\mathbf{b}^* = \mathbf{p}$.

Proof: We rewrite the function $W(\mathbf{b}, \mathbf{p})$ in a form in which the maximum is obvious:

$$W(\mathbf{b}, \mathbf{p}) = \sum p_i \log b_i o_i \tag{6.11}$$

$$= \sum p_i \log \left(\frac{b_i}{p_i} p_i o_i \right) \tag{6.12}$$

$$= \sum p_i \log o_i - H(\mathbf{p}) - D(\mathbf{p}||\mathbf{b}) \tag{6.13}$$

$$\leq \sum p_i \log o_i - H(\mathbf{p}), \tag{6.14}$$

with equality iff $\mathbf{p} = \mathbf{b}$ (i.e., the gambler bets on each horse in proportion to its probability of winning). □

Example 6.1.1 Consider a case with two horses, where horse 1 wins with probability p_1 and horse 2 wins with probability p_2. Assume even odds (2-for-1 on both horses). Then the optimal bet is proportional betting (i.e., $b_1 = p_1$, $b_2 = p_2$). The optimal doubling rate is $W^*(\mathbf{p}) = \sum p_i \log o_i - H(\mathbf{p}) = 1 - H(\mathbf{p})$, and the resulting wealth grows to infinity at this rate:

$$S_n \doteq 2^{n(1-H(\mathbf{p}))}. \tag{6.15}$$

Thus, we have shown that proportional betting is growth rate optimal for a sequence of i.i.d. horse races if the gambler can reinvest his wealth and if there is no alternative of keeping some of the wealth in cash.

We now consider a special case when the odds are fair with respect to some distribution (i.e., there is no track take and $\sum \frac{1}{o_i} = 1$). In this case, we write $r_i = \frac{1}{o_i}$, where r_i can be interpreted as a probability mass function over the horses. (This is the bookie's estimate of the win probabilities.) With this definition, we can write the doubling rate as

$$W(\mathbf{b}, \mathbf{p}) = \sum p_i \log b_i o_i \tag{6.16}$$

$$= \sum p_i \log \left(\frac{b_i}{p_i} \frac{p_i}{r_i} \right) \tag{6.17}$$

$$= D(\mathbf{p}||\mathbf{r}) - D(\mathbf{p}||\mathbf{b}). \tag{6.18}$$

This equation gives another interpretation for the relative entropy *distance*: The doubling rate is the difference between the distance of the bookie's estimate from the true distribution and the distance of the gambler's estimate from the true distribution. Hence, the gambler can make money only if his estimate (as expressed by $\mathbf{b}$) is better than the bookie's.

An even more special case is when the odds are m-for-1 on each horse. In this case, the odds are fair with respect to the uniform distribution and the optimum doubling rate is

$$W^*(\mathbf{p}) = D\left(\mathbf{p}||\frac{1}{m}\right) = \log m - H(\mathbf{p}). \tag{6.19}$$

In this case we can clearly see the duality between data compression and the doubling rate.

Theorem 6.1.3 (*Conservation theorem*) *For uniform fair odds,*

$$W^*(\mathbf{p}) + H(\mathbf{p}) = \log m. \tag{6.20}$$

Thus, the sum of the doubling rate and the entropy rate is a constant.

Every bit of entropy decrease doubles the gambler's wealth. Low entropy races are the most profitable.

In the analysis above, we assumed that the gambler was fully invested. In general, we should allow the gambler the option of retaining some of his wealth as cash. Let $b(0)$ be the proportion of wealth held out as cash, and $b(1), b(2), \ldots, b(m)$ be the proportions bet on the various horses. Then at the end of a race, the ratio of final wealth to initial wealth (the *wealth relative*) is

$$S(X) = b(0) + b(X)o(X). \tag{6.21}$$

Now the optimum strategy may depend on the odds and will not necessarily have the simple form of proportional gambling. We distinguish three subcases:

1. *Fair odds with respect to some distribution*: $\sum \frac{1}{o_i} = 1$. For fair odds, the option of withholding cash does not change the analysis. This is because we can get the effect of withholding cash by betting $b_i = \frac{1}{o_i}$ on the ith horse, $i = 1, 2, \ldots, m$. Then $S(X) = 1$ irrespective of which horse wins. Thus, whatever money the gambler keeps aside as cash can equally well be distributed over the horses, and the assumption that the gambler must invest all his money does not change the analysis. Proportional betting is optimal.
2. *Superfair odds*: $\sum \frac{1}{o_i} < 1$. In this case, the odds are even better than fair odds, so one would always want to put all one's wealth into the race rather than leave it as cash. In this race, too, the optimum strategy is proportional betting. However, it is possible to choose $\mathbf{b}$ so as to form a *Dutch book* by choosing $b_i = c\frac{1}{o_i}$, where $c = 1/\sum \frac{1}{c_i}$, to get $o_i b_i = c$, irrespective of which horse wins. With this allotment, one has wealth $S(X) = 1/\sum \frac{1}{o_i} > 1$ with probability 1 (i.e., no risk). Needless to say, one seldom finds such odds in real life. Incidentally, a Dutch book, although risk-free, does not optimize the doubling rate.
3. *Subfair odds*: $\sum \frac{1}{o_i} > 1$. This is more representative of real life. The organizers of the race track take a cut of all the bets. In this case it is optimal to bet only some of the money and leave the rest aside as cash. Proportional gambling is no longer log-optimal. A parametric form for the optimal strategy can be found using Kuhn–Tucker conditions (Problem 6.6.2); it has a simple "water-filling" interpretation.

6.2 GAMBLING AND SIDE INFORMATION

Suppose the gambler has some information that is relevant to the outcome of the gamble. For example, the gambler may have some information about the performance of the horses in previous races. What is the value of this side information?

One definition of the financial value of such information is the increase in wealth that results from that information. In the setting described in Section 6.1 the measure of the value of information is the increase in the doubling rate due to that information. We will now derive a connection between mutual information and the increase in the doubling rate.

To formalize the notion, let horse $X \in \{1, 2, \ldots, m\}$ win the race with probability $p(x)$ and pay odds of $o(x)$ for 1. Let (X, Y) have joint probability mass function $p(x, y)$. Let $b(x|y) \geq 0$, $\sum_x b(x|y) = 1$ be an arbitrary conditional betting strategy depending on the side information Y, where $b(x|y)$ is the proportion of wealth bet on horse x when y is observed. As before, let $b(x) \geq 0$, $\sum b(x) = 1$ denote the unconditional betting scheme.

Let the unconditional and the conditional doubling rates be

$$W^*(X) = \max_{\mathbf{b}(x)} \sum_x p(x) \log b(x)o(x), \tag{6.22}$$

$$W^*(X|Y) = \max_{\mathbf{b}(x|y)} \sum_{x,y} p(x, y) \log b(x|y)o(x) \tag{6.23}$$

and let

$$\Delta W = W^*(X|Y) - W^*(X). \tag{6.24}$$

We observe that for (X_i, Y_i) i.i.d. horse races, wealth grows like $2^{nW^*(X|Y)}$ with side information and like $2^{nW^*(X)}$ without side information.

Theorem 6.2.1 *The increase ΔW in doubling rate due to side information Y for a horse race X is*

$$\Delta W = I(X; Y). \tag{6.25}$$

Proof: With side information, the maximum value of $W^*(X|Y)$ with side information Y is achieved by conditionally proportional gambling [i.e., $b^*(x|y) = p(x|y)$]. Thus,

$$W^*(X|Y) = \max_{\mathbf{b}(x|y)} E\left[\log S\right] = \max_{\mathbf{b}(x|y)} \sum p(x, y) \log o(x)b(x|y) \tag{6.26}$$

$$= \sum p(x, y) \log o(x)p(x|y) \tag{6.27}$$

$$= \sum p(x) \log o(x) - H(X|Y). \tag{6.28}$$

Without side information, the optimal doubling rate is

$$W^*(X) = \sum p(x) \log o(x) - H(X). \tag{6.29}$$

Thus, the increase in doubling rate due to the presence of side information Y is

$$\Delta W = W^*(X|Y) - W^*(X) = H(X) - H(X|Y) = I(X; Y). \quad \square \tag{6.30}$$

Hence, the increase in doubling rate is equal to the mutual information between the side information and the horse race. Not surprisingly, independent side information does not increase the doubling rate.

This relationship can also be extended to the general stock market (Chapter 16). In this case, however, one can only show the inequality $\Delta W \leq I$, with equality if and only if the market is a horse race.

6.3 DEPENDENT HORSE RACES AND ENTROPY RATE

The most common example of side information for a horse race is the past performance of the horses. If the horse races are independent, this information will be useless. If we assume that there is dependence among the races, we can calculate the effective doubling rate if we are allowed to use the results of previous races to determine the strategy for the next race.

Suppose that the sequence $\{X_k\}$ of horse race outcomes forms a stochastic process. Let the strategy for each race depend on the results of previous races. In this case, the optimal doubling rate for uniform fair odds is

$$\begin{aligned} & W^*(X_k|X_{k-1}, X_{k-2}, \ldots, X_1) \\ & \quad = E\left[\max_{\mathbf{b}(\cdot|X_{k-1}, X_{k-2}, \ldots, X_1)} E[\log S(X_k)|X_{k-1}, X_{k-2}, \ldots, X_1] \right] \\ & \quad = \log m - H(X_k|X_{k-1}, X_{k-2}, \ldots, X_1), \end{aligned} \tag{6.31}$$

which is achieved by $b^*(x_k|x_{k-1}, \ldots, x_1) = p(x_k|x_{k-1}, \ldots, x_1)$.

At the end of n races, the gambler's wealth is

$$S_n = \prod_{i=1}^{n} S(X_i), \tag{6.32}$$

and the exponent in the growth rate (assuming m for 1 odds) is

$$\frac{1}{n} E \log S_n = \frac{1}{n} \sum E \log S(X_i) \tag{6.33}$$

$$= \frac{1}{n} \sum (\log m - H(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)) \tag{6.34}$$

$$= \log m - \frac{H(X_1, X_2, \ldots, X_n)}{n}. \tag{6.35}$$

The quantity $\frac{1}{n} H(X_1, X_2, \ldots, X_n)$ is the average entropy per race. For a stationary process with entropy rate $H(\mathcal{X})$, the limit in (6.35) yields

$$\lim_{n\to\infty} \frac{1}{n} E \log S_n + H(\mathcal{X}) = \log m. \tag{6.36}$$

Again, we have the result that the entropy rate plus the doubling rate is a constant.

The expectation in (6.36) can be removed if the process is ergodic. It will be shown in Chapter 16 that for an ergodic sequence of horse races,

$$S_n \doteq 2^{nW} \quad \text{with probability 1}, \tag{6.37}$$

where $W = \log m - H(\mathcal{X})$ and

$$H(\mathcal{X}) = \lim \frac{1}{n} H(X_1, X_2, \ldots, X_n). \tag{6.38}$$

Example 6.3.1 *(Red and black)* In this example, cards replace horses and the outcomes become more predictable as time goes on. Consider the case of betting on the color of the next card in a deck of 26 red and 26 black cards. Bets are placed on whether the next card will be red or black, as we go through the deck. We also assume that the game pays 2-for-1; that is, the gambler gets back twice what he bets on the right color. These are fair odds if red and black are equally probable.

We consider two alternative betting schemes:

1. If we bet sequentially, we can calculate the conditional probability of the next card and bet proportionally. Thus, we should bet $(\frac{1}{2}, \frac{1}{2})$ on (red, black) for the first card, $(\frac{26}{51}, \frac{25}{51})$ for the second card if the first card is black, and so on.
2. Alternatively, we can bet on the entire sequence of 52 cards at once. There are $\binom{52}{26}$ possible sequences of 26 red and 26 black cards, all of them equally likely. Thus, proportional betting implies that we put $1/\binom{52}{26}$ of our money on each of these sequences and let each bet "ride."

We will argue that these procedures are equivalent. For example, half the sequences of 52 cards start with red, and so the proportion of money bet on sequences that start with red in scheme 2 is also one-half, agreeing with the proportion used in the first scheme. In general, we can verify that betting $1/\binom{52}{26}$ of the money on each of the possible outcomes will at each stage give bets that are proportional to the probability of red and black at that stage. Since we bet $1/\binom{52}{26}$ of the wealth on each possible output sequence, and a bet on a sequence increases wealth by a factor of 2^{52} on the sequence observed and 0 on all the others, the resulting wealth is

$$S^*_{52} = \frac{2^{52}}{\binom{52}{26}} = 9.08. \qquad (6.39)$$

Rather interestingly, the return does not depend on the actual sequence. This is like the AEP in that the return is the same for all sequences. All sequences are typical in this sense.

6.4 THE ENTROPY OF ENGLISH

An important example of an information source is English text. It is not immediately obvious whether English is a stationary ergodic process. Probably not! Nonetheless, we will be interested in the entropy rate of English. We discuss various stochastic approximations to English. As we increase the complexity of the model, we can generate text that looks like English. The stochastic models can be used to compress English text. The better the stochastic approximation, the better the compression.

For the purposes of discussion, we assume that the alphabet of English consists of 26 letters and the space symbol. We therefore ignore punctuation and the difference between upper- and lowercase letters. We construct models for English using empirical distributions collected from samples of text. The frequency of letters in English is far from uniform. The most common letter, E, has a frequency of about 13%, and the least common letters, Q and Z, occur with a frequency of about 0.1%. The letter E is so common that it is rare to find a sentence of any length that does not contain the letter. [A surprising exception to this is the 267-page novel, *Gadsby*, by Ernest Vincent Wright (Lightyear Press, Boston, 1997; original publication in 1939), in which the author deliberately makes no use of the letter E.]

The frequency of pairs of letters is also far from uniform. For example, the letter Q is always followed by a U. The most frequent pair is TH, which occurs normally with a frequency of about 3.7%. We can use the frequency of the pairs to estimate the probability that a letter follows any other letter. Proceeding this way, we can also estimate higher-order conditional probabilities and build more complex models for the language. However, we soon run out of data. For example, to build a third-order Markov approximation, we must estimate the values of

$p(x_i|x_{i-1}, x_{i-2}, x_{i-3})$. There are $27^4 = 531,441$ entries in this table, and we would need to process millions of letters to make accurate estimates of these probabilities.

The conditional probability estimates can be used to generate random samples of letters drawn according to these distributions (using a random number generator). But there is a simpler method to simulate randomness using a sample of text (a book, say). For example, to construct the second-order model, open the book at random and choose a letter at random on the page. This will be the first letter. For the next letter, again open the book at random and starting at a random point, read until the first letter is encountered again. Then take the letter after that as the second letter. We repeat this process by opening to another page, searching for the second letter, and taking the letter after that as the third letter. Proceeding this way, we can generate text that simulates the second-order statistics of the English text.

Here are some examples of Markov approximations to English from Shannon's original paper [472]:

1. *Zero-order approximation*. (The symbols are independent and equiprobable.)

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ
   FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD

2. *First-order approximation*. (The symbols are independent. The frequency of letters matches English text.)

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI
   ALHENHTTPA OOBTTVA NAH BRL

3. *Second-order approximation*. (The frequency of pairs of letters matches English text.)

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY
   ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO
   TIZIN ANDY TOBE SEACE CTISBE

4. *Third-order approximation*. (The frequency of triplets of letters matches English text.)

   IN NO IST LAT WHEY CRATICT FROURE BERS GROCID
   PONDENOME OF DEMONSTURES OF THE REPTAGIN IS
   REGOACTIONA OF CRE

5. *Fourth-order approximation*. (The frequency of quadruplets of letters matches English text. Each letter depends on the previous three letters. This sentence is from Lucky's book, *Silicon Dreams* [366].)

   THE GENERATED JOB PROVIDUAL BETTER TRAND THE DISPLAYED CODE, ABOVERY UPONDULTS WELL THE CODERST IN THESTICAL IT DO HOCK BOTHE MERG. (INSTATES CONS ERATION. NEVER ANY OF PUBLE AND TO THEORY. EVENTIAL CALLEGAND TO ELAST BENERATED IN WITH PIES AS IS WITH THE)

   Instead of continuing with the letter models, we jump to word models.

6. *First-order word model*. (The words are chosen independently but with frequencies as in English.)

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

7. *Second-order word model*. (The word transition probabilities match English text.)

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

The approximations get closer and closer to resembling English. For example, long phrases of the last approximation could easily have occurred in a real English sentence. It appears that we could get a very good approximation by using a more complex model. These approximations could be used to estimate the entropy of English. For example, the entropy of the zeroth-order model is $\log 27 = 4.76$ bits per letter. As we increase the complexity of the model, we capture more of the structure of English, and the conditional uncertainty of the next letter is reduced. The first-order model gives an estimate of the entropy of 4.03 bits per letter, while the fourth-order model gives an estimate of 2.8 bits per letter. But even the fourth-order model does not capture all the structure of English. In Section 6.6 we describe alternative methods for estimating the entropy of English.

The distribution of English is useful in decoding encrypted English text. For example, a simple substitution cipher (where each letter is replaced by some other letter) can be solved by looking for the most frequent letter and guessing that it is the substitute for E, and so on. The redundancy in English can be used to fill in some of the missing letters after the other letters are decrypted: for example,

TH_R_ _S _NLY _N_ W_Y T_ F_LL _N TH_ V_W_LS _N TH_S S_NT_NC_.

Some of the inspiration for Shannon's original work on information theory came out of his work in cryptography during World War II. The mathematical theory of cryptography and its relationship to the entropy of language is developed in Shannon [481].

Stochastic models of language also play a key role in some speech recognition systems. A commonly used model is the trigram (second-order Markov) word model, which estimates the probability of the next word given the preceding two words. The information from the speech signal is combined with the model to produce an estimate of the most likely word that could have produced the observed speech. Random models do surprisingly well in speech recognition, even when they do not explicitly incorporate the complex rules of grammar that govern natural languages such as English.

We can apply the techniques of this section to estimate the entropy rate of other information sources, such as speech and images. A fascinating nontechnical introduction to these issues may be found in the book by Lucky [366].

6.5 DATA COMPRESSION AND GAMBLING

We now show a direct connection between gambling and data compression, by showing that a good gambler is also a good data compressor. Any sequence on which a gambler makes a large amount of money is also a sequence that can be compressed by a large factor. The idea of using the gambler as a data compressor is based on the fact that the gambler's bets can be considered to be his estimate of the probability distribution of the data. A good gambler will make a good estimate of the probability distribution. We can use this estimate of the distribution to do arithmetic coding (Section 13.3). This is the essential idea of the scheme described below.

We assume that the gambler has a mechanically identical twin, who will be used for the data decompression. The identical twin will place the same bets on possible sequences of outcomes as the original gambler (and will therefore make the same amount of money). The cumulative amount of money that the gambler would have made on all sequences that are lexicographically less than the given sequence will be used as a code for the sequence. The decoder will use the identical twin to gamble on all sequences, and look for the sequence for which the same cumulative amount of money is made. This sequence will be chosen as the decoded sequence.

Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables that we wish to compress. Without loss of generality, we will assume that the random variables are binary. Gambling on this sequence will be defined by a sequence of bets

$$b(x_{k+1} \mid x_1, x_2, \ldots, x_k) \geq 0, \quad \sum_{x_{k+1}} b(x_{k+1} \mid x_1, x_2, \ldots, x_k) = 1, \tag{6.40}$$

where $b(x_{k+1} \mid x_1, x_2, \ldots, x_k)$ is the proportion of money bet at time k on the event that $X_{k+1} = x_{k+1}$ given the observed past $x_1, x_2, \ldots, x_k$. Bets are paid at uniform odds (2-for-1). Thus, the wealth S_n at the end of the sequence is given by

$$S_n = 2^n \prod_{k=1}^{n} b(x_k \mid x_1, \ldots, x_{k-1}) \tag{6.41}$$

$$= 2^n b(x_1, x_2, \ldots, x_n), \tag{6.42}$$

where

$$b(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} b(x_k|x_{k-1}, \ldots, x_1). \tag{6.43}$$

So sequential gambling can also be considered as an assignment of probabilities (or bets) $b(x_1, x_2, \ldots, x_n) \geq 0$, $\sum_{x_1,\ldots,x_n} b(x_1, \ldots, x_n) = 1$, on the 2^n possible sequences.

This gambling elicits both an estimate of the true probability of the text sequence $(\hat{p}(x_1, \ldots, x_n) = S_n/2^n)$ as well as an estimate of the entropy $\left[\hat{H} = -\frac{1}{n}\log \hat{p}\right]$ of the text from which the sequence was drawn. We now wish to show that high values of wealth S_n lead to high data compression. Specifically, we argue that if the text in question results in wealth S_n, then $\log S_n$ bits can be saved in a naturally associated deterministic data compression scheme. We further assert that if the gambling is log optimal, the data compression achieves the Shannon limit H.

Consider the following data compression algorithm that maps the text $\mathbf{x} = x_1 x_2 \cdots x_n \in \{0, 1\}^n$ into a code sequences $c_1 c_2 \cdots c_k$, $c_i \in \{0, 1\}$. Both the compressor and the decompressor know n. Let the 2^n text sequences be arranged in lexicographical order: for example, $0100101 < 0101101$. The encoder observes the sequence $x^n = (x_1, x_2, \ldots, x_n)$. He then calculates what his wealth $S_n(x^{'}(n))$ would have been on all sequences $x^{'}(n) \leq x(n)$ and calculates $F(x(n)) = \sum_{x^{'}(n) \leq x(n)} 2^{-n} S_n(x^{'}(n))$. Clearly, $F(x(n)) \in [0, 1]$. Let $k = \lceil n - \log\ S_n(x(n))\rceil$. Now express $F(x(n))$ as a binary decimal to k-place accuracy: $\lfloor F(x(n))\rfloor = .c_1 c_2 \cdots c_k$. The sequence $c(k) = (c_1, c_2, \ldots, c_k)$ is transmitted to the decoder.

The decoder twin can calculate the precise value $S(x^{'}(n))$ associated with each of the 2^n sequences $x^{'}(n)$. He thus knows the cumulative sum of $2^{-n} S(x^{'}(n))$ up through any sequence $x(n)$. He tediously calculates this sum until it first exceeds $.c(k)$. The first sequence $x(n)$ such that the cumulative sum falls in the interval $[.c_1 \cdots c_k, .c_1 \ldots c_k + (1/2)^k]$ is defined uniquely, and the size of $S(x(n))/2^n$ guarantees that this sequence will be precisely the encoded $x(n)$.

Thus, the twin uniquely recovers $x(n)$. The number of bits required is $k = \lceil n - \log S(x(n))\rceil$. The number of bits saved is $n - k = \lfloor \log S(x(n))\rfloor$. For proportional gambling, $S(x(n)) = 2^n p(x(n))$. Thus, the expected number of bits is $Ek = \sum p(x(n))\lceil -\log p(x(n))\rceil \leq H(X_1, \ldots, X_n) + 1$.

We see that if the betting operation is deterministic and is known both to the encoder and the decoder, the number of bits necessary to encode $x_1, \ldots, x_n$ is less than $n - \log S_n + 1$. Moreover, if $p(x)$ is known, and if proportional gambling is used, the description length expected is $E(n - \log S_n) \leq H(X_1, \ldots, X_n) + 1$. Thus, the gambling results correspond precisely to the data compression that would have been achieved by the given human encoder–decoder identical twin pair.

The data compression scheme using a gambler is similar to the idea of arithmetic coding (Section 13.3) using a distribution $b(x_1, x_2, \ldots, x_n)$ rather than the true distribution. The procedure above brings out the duality between gambling and data compression. Both involve estimation of the true distribution. The better the estimate, the greater the growth rate of the gambler's wealth and the better the data compression.

6.6 GAMBLING ESTIMATE OF THE ENTROPY OF ENGLISH

We now estimate the entropy rate for English using a human gambler to estimate probabilities. We assume that English consists of 27 characters

(26 letters and a space symbol). We therefore ignore punctuation and case of letters. Two different approaches have been proposed to estimate the entropy of English.

1. *Shannon guessing game*. In this approach the human subject is given a sample of English text and asked to guess the next letter. An optimal subject will estimate the probabilities of the next letter and guess the most probable letter first, then the second most probable letter next, and so on. The experimenter records the number of guesses required to guess the next letter. The subject proceeds this way through a fairly large sample of text. We can then calculate the empirical frequency distribution of the number of guesses required to guess the next letter. Many of the letters will require only one guess; but a large number of guesses will usually be needed at the beginning of words or sentences.

   Now let us assume that the subject can be modeled as a computer making a deterministic choice of guesses given the past text. Then if we have the same machine and the sequence of guess numbers, we can reconstruct the English text. Just let the machine run, and if the number of guesses at any position is k, choose the kth guess of the machine as the next letter. Hence the amount of information in the sequence of guess numbers is the same as in the English text. The entropy of the guess sequence is the entropy of English text. We can bound the entropy of the guess sequence by assuming that the samples are independent. Hence, the entropy of the guess sequence is bounded above by the entropy of the histogram in the experiment. The experiment was conducted in 1950 by Shannon [482], who obtained a value of 1.3 bits per symbol for the entropy of English.
2. *Gambling estimate*. In this approach we let a human subject gamble on the next letter in a sample of English text. This allows finer gradations of judgment than does guessing. As in the case of a horse race, the optimal bet is proportional to the conditional probability of the next letter. The payoff is 27-for-1 on the correct letter.

   Since sequential betting is equivalent to betting on the entire sequence, we can write the payoff after n letters as

$$S_n = (27)^n b(X_1, X_2, \ldots, X_n). \tag{6.44}$$

   Thus, after n rounds of betting, the expected log wealth satisfies

$$E\frac{1}{n}\log S_n = \log 27 + \frac{1}{n}E\log b(X_1, X_2, \ldots, X_n) \tag{6.45}$$

$$
\begin{aligned}
&= \log 27 + \frac{1}{n}\sum_{x^n} p(x^n)\log b(x^n) && (6.46)\\
&= \log 27 - \frac{1}{n}\sum_{x^n} p(x^n)\log\frac{p(x^n)}{b(x^n)}\\
&\quad + \frac{1}{n}\sum_{x^n} p(x^n)\log p(x^n) && (6.47)\\
&= \log 27 - \frac{1}{n}D(p(x^n)||b(x^n)) - \frac{1}{n}H(X_1, X_2, \ldots, X_n)\\
& && (6.48)\\
&\le \log 27 - \frac{1}{n}H(X_1, X_2, \ldots, X_n) && (6.49)\\
&\le \log 27 - H(\mathcal{X}), && (6.50)
\end{aligned}
$$

where $H(\mathcal{X})$ is the entropy rate of English. Thus, $\log 27 - E\frac{1}{n}\log S_n$ is an upper bound on the entropy rate of English. The upper bound estimate, $\hat{H}(\mathcal{X}) = \log 27 - \frac{1}{n}\log S_n$, converges to $H(\mathcal{X})$ with probability 1 if English is ergodic and the gambler uses $b(x^n) = p(x^n)$. An experiment [131] with 12 subjects and a sample of 75 letters from the book *Jefferson the Virginian* by Dumas Malone (Little, Brown, Boston, 1948; the source used by Shannon) resulted in an estimate of 1.34 bits per letter for the entropy of English.

SUMMARY

Doubling rate. $W(\mathbf{b}, \mathbf{p}) = E(\log S(X)) = \sum_{k=1}^{m} p_k \log b_k o_k$.

Optimal doubling rate. $W^*(\mathbf{p}) = \max_{\mathbf{b}} W(\mathbf{b}, \mathbf{p})$.

Proportional gambling is log-optimal

$$W^*(\mathbf{p}) = \max_{\mathbf{b}} W(\mathbf{b}, \mathbf{p}) = \sum p_i \log o_i - H(\mathbf{p}) \tag{6.51}$$

is achieved by $\mathbf{b}^* = \mathbf{p}$.

Growth rate. Wealth grows as $S_n \doteq 2^{nW^*(\mathbf{p})}$.

Conservation law. For uniform fair odds,

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

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

$$\Delta W = I(X;Y). \qquad (6.53)$$

PROBLEMS

6.1 *Horse race.* Three horses run a race. A gambler offers 3-for-1 odds on each horse. These are fair odds under the assumption that all horses are equally likely to win the race. The true win probabilities are known to be

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

Let $\mathbf{b} = (b_1, b_2, b_3)$, $b_i \geq 0$, $\sum b_i = 1$, be the amount invested on each of the horses. The expected log wealth is thus

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

(a) Maximize this over $\mathbf{b}$ to find $\mathbf{b}^*$ and W^*. Thus, the wealth achieved in repeated horse races should grow to infinity like 2^{nW^*} with probability 1.

(b) Show that if instead we put all of our money on horse 1, the most likely winner, we will eventually go broke with probability 1.

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

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

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

6.3 *Cards.* An ordinary deck of cards containing 26 red cards and 26 black cards is shuffled and dealt out one card at time without replacement. Let X_i be the color of the ith card.

(a) Determine $H(X_1)$.

(b) Determine $H(X_2)$.

(c) Does $H(X_k \mid X_1, X_2, \ldots, X_{k-1})$ increase or decrease?

(d) Determine $H(X_1, X_2, \ldots, X_{52})$.

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

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

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

and fair odds with respect to the (false) distribution

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

Thus, the odds are

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

(a) What is the entropy of the race?

(b) Find the set of bets (b_1, b_2, b_3) such that the compounded wealth in repeated plays will grow to infinity.

6.6 *Horse race.* A three-horse race has win probabilities $\mathbf{p} = (p_1, p_2, p_3)$, and odds $\mathbf{o} = (1, 1, 1)$. The gambler places bets $\mathbf{b} = (b_1, b_2, b_3)$, $b_i \geq 0$, $\sum b_i = 1$, where b_i denotes the proportion on wealth bet on horse i. These odds are very bad. The gambler gets his money back on the winning horse and loses the other bets. Thus, the wealth S_n at time n resulting from independent gambles goes exponentially to zero.

(a) Find the exponent.

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

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

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

6.8 *Lotto.* The following analysis is a crude approximation to the games of Lotto conducted by various states. Assume that the player of the game is required to pay \$1 to play and is asked to choose one number from a range 1 to 8. At the end of every day, the state lottery commission picks a number uniformly over the same range. The jackpot (i.e., all the money collected that day) is split among all the people who chose the same number as the one chosen by the state. For example, if 100 people played today, 10 of them chose the number 2, and the drawing at the end of the day picked 2, the \$100 collected is split among the 10 people (i.e., each person who picked 2 will receive \$10, and the others will receive nothing). The general population does not choose numbers uniformly—numbers such as 3 and 7 are supposedly lucky and are more popular than 4 or 8. Assume that the fraction of people choosing the various numbers $1, 2, \ldots, 8$ is $(f_1, f_2, \ldots, f_8)$, and assume that n people play every day. Also assume that n is very large, so that any single person's choice does not change the proportion of people betting on any number.

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

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

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

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

6.10 *Horse race with probability estimates.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The odds are superfair.

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

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

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

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

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

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

(c) Suppose that

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

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

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

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

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

(b) What is the optimal growth rate?

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

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

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

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

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

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

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

Let

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

We have

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

Let

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

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

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

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

(c) How does $W^*(c)$ fall off with c?

Note that since $W^*(c) > 0$, for all c, we can conclude that any entry fee c is fair.

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

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

and interpret the answer.

HISTORICAL NOTES

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

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

CHAPTER 7

CHANNEL CAPACITY

What do we mean when we say that A communicates with B? We mean that the physical acts of A have induced a desired physical state in B. This transfer of information is a physical process and therefore is subject to the uncontrollable ambient noise and imperfections of the physical signaling process itself. The communication is successful if the receiver B and the transmitter A agree on what was sent.

In this chapter we find the maximum number of distinguishable signals for n uses of a communication channel. This number grows exponentially with n, and the exponent is known as the channel capacity. The characterization of the channel capacity (the logarithm of the number of distinguishable signals) as the maximum mutual information is the central and most famous success of information theory.

The mathematical analog of a physical signaling system is shown in Figure 7.1. Source symbols from some finite alphabet are mapped into some sequence of channel symbols, which then produces the output sequence of the channel. The output sequence is random but has a distribution that depends on the input sequence. From the output sequence, we attempt to recover the transmitted message.

Each of the possible input sequences induces a probability distribution on the output sequences. Since two different input sequences may give rise to the same output sequence, the inputs are confusable. In the next few sections, we show that we can choose a "nonconfusable" subset of input sequences so that with high probability there is only one highly likely input that could have caused the particular output. We can then reconstruct the input sequences at the output with a negligible probability of error. By mapping the source into the appropriate "widely spaced" input sequences to the channel, we can transmit a message with very low probability of error and reconstruct the source message at the output. The maximum rate at which this can be done is called the *capacity* of the channel.

Definition We define a *discrete channel* to be a system consisting of an input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$ and a probability transition matrix

Elements of Information Theory, Second Edition, By Thomas M. Cover and Joy A. Thomas
Copyright © 2006 John Wiley & Sons, Inc.