$$= \log 27 + \frac{1}{n} \sum_{x^n} p(x^n) \log b(x^n) \tag{6.46}$$

$$= \log 27 - \frac{1}{n} \sum_{x^n} p(x^n) \log \frac{p(x^n)}{b(x^n)}$$

$$+ \frac{1}{n} \sum_{x^n} p(x^n) \log p(x^n) \tag{6.47}$$

$$= \log 27 - \frac{1}{n} D(p(x^n)||b(x^n)) - \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{6.48}$$

$$\leq \log 27 - \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{6.49}$$

$$\leq \log 27 - H(\mathcal{X}), \tag{6.50}$$

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,

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

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

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

PROBLEMS

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

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

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

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

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

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

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

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

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

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

(a) Determine $H(X_1)$.

(b) Determine $H(X_2)$.

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

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

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

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

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

and fair odds with respect to the (false) distribution

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

Thus, the odds are

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

(a) What is the entropy of the race?

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

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

(a) Find the exponent.

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

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

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

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

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

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

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

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

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

6.10 *Horse race with probability estimates.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The odds are superfair.

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

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

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

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

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

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

(c) Suppose that

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

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

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

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

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

(b) What is the optimal growth rate?

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

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

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

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

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

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

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

Let

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

We have

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

Let

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

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

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

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

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

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

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

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

and interpret the answer.

HISTORICAL NOTES

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

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

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

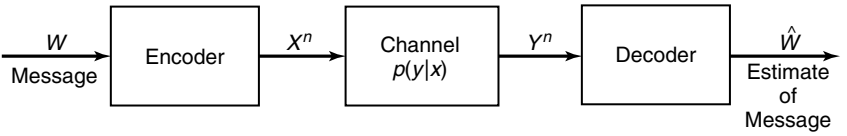


FIGURE 7.1. Communication system.

$p(y|x)$ that expresses the probability of observing the output symbol y given that we send the symbol x. The channel is said to be *memoryless* if the probability distribution of the output depends only on the input at that time and is conditionally independent of previous channel inputs or outputs.

Definition We define the *"information" channel capacity* of a discrete memoryless channel as

$$C = \max_{p(x)} I(X; Y), \tag{7.1}$$

where the maximum is taken over all possible input distributions $p(x)$.

We shall soon give an operational definition of channel capacity as the highest rate in bits per channel use at which information can be sent with arbitrarily low probability of error. Shannon's second theorem establishes that the information channel capacity is equal to the operational channel capacity. Thus, we drop the word *information* in most discussions of channel capacity.

There is a duality between the problems of data compression and data transmission. During compression, we remove all the redundancy in the data to form the most compressed version possible, whereas during data transmission, we add redundancy in a controlled fashion to combat errors in the channel. In Section 7.13 we show that a general communication system can be broken into two parts and that the problems of data compression and data transmission can be considered separately.

7.1 EXAMPLES OF CHANNEL CAPACITY

7.1.1 Noiseless Binary Channel

Suppose that we have a channel whose the binary input is reproduced exactly at the output (Figure 7.2).

In this case, any transmitted bit is received without error. Hence, one error-free bit can be transmitted per use of the channel, and the capacity is

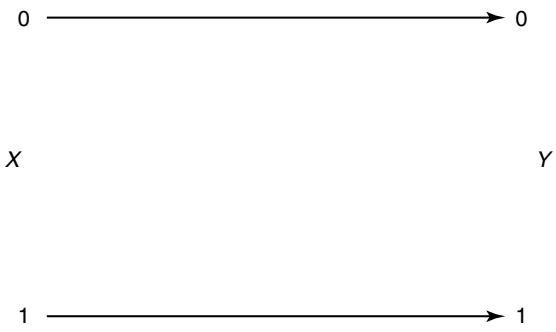


FIGURE 7.2. Noiseless binary channel. $C = 1$ bit.

1 bit. We can also calculate the information capacity $C = \max I(X; Y) = 1$ bit, which is achieved by using $p(x) = (\frac{1}{2}, \frac{1}{2})$.

7.1.2 Noisy Channel with Nonoverlapping Outputs

This channel has two possible outputs corresponding to each of the two inputs (Figure 7.3). The channel appears to be noisy, but really is not. Even though the output of the channel is a random consequence of the input, the input can be determined from the output, and hence every transmitted bit can be recovered without error. The capacity of this channel is also 1 bit per transmission. We can also calculate the information capacity $C = \max I(X; Y) = 1$ bit, which is achieved by using $p(x) = (\frac{1}{2}, \frac{1}{2})$.

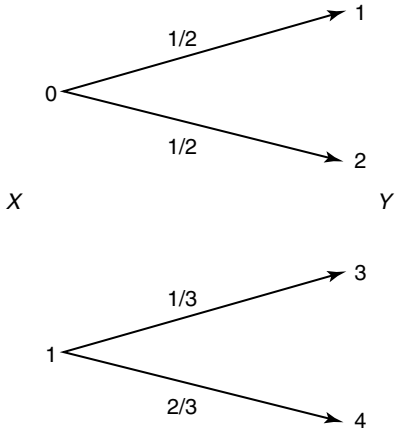


FIGURE 7.3. Noisy channel with nonoverlapping outputs. $C = 1$ bit.

7.1.3 Noisy Typewriter

In this case the channel input is either received unchanged at the output with probability $\frac{1}{2}$ or is transformed into the next letter with probability $\frac{1}{2}$ (Figure 7.4). If the input has 26 symbols and we use every alternate input symbol, we can transmit one of 13 symbols without error with each transmission. Hence, the capacity of this channel is $\log 13$ bits per transmission. We can also calculate the information capacity $C = \max I(X; Y) = \max (H(Y) - H(Y|X)) = \max H(Y) - 1 = \log 26 - 1 = \log 13$, achieved by using $p(x)$ distributed uniformly over all the inputs.

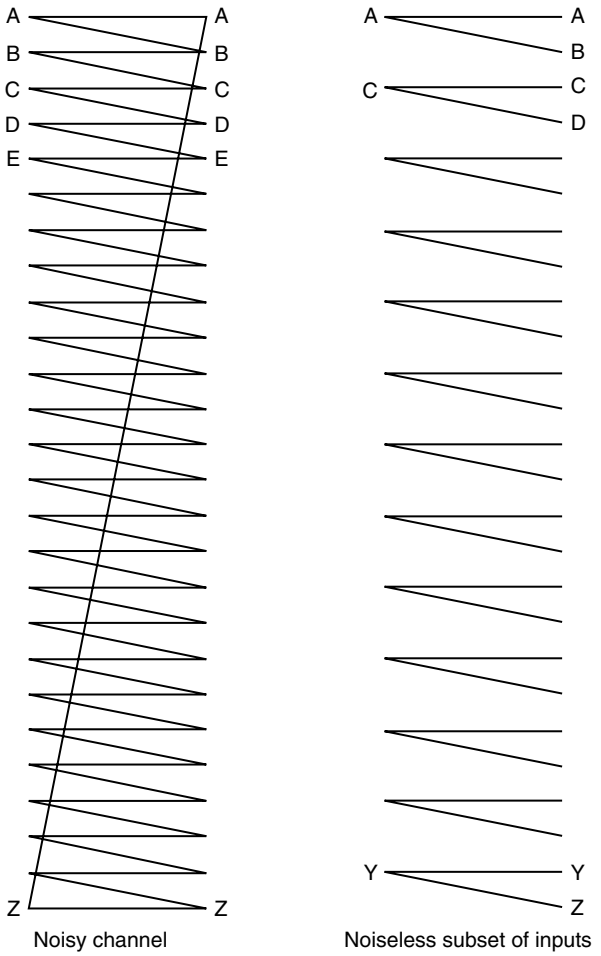


FIGURE 7.4. Noisy Typewriter. $C = \log 13$ bits.

7.1.4 Binary Symmetric Channel

Consider the binary symmetric channel (BSC), which is shown in Fig. 7.5. This is a binary channel in which the input symbols are complemented with probability p. This is the simplest model of a channel with errors, yet it captures most of the complexity of the general problem.

When an error occurs, a 0 is received as a 1, and vice versa. The bits received do not reveal where the errors have occurred. In a sense, all the bits received are unreliable. Later we show that we can still use such a communication channel to send information at a nonzero rate with an arbitrarily small probability of error.

We bound the mutual information by

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) && (7.2)\\ &= H(Y) - \sum p(x)H(Y|X=x) && (7.3)\\ &= H(Y) - \sum p(x)H(p) && (7.4)\\ &= H(Y) - H(p) && (7.5)\\ &\leq 1 - H(p), && (7.6) \end{aligned}$$

where the last inequality follows because Y is a binary random variable. Equality is achieved when the input distribution is uniform. Hence, the information capacity of a binary symmetric channel with parameter p is

$$C = 1 - H(p) \qquad \text{bits.} \tag{7.7}$$

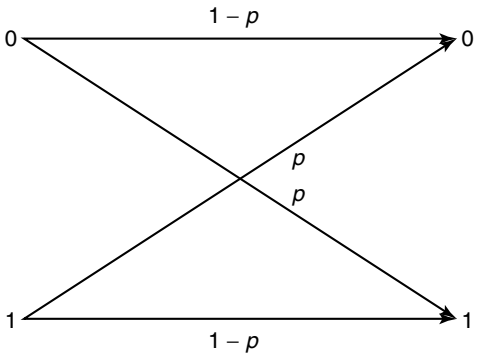


FIGURE 7.5. Binary symmetric channel. $C = 1 - H(p)$ bits.

7.1.5 Binary Erasure Channel

The analog of the binary symmetric channel in which some bits are lost (rather than corrupted) is the *binary erasure channel*. In this channel, a fraction α of the bits are erased. The receiver knows which bits have been erased. The binary erasure channel has two inputs and three outputs (Figure 7.6).

We calculate the capacity of the binary erasure channel as follows:

$$C = \max_{p(x)} I(X; Y) \tag{7.8}$$

$$= \max_{p(x)} (H(Y) - H(Y|X)) \tag{7.9}$$

$$= \max_{p(x)} H(Y) - H(\alpha). \tag{7.10}$$

The first guess for the maximum of $H(Y)$ would be $\log 3$, but we cannot achieve this by any choice of input distribution $p(x)$. Letting E be the event $\{Y = e\}$, using the expansion

$$H(Y) = H(Y, E) = H(E) + H(Y|E), \tag{7.11}$$

and letting $\Pr(X = 1) = \pi$, we have

$$H(Y) = H((1 - \pi)(1 - \alpha), \alpha, \pi(1 - \alpha)) = H(\alpha) + (1 - \alpha)H(\pi). \tag{7.12}$$

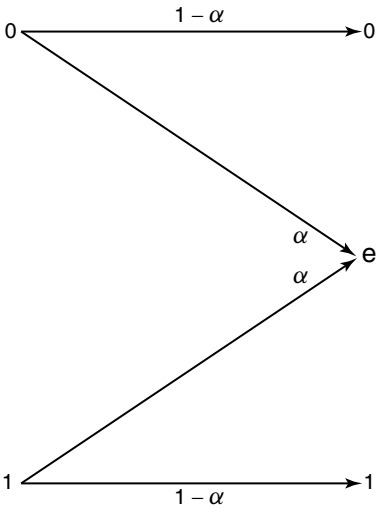


FIGURE 7.6. Binary erasure channel.

Hence

$$C = \max_{p(x)} H(Y) - H(\alpha) \tag{7.13}$$

$$= \max_{\pi} (1-\alpha)H(\pi) + H(\alpha) - H(\alpha) \tag{7.14}$$

$$= \max_{\pi} (1-\alpha)H(\pi) \tag{7.15}$$

$$= 1 - \alpha, \tag{7.16}$$

where capacity is achieved by $\pi = \frac{1}{2}$.

The expression for the capacity has some intuitive meaning: Since a proportion α of the bits are lost in the channel, we can recover (at most) a proportion $1 - \alpha$ of the bits. Hence the capacity is at most $1 - \alpha$. It is not immediately obvious that it is possible to achieve this rate. This will follow from Shannon's second theorem.

In many practical channels, the sender receives some feedback from the receiver. If feedback is available for the binary erasure channel, it is very clear what to do: If a bit is lost, retransmit it until it gets through. Since the bits get through with probability $1 - \alpha$, the effective rate of transmission is $1 - \alpha$. In this way we are easily able to achieve a capacity of $1 - \alpha$ with feedback.

Later in the chapter we prove that the rate $1 - \alpha$ is the best that can be achieved both with and without feedback. This is one of the consequences of the surprising fact that feedback does not increase the capacity of discrete memoryless channels.

7.2 SYMMETRIC CHANNELS

The capacity of the binary symmetric channel is $C = 1 - H(p)$ bits per transmission, and the capacity of the binary erasure channel is $C = 1 - \alpha$ bits per transmission. Now consider the channel with transition matrix:

$$p(y|x) = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.5 & 0.3 & 0.2 \\ 0.2 & 0.5 & 0.3 \end{bmatrix}. \tag{7.17}$$

Here the entry in the xth row and the yth column denotes the conditional probability $p(y|x)$ that y is received when x is sent. In this channel, all the rows of the probability transition matrix are permutations of each other and so are the columns. Such a channel is said to be *symmetric*. Another example of a symmetric channel is one of the form

$$Y = X + Z \pmod{c}, \tag{7.18}$$

where Z has some distribution on the integers $\{0, 1, 2, \ldots, c-1\}$, X has the same alphabet as Z, and Z is independent of X.

In both these cases, we can easily find an explicit expression for the capacity of the channel. Letting $\mathbf{r}$ be a row of the transition matrix, we have

$$I(X;Y) = H(Y) - H(Y|X) \tag{7.19}$$

$$= H(Y) - H(\mathbf{r}) \tag{7.20}$$

$$\leq \log |\mathcal{Y}| - H(\mathbf{r}) \tag{7.21}$$

with equality if the output distribution is uniform. But $p(x) = 1/|\mathcal{X}|$ achieves a uniform distribution on Y, as seen from

$$p(y) = \sum_{x \in \mathcal{X}} p(y|x)p(x) = \frac{1}{|\mathcal{X}|} \sum p(y|x) = c\frac{1}{|\mathcal{X}|} = \frac{1}{|\mathcal{Y}|}, \tag{7.22}$$

where c is the sum of the entries in one column of the probability transition matrix.

Thus, the channel in (7.17) has the capacity

$$C = \max_{p(x)} I(X;Y) = \log 3 - H(0.5, 0.3, 0.2), \tag{7.23}$$

and C is achieved by a uniform distribution on the input.

The transition matrix of the symmetric channel defined above is doubly stochastic. In the computation of the capacity, we used the facts that the rows were permutations of one another and that all the column sums were equal.

Considering these properties, we can define a generalization of the concept of a symmetric channel as follows:

Definition A channel is said to be *symmetric* if the rows of the channel transition matrix $p(y|x)$ are permutations of each other and the columns are permutations of each other. A channel is said to be *weakly symmetric* if every row of the transition matrix $p(\cdot|x)$ is a permutation of every other row and all the column sums $\sum_x p(y|x)$ are equal.

For example, the channel with transition matrix

$$p(y|x) = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{6} \end{pmatrix} \tag{7.24}$$

is weakly symmetric but not symmetric.

The above derivation for symmetric channels carries over to weakly symmetric channels as well. We have the following theorem for weakly symmetric channels:

Theorem 7.2.1 *For a weakly symmetric channel,*

$$C = \log |\mathcal{Y}| - H(\textit{row of transition matrix}), \tag{7.25}$$

and this is achieved by a uniform distribution on the input alphabet.

7.3 PROPERTIES OF CHANNEL CAPACITY

1. $C \geq 0$ since $I(X; Y) \geq 0$.
2. $C \leq \log |\mathcal{X}|$ since $C = \max I(X; Y) \leq \max H(X) = \log |\mathcal{X}|$.
3. $C \leq \log |\mathcal{Y}|$ for the same reason.
4. $I(X; Y)$ is a continuous function of $p(x)$.
5. $I(X; Y)$ is a concave function of $p(x)$ (Theorem 2.7.4). Since $I(X; Y)$ is a concave function over a closed convex set, a local maximum is a global maximum. From properties 2 and 3, the maximum is finite, and we are justified in using the term *maximum* rather than *supremum* in the definition of capacity. The maximum can then be found by standard nonlinear optimization techniques such as gradient search. Some of the methods that can be used include the following:

    - Constrained maximization using calculus and the Kuhn–Tucker conditions.
    - The Frank–Wolfe gradient search algorithm.
    - An iterative algorithm developed by Arimoto [25] and Blahut [65]. We describe the algorithm in Section 10.8.

In general, there is no closed-form solution for the capacity. But for many simple channels it is possible to calculate the capacity using properties such as symmetry. Some of the examples considered earlier are of this form.

7.4 PREVIEW OF THE CHANNEL CODING THEOREM

So far, we have defined the information capacity of a discrete memoryless channel. In the next section we prove Shannon's second theorem, which

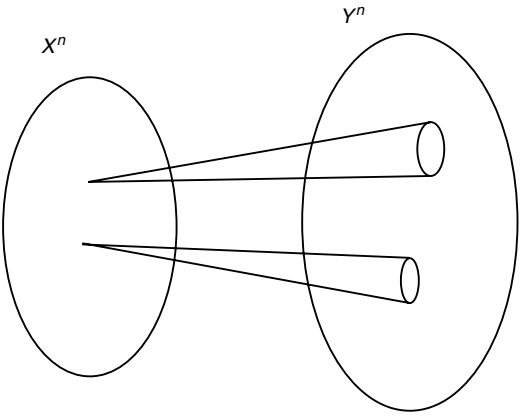


FIGURE 7.7. Channels after n uses.

gives an operational meaning to the definition of capacity as the number of bits we can transmit reliably over the channel. But first we will try to give an intuitive idea as to why we can transmit C bits of information over a channel. The basic idea is that for large block lengths, every channel looks like the noisy typewriter channel (Figure 7.4) and the channel has a subset of inputs that produce essentially disjoint sequences at the output.

For each (typical) input n-sequence, there are approximately $2^{nH(Y|X)}$ possible Y sequences, all of them equally likely (Figure 7.7). We wish to ensure that no two X sequences produce the same Y output sequence. Otherwise, we will not be able to decide which X sequence was sent.

The total number of possible (typical) Y sequences is $\approx 2^{nH(Y)}$. This set has to be divided into sets of size $2^{nH(Y|X)}$ corresponding to the different input X sequences. The total number of disjoint sets is less than or equal to $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$. Hence, we can send at most $\approx 2^{nI(X;Y)}$ distinguishable sequences of length n.

Although the above derivation outlines an upper bound on the capacity, a stronger version of the above argument will be used in the next section to prove that this rate I is achievable with an arbitrarily low probability of error.

Before we proceed to the proof of Shannon's second theorem, we need a few definitions.

7.5 DEFINITIONS

We analyze a communication system as shown in Figure 7.8.

A message W, drawn from the index set $\{1, 2, \ldots, M\}$, results in the signal $X^n(W)$, which is received by the receiver as a random sequence

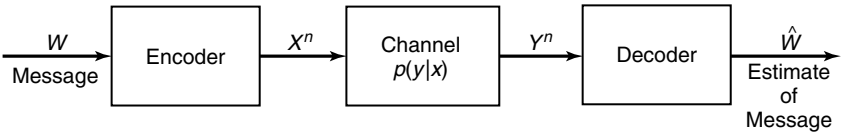


FIGURE 7.8. Communication channel.

$Y^n \sim p(y^n|x^n)$. The receiver then guesses the index W by an appropriate decoding rule $\hat{W} = g(Y^n)$. The receiver makes an error if $\hat{W}$ is not the same as the index W that was transmitted. We now define these ideas formally.

Definition A *discrete channel*, denoted by $(\mathcal{X}, p(y|x), \mathcal{Y})$, consists of two finite sets $\mathcal{X}$ and $\mathcal{Y}$ and a collection of probability mass functions $p(y|x)$, one for each $x \in \mathcal{X}$, such that for every x and y, $p(y|x) \geq 0$, and for every x, $\sum_y p(y|x) = 1$, with the interpretation that X is the input and Y is the output of the channel.

Definition The *nth extension of the discrete memoryless channel* (DMC) is the channel $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$, where

$$p(y_k|x^k, y^{k-1}) = p(y_k|x_k), \qquad k = 1, 2, \ldots, n. \tag{7.26}$$

Remark If the channel is used *without feedback* [i.e., if the input symbols do not depend on the past output symbols, namely, $p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$], the channel transition function for the nth extension of the discrete memoryless channel reduces to

$$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i). \tag{7.27}$$

When we refer to the discrete memoryless channel, we mean the discrete memoryless channel without feedback unless we state explicitly otherwise.

Definition An (M, n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of the following:

1. An index set $\{1, 2, \ldots, M\}$.
2. An encoding function $X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2), \ldots, x^n(M)$. The set of codewords is called the *codebook*.

3. A decoding function

$$g: \mathcal{Y}^n \to \{1, 2, \ldots, M\}, \tag{7.28}$$

which is a deterministic rule that assigns a guess to each possible received vector.

Definition (*Conditional probability of error*) Let

$$\lambda_i = \Pr(g(Y^n) \neq i | X^n = x^n(i)) = \sum_{y^n} p(y^n|x^n(i)) I(g(y^n) \neq i) \tag{7.29}$$

be the *conditional probability of error* given that index i was sent, where $I(\cdot)$ is the indicator function.

Definition The *maximal probability of error* $\lambda^{(n)}$ for an (M, n) code is defined as

$$\lambda^{(n)} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i. \tag{7.30}$$

Definition The (*arithmetic*) *average probability of error* $P_e^{(n)}$ for an (M, n) code is defined as

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i. \tag{7.31}$$

Note that if the index W is chosen according to a uniform distribution over the set $\{1, 2, \ldots, M\}$, and $X^n = x^n(W)$, then

$$P_e^{(n)} \triangleq \Pr(W \neq g(Y^n)), \tag{7.32}$$

(i.e., $P_e^{(n)}$ is the probability of error). Also, obviously,

$$P_e^{(n)} \leq \lambda^{(n)}. \tag{7.33}$$

One would expect the maximal probability of error to behave quite differently from the average probability. But in the next section we prove that a small average probability of error implies a small maximal probability of error at essentially the same rate.

It is worth noting that $P_e^{(n)}$ defined in (7.32) is only a mathematical construct of the conditional probabilities of error λ_i and is itself a probability of error only if the message is chosen uniformly over the message set $\{1, 2, \ldots, 2^M\}$. However, both in the proof of achievability and the converse, we choose a uniform distribution on W to bound the probability of error. This allows us to establish the behavior of $P_e^{(n)}$ and the maximal probability of error $\lambda^{(n)}$ and thus characterize the behavior of the channel regardless of how it is used (i.e., no matter what the distribution of W).

Definition The *rate* R of an (M, n) code is

$$R = \frac{\log M}{n} \qquad \text{bits per transmission.} \tag{7.34}$$

Definition A rate R is said to be *achievable* if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the maximal probability of error $\lambda^{(n)}$ tends to 0 as $n \to \infty$.

Later, we write $(2^{nR}, n)$ codes to mean $(\lceil 2^{nR} \rceil, n)$ codes. This will simplify the notation.

Definition The *capacity* of a channel is the supremum of all achievable rates.

Thus, rates less than capacity yield arbitrarily small probability of error for sufficiently large block lengths.

7.6 JOINTLY TYPICAL SEQUENCES

Roughly speaking, we decode a channel output Y^n as the ith index if the codeword $X^n(i)$ is "jointly typical" with the received signal Y^n. We now define the important idea of joint typicality and find the probability of joint typicality when $X^n(i)$ is the true cause of Y^n and when it is not.

Definition The set $A_\epsilon^{(n)}$ of *jointly typical* sequences $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is the set of n-sequences with empirical entropies ϵ-close to the true entropies:

$$A_\epsilon^{(n)} = \Big\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \tag{7.35}$$

$$\left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \tag{7.36}$$

$$\left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon \Bigg\}, \tag{7.37}$$

where

$$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i). \tag{7.38}$$

Theorem 7.6.1 (*Joint AEP*) *Let (X^n, Y^n) be sequences of length n drawn i.i.d. according to $p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$. Then:*

1. $\Pr((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ *as* $n \to \infty$.
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.
3. *If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$ [i.e., $\tilde{X}^n$ and $\tilde{Y}^n$ are independent with the same marginals as $p(x^n, y^n)$], then*

$$\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \leq 2^{-n(I(X;Y)-3\epsilon)}. \tag{7.39}$$

Also, for sufficiently large n,

$$\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}. \tag{7.40}$$

Proof

1. We begin by showing that with high probability, the sequence is in the typical set. By the weak law of large numbers,

$$-\frac{1}{n} \log p(X^n) \to -E[\log p(X)] = H(X) \quad \text{in probability.} \tag{7.41}$$

Hence, given $\epsilon > 0$, there exists n_1, such that for all $n > n_1$,

$$\Pr\left(\left| -\frac{1}{n} \log p(X^n) - H(X) \right| \geq \epsilon\right) < \frac{\epsilon}{3}. \tag{7.42}$$

Similarly, by the weak law,

$$-\frac{1}{n} \log p(Y^n) \to -E[\log p(Y)] = H(Y) \quad \text{in probability} \tag{7.43}$$

and

$$-\frac{1}{n}\log p(X^n, Y^n) \to -E[\log p(X, Y)] = H(X, Y) \quad \text{in probability,} \tag{7.44}$$

and there exist n_2 and n_3, such that for all $n \geq n_2$,

$$\Pr\left(\left|-\frac{1}{n}\log p(Y^n) - H(Y)\right| \geq \epsilon\right) < \frac{\epsilon}{3} \tag{7.45}$$

and for all $n \geq n_3$,

$$\Pr\left(\left|-\frac{1}{n}\log p(X^n, Y^n) - H(X, Y)\right| \geq \epsilon\right) < \frac{\epsilon}{3}. \tag{7.46}$$

Choosing $n > \max\{n_1, n_2, n_3\}$, the probability of the union of the sets in (7.42), (7.45), and (7.46) must be less than ϵ. Hence for n sufficiently large, the probability of the set $A_\epsilon^{(n)}$ is greater than $1 - \epsilon$, establishing the first part of the theorem.

2. To prove the second part of the theorem, we have

$$1 = \sum p(x^n, y^n) \tag{7.47}$$

$$\geq \sum_{A_\epsilon^{(n)}} p(x^n, y^n) \tag{7.48}$$

$$\geq |A_\epsilon^{(n)}| 2^{-n(H(X,Y)+\epsilon)}, \tag{7.49}$$

and hence

$$|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}. \tag{7.50}$$

3. Now if $\tilde{X}^n$ and $\tilde{Y}^n$ are independent but have the same marginals as X^n and Y^n, then

$$\Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) = \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n) p(y^n) \tag{7.51}$$

$$\leq 2^{n(H(X,Y)+\epsilon)} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} \tag{7.52}$$

$$= 2^{-n(I(X;Y)-3\epsilon)}. \tag{7.53}$$

For sufficiently large n, $\Pr(A_\epsilon^{(n)}) \geq 1 - \epsilon$, and therefore

$$1 - \epsilon \leq \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n, y^n) \tag{7.54}$$

$$\leq |A_\epsilon^{(n)}| 2^{-n(H(X,Y)-\epsilon)} \tag{7.55}$$

and

$$|A_\epsilon^{(n)}| \geq (1 - \epsilon) 2^{n(H(X,Y)-\epsilon)}. \tag{7.56}$$

By similar arguments to the upper bound above, we can also show that for n sufficiently large,

$$\Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) = \sum_{A_\epsilon^{(n)}} p(x^n) p(y^n) \tag{7.57}$$

$$\geq (1 - \epsilon) 2^{n(H(X,Y)-\epsilon)} 2^{-n(H(X)+\epsilon)} 2^{-n(H(Y)+\epsilon)} \tag{7.58}$$

$$= (1 - \epsilon) 2^{-n(I(X;Y)+3\epsilon)}. \qquad \square \tag{7.59}$$

The jointly typical set is illustrated in Figure 7.9. There are about $2^{nH(X)}$ typical X sequences and about $2^{nH(Y)}$ typical Y sequences. However, since there are only $2^{nH(X,Y)}$ jointly typical sequences, not all pairs of typical X^n and typical Y^n are also jointly typical. The probability that

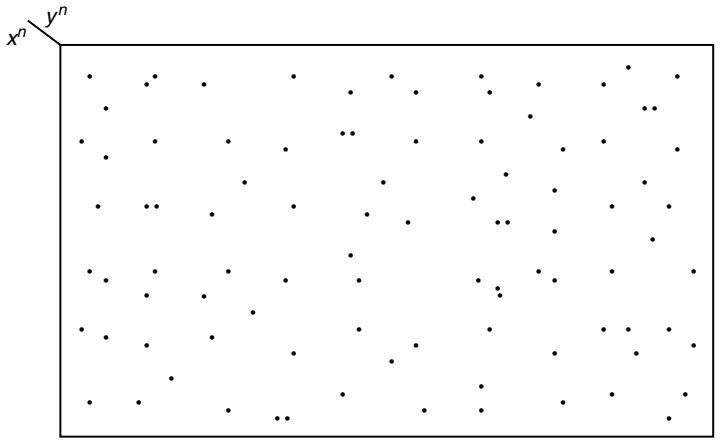


FIGURE 7.9. Jointly typical sequences.

any randomly chosen pair is jointly typical is about $2^{-nI(X;Y)}$. Hence, we can consider about $2^{nI(X;Y)}$ such pairs before we are likely to come across a jointly typical pair. This suggests that there are about $2^{nI(X;Y)}$ distinguishable signals X^n.

Another way to look at this is in terms of the set of jointly typical sequences for a fixed output sequence Y^n, presumably the output sequence resulting from the true input signal X^n. For this sequence Y^n, there are about $2^{nH(X|Y)}$ conditionally typical input signals. The probability that some randomly chosen (other) input signal X^n is jointly typical with Y^n is about $2^{nH(X|Y)}/2^{nH(X)} = 2^{-nI(X;Y)}$. This again suggests that we can choose about $2^{nI(X;Y)}$ codewords $X^n(W)$ before one of these codewords will get confused with the codeword that caused the output Y^n.

7.7 CHANNEL CODING THEOREM

We now prove what is perhaps the basic theorem of information theory, the achievability of channel capacity, first stated and essentially proved by Shannon in his original 1948 paper. The result is rather counterintuitive; if the channel introduces errors, how can one correct them all? Any correction process is also subject to error, ad infinitum.

Shannon used a number of new ideas to prove that information can be sent reliably over a channel at all rates up to the channel capacity. These ideas include:

- Allowing an arbitrarily small but nonzero probability of error
- Using the channel many times in succession, so that the law of large numbers comes into effect
- Calculating the average of the probability of error over a random choice of codebooks, which symmetrizes the probability, and which can then be used to show the existence of at least one good code

Shannon's outline of the proof was based on the idea of typical sequences, but the proof was not made rigorous until much later. The proof given below makes use of the properties of typical sequences and is probably the simplest of the proofs developed so far. As in all the proofs, we use the same essential ideas—random code selection, calculation of the average probability of error for a random choice of codewords, and so on. The main difference is in the decoding rule. In the proof, we decode by joint typicality; we look for a codeword that is jointly typical with the received sequence. If we find a unique codeword satisfying this property, we declare that word to be the transmitted codeword. By the properties

of joint typicality stated previously, with high probability the transmitted codeword and the received sequence are jointly typical, since they are probabilistically related. Also, the probability that any other codeword looks jointly typical with the received sequence is 2^{-nI}. Hence, if we have fewer then 2^{nI} codewords, then with high probability there will be no other codewords that can be confused with the transmitted codeword, and the probability of error is small.

Although jointly typical decoding is suboptimal, it is simple to analyze and still achieves all rates below capacity.

We now give the complete statement and proof of Shannon's second theorem:

Theorem 7.7.1 (*Channel coding theorem*) *For a discrete memoryless channel, all rates below capacity C are achievable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$.*

Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

Proof: We prove that rates $R < C$ are achievable and postpone proof of the converse to Section 7.9.

Achievability: Fix $p(x)$. Generate a $(2^{nR}, n)$ code at random according to the distribution $p(x)$. Specifically, we generate 2^{nR} codewords independently according to the distribution

$$p(x^n) = \prod_{i=1}^{n} p(x_i). \tag{7.60}$$

We exhibit the 2^{nR} codewords as the rows of a matrix:

$$\mathcal{C} = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{bmatrix}. \tag{7.61}$$

Each entry in this matrix is generated i.i.d. according to $p(x)$. Thus, the probability that we generate a particular code $\mathcal{C}$ is

$$\Pr(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(w)). \tag{7.62}$$

Consider the following sequence of events:

1. A random code $\mathcal{C}$ is generated as described in (7.62) according to $p(x)$.
2. The code $\mathcal{C}$ is then revealed to both sender and receiver. Both sender and receiver are also assumed to know the channel transition matrix $p(y|x)$ for the channel.
3. A message W is chosen according to a uniform distribution

$$\Pr(W = w) = 2^{-nR}, \qquad w = 1, 2, \ldots, 2^{nR}. \tag{7.63}$$

4. The wth codeword $X^n(w)$, corresponding to the wth row of $\mathcal{C}$, is sent over the channel.
5. The receiver receives a sequence Y^n according to the distribution

$$P(y^n|x^n(w)) = \prod_{i=1}^{n} p(y_i|x_i(w)). \tag{7.64}$$

6. The receiver guesses which message was sent. (The optimum procedure to minimize probability of error is maximum likelihood decoding (i.e., the receiver should choose the *a posteriori* most likely message). But this procedure is difficult to analyze. Instead, we will use *jointly typical decoding*, which is described below. Jointly typical decoding is easier to analyze and is asymptotically optimal.) In jointly typical decoding, the receiver declares that the index $\hat{W}$ was sent if the following conditions are satisfied:

   - $(X^n(\hat{W}), Y^n)$ is jointly typical.
   - There is no other index $W' \neq \hat{W}$ such that $(X^n(W'), Y^n) \in A_\epsilon^{(n)}$.

   If no such $\hat{W}$ exists or if there is more than one such, an error is declared. (We may assume that the receiver outputs a dummy index such as 0 in this case.)
7. There is a decoding error if $\hat{W} \neq W$. Let $\mathcal{E}$ be the event $\{\hat{W} \neq W\}$.

Analysis of the probability of error

Outline: We first outline the analysis. Instead of calculating the probability of error for a single code, we calculate the average over all codes generated at random according to the distribution (7.62). By the symmetry of the code construction, the average probability of error does not depend

on the particular index that was sent. For a typical codeword, there are two different sources of error when we use jointly typical decoding: Either the output Y^n is not jointly typical with the transmitted codeword or there is some other codeword that is jointly typical with Y^n. The probability that the transmitted codeword and the received sequence are jointly typical goes to 1, as shown by the joint AEP. For any rival codeword, the probability that it is jointly typical with the received sequence is approximately 2^{-nI}, and hence we can use about 2^{nI} codewords and still have a low probability of error. We will later extend the argument to find a code with a low maximal probability of error.

Detailed calculation of the probability of error: We let W be drawn according to a uniform distribution over $\{1, 2, \ldots, 2^{nR}\}$ and use jointly typical decoding $\hat{W}(y^n)$ as described in step 6. Let $\mathcal{E} = \{\hat{W}(Y^n) \neq W\}$ denote the error event. We will calculate the average probability of error, averaged over all codewords in the codebook, and averaged over all codebooks; that is, we calculate

$$\Pr(\mathcal{E}) = \sum_{\mathcal{C}} \Pr(\mathcal{C}) P_e^{(n)}(\mathcal{C}) \tag{7.65}$$

$$= \sum_{\mathcal{C}} \Pr(\mathcal{C}) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(\mathcal{C}) \tag{7.66}$$

$$= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_w(\mathcal{C}), \tag{7.67}$$

where $P_e^{(n)}(\mathcal{C})$ is defined for jointly typical decoding. By the symmetry of the code construction, the average probability of error averaged over all codes does not depend on the particular index that was sent [i.e., $\sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_w(\mathcal{C})$ does not depend on w]. Thus, we can assume without loss of generality that the message $W = 1$ was sent, since

$$\Pr(\mathcal{E}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_w(\mathcal{C}) \tag{7.68}$$

$$= \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_1(\mathcal{C}) \tag{7.69}$$

$$= \Pr(\mathcal{E} | W = 1). \tag{7.70}$$

Define the following events:

$$E_i = \{ (X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)} \}, \quad i \in \{1, 2, \ldots, 2^{nR}\}, \tag{7.71}$$

where E_i is the event that the ith codeword and Y^n are jointly typical. Recall that Y^n is the result of sending the first codeword $X^n(1)$ over the channel.

Then an error occurs in the decoding scheme if either E_1^c occurs (when the transmitted codeword and the received sequence are not jointly typical) or $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$ occurs (when a wrong codeword is jointly typical with the received sequence). Hence, letting $P(\mathcal{E})$ denote $\Pr(\mathcal{E}|W = 1)$, we have

$$\Pr(\mathcal{E}|W = 1) = P\left(E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}|W = 1\right) \tag{7.72}$$

$$\leq P(E_1^c|W = 1) + \sum_{i=2}^{2^{nR}} P(E_i|W = 1), \tag{7.73}$$

by the union of events bound for probabilities. Now, by the joint AEP, $P(E_1^c|W = 1) \to 0$, and hence

$$P(E_1^c|W = 1) \leq \epsilon \qquad \text{for } n \text{ sufficiently large.} \tag{7.74}$$

Since by the code generation process, $X^n(1)$ and $X^n(i)$ are independent for $i \neq 1$, so are Y^n and $X^n(i)$. Hence, the probability that $X^n(i)$ and Y^n are jointly typical is $\leq 2^{-n(I(X;Y)-3\epsilon)}$ by the joint AEP. Consequently,

$$\Pr(\mathcal{E}) = \Pr(\mathcal{E}|W = 1) \leq P(E_1^c|W = 1) + \sum_{i=2}^{2^{nR}} P(E_i|W = 1) \tag{7.75}$$

$$\leq \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \tag{7.76}$$

$$= \epsilon + \left(2^{nR} - 1\right) 2^{-n(I(X;Y)-3\epsilon)} \tag{7.77}$$

$$\leq \epsilon + 2^{3n\epsilon} 2^{-n(I(X;Y)-R)} \tag{7.78}$$

$$\leq 2\epsilon \tag{7.79}$$

if n is sufficiently large and $R < I(X; Y) - 3\epsilon$. Hence, if $R < I(X; Y)$, we can choose ϵ and n so that the average probability of error, averaged over codebooks and codewords, is less than 2ϵ.

To finish the proof, we will strengthen this conclusion by a series of code selections.

1. Choose $p(x)$ in the proof to be $p^*(x)$, the distribution on X that achieves capacity. Then the condition $R < I(X; Y)$ can be replaced by the achievability condition $R < C$.

2. Get rid of the average over codebooks. Since the average probability of error over codebooks is small ($\leq 2\epsilon$), there exists at least one codebook $\mathcal{C}^*$ with a small average probability of error. Thus, $\Pr(\mathcal{E}|\mathcal{C}^*) \leq 2\epsilon$. Determination of $\mathcal{C}^*$ can be achieved by an exhaustive search over all $(2^{nR}, n)$ codes. Note that

$$\Pr(\mathcal{E}|\mathcal{C}^*) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}^*), \tag{7.80}$$

since we have chosen $\hat{W}$ according to a uniform distribution as specified in (7.63).

3. Throw away the worst half of the codewords in the best codebook $\mathcal{C}^*$. Since the arithmetic average probability of error $P_e^{(n)}(\mathcal{C}^*)$ for this code is less then 2ϵ, we have

$$\Pr(\mathcal{E}|\mathcal{C}^*) \leq \frac{1}{2^{nR}} \sum \lambda_i(\mathcal{C}^*) \leq 2\epsilon, \tag{7.81}$$

which implies that at least half the indices i and their associated codewords $X^n(i)$ must have conditional probability of error λ_i less than 4ϵ (otherwise, these codewords themselves would contribute more than 2ϵ to the sum). Hence the best half of the codewords have a maximal probability of error less than 4ϵ. If we reindex these codewords, we have 2^{nR-1} codewords. Throwing out half the codewords has changed the rate from R to $R - \frac{1}{n}$, which is negligible for large n.

Combining all these improvements, we have constructed a code of rate $R' = R - \frac{1}{n}$, with maximal probability of error $\lambda^{(n)} \leq 4\epsilon$. This proves the achievability of any rate below capacity. □

Random coding is the method of proof for Theorem 7.7.1, not the method of signaling. Codes are selected at random in the proof merely to symmetrize the mathematics and to show the existence of a good deterministic code. We proved that the average over all codes of block length n has a small probability of error. We can find the best code within this set by an exhaustive search. Incidentally, this shows that the Kolmogorov complexity (Chapter 14) of the best code is a small constant. This means that the revelation (in step 2) to the sender and receiver of the best code $\mathcal{C}^*$ requires no channel. The sender and receiver merely agree to use the best $(2^{nR}, n)$ code for the channel.

Although the theorem shows that there exist good codes with arbitrarily small probability of error for long block lengths, it does not provide a way of constructing the best codes. If we used the scheme suggested

by the proof and generate a code at random with the appropriate distribution, the code constructed is likely to be good for long block lengths. However, without some structure in the code, it is very difficult to decode (the simple scheme of table lookup requires an exponentially large table). Hence the theorem does not provide a practical coding scheme. Ever since Shannon's original paper on information theory, researchers have tried to develop structured codes that are easy to encode and decode. In Section 7.11, we discuss Hamming codes, the simplest of a class of algebraic error correcting codes that can correct one error in a block of bits. Since Shannon's paper, a variety of techniques have been used to construct error correcting codes, and with turbo codes have come close to achieving capacity for Gaussian channels.

7.8 ZERO-ERROR CODES

The outline of the proof of the converse is most clearly motivated by going through the argument when absolutely no errors are allowed. We will now prove that $P_e^{(n)} = 0$ implies that $R \leq C$. Assume that we have a $(2^{nR}, n)$ code with zero probability of error [i.e., the decoder output $g(Y^n)$ is equal to the input index W with probability 1]. Then the input index W is determined by the output sequence [i.e., $H(W|Y^n) = 0$]. Now, to obtain a strong bound, we arbitrarily assume that W is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$. Thus, $H(W) = nR$. We can now write the string of inequalities:

$$nR = H(W) = \underbrace{H(W|Y^n)}_{=0} + I(W; Y^n) \tag{7.82}$$

$$= I(W; Y^n) \tag{7.83}$$

$$\overset{(a)}{\leq} I(X^n; Y^n) \tag{7.84}$$

$$\overset{(b)}{\leq} \sum_{i=1}^{n} I(X_i; Y_i) \tag{7.85}$$

$$\overset{(c)}{\leq} nC, \tag{7.86}$$

where (a) follows from the data-processing inequality (since $W \to X^n(W) \to Y^n$ forms a Markov chain), (b) will be proved in Lemma 7.9.2 using the discrete memoryless assumption, and (c) follows from the definition of (information) capacity. Hence, for any zero-error $(2^{nR}, n)$ code, for all n,

$$R \leq C. \tag{7.87}$$

7.9 FANO'S INEQUALITY AND THE CONVERSE TO THE CODING THEOREM

We now extend the proof that was derived for zero-error codes to the case of codes with very small probabilities of error. The new ingredient will be Fano's inequality, which gives a lower bound on the probability of error in terms of the conditional entropy. Recall the proof of Fano's inequality, which is repeated here in a new context for reference.

Let us define the setup under consideration. The index W is uniformly distributed on the set $\mathcal{W} = \{1, 2, \ldots, 2^{nR}\}$, and the sequence Y^n is related probabilistically to W. From Y^n, we estimate the index W that was sent. Let the estimate be $\hat{W} = g(Y^n)$. Thus, $W \to X^n(W) \to Y^n \to \hat{W}$ forms a Markov chain. Note that the probability of error is

$$\Pr\left(\hat{W} \neq W\right) = \frac{1}{2^{nR}} \sum_i \lambda_i = P_e^{(n)}. \tag{7.88}$$

We begin with the following lemma, which has been proved in Section 2.10:

Lemma 7.9.1 (*Fano's inequality*) *For a discrete memoryless channel with a codebook $\mathcal{C}$ and the input message W uniformly distributed over 2^{nR}, we have*

$$H(W|\hat{W}) \leq 1 + P_e^{(n)} nR. \tag{7.89}$$

Proof: Since W is uniformly distributed, we have $P_e^{(n)} = \Pr(W \neq \hat{W})$. We apply Fano's inequality (Theorem 2.10.1) for W in an alphabet of size 2^{nR}. □

We will now prove a lemma which shows that the capacity per transmission is not increased if we use a discrete memoryless channel many times.

Lemma 7.9.2 *Let Y^n be the result of passing X^n through a discrete memoryless channel of capacity C. Then*

$$I(X^n; Y^n) \leq nC \qquad \text{for all } p(x^n). \tag{7.90}$$

Proof

$$I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n) \tag{7.91}$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1}, X^n) \tag{7.92}$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i), \tag{7.93}$$

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

$$I(X^n; Y^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \tag{7.94}$$

$$\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \tag{7.95}$$

$$= \sum_{i=1}^{n} I(X_i; Y_i) \tag{7.96}$$

$$\leq nC, \tag{7.97}$$

where (7.95) follows from the fact that the entropy of a collection of random variables is less than the sum of their individual entropies, and (7.97) follows from the definition of capacity. Thus, we have proved that using the channel many times does not increase the information capacity in bits per transmission. □

We are now in a position to prove the converse to the channel coding theorem.

Proof: *Converse to Theorem 7.7.1* (*Channel coding theorem*). We have to show that any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$. If the maximal probability of error tends to zero, the average probability of error for the sequence of codes also goes to zero [i.e., $\lambda^{(n)} \to 0$ implies $P_e^{(n)} \to 0$, where $P_e^{(n)}$ is defined in (7.32)]. For a fixed encoding rule $X^n(\cdot)$ and a fixed decoding rule $\hat{W} = g(Y^n)$, we have $W \to X^n(W) \to Y^n \to \hat{W}$. For each n, let W be drawn according to a uniform distribution over $\{1, 2, \ldots, 2^{nR}\}$. Since W has a uniform distribution, $\Pr(\hat{W} \neq W) = P_e^{(n)} = \frac{1}{2^{nR}} \sum_i \lambda_i$. Hence,

$$nR \overset{(a)}{=} H(W) \tag{7.98}$$

$$\overset{(b)}{=} H(W|\hat{W}) + I(W; \hat{W}) \tag{7.99}$$

$$\overset{(c)}{\leq} 1 + P_e^{(n)} nR + I(W; \hat{W}) \tag{7.100}$$

$$\overset{(d)}{\leq} 1 + P_e^{(n)} nR + I(X^n; Y^n) \tag{7.101}$$

$$\overset{(e)}{\leq} 1 + P_e^{(n)} nR + nC, \tag{7.102}$$

where (a) follows from the assumption that W is uniform over $\{1, 2, \ldots, 2^{nR}\}$, (b) is an identity, (c) is Fano's inequality for W taking on at most 2^{nR} values, (d) is the data-processing inequality, and (e) is from Lemma 7.9.2. Dividing by n, we obtain

$$R \leq P_e^{(n)} R + \frac{1}{n} + C. \tag{7.103}$$

Now letting $n \to \infty$, we see that the first two terms on the right-hand side tend to 0, and hence

$$R \leq C. \tag{7.104}$$

We can rewrite (7.103) as

$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}. \tag{7.105}$$

This equation shows that if $R > C$, the probability of error is bounded away from 0 for sufficiently large n (and hence for all n, since if $P_e^{(n)} = 0$ for small n, we can construct codes for large n with $P_e^{(n)} = 0$ by concatenating these codes). Hence, we cannot achieve an arbitrarily low probability of error at rates above capacity. □

This converse is sometimes called the *weak converse* to the channel coding theorem. It is also possible to prove a *strong converse*, which states that for rates above capacity, the probability of error goes exponentially to 1. Hence, the capacity is a very clear dividing point—at rates below capacity, $P_e^{(n)} \to 0$ exponentially, and at rates above capacity, $P_e^{(n)} \to 1$ exponentially.

7.10 EQUALITY IN THE CONVERSE TO THE CHANNEL CODING THEOREM

We have proved the channel coding theorem and its converse. In essence, these theorems state that when $R < C$, it is possible to send information with an arbitrarily low probability of error, and when $R > C$, the probability of error is bounded away from zero.

It is interesting and rewarding to examine the consequences of equality in the converse; hopefully, it will give some ideas as to the kinds of codes that achieve capacity. Repeating the steps of the converse in the case when $P_e = 0$, we have

$$nR = H(W) \tag{7.106}$$

$$= H(W|\hat{W}) + I(W; \hat{W}) \tag{7.107}$$

$$= I(W; \hat{W})) \tag{7.108}$$

$$\overset{(a)}{\leq} I(X^n(W); Y^n) \tag{7.109}$$

$$= H(Y^n) - H(Y^n|X^n) \tag{7.110}$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \tag{7.111}$$

$$\overset{(b)}{\leq} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \tag{7.112}$$

$$= \sum_{i=1}^{n} I(X_i; Y_i) \tag{7.113}$$

$$\overset{(c)}{\leq} nC. \tag{7.114}$$

We have equality in (a), the data-processing inequality, only if $I(Y^n; X^n(W)|W) = 0$ and $I(X^n; Y^n|\hat{W}) = 0$, which is true if all the codewords are distinct and if $\hat{W}$ is a sufficient statistic for decoding. We have equality in (b) only if the Y_i's are independent, and equality in (c) only if the distribution of X_i is $p^*(x)$, the distribution on X that achieves capacity. We have equality in the converse only if these conditions are satisfied. This indicates that a capacity-achieving zero-error code has distinct codewords and the distribution of the Y_i's must be i.i.d. with

$$p^*(y) = \sum_x p^*(x)p(y|x), \tag{7.115}$$

the distribution on Y induced by the optimum distribution on X. The distribution referred to in the converse is the empirical distribution on X and Y induced by a uniform distribution over codewords, that is,

$$p(x_i, y_i) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} I(X_i(w) = x_i)p(y_i|x_i). \tag{7.116}$$

We can check this result in examples of codes that achieve capacity:

1. *Noisy typewriter*. In this case we have an input alphabet of 26 letters, and each letter is either printed out correctly or changed to the next letter with probability $\frac{1}{2}$. A simple code that achieves capacity (log 13) for this channel is to use every alternate input letter so that

no two letters can be confused. In this case, there are 13 codewords of block length 1. If we choose the codewords i.i.d. according to a uniform distribution on $\{1, 3, 5, 7, \ldots, 25\}$, the output of the channel is also i.i.d. and uniformly distributed on $\{1, 2, \ldots, 26\}$, as expected.

2. *Binary symmetric channel*. Since given any input sequence, every possible output sequence has some positive probability, it will not be possible to distinguish even two codewords with zero probability of error. Hence the zero-error capacity of the BSC is zero. However, even in this case, we can draw some useful conclusions. The efficient codes will still induce a distribution on Y that looks i.i.d. $\sim$ Bernoulli($\frac{1}{2}$). Also, from the arguments that lead up to the converse, we can see that at rates close to capacity, we have almost entirely covered the set of possible output sequences with decoding sets corresponding to the codewords. At rates above capacity, the decoding sets begin to overlap, and the probability of error can no longer be made arbitrarily small.

7.11 HAMMING CODES

The channel coding theorem promises the existence of block codes that will allow us to transmit information at rates below capacity with an arbitrarily small probability of error if the block length is large enough. Ever since the appearance of Shannon's original paper [471], people have searched for such codes. In addition to achieving low probabilities of error, useful codes should be "simple," so that they can be encoded and decoded efficiently.

The search for simple good codes has come a long way since the publication of Shannon's original paper in 1948. The entire field of coding theory has been developed during this search. We will not be able to describe the many elegant and intricate coding schemes that have been developed since 1948. We will only describe the simplest such scheme developed by Hamming [266]. It illustrates some of the basic ideas underlying most codes.

The object of coding is to introduce redundancy so that even if some of the information is lost or corrupted, it will still be possible to recover the message at the receiver. The most obvious coding scheme is to repeat information. For example, to send a 1, we send 11111, and to send a 0, we send 00000. This scheme uses five symbols to send 1 bit, and therefore has a *rate* of $\frac{1}{5}$ bit per symbol. If this code is used on a binary symmetric channel, the optimum decoding scheme is to take the majority vote of each block of five received bits. If three or more bits are 1, we decode

the block as a 1; otherwise, we decode it as 0. An error occurs if and only if more than three of the bits are changed. By using longer repetition codes, we can achieve an arbitrarily low probability of error. But the rate of the code also goes to zero with block length, so even though the code is "simple," it is really not a very useful code.

Instead of simply repeating the bits, we can combine the bits in some intelligent fashion so that each extra bit checks whether there is an error in some subset of the information bits. A simple example of this is a parity check code. Starting with a block of $n-1$ information bits, we choose the nth bit so that the parity of the entire block is 0 (the number of 1's in the block is even). Then if there is an odd number of errors during the transmission, the receiver will notice that the parity has changed and detect the error. This is the simplest example of an *error-detecting code*. The code does not detect an even number of errors and does not give any information about how to correct the errors that occur.

We can extend the idea of parity checks to allow for more than one parity check bit and to allow the parity checks to depend on various subsets of the information bits. The Hamming code that we describe below is an example of a parity check code. We describe it using some simple ideas from linear algebra.

To illustrate the principles of Hamming codes, we consider a binary code of block length 7. All operations will be done modulo 2. Consider the set of all nonzero binary vectors of length 3. Arrange them in columns to form a matrix:

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}. \tag{7.117}$$

Consider the set of vectors of length 7 in the null space of H (the vectors which when multiplied by H give 000). From the theory of linear spaces, since H has rank 3, we expect the null space of H to have dimension 4. These 2^4 codewords are

0000000	0100101	1000011	1100110
0001111	0101010	1001100	1101001
0010110	0110011	1010101	1110000
0011001	0111100	1011010	1111111

Since the set of codewords is the null space of a matrix, it is *linear* in the sense that the sum of any two codewords is also a codeword. The set of codewords therefore forms a linear subspace of dimension 4 in the vector space of dimension 7.

Looking at the codewords, we notice that other than the all-0 codeword, the minimum number of 1's in any codeword is 3. This is called the *minimum weight* of the code. We can see that the minimum weight of a code has to be at least 3 since all the columns of H are different, so no two columns can add to 000. The fact that the minimum distance is exactly 3 can be seen from the fact that the sum of any two columns must be one of the columns of the matrix.

Since the code is linear, the difference between any two codewords is also a codeword, and hence any two codewords differ in at least three places. The minimum number of places in which two codewords differ is called the *minimum distance* of the code. The minimum distance of the code is a measure of how far apart the codewords are and will determine how distinguishable the codewords will be at the output of the channel. The minimum distance is equal to the minimum weight for a linear code. We aim to develop codes that have a large minimum distance.

For the code described above, the minimum distance is 3. Hence if a codeword $\mathbf{c}$ is corrupted in only one place, it will differ from any other codeword in at least two places and therefore be closer to $\mathbf{c}$ than to any other codeword. But can we discover which is the closest codeword without searching over all the codewords?

The answer is yes. We can use the structure of the matrix H for decoding. The matrix H, called the *parity check matrix*, has the property that for every codeword $\mathbf{c}$, $H\mathbf{c} = 0$. Let $\mathbf{e}_i$ be a vector with a 1 in the ith position and 0's elsewhere. If the codeword is corrupted at position i, the received vector $\mathbf{r} = \mathbf{c} + \mathbf{e}_i$. If we multiply this vector by the matrix H, we obtain

$$H\mathbf{r} = H(\mathbf{c} + \mathbf{e}_i) = H\mathbf{c} + H\mathbf{e}_i = H\mathbf{e}_i, \tag{7.118}$$

which is the vector corresponding to the ith column of H. Hence looking at $H\mathbf{r}$, we can find which position of the vector was corrupted. Reversing this bit will give us a codeword. This yields a simple procedure for correcting one error in the received sequence. We have constructed a codebook with 16 codewords of block length 7, which can correct up to one error. This code is called a *Hamming code*.

We have not yet identified a simple encoding procedure; we could use any mapping from a set of 16 messages into the codewords. But if we examine the first 4 bits of the codewords in the table, we observe that they cycle through all 2^4 combinations of 4 bits. Thus, we could use these 4 bits to be the 4 bits of the message we want to send; the other 3 bits are then determined by the code. In general, it is possible to modify a linear code so that the mapping is explicit, so that the first k bits in each

codeword represent the message, and the last $n - k$ bits are parity check bits. Such a code is called a *systematic code*. The code is often identified by its block length n, the number of information bits k and the minimum distance d. For example, the above code is called a (7,4,3) Hamming code (i.e., $n = 7$, $k = 4$, and $d = 3$).

An easy way to see how Hamming codes work is by means of a Venn diagram. Consider the following Venn diagram with three circles and with four intersection regions as shown in Figure 7.10. To send the information sequence 1101, we place the 4 information bits in the four intersection regions as shown in the figure. We then place a parity bit in each of the three remaining regions so that the parity of each circle is even (i.e., there are an even number of 1's in each circle). Thus, the parity bits are as shown in Figure 7.11.

Now assume that one of the bits is changed; for example one of the information bits is changed from 1 to 0 as shown in Figure 7.12. Then the parity constraints are violated for two of the circles (highlighted in the figure), and it is not hard to see that given these violations, the only single bit error that could have caused it is at the intersection of the two circles (i.e., the bit that was changed). Similarly working through the other error cases, it is not hard to see that this code can detect and correct any single bit error in the received codeword.

We can easily generalize this procedure to construct larger matrices H. In general, if we use l rows in H, the code that we obtain will have block length $n = 2^l - 1$, $k = 2^l - l - 1$ and minimum distance 3. All these codes are called Hamming codes and can correct one error.

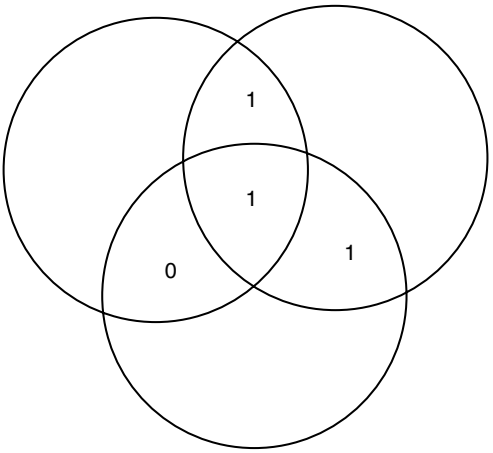


FIGURE 7.10. Venn diagram with information bits.

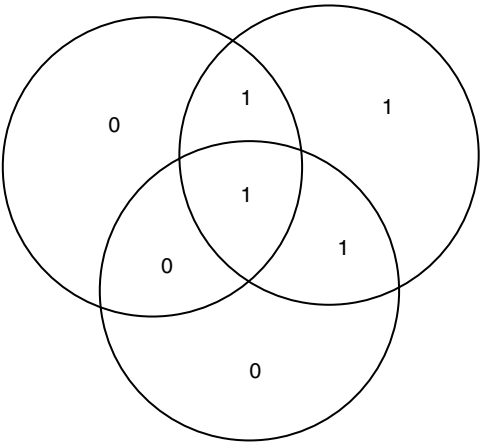


FIGURE 7.11. Venn diagram with information bits and parity bits with even parity for each circle.

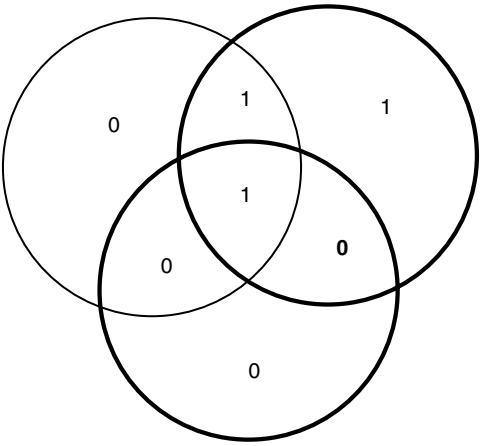


FIGURE 7.12. Venn diagram with one of the information bits changed.

Hamming codes are the simplest examples of linear parity check codes. They demonstrate the principle that underlies the construction of other linear codes. But with large block lengths it is likely that there will be more than one error in the block. In the early 1950s, Reed and Solomon found a class of multiple error-correcting codes for nonbinary channels. In the late 1950s, Bose and Ray-Chaudhuri [72] and Hocquenghem [278] generalized the ideas of Hamming codes using Galois field theory to construct t-error correcting codes (called *BCH codes*) for any t. Since then, various authors have developed other codes and also developed efficient

decoding algorithms for these codes. With the advent of integrated circuits, it has become feasible to implement fairly complex codes in hardware and realize some of the error-correcting performance promised by Shannon's channel capacity theorem. For example, all compact disc players include error-correction circuitry based on two interleaved (32, 28, 5) and (28, 24, 5) Reed–Solomon codes that allow the decoder to correct bursts of up to 4000 errors.

All the codes described above are *block codes*—they map a block of information bits onto a channel codeword and there is no dependence on past information bits. It is also possible to design codes where each output block depends not only on the current input block, but also on some of the past inputs as well. A highly structured form of such a code is called a *convolutional code*. The theory of convolutional codes has developed considerably over the last 40 years. We will not go into the details, but refer the interested reader to textbooks on coding theory [69, 356].

For many years, none of the known coding algorithms came close to achieving the promise of Shannon's channel capacity theorem. For a binary symmetric channel with crossover probability p, we would need a code that could correct up to np errors in a block of length n and have $n(1 - H(p))$ information bits. For example, the repetition code suggested earlier corrects up to $n/2$ errors in a block of length n, but its rate goes to 0 with n. Until 1972, all known codes that could correct $n\alpha$ errors for block length n had asymptotic rate 0. In 1972, Justesen [301] described a class of codes with positive asymptotic rate and positive asymptotic minimum distance as a fraction of the block length.

In 1993, a paper by Berrou et al. [57] introduced the notion that the combination of two interleaved convolution codes with a parallel cooperative decoder achieved much better performance than any of the earlier codes. Each decoder feeds its "opinion" of the value of each bit to the other decoder and uses the opinion of the other decoder to help it decide the value of the bit. This iterative process is repeated until both decoders agree on the value of the bit. The surprising fact is that this iterative procedure allows for efficient decoding at rates close to capacity for a variety of channels. There has also been a renewed interest in the theory of low-density parity check (LDPC) codes that were introduced by Robert Gallager in his thesis [231, 232]. In 1997, MacKay and Neal [368] showed that an iterative message-passing algorithm similar to the algorithm used for decoding turbo codes could achieve rates close to capacity with high probability for LDPC codes. Both Turbo codes and LDPC codes remain active areas of research and have been applied to wireless and satellite communication channels.

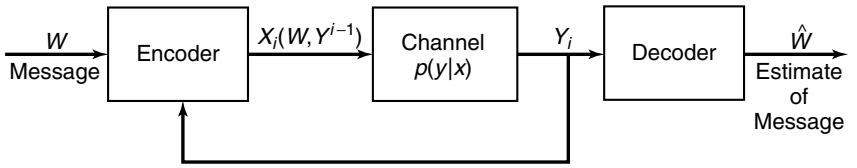


FIGURE 7.13. Discrete memoryless channel with feedback.

7.12 FEEDBACK CAPACITY

A channel with feedback is illustrated in Figure 7.13. We assume that all the received symbols are sent back immediately and noiselessly to the transmitter, which can then use them to decide which symbol to send next. Can we do better with feedback? The surprising answer is no, which we shall now prove. We define a $(2^{nR}, n)$ *feedback code* as a sequence of mappings $x_i(W, Y^{i-1})$, where each x_i is a function only of the message $W \in 2^{nR}$ and the previous received values, $Y_1, Y_2, \ldots, Y_{i-1}$, and a sequence of decoding functions $g : \mathcal{Y}^n \to \{1, 2, \ldots, 2^{nR}\}$. Thus,

$$P_e^{(n)} = \Pr\left\{g(Y^n) \neq W\right\}, \tag{7.119}$$

when W is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$.

Definition The *capacity with feedback*, C_{FB}, of a discrete memoryless channel is the supremum of all rates achievable by feedback codes.

Theorem 7.12.1 (*Feedback capacity*)

$$C_{FB} = C = \max_{p(x)} I(X; Y). \tag{7.120}$$

Proof: Since a nonfeedback code is a special case of a feedback code, any rate that can be achieved without feedback can be achieved with feedback, and hence

$$C_{FB} \geq C. \tag{7.121}$$

Proving the inequality the other way is slightly more tricky. We cannot use the same proof that we used for the converse to the coding theorem without feedback. Lemma 7.9.2 is no longer true, since X_i depends on the past received symbols, and it is no longer true that Y_i depends only on X_i and is conditionally independent of the future X's in (7.93).

There is a simple change that will fix the problem with the proof. Instead of using X^n, we will use the index W and prove a similar series of inequalities. Let W be uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$. Then $\Pr(W \neq \hat{W}) = P_e^{(n)}$ and

$$nR = H(W) = H(W|\hat{W}) + I(W; \hat{W}) \tag{7.122}$$

$$\leq 1 + P_e^{(n)} nR + I(W; \hat{W}) \tag{7.123}$$

$$\leq 1 + P_e^{(n)} nR + I(W; Y^n), \tag{7.124}$$

by Fano's inequality and the data-processing inequality. Now we can bound $I(W; Y^n)$ as follows:

$$I(W; Y^n) = H(Y^n) - H(Y^n|W) \tag{7.125}$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, Y_2, \ldots, Y_{i-1}, W) \tag{7.126}$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, Y_2, \ldots, Y_{i-1}, W, X_i) \tag{7.127}$$

$$= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i), \tag{7.128}$$

since X_i is a function of $Y_1, \ldots, Y_{i-1}$ and W; and conditional on X_i, Y_i is independent of W and past samples of Y. Continuing, we have

$$I(W; Y^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \tag{7.129}$$

$$\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \tag{7.130}$$

$$= \sum_{i=1}^{n} I(X_i; Y_i) \tag{7.131}$$

$$\leq nC \tag{7.132}$$

from the definition of capacity for a discrete memoryless channel. Putting these together, we obtain

$$nR \leq P_e^{(n)} nR + 1 + nC, \tag{7.133}$$

and dividing by n and letting $n \to \infty$, we conclude that

$$R \leq C. \tag{7.134}$$

Thus, we cannot achieve any higher rates with feedback than we can without feedback, and

$$C_{FB} = C. \quad \square \tag{7.135}$$

As we have seen in the example of the binary erasure channel, feedback can help enormously in simplifying encoding and decoding. However, it cannot increase the capacity of the channel.

7.13 SOURCE–CHANNEL SEPARATION THEOREM

It is now time to combine the two main results that we have proved so far: data compression ($R > H$: Theorem 5.4.2) and data transmission ($R < C$: Theorem 7.7.1). Is the condition $H < C$ necessary and sufficient for sending a source over a channel? For example, consider sending digitized speech or music over a discrete memoryless channel. We could design a code to map the sequence of speech samples directly into the input of the channel, or we could compress the speech into its most efficient representation, then use the appropriate channel code to send it over the channel. It is not immediately clear that we are not losing something by using the two-stage method, since data compression does not depend on the channel and the channel coding does not depend on the source distribution.

We will prove in this section that the two-stage method is as good as any other method of transmitting information over a noisy channel. This result has some important practical implications. It implies that we can consider the design of a communication system as a combination of two parts, source coding and channel coding. We can design source codes for the most efficient representation of the data. We can, separately and independently, design channel codes appropriate for the channel. The combination will be as efficient as anything we could design by considering both problems together.

The common representation for all kinds of data uses a binary alphabet. Most modern communication systems are digital, and data are reduced to a binary representation for transmission over the common channel. This offers an enormous reduction in complexity. Networks like, ATM networks and the Internet use the common binary representation to allow speech, video, and digital data to use the same communication channel.

The result—that a two-stage process is as good as any one-stage process—seems so obvious that it may be appropriate to point out that it is not always true. There are examples of multiuser channels where the decomposition breaks down. We also consider two simple situations where the theorem appears to be misleading. A simple example is that of sending English text over an erasure channel. We can look for the most efficient binary representation of the text and send it over the channel. But the errors will be very difficult to decode. If, however, we send the English text directly over the channel, we can lose up to about half the letters and yet be able to make sense out of the message. Similarly, the human ear has some unusual properties that enable it to distinguish speech under very high noise levels if the noise is white. In such cases, it may be appropriate to send the uncompressed speech over the noisy channel rather than the compressed version. Apparently, the redundancy in the source is suited to the channel.

Let us define the setup under consideration. We have a source V that generates symbols from an alphabet $\mathcal{V}$. We will not make any assumptions about the kind of stochastic process produced by V other than that it is from a finite alphabet and satisfies the AEP. Examples of such processes include a sequence of i.i.d. random variables and the sequence of states of a stationary irreducible Markov chain. Any stationary ergodic source satisfies the AEP, as we show in Section 16.8.

We want to send the sequence of symbols $V^n = V_1, V_2, \ldots, V_n$ over the channel so that the receiver can reconstruct the sequence. To do this, we map the sequence onto a codeword $X^n(V^n)$ and send the codeword over the channel. The receiver looks at his received sequence Y^n and makes an estimate $\hat{V}^n$ of the sequence V^n that was sent. The receiver makes an error if $V^n \neq \hat{V}^n$. We define the probability of error as

$$\Pr(V^n \neq \hat{V}^n) = \sum_{y^n} \sum_{v^n} p(v^n) p(y^n | x^n(v^n)) I(g(y^n) \neq v^n), \qquad (7.136)$$

where I is the indicator function and $g(y^n)$ is the decoding function. The system is illustrated in Figure 7.14.

We can now state the joint source–channel coding theorem:

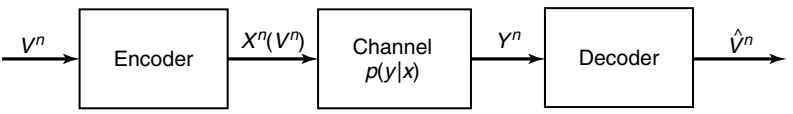


FIGURE 7.14. Joint source and channel coding.

Theorem 7.13.1 (*Source–channel coding theorem*) *If $V_1, V_2, \ldots V^n$ is a finite alphabet stochastic process that satisfies the AEP and $H(\mathcal{V}) < C$, there exists a source–channel code with probability of error* $\Pr(\hat{V}^n \neq V^n) \to 0$. *Conversely, for any stationary stochastic process, if $H(\mathcal{V}) > C$, the probability of error is bounded away from zero, and it is not possible to send the process over the channel with arbitrarily low probability of error.*

Proof: *Achievability*. The essence of the forward part of the proof is the two-stage encoding described earlier. Since we have assumed that the stochastic process satisfies the AEP, it implies that there exists a typical set $A_\epsilon^{(n)}$ of size $\leq 2^{n(H(\mathcal{V})+\epsilon)}$ which contains most of the probability. We will encode only the source sequences belonging to the typical set; all other sequences will result in an error. This will contribute at most ϵ to the probability of error.

We index all the sequences belonging to $A_\epsilon^{(n)}$. Since there are at most $2^{n(H+\epsilon)}$ such sequences, $n(H+\epsilon)$ bits suffice to index them. We can transmit the desired index to the receiver with probability of error less than ϵ if

$$H(\mathcal{V}) + \epsilon = R < C. \tag{7.137}$$

The receiver can reconstruct V^n by enumerating the typical set $A_\epsilon^{(n)}$ and choosing the sequence corresponding to the estimated index. This sequence will agree with the transmitted sequence with high probability. To be precise,

$$P(V^n \neq \hat{V}^n) \leq P(V^n \notin A_\epsilon^{(n)}) + P(g(Y^n) \neq V^n | V^n \in A_\epsilon^{(n)}) \tag{7.138}$$

$$\leq \epsilon + \epsilon = 2\epsilon \tag{7.139}$$

for n sufficiently large. Hence, we can reconstruct the sequence with low probability of error for n sufficiently large if

$$H(\mathcal{V}) < C. \tag{7.140}$$

Converse: We wish to show that $\Pr(\hat{V}^n \neq V^n) \to 0$ implies that $H(\mathcal{V}) \leq C$ for any sequence of source-channel codes

$$X^n(V^n) : \mathcal{V}^n \to \mathcal{X}^n, \tag{7.141}$$

$$g_n(Y^n) : \mathcal{Y}^n \to \mathcal{V}^n. \tag{7.142}$$

Thus $X^n(\cdot)$ is an arbitrary (perhaps random) assignment of codewords to data sequences V^n, and $g_n(\cdot)$ is any decoding function (assignment of estimates $\hat{V}^n$ to output sequences Y^n. By Fano's inequality, we must have

$$H(V^n|\hat{V}^n) \le 1 + \Pr(\hat{V}^n \neq V^n) \log |\mathcal{V}^n| = 1 + \Pr(\hat{V}^n \neq V^n) n \log |\mathcal{V}|. \tag{7.143}$$

Hence for the code,

$$H(\mathcal{V}) \overset{(a)}{\le} \frac{H(V_1, V_2, \ldots, V_n)}{n} \tag{7.144}$$

$$= \frac{H(V^n)}{n} \tag{7.145}$$

$$= \frac{1}{n} H(V^n|\hat{V}^n) + \frac{1}{n} I(V^n; \hat{V}^n) \tag{7.146}$$

$$\overset{(b)}{\le} \frac{1}{n}(1 + \Pr(\hat{V}^n \neq V^n) n \log |\mathcal{V}|) + \frac{1}{n} I(V^n; \hat{V}^n) \tag{7.147}$$

$$\overset{(c)}{\le} \frac{1}{n}(1 + \Pr(\hat{V}^n \neq V^n) n \log |\mathcal{V}|) + \frac{1}{n} I(X^n; Y^n) \tag{7.148}$$

$$\overset{(d)}{\le} \frac{1}{n} + \Pr(\hat{V}^n \neq V^n) \log |\mathcal{V}| + C, \tag{7.149}$$

where (a) follows from the definition of entropy rate of a stationary process, (b) follows from Fano's inequality, (c) follows from the data-processing inequality (since $V^n \to X^n \to Y^n \to \hat{V}^n$ forms a Markov chain) and (d) follows from the memorylessness of the channel. Now letting $n \to \infty$, we have $\Pr(\hat{V}^n \neq V^n) \to 0$ and hence

$$H(\mathcal{V}) \le C. \tag{7.150}$$

□

Hence, we can transmit a stationary ergodic source over a channel if and only if its entropy rate is less than the capacity of the channel. The joint source–channel separation theorem enables us to consider the problem of source coding separately from the problem of channel coding. The source coder tries to find the most efficient representation of the source, and the channel coder encodes the message to combat the noise and errors introduced by the channel. The separation theorem says that the separate encoders (Figure 7.15) can achieve the same rates as the joint encoder (Figure 7.14).

With this result, we have tied together the two basic theorems of information theory: data compression and data transmission. We will try to summarize the proofs of the two results in a few words. The data

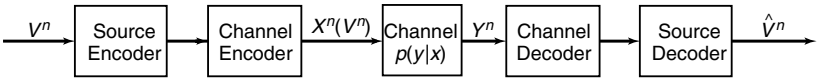


FIGURE 7.15. Separate source and channel coding.

compression theorem is a consequence of the AEP, which shows that there exists a "small" subset (of size 2^{nH}) of all possible source sequences that contain most of the probability and that we can therefore represent the source with a small probability of error using H bits per symbol. The data transmission theorem is based on the joint AEP; it uses the fact that for long block lengths, the output sequence of the channel is very likely to be jointly typical with the input codeword, while any other codeword is jointly typical with probability $\approx 2^{-nI}$. Hence, we can use about 2^{nI} codewords and still have negligible probability of error. The source–channel separation theorem shows that we can design the source code and the channel code separately and combine the results to achieve optimal performance.

SUMMARY

Channel capacity. The logarithm of the number of distinguishable inputs is given by

$$C = \max_{p(x)} I(X; Y).$$

Examples

- Binary symmetric channel: $C = 1 - H(p)$.
- Binary erasure channel: $C = 1 - \alpha$.
- Symmetric channel: $C = \log |\mathcal{Y}| - H(\text{row of transition matrix})$.

Properties of C

1. $0 \leq C \leq \min\{\log |\mathcal{X}|, \log |\mathcal{Y}|\}$.
2. $I(X; Y)$ is a continuous concave function of $p(x)$.

Joint typicality. The set $A_\epsilon^{(n)}$ of *jointly typical* sequences $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is given by

$$A_\epsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \tag{7.151}$$

$$\left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \tag{7.152}$$

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

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

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

Joint AEP. Let (X^n, Y^n) be sequences of length n drawn i.i.d. according to $p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$. Then:

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

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

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

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

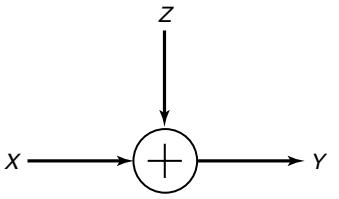
PROBLEMS

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

(a) Show that he is wrong.

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

7.2 *Additive noise channel.* Find the channel capacity of the following discrete memoryless channel:



where $\Pr\{Z = 0\} = \Pr\{Z = a\} = \frac{1}{2}$. The alphabet for x is $\mathbf{X} = \{0, 1\}$. Assume that Z is independent of X. Observe that the channel capacity depends on the value of a.

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

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

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

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

(a) Find the capacity.

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

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

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

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

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

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

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

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

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

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

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

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

7.9 *Suboptimal codes.* For the Z-channel of Problem 7.8, assume that we choose a $(2^{nR}, n)$ code at random, where each codeword is a sequence of *fair* coin tosses. This will not achieve capacity. Find the maximum rate R such that the probability of error $P_e^{(n)}$, averaged over the randomly generated codes, tends to zero as the block length n tends to infinity.

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

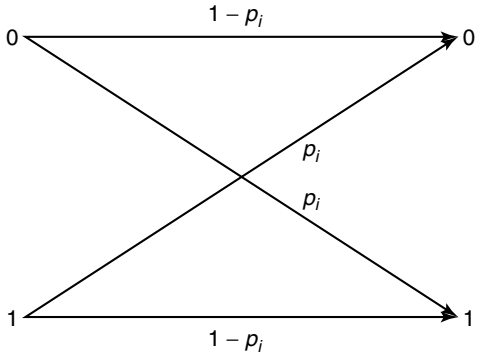
$$p(y|x) = \begin{cases} 1/2 & \text{if } y = x \pm 1 \bmod 5 \\ 0 & \text{otherwise.} \end{cases}$$

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

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

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

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



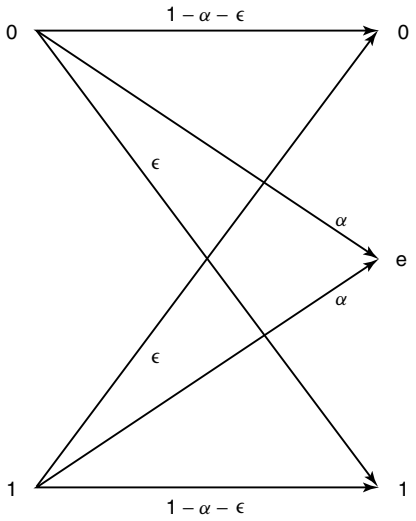
7.12 *Unused symbols.* Show that the capacity of the channel with probability transition matrix

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

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

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

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



(a) Find the capacity of this channel.

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

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

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

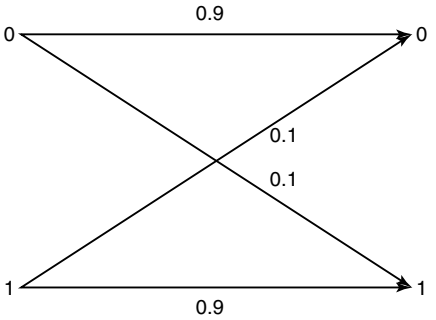
(a) Calculate the mutual information $I(X_1, X_2; Y_1, Y_2)$ as a function of the input distribution on the four possible pairs of inputs.

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

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

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

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



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

X \ Y	0	1
0	0.45	0.05
1	0.05	0.45

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(a) Calculate the crossover probability of this channel.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

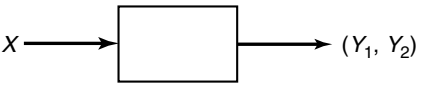
(c) Now assume that the enemy is more cunning and that every time they shoot down a pigeon, they send out a dummy pigeon carrying a random letter (chosen uniformly from all 8-bit letters). What is the capacity of this link in bits/hour?

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

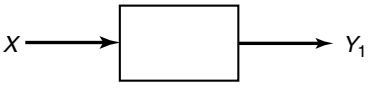
7.20 *Channel with two independent looks at Y.* Let Y_1 and Y_2 be conditionally independent and conditionally identically distributed given X.

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

(b) Conclude that the capacity of the channel



is less than twice the capacity of the channel



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

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

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

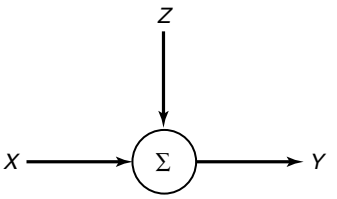
7.22 *Can signal alternatives lower capacity?* Show that adding a row to a channel transition matrix does not decrease capacity.

7.23 *Binary multiplier channel*

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

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

7.24 *Noise alphabets.* Consider the channel

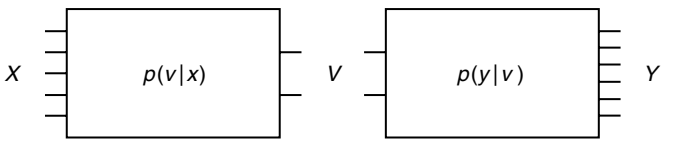


$\mathcal{X} = \{0, 1, 2, 3\}$, where $Y = X + Z$, and Z is uniformly distributed over three distinct integer values $\mathcal{Z} = \{z_1, z_2, z_3\}$.

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

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

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



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

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

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

(a) Find the capacity of this channel.

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

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

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

(i)

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

(ii)

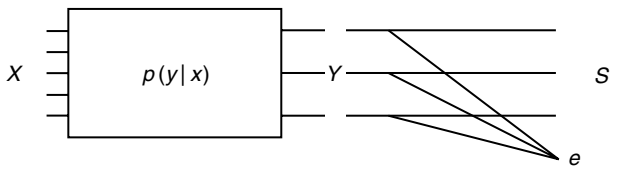
$$p(x) = \begin{cases} 0 & \text{if } x \in \{1, 3\} \\ \frac{1}{2} & \text{if } x \in \{0, 2\}. \end{cases}$$

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

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

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

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



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

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

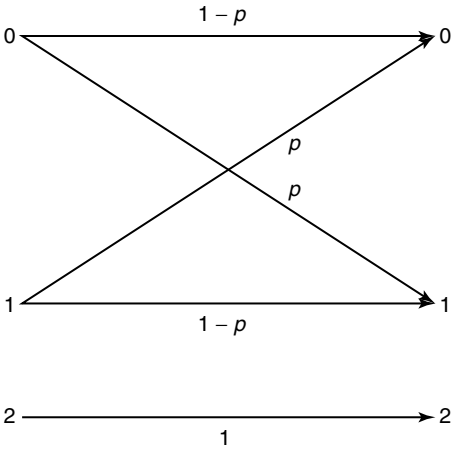
Determine the capacity of this channel.

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

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

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

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

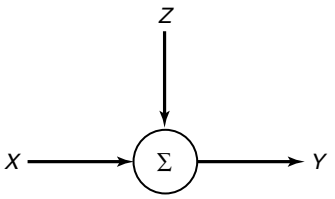


7.29 *Binary multiplier channel*

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

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

7.30 *Noise alphabets.* Consider the channel

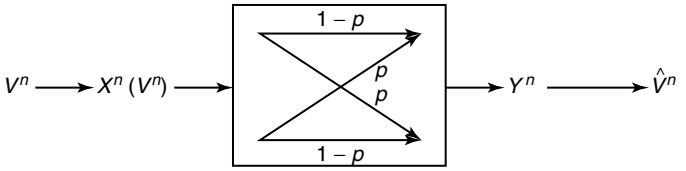


$\mathcal{X} = \{0, 1, 2, 3\}$, where $Y = X + Z$, and Z is uniformly distributed over three distinct integer values $\mathcal{Z} = \{z_1, z_2, z_3\}$.

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

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

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



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

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

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

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

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

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

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

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

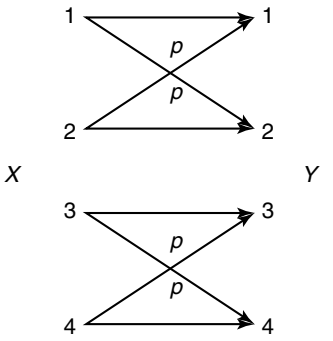
(a) Find $\lim_{n\to\infty} \frac{1}{n} I(X^n; Y^n)$.

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

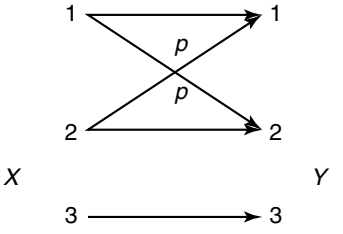
(c) Using this feedback transmission scheme, $X^n(W, Y^n) = (X_1(W), Y_1, Y_2, \ldots, Y_{m-1})$, what is the asymptotic communication rate achieved; that is, what is $\lim_{n\to\infty} \frac{1}{n} I(W; Y^n)$?

7.34 *Capacity.* Find the capacity of

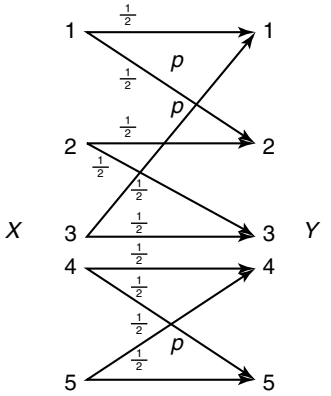
(a) Two parallel BSCs:



(b) BSC and a single symbol:



(c) BSC and a ternary channel:



(d) Ternary channel:

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

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

(a) What is the capacity of

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

(b) What about the capacity of

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

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

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

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

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

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

(b) What is the capacity if

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

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

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

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

HISTORICAL NOTES

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

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

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

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

CHAPTER 8

DIFFERENTIAL ENTROPY

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

8.1 DEFINITIONS

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

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

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

where S is the support set of the random variable.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Changing the base of the logarithm, we have

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

8.2 AEP FOR CONTINUOUS RANDOM VARIABLES

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

We can do the same for a continuous random variable.

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

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

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

This leads to the following definition of the typical set.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

8.3 RELATION OF DIFFERENTIAL ENTROPY TO DISCRETE ENTROPY

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

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

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

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

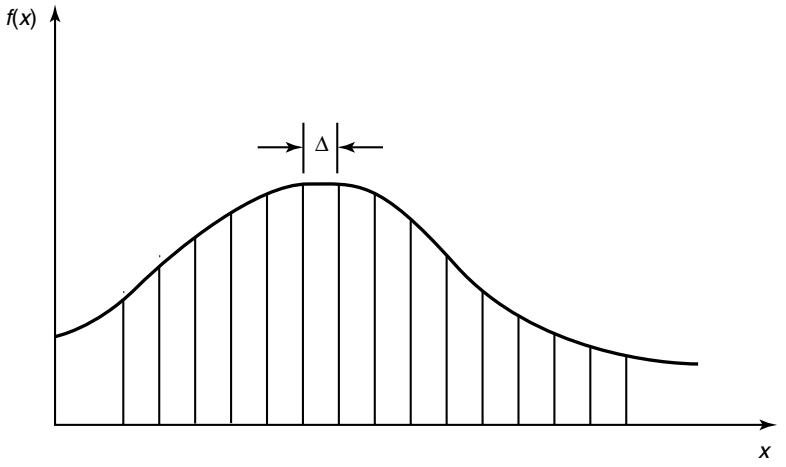


FIGURE 8.1. Quantization of a continuous random variable.

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

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

The entropy of the quantized version is

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

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

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

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

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

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

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

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

Example 8.3.1

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

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

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

8.4 JOINT AND CONDITIONAL DIFFERENTIAL ENTROPY

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

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

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

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

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

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

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

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

The next entropy evaluation is used frequently in the text.

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

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

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

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

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

Then

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

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

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

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

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

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

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

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

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

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

8.5 RELATIVE ENTROPY AND MUTUAL INFORMATION

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

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

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

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

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

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x) f(y)} \, dx \, dy. \tag{8.47}$$

From the definition it is clear that

$$I(X; Y) = h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X, Y) \tag{8.48}$$

and

$$I(X; Y) = D(f(x, y)||f(x) f(y)). \tag{8.49}$$

The properties of $D(f||g)$ and $I(X; Y)$ are the same as in the discrete case. In particular, the mutual information between two random variables is the limit of the mutual information between their quantized versions, since

$$I(X^\Delta; Y^\Delta) = H(X^\Delta) - H(X^\Delta|Y^\Delta) \tag{8.50}$$

$$\approx h(X) - \log \Delta - (h(X|Y) - \log \Delta) \tag{8.51}$$

$$= I(X; Y). \tag{8.52}$$

More generally, we can define mutual information in terms of finite partitions of the range of the random variable. Let $\mathcal{X}$ be the range of a random variable X. A partition $\mathcal{P}$ of $\mathcal{X}$ is a finite collection of disjoint sets P_i such that $\cup_i P_i = \mathcal{X}$. The quantization of X by $\mathcal{P}$ (denoted $[X]_\mathcal{P}$) is the discrete random variable defined by

$$\Pr([X]_\mathcal{P} = i) = \Pr(X \in P_i) = \int_{P_i} dF(x). \tag{8.53}$$

For two random variables X and Y with partitions $\mathcal{P}$ and $\mathcal{Q}$, we can calculate the mutual information between the quantized versions of X and Y using (2.28). Mutual information can now be defined for arbitrary pairs of random variables as follows:

Definition The *mutual information* between two random variables X and Y is given by

$$I(X;Y) = \sup_{\mathcal{P},\mathcal{Q}} I([X]_{\mathcal{P}};[Y]_{\mathcal{Q}}), \tag{8.54}$$

where the supremum is over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$.

This is the master definition of mutual information that always applies, even to joint distributions with atoms, densities, and singular parts. Moreover, by continuing to refine the partitions P and Q, one finds a monotonically increasing sequence $I([X]_P;[Y]_Q) \nearrow I$.

By arguments similar to (8.52), we can show that this definition of mutual information is equivalent to (8.47) for random variables that have a density. For discrete random variables, this definition is equivalent to the definition of mutual information in (2.28).

Example 8.5.1 (*Mutual information between correlated Gaussian random variables with correlation* ρ) Let $(X,Y) \sim \mathcal{N}(0,K)$, where

$$K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}. \tag{8.55}$$

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

$$I(X;Y) = h(X) + h(Y) - h(X,Y) = -\frac{1}{2}\log(1-\rho^2). \tag{8.56}$$

If $\rho = 0$, X and Y are independent and the mutual information is 0. If $\rho = \pm 1$, X and Y are perfectly correlated and the mutual information is infinite.

8.6 PROPERTIES OF DIFFERENTIAL ENTROPY, RELATIVE ENTROPY, AND MUTUAL INFORMATION

Theorem 8.6.1

$$D(f||g) \geq 0 \tag{8.57}$$

with equality iff $f = g$ *almost everywhere (a.e.).*

Proof: Let S be the support set of f. Then

$$-D(f||g) = \int_S f \log \frac{g}{f} \tag{8.58}$$

$$\leq \log \int_S f \frac{g}{f} \qquad \text{(by Jensen's inequality)} \tag{8.59}$$

$$= \log \int_S g \tag{8.60}$$

$$\leq \log 1 = 0. \tag{8.61}$$

We have equality iff we have equality in Jensen's inequality, which occurs iff $f = g$ a.e. □

Corollary *$I(X; Y) \geq 0$ with equality iff X and Y are independent.*

Corollary *$h(X|Y) \leq h(X)$ with equality iff X and Y are independent.*

Theorem 8.6.2 (*Chain rule for differential entropy*)

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, X_2, \ldots, X_{i-1}). \tag{8.62}$$

Proof: Follows directly from the definitions. □

Corollary

$$h(X_1, X_2, \ldots, X_n) \leq \sum h(X_i), \tag{8.63}$$

with equality iff $X_1, X_2, \ldots, X_n$ are independent.

Proof: Follows directly from Theorem 8.6.2 and the corollary to Theorem 8.6.1. □

Application (*Hadamard's inequality*) *If we let $\mathbf{X} \sim \mathcal{N}(0, K)$ be a multivariate normal random variable, calculating the entropy in the above inequality gives us*

$$|K| \leq \prod_{i=1}^{n} K_{ii}, \tag{8.64}$$

which is Hadamard's inequality. A number of determinant inequalities can be derived in this fashion from information-theoretic inequalities (Chapter 17).

Theorem 8.6.3

$$h(X + c) = h(X). \tag{8.65}$$

Translation does not change the differential entropy.

Proof: Follows directly from the definition of differential entropy. □

Theorem 8.6.4

$$h(aX) = h(X) + \log |a|. \tag{8.66}$$

Proof: Let $Y = aX$. Then $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$, and

$$h(aX) = -\int f_Y(y) \log f_Y(y)\, dy \tag{8.67}$$

$$= -\int \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right) dy \tag{8.68}$$

$$= -\int f_X(x) \log f_X(x)\, dx + \log |a| \tag{8.69}$$

$$= h(X) + \log |a|, \tag{8.70}$$

after a change of variables in the integral. □

Similarly, we can prove the following corollary for vector-valued random variables.

Corollary

$$h(A\mathbf{X}) = h(\mathbf{X}) + \log |\det(A)|. \tag{8.71}$$

We now show that the multivariate normal distribution maximizes the entropy over all distributions with the same covariance.

Theorem 8.6.5 *Let the random vector $\mathbf{X} \in \mathbf{R}^n$ have zero mean and covariance $K = E\mathbf{X}\mathbf{X}^t$ (i.e., $K_{ij} = EX_iX_j$, $1 \le i, j \le n$). Then $h(\mathbf{X}) \le \frac{1}{2} \log(2\pi e)^n |K|$, with equality iff $\mathbf{X} \sim \mathcal{N}(0, K)$.*

Proof: Let $g(\mathbf{x})$ be any density satisfying $\int g(\mathbf{x}) x_i x_j d\mathbf{x} = K_{ij}$ for all i, j. Let ϕ_K be the density of a $\mathcal{N}(0, K)$ vector as given in (8.35), where we set $\mu = 0$. Note that $\log \phi_K(\mathbf{x})$ is a quadratic form and $\int x_i x_j \phi_K(\mathbf{x})\, d\mathbf{x} = K_{ij}$. Then

$$0 \le D(g||\phi_K) \tag{8.72}$$

$$= \int g \log(g/\phi_K) \tag{8.73}$$

$$= -h(g) - \int g \log \phi_K \tag{8.74}$$

$$= -h(g) - \int \phi_K \log \phi_K \tag{8.75}$$

$$= -h(g) + h(\phi_K), \tag{8.76}$$

where the substitution $\int g \log \phi_K = \int \phi_K \log \phi_K$ follows from the fact that g and ϕ_K yield the same moments of the quadratic form $\log \phi_K(\mathbf{x})$. □

In particular, the Gaussian distribution maximizes the entropy over all distributions with the same variance. This leads to the estimation counterpart to Fano's inequality. Let X be a random variable with differential entropy $h(X)$. Let $\hat{X}$ be an estimate of X, and let $E(X - \hat{X})^2$ be the expected prediction error. Let $h(X)$ be in nats.

Theorem 8.6.6 (*Estimation error and differential entropy*) *For any random variable X and estimator $\hat{X}$,*

$$E(X - \hat{X})^2 \geq \frac{1}{2\pi e} e^{2h(X)},$$

with equality if and only if X is Gaussian and $\hat{X}$ is the mean of X.

Proof: Let $\hat{X}$ be any estimator of X; then

$$E(X - \hat{X})^2 \geq \min_{\hat{X}} E(X - \hat{X})^2 \tag{8.77}$$

$$= E\,(X - E(X))^2 \tag{8.78}$$

$$= \mathrm{var}(X) \tag{8.79}$$

$$\geq \frac{1}{2\pi e} e^{2h(X)}, \tag{8.80}$$

where (8.78) follows from the fact that the mean of X is the best estimator for X and the last inequality follows from the fact that the Gaussian distribution has the maximum entropy for a given variance. We have equality only in (8.78) only if $\hat{X}$ is the best estimator (i.e., $\hat{X}$ is the mean of X and equality in (8.80) only if X is Gaussian). □

Corollary *Given side information Y and estimator $\hat{X}(Y)$, it follows that*

$$E(X - \hat{X}(Y))^2 \geq \frac{1}{2\pi e} e^{2h(X|Y)}.$$

SUMMARY

$$h(X) = h(f) = -\int_S f(x) \log f(x)\, dx \tag{8.81}$$

$$f(X^n) \doteq 2^{-nh(X)} \tag{8.82}$$

$$\text{Vol}(A_\epsilon^{(n)}) \doteq 2^{nh(X)}. \tag{8.83}$$

$$H([X]_{2^{-n}}) \approx h(X) + n. \tag{8.84}$$

$$h(\mathcal{N}(0, \sigma^2)) = \frac{1}{2} \log 2\pi e \sigma^2. \tag{8.85}$$

$$h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log (2\pi e)^n |K|. \tag{8.86}$$

$$D(f||g) = \int f \log \frac{f}{g} \geq 0. \tag{8.87}$$

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, X_2, \ldots, X_{i-1}). \tag{8.88}$$

$$h(X|Y) \leq h(X). \tag{8.89}$$

$$h(aX) = h(X) + \log |a|. \tag{8.90}$$

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x) f(y)} \geq 0. \tag{8.91}$$

$$\max_{E\mathbf{X}\mathbf{X}^t = K} h(\mathbf{X}) = \frac{1}{2} \log (2\pi e)^n |K|. \tag{8.92}$$

$$E(X - \hat{X}(Y))^2 \geq \frac{1}{2\pi e} e^{2h(X|Y)}.$$

$2^{nH(X)}$ is the effective alphabet size for a discrete random variable.
$2^{nh(X)}$ is the effective support set size for a continuous random variable.
2^C is the effective alphabet size of a channel of capacity C.

PROBLEMS

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

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

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

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

8.2 *Concavity of determinants*. Let K_1 and K_2 be two symmetric nonnegative definite $n \times n$ matrices. Prove the result of Ky Fan [199]:

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

where $| K |$ denotes the determinant of K. [*Hint*: Let $\mathbf{Z} = \mathbf{X}_\theta$, where $\mathbf{X}_1 \sim N(0, K_1)$, $\mathbf{X}_2 \sim N(0, K_2)$ and $\theta = \text{Bernoulli}(\lambda)$. Then use $h(\mathbf{Z} \mid \theta) \leq h(\mathbf{Z})$.]

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

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

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

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

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

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

8.6 *Variational inequality*. Verify for positive random variables X that

$$\log E_P(X) = \sup_Q \left[E_Q(\log X) - D(Q||P) \right], \tag{8.93}$$

where $E_P(X) = \sum_x x P(x)$ and $D(Q||P) = \sum_x Q(x) \log \frac{Q(x)}{P(x)}$, and the supremum is over all $Q(x) \geq 0$, $\sum Q(x) = 1$. It is enough to extremize $J(Q) = E_Q \ln X - D(Q||P) + \lambda(\sum Q(x) - 1)$.

8.7 *Differential entropy bound on discrete entropy*. Let X be a discrete random variable on the set $\mathcal{X} = \{a_1, a_2, \ldots\}$ with $\Pr(X = a_i) = p_i$. Show that

$$H(p_1, p_2, \ldots) \le \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{\infty} p_i i^2 - \left(\sum_{i=1}^{\infty} i p_i\right)^2 + \frac{1}{12}\right). \tag{8.94}$$

Moreover, for every permutation σ,

$$H(p_1, p_2, \ldots) \le \frac{1}{2}\log(2\pi e)\left(\sum_{i=1}^{\infty} p_{\sigma(i)} i^2 - \left(\sum_{i=1}^{\infty} i p_{\sigma(i)}\right)^2 + \frac{1}{12}\right). \tag{8.95}$$

[*Hint:* Construct a random variable X' such that $\Pr(X' = i) = p_i$. Let U be a uniform (0,1] random variable and let $Y = X' + U$, where X' and U are independent. Use the maximum entropy bound on Y to obtain the bounds in the problem. This bound is due to Massey (unpublished) and Willems (unpublished).]

8.8 *Channel with uniformly distributed noise*. Consider a additive channel whose input alphabet $\mathcal{X} = \{0, \pm 1, \pm 2\}$ and whose output $Y = X + Z$, where Z is distributed uniformly over the interval $[-1, 1]$. Thus, the input of the channel is a discrete random variable, whereas the output is continuous. Calculate the capacity $C = \max_{p(x)} I(X; Y)$ of this channel.

8.9 *Gaussian mutual information*. Suppose that (X, Y, Z) are jointly Gaussian and that $X \to Y \to Z$ forms a Markov chain. Let X and Y have correlation coefficient ρ_1 and let Y and Z have correlation coefficient ρ_2. Find $I(X; Z)$.

8.10 *Shape of the typical set*. Let X_i be i.i.d. $\sim f(x)$, where

$$f(x) = ce^{-x^4}.$$

Let $h = -\int f \ln f$. Describe the shape (or form) or the typical set $A_\epsilon^{(n)} = \{x^n \in \mathcal{R}^n : f(x^n) \in 2^{-n(h \pm \epsilon)}\}$.

8.11 *Nonergodic Gaussian process*. Consider a constant signal V in the presence of iid observational noise $\{Z_i\}$. Thus, $X_i = V + Z_i$, where $V \sim N(0, S)$ and Z_i are iid $\sim N(0, N)$. Assume that V and $\{Z_i\}$ are independent.

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

(b) Find $\lim_{n\longrightarrow\infty} \frac{1}{n}\sum_{i=1}^{n} X_i$. Is the limit random?

(c) What is the entropy rate h of $\{X_i\}$?

(d) Find the least-mean-squared error predictor $\hat{X}_{n+1}(X^n)$, and find $\sigma_\infty^2 = \lim_{n\longrightarrow\infty} E(\hat{X}_n - X_n)^2$.

(e) Does $\{X_i\}$ have an AEP? That is, does $-\frac{1}{n}\log f(X^n) \longrightarrow h$?

HISTORICAL NOTES

Differential entropy and discrete entropy were introduced in Shannon's original paper [472]. The general rigorous definition of relative entropy and mutual information for arbitrary random variables was developed by Kolmogorov [319] and Pinsker [425], who defined mutual information as $\sup_{\mathcal{P},\mathcal{Q}} I([X]_{\mathcal{P}}; [Y]_{\mathcal{Q}})$, where the supremum is over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$.

CHAPTER 9

GAUSSIAN CHANNEL

The most important continuous alphabet channel is the Gaussian channel depicted in Figure 9.1. This is a time-discrete channel with output Y_i at time i, where Y_i is the sum of the input X_i and the noise Z_i. The noise Z_i is drawn i.i.d. from a Gaussian distribution with variance N. Thus,

$$Y_i = X_i + Z_i, \qquad Z_i \sim \mathcal{N}(0, N). \tag{9.1}$$

The noise Z_i is assumed to be independent of the signal X_i. This channel is a model for some common communication channels, such as wired and wireless telephone channels and satellite links. Without further conditions, the capacity of this channel may be infinite. If the noise variance is zero, the receiver receives the transmitted symbol perfectly. Since X can take on any real value, the channel can transmit an arbitrary real number with no error.

If the noise variance is nonzero and there is no constraint on the input, we can choose an infinite subset of inputs arbitrarily far apart, so that they are distinguishable at the output with arbitrarily small probability of error. Such a scheme has an infinite capacity as well. Thus if the noise variance is zero or the input is unconstrained, the capacity of the channel is infinite.

The most common limitation on the input is an energy or power constraint. We assume an average power constraint. For any codeword $(x_1, x_2, \ldots, x_n)$ transmitted over the channel, we require that

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P. \tag{9.2}$$

This communication channel models many practical channels, including radio and satellite links. The additive noise in such channels may be due to a variety of causes. However, by the central limit theorem, the

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

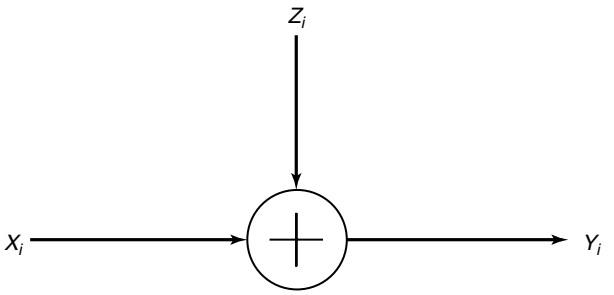


FIGURE 9.1. Gaussian channel.

cumulative effect of a large number of small random effects will be approximately normal, so the Gaussian assumption is valid in a large number of situations.

We first analyze a simple suboptimal way to use this channel. Assume that we want to send 1 bit over the channel in one use of the channel. Given the power constraint, the best that we can do is to send one of two levels, $+\sqrt{P}$ or $-\sqrt{P}$. The receiver looks at the corresponding Y received and tries to decide which of the two levels was sent. Assuming that both levels are equally likely (this would be the case if we wish to send exactly 1 bit of information), the optimum decoding rule is to decide that $+\sqrt{P}$ was sent if $Y > 0$ and decide $-\sqrt{P}$ was sent if $Y < 0$. The probability of error with such a decoding scheme is

$$P_e = \frac{1}{2}\Pr(Y < 0|X = +\sqrt{P}) + \frac{1}{2}\Pr(Y > 0|X = -\sqrt{P}) \tag{9.3}$$

$$= \frac{1}{2}\Pr(Z < -\sqrt{P}|X = +\sqrt{P}) + \frac{1}{2}\Pr(Z > \sqrt{P}|X = -\sqrt{P}) \tag{9.4}$$

$$= \Pr(Z > \sqrt{P}) \tag{9.5}$$

$$= 1 - \Phi\left(\sqrt{P/N}\right), \tag{9.6}$$

where $\Phi(x)$ is the cumulative normal function

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}\, dt. \tag{9.7}$$

Using such a scheme, we have converted the Gaussian channel into a discrete binary symmetric channel with crossover probability P_e. Similarly, by using a four-level input signal, we can convert the Gaussian channel

into a discrete four-input channel. In some practical modulation schemes, similar ideas are used to convert the continuous channel into a discrete channel. The main advantage of a discrete channel is ease of processing of the output signal for error correction, but some information is lost in the quantization.

9.1 GAUSSIAN CHANNEL: DEFINITIONS

We now define the (information) capacity of the channel as the maximum of the mutual information between the input and output over all distributions on the input that satisfy the power constraint.

Definition The *information capacity* of the Gaussian channel with power constraint P is

$$C = \max_{f(x):EX^2 \le P} I(X;Y). \tag{9.8}$$

We can calculate the information capacity as follows: Expanding $I(X;Y)$, we have

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) && (9.9)\\ &= h(Y) - h(X+Z|X) && (9.10)\\ &= h(Y) - h(Z|X) && (9.11)\\ &= h(Y) - h(Z), && (9.12) \end{aligned}$$

since Z is independent of X. Now, $h(Z) = \frac{1}{2}\log 2\pi eN$. Also,

$$EY^2 = E(X+Z)^2 = EX^2 + 2EXEZ + EZ^2 = P + N, \tag{9.13}$$

since X and Z are independent and $EZ = 0$. Given $EY^2 = P + N$, the entropy of Y is bounded by $\frac{1}{2}\log 2\pi e(P+N)$ by Theorem 8.6.5 (the normal maximizes the entropy for a given variance).

Applying this result to bound the mutual information, we obtain

$$\begin{aligned} I(X;Y) &= h(Y) - h(Z) && (9.14)\\ &\le \frac{1}{2}\log 2\pi e(P+N) - \frac{1}{2}\log 2\pi eN && (9.15)\\ &= \frac{1}{2}\log\left(1 + \frac{P}{N}\right). && (9.16) \end{aligned}$$

Hence, the information capacity of the Gaussian channel is

$$C = \max_{EX^2 \le P} I(X;Y) = \frac{1}{2}\log\left(1 + \frac{P}{N}\right), \tag{9.17}$$

and the maximum is attained when $X \sim \mathcal{N}(0, P)$.

We will now show that this capacity is also the supremum of the rates achievable for the channel. The arguments are similar to the arguments for a discrete channel. We will begin with the corresponding definitions.

Definition An (M, n) code for the Gaussian channel with power constraint P consists of the following:

1. An index set $\{1, 2, \ldots, M\}$.
2. An encoding function $x : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2), \ldots, x^n(M)$, satisfying the power constraint P; that is, for every codeword

$$\sum_{i=1}^{n} x_i^2(w) \le nP, \qquad w = 1, 2, \ldots, M. \tag{9.18}$$

3. A decoding function

$$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}. \tag{9.19}$$

The rate and probability of error of the code are defined as in Chapter 7 for the discrete case. The arithmetic average of the probability of error is defined by

$$P_e^{(n)} = \frac{1}{2^{nR}} \sum \lambda_i. \tag{9.20}$$

Definition A rate R is said to be *achievable* for a Gaussian channel with a power constraint P if there exists a sequence of $(2^{nR}, n)$ codes with codewords satisfying the power constraint such that the maximal probability of error $\lambda^{(n)}$ tends to zero. The capacity of the channel is the supremum of the achievable rates.

Theorem 9.1.1 *The capacity of a Gaussian channel with power constraint P and noise variance N is*

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) \qquad \textit{bits per transmission.} \tag{9.21}$$

Remark We first present a plausibility argument as to why we may be able to construct $(2^{nC}, n)$ codes with a low probability of error. Consider any codeword of length n. The received vector is normally distributed with mean equal to the true codeword and variance equal to the noise variance. With high probability, the received vector is contained in a sphere of radius $\sqrt{n(N+\epsilon)}$ around the true codeword. If we assign everything within this sphere to the given codeword, when this codeword is sent there will be an error only if the received vector falls outside the sphere, which has low probability.

Similarly, we can choose other codewords and their corresponding decoding spheres. How many such codewords can we choose? The volume of an n-dimensional sphere is of the form $C_n r^n$, where r is the radius of the sphere. In this case, each decoding sphere has radius $\sqrt{nN}$. These spheres are scattered throughout the space of received vectors. The received vectors have energy no greater than $n(P+N)$, so they lie in a sphere of radius $\sqrt{n(P+N)}$. The maximum number of nonintersecting decoding spheres in this volume is no more than

$$\frac{C_n(n(P+N))^{\frac{n}{2}}}{C_n(nN)^{\frac{n}{2}}} = 2^{\frac{n}{2}\log\left(1+\frac{P}{N}\right)} \tag{9.22}$$

and the rate of the code is $\frac{1}{2}\log(1+\frac{P}{N})$. This idea is illustrated in Figure 9.2.

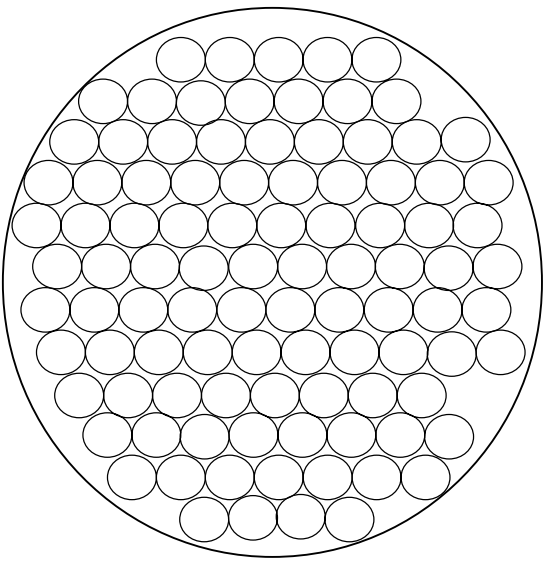


FIGURE 9.2. Sphere packing for the Gaussian channel.

This sphere-packing argument indicates that we cannot hope to send at rates greater than C with low probability of error. However, we can actually do almost as well as this, as is proved next.

Proof: (*Achievability*). We will use the same ideas as in the proof of the channel coding theorem in the case of discrete channels: namely, random codes and joint typicality decoding. However, we must make some modifications to take into account the power constraint and the fact that the variables are continuous and not discrete.

1. *Generation of the codebook*. We wish to generate a codebook in which all the codewords satisfy the power constraint. To ensure this, we generate the codewords with each element i.i.d. according to a normal distribution with variance $P - \epsilon$. Since for large n, $\frac{1}{n}\sum X_i^2 \to P - \epsilon$, the probability that a codeword does not satisfy the power constraint will be small. Let $X_i(w)$, $i = 1, 2, \ldots, n$, $w = 1, 2, \ldots, 2^{nR}$ be i.i.d. $\sim \mathcal{N}(0, P - \epsilon)$, forming codewords $X^n(1), X^n(2), \ldots, X^n(2^{nR}) \in \mathcal{R}^n$.
2. *Encoding*. After the generation of the codebook, the codebook is revealed to both the sender and the receiver. To send the message index w, the transmitter sends the wth codeword $X^n(w)$ in the codebook.
3. *Decoding*. The receiver looks down the list of codewords $\{X^n(w)\}$ and searches for one that is jointly typical with the received vector. If there is one and only one such codeword $X^n(w)$, the receiver declares $\hat{W} = w$ to be the transmitted codeword. Otherwise, the receiver declares an error. The receiver also declares an error if the chosen codeword does not satisfy the power constraint.
4. *Probability of error*. Without loss of generality, assume that codeword 1 was sent. Thus, $Y^n = X^n(1) + Z^n$. Define the following events:

$$E_0 = \left\{ \frac{1}{n} \sum_{j=1}^{n} X_j^2(1) > P \right\} \tag{9.23}$$

and

$$E_i = \left\{ (X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)} \right\}. \tag{9.24}$$

Then an error occurs if E_0 occurs (the power constraint is violated) or E_1^c occurs (the transmitted codeword and the received sequence are not jointly typical) or $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$ occurs (some wrong

codeword is jointly typical with the received sequence). Let $\mathcal{E}$ denote the event $\hat{W} \neq W$ and let P denote the conditional probability given that $W = 1$. Hence,

$$\Pr(\mathcal{E}|W=1) = P(\mathcal{E}) = P\left(E_0 \cup E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}\right) \tag{9.25}$$

$$\leq P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i), \tag{9.26}$$

by the union of events bound for probabilities. By the law of large numbers, $P(E_0) \to 0$ as $n \to \infty$. Now, by the joint AEP (which can be proved using the same argument as that used in the discrete case), $P(E_1^c) \to 0$, and hence

$$P(E_1^c) \leq \epsilon \quad \text{for } n \text{ sufficiently large.} \tag{9.27}$$

Since by the code generation process, $X^n(1)$ and $X^n(i)$ are independent, so are Y^n and $X^n(i)$. Hence, the probability that $X^n(i)$ and Y^n will be jointly typical is $\leq 2^{-n(I(X;Y)-3\epsilon)}$ by the joint AEP.
Now let W be uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$, and consequently,

$$\Pr(\mathcal{E}) = \frac{1}{2^{nR}} \sum \lambda_i = P_e^{(n)}. \tag{9.28}$$

Then

$$P_e^{(n)} = \Pr(\mathcal{E}) = \Pr(\mathcal{E}|W=1) \tag{9.29}$$

$$\leq P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \tag{9.30}$$

$$\leq \epsilon + \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \tag{9.31}$$

$$= 2\epsilon + \left(2^{nR} - 1\right) 2^{-n(I(X;Y)-3\epsilon)} \tag{9.32}$$

$$\leq 2\epsilon + 2^{3n\epsilon} 2^{-n(I(X;Y)-R)} \tag{9.33}$$

$$\leq 3\epsilon \tag{9.34}$$

for n sufficiently large and $R < I(X;Y) - 3\epsilon$. This proves the existence of a good $(2^{nR}, n)$ code.

Now choosing a good codebook and deleting the worst half of the codewords, we obtain a code with low maximal probability of error. In particular, the power constraint is satisfied by each of the remaining codewords (since the codewords that do not satisfy the power constraint have probability of error 1 and must belong to the worst half of the codewords). Hence we have constructed a code that achieves a rate arbitrarily close to capacity. The forward part of the theorem is proved. In the next section we show that the achievable rate cannot exceed the capacity. □

9.2 CONVERSE TO THE CODING THEOREM FOR GAUSSIAN CHANNELS

In this section we complete the proof that the capacity of a Gaussian channel is $C = \frac{1}{2}\log(1 + \frac{P}{N})$ by proving that rates $R > C$ are not achievable. The proof parallels the proof for the discrete channel. The main new ingredient is the power constraint.

Proof: (*Converse to Theorem 9.1.1*). We must show that if $P_e^{(n)} \to 0$ for a sequence of $(2^{nR}, n)$ codes for a Gaussian channel with power constraint P, then

$$R \le C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right). \tag{9.35}$$

Consider any $(2^{nR}, n)$ code that satisfies the power constraint, that is,

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2(w) \le P, \tag{9.36}$$

for $w = 1, 2, \ldots, 2^{nR}$. Proceeding as in the converse for the discrete case, let W be distributed uniformly over $\{1, 2, \ldots, 2^{nR}\}$. The uniform distribution over the index set $W \in \{1, 2, \ldots, 2^{nR}\}$ induces a distribution on the input codewords, which in turn induces a distribution over the input alphabet. This specifies a joint distribution on $W \to X^n(W) \to Y^n \to \hat{W}$. To relate probability of error and mutual information, we can apply Fano's inequality to obtain

$$H(W|\hat{W}) \le 1 + nRP_e^{(n)} = n\epsilon_n, \tag{9.37}$$

where $\epsilon_n \to 0$ as $P_e^{(n)} \to 0$. Hence,

$$nR = H(W) = I(W; \hat{W}) + H(W|\hat{W}) \tag{9.38}$$

$$\leq I(W; \hat{W}) + n\epsilon_n \tag{9.39}$$

$$\leq I(X^n; Y^n) + n\epsilon_n \tag{9.40}$$

$$= h(Y^n) - h(Y^n|X^n) + n\epsilon_n \tag{9.41}$$

$$= h(Y^n) - h(Z^n) + n\epsilon_n \tag{9.42}$$

$$\leq \sum_{i=1}^{n} h(Y_i) - h(Z^n) + n\epsilon_n \tag{9.43}$$

$$= \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Z_i) + n\epsilon_n \tag{9.44}$$

$$= \sum_{i=1}^{n} I(X_i; Y_i) + n\epsilon_n. \tag{9.45}$$

Here $X_i = x_i(W)$, where W is drawn according to the uniform distribution on $\{1, 2, \ldots, 2^{nR}\}$. Now let P_i be the average power of the ith column of the codebook, that is,

$$P_i = \frac{1}{2^{nR}} \sum_{w} x_i^2(w). \tag{9.46}$$

Then, since $Y_i = X_i + Z_i$ and since X_i and Z_i are independent, the average power EY_i^2 of Y_i is $P_i + N$. Hence, since entropy is maximized by the normal distribution,

$$h(Y_i) \leq \frac{1}{2} \log 2\pi e(P_i + N). \tag{9.47}$$

Continuing with the inequalities of the converse, we obtain

$$nR \leq \sum (h(Y_i) - h(Z_i)) + n\epsilon_n \tag{9.48}$$

$$\leq \sum \left(\frac{1}{2} \log(2\pi e(P_i + N)) - \frac{1}{2} \log 2\pi eN \right) + n\epsilon_n \tag{9.49}$$

$$= \sum \frac{1}{2} \log \left(1 + \frac{P_i}{N} \right) + n\epsilon_n. \tag{9.50}$$

Since each of the codewords satisfies the power constraint, so does their average, and hence

$$\frac{1}{n}\sum_i P_i \le P. \tag{9.51}$$

Since $f(x) = \frac{1}{2}\log(1+x)$ is a concave function of x, we can apply Jensen's inequality to obtain

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\log\left(1+\frac{P_i}{N}\right) \le \frac{1}{2}\log\left(1+\frac{1}{n}\sum_{i=1}^{n}\frac{P_i}{N}\right) \tag{9.52}$$

$$\le \frac{1}{2}\log\left(1+\frac{P}{N}\right). \tag{9.53}$$

Thus $R \le \frac{1}{2}\log(1+\frac{P}{N}) + \epsilon_n$, $\epsilon_n \to 0$, and we have the required converse.

Note that the power constraint enters the standard proof in (9.46).

9.3 BANDLIMITED CHANNELS

A common model for communication over a radio network or a telephone line is a bandlimited channel with white noise. This is a continuous-time channel. The output of such a channel can be described as the convolution

$$Y(t) = (X(t) + Z(t)) * h(t), \tag{9.54}$$

where $X(t)$ is the signal waveform, $Z(t)$ is the waveform of the white Gaussian noise, and $h(t)$ is the impulse response of an ideal bandpass filter, which cuts out all frequencies greater than W. In this section we give simplified arguments to calculate the capacity of such a channel.

We begin with a representation theorem due to Nyquist [396] and Shannon [480], which shows that sampling a bandlimited signal at a sampling rate $\frac{1}{2W}$ is sufficient to reconstruct the signal from the samples. Intuitively, this is due to the fact that if a signal is bandlimited to W, it cannot change by a substantial amount in a time less than half a cycle of the maximum frequency in the signal, that is, the signal cannot change very much in time intervals less than $\frac{1}{2W}$ seconds.

Theorem 9.3.1 *Suppose that a function $f(t)$ is bandlimited to W, namely, the spectrum of the function is 0 for all frequencies greater than W. Then the function is completely determined by samples of the function spaced $\frac{1}{2W}$ seconds apart.*

Proof: Let $F(\omega)$ be the Fourier transform of $f(t)$. Then

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t}\, d\omega \tag{9.55}$$

$$= \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} F(\omega) e^{i\omega t}\, d\omega, \tag{9.56}$$

since $F(\omega)$ is zero outside the band $-2\pi W \le \omega \le 2\pi W$. If we consider samples spaced $\frac{1}{2W}$ seconds apart, the value of the signal at the sample points can be written

$$f\left(\frac{n}{2W}\right) = \frac{1}{2\pi} \int_{-2\pi W}^{2\pi W} F(\omega) e^{i\omega \frac{n}{2W}}\, d\omega. \tag{9.57}$$

The right-hand side of this equation is also the definition of the coefficients of the Fourier series expansion of the periodic extension of the function $F(\omega)$, taking the interval $-2\pi W$ to $2\pi W$ as the fundamental period. Thus, the sample values $f(\frac{n}{2W})$ determine the Fourier coefficients and, by extension, they determine the value of $F(\omega)$ in the interval $(-2\pi W, 2\pi W)$. Since a function is uniquely specified by its Fourier transform, and since $F(\omega)$ is zero outside the band W, we can determine the function uniquely from the samples.

Consider the function

$$\text{sinc}(t) = \frac{\sin(2\pi W t)}{2\pi W t}. \tag{9.58}$$

This function is 1 at $t = 0$ and is 0 for $t = n/2W$, $n \neq 0$. The spectrum of this function is constant in the band $(-W, W)$ and is zero outside this band. Now define

$$g(t) = \sum_{n=-\infty}^{\infty} f\left(\frac{n}{2W}\right) \text{sinc}\left(t - \frac{n}{2W}\right). \tag{9.59}$$

From the properties of the sinc function, it follows that $g(t)$ is bandlimited to W and is equal to $f(n/2W)$ at $t = n/2W$. Since there is only

one function satisfying these constraints, we must have $g(t) = f(t)$. This provides an explicit representation of $f(t)$ in terms of its samples. □

A general function has an infinite number of degrees of freedom—the value of the function at every point can be chosen independently. The Nyquist–Shannon sampling theorem shows that a bandlimited function has only $2W$ degrees of freedom per second. The values of the function at the sample points can be chosen independently, and this specifies the entire function.

If a function is bandlimited, it cannot be limited in time. But we can consider functions that have most of their energy in bandwidth W and have most of their energy in a finite time interval, say $(0, T)$. We can describe these functions using a basis of *prolate spheroidal functions*. We do not go into the details of this theory here; it suffices to say that there are about $2TW$ orthonormal basis functions for the set of almost time-limited, almost bandlimited functions, and we can describe any function within the set by its coordinates in this basis. The details can be found in a series of papers by Landau, Pollak, and Slepian [340, 341, 500]. Moreover, the projection of white noise on these basis vectors forms an i.i.d. Gaussian process. The above arguments enable us to view the bandlimited, time-limited functions as vectors in a vector space of $2TW$ dimensions.

Now we return to the problem of communication over a bandlimited channel. Assuming that the channel has bandwidth W, we can represent both the input and the output by samples taken $1/2W$ seconds apart. Each of the input samples is corrupted by noise to produce the corresponding output sample. Since the noise is white and Gaussian, it can be shown that each noise sample is an independent, identically distributed Gaussian random variable.

If the noise has power spectral density $N_0/2$ watts/hertz and bandwidth W hertz, the noise has power $\frac{N_0}{2}2W = N_0W$ and each of the $2WT$ noise samples in time T has variance $N_0WT/2WT = N_0/2$. Looking at the input as a vector in the $2TW$-dimensional space, we see that the received signal is spherically normally distributed about this point with covariance $\frac{N_0}{2}I$.

Now we can use the theory derived earlier for discrete-time Gaussian channels, where it was shown that the capacity of such a channel is

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) \qquad \text{bits per transmission.} \tag{9.60}$$

Let the channel be used over the time interval $[0, T]$. In this case, the energy per sample is $PT/2WT = P/2W$, the noise variance per sample

is $\frac{N_0}{2} 2W \frac{T}{2WT} = N_0/2$, and hence the capacity per sample is

$$C = \frac{1}{2} \log \left(\frac{1 + \frac{P}{2W}}{\frac{N_0}{2}} \right) = \frac{1}{2} \log \left(1 + \frac{P}{N_0 W} \right) \qquad \text{bits per sample.} \tag{9.61}$$

Since there are $2W$ samples each second, the capacity of the channel can be rewritten as

$$C = W \log \left(1 + \frac{P}{N_0 W} \right) \qquad \text{bits per second.} \tag{9.62}$$

This equation is one of the most famous formulas of information theory. It gives the capacity of a bandlimited Gaussian channel with noise spectral density $N_0/2$ watts/Hz and power P watts.

A more precise version of the capacity argument [576] involves consideration of signals with a small fraction of their energy outside the bandwidth W of the channel and a small fraction of their energy outside the time interval $(0, T)$. The capacity above is then obtained as a limit as the fraction of energy outside the band goes to zero.

If we let $W \to \infty$ in (9.62), we obtain

$$C = \frac{P}{N_0} \log_2 e \qquad \text{bits per second} \tag{9.63}$$

as the capacity of a channel with an infinite bandwidth, power P, and noise spectral density $N_0/2$. Thus, for infinite bandwidth channels, the capacity grows linearly with the power.

Example 9.3.1 (Telephone line) To allow multiplexing of many channels, telephone signals are bandlimited to 3300 Hz. Using a bandwidth of 3300 Hz and a SNR (signal-to-noise ratio) of 33 dB (i.e., $P/N_0 W = 2000$) in (9.62), we find the capacity of the telephone channel to be about 36,000 bits per second. Practical modems achieve transmission rates up to 33,600 bits per second in both directions over a telephone channel. In real telephone channels, there are other factors, such as crosstalk, interference, echoes, and nonflat channels which must be compensated for to achieve this capacity.

The V.90 modems that achieve 56 kb/s over the telephone channel achieve this rate in only one direction, taking advantage of a purely digital channel from the server to final telephone switch in the network. In this case, the only impairments are due to the digital-to-analog conversion at this switch and the noise in the copper link from the switch to the home;

these impairments reduce the maximum bit rate from the 64 kb/s for the digital signal in the network to the 56 kb/s in the best of telephone lines.

The actual bandwidth available on the copper wire that links a home to a telephone switch is on the order of a few megahertz; it depends on the length of the wire. The frequency response is far from flat over this band. If the entire bandwidth is used, it is possible to send a few megabits per second through this channel; schemes such at DSL (Digital Subscriber Line) achieve this using special equipment at both ends of the telephone line (unlike modems, which do not require modification at the telephone switch).

9.4 PARALLEL GAUSSIAN CHANNELS

In this section we consider k independent Gaussian channels in parallel with a common power constraint. The objective is to distribute the total power among the channels so as to maximize the capacity. This channel models a nonwhite additive Gaussian noise channel where each parallel component represents a different frequency.

Assume that we have a set of Gaussian channels in parallel as illustrated in Figure 9.3. The output of each channel is the sum of the input and Gaussian noise. For channel j,

$$Y_j = X_j + Z_j, \quad j = 1, 2, \ldots, k, \tag{9.64}$$

with

$$Z_j \sim \mathcal{N}(0, N_j), \tag{9.65}$$

and the noise is assumed to be independent from channel to channel. We assume that there is a common power constraint on the total power used, that is,

$$E \sum_{j=1}^{k} X_j^2 \leq P. \tag{9.66}$$

We wish to distribute the power among the various channels so as to maximize the total capacity.

The information capacity of the channel C is

$$C = \max_{f(x_1, x_2, \ldots, x_k): \sum E X_i^2 \leq P} I(X_1, X_2, \ldots, X_k; Y_1, Y_2, \ldots, Y_k). \tag{9.67}$$