Elements of Information Theory
Thomas M. Cover, Joy A. Thomas
Copyright © 1991 John Wiley & Sons, Inc.
Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1

Chapter 14

Network Information Theory

A system with many senders and receivers contains many new elements in the communication problem: interference, cooperation and feedback. These are the issues that are the domain of network information theory. The general problem is easy to state. Given many senders and receivers and a channel transition matrix which describes the effects of the interference and the noise in the network, decide whether or not the sources can be transmitted over the channel. This problem involves distributed source coding (data compression) as well as distributed communication (finding the capacity region of the network). This general problem has not yet been solved, so we consider various special cases in this chapter.

Examples of large communication networks include computer networks, satellite networks and the phone system. Even within a single computer, there are various components that talk to each other. A complete theory of network information would have wide implications for the design of communication and computer networks.

Suppose that m stations wish to communicate with a common satellite over a common channel, as shown in Figure 14.1. This is known as a multiple access channel. How do the various senders cooperate with each other to send information to the receiver? What rates of communication are simultaneously achievable? What limitations does interference among the senders put on the total rate of communication? This is the best understood multi-user channel, and the above questions have satisfying answers.

In contrast, we can reverse the network and consider one TV station sending information to m TV receivers, as in Figure 14.2. How does the sender encode information meant for different receivers in a common

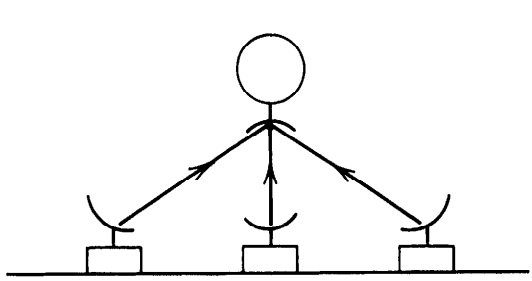


Figure 14.1. A multiple access channel.

signal? What are the rates at which information can be sent to the different receivers? For this channel, the answers are known only in special cases.

There are other channels such as the relay channel (where there is one source and one destination, but one or more intermediate sender-receiver pairs that act as relays to facilitate the communication between the source and the destination), the interference channel (two senders and two receivers with crosstalk) or the two-way channel (two sender-receiver pairs sending information to each other). For all these channels, we only have some of the answers to questions about achievable communication rates and the appropriate coding strategies.

All these channels can be considered special cases of a general communication network that consists of m nodes trying to communicate with each other, as shown in Figure 14.3. At each instant of time, the ith node sends a symbol x_i that depends on the messages that it wants to send and on past received symbols at the node. The simultaneous transmission of the symbols $(x_1, x_2, \ldots, x_m)$ results in random received symbols $(Y_1, Y_2, \ldots, Y_m)$ drawn according to the conditional probability distribution $p(y^{(1)}, y^{(2)}, \ldots, y^{(m)} | x^{(1)}, x^{(2)}, \ldots, x^{(m)})$. Here $p(\cdot \mid \cdot)$ expresses the effects of the noise and interference present in the network. If $p(\cdot \mid \cdot)$ takes on only the values 0 and 1, the network is deterministic.

Associated with some of the nodes in the network are stochastic data sources, which are to be communicated to some of the other nodes in the network. If the sources are independent, the messages sent by the nodes

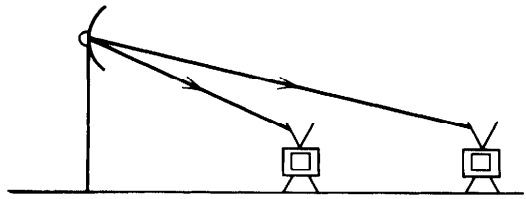


Figure 14.2. A broadcast channel.

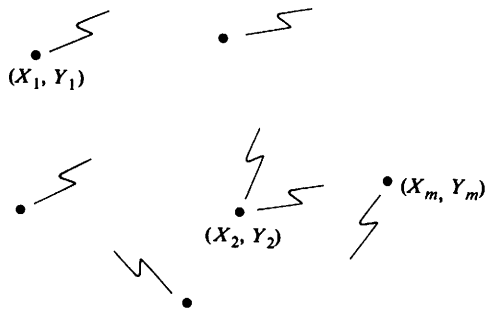
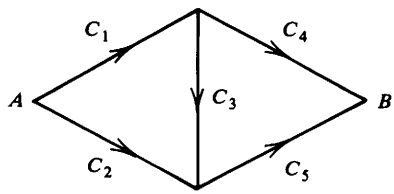


Figure 14.3. A communication network.

are also independent. However, for full generality, we must allow the sources to be dependent. How does one take advantage of the dependence to reduce the amount of information transmitted? Given the probability distribution of the sources and the channel transition function, can one transmit these sources over the channel and recover the sources at the destinations with the appropriate distortion?

We consider various special cases of network communication. We consider the problem of source coding when the channels are noiseless and without interference. In such cases, the problem reduces to finding the set of rates associated with each source such that the required sources can be decoded at the destination with low probability of error (or appropriate distortion). The simplest case for distributed source coding is the Slepian-Wolf source coding problem, where we have two sources which must be encoded separately, but decoded together at a common node. We consider extensions to this theory when only one of the two sources needs to be recovered at the destination.

The theory of flow in networks has satisfying answers in domains like circuit theory and the flow of water in pipes. For example, for the single-source single-sink network of pipes shown in Figure 14.4, the maximum flow from A to B can be easily computed from the Ford-Fulkerson theorem. Assume that the edges have capacities C_i as shown. Clearly, the maximum flow across any cut-set cannot be greater than



$$C = \min\{C_1 + C_2, C_2 + C_3 + C_4, C_4 + C_5, C_1 + C_3 + C_5\}$$

Figure 14.4. Network of water pipes.

the sum of the capacities of the cut edges. Thus minimizing the maximum flow across cut-sets yields an upper bound on the capacity of the network. The Ford-Fulkerson [113] theorem shows that this capacity can be achieved.

The theory of information flow in networks does not have the same simple answers as the theory of flow of water in pipes. Although we prove an upper bound on the rate of information flow across any cut-set, these bounds are not achievable in general. However, it is gratifying that some problems like the relay channel and the cascade channel admit a simple max flow min cut interpretation. Another subtle problem in the search for a general theory is the absence of a source-channel separation theorem, which we will touch on briefly in the last section of this chapter. A complete theory combining distributed source coding and network channel coding is still a distant goal.

In the next section, we consider Gaussian examples of some of the basic channels of network information theory. The physically motivated Gaussian channel lends itself to concrete and easily interpreted answers. Later we prove some of the basic results about joint typicality that we use to prove the theorems of multiuser information theory. We then consider various problems in detail—the multiple access channel, the coding of correlated sources (Slepian-Wolf data compression), the broadcast channel, the relay channel, the coding of a random variable with side information and the rate distortion problem with side information. We end with an introduction to the general theory of information flow in networks. There are a number of open problems in the area, and there does not yet exist a comprehensive theory of information networks. Even if such a theory is found, it may be too complex for easy implementation. But the theory will be able to tell communication designers how close they are to optimality and perhaps suggest some means of improving the communication rates.

14.1 GAUSSIAN MULTIPLE USER CHANNELS

Gaussian multiple user channels illustrate some of the important features of network information theory. The intuition gained in Chapter 10 on the Gaussian channel should make this section a useful introduction. Here the key ideas for establishing the capacity regions of the Gaussian multiple access, broadcast, relay and two-way channels will be given without proof. The proofs of the coding theorems for the discrete memoryless counterparts to these theorems will be given in later sections of this chapter.

The basic discrete time additive white Gaussian noise channel with input power P and noise variance N is modeled by

$$Y_i = X_i + Z_i, \qquad i = 1, 2, \ldots \tag{14.1}$$

where the Z_i are i.i.d. Gaussian random variables with mean 0 and variance N. The signal $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ has a power constraint

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 \leq P. \tag{14.2}$$

The Shannon capacity C is obtained by maximizing $I(X; Y)$ over all random variables X such that $EX^2 \leq P$, and is given (Chapter 10) by

$$C = \frac{1}{2} \log\left(1 + \frac{P}{N}\right) \text{ bits per transmission}. \tag{14.3}$$

In this chapter we will restrict our attention to discrete-time memoryless channels; the results can be extended to continuous time Gaussian channels.

14.1.1 Single User Gaussian Channel

We first review the single user Gaussian channel studied in Chapter 10. Here $Y = X + Z$. Choose a rate $R < \frac{1}{2} \log(1 + \frac{P}{N})$. Fix a good $(2^{nR}, n)$ codebook of power P. Choose an index i in the set 2^{nR}. Send the ith codeword $\mathbf{X}(i)$ from the codebook generated above. The receiver observes $\mathbf{Y} = \mathbf{X}(i) + \mathbf{Z}$ and then finds the index $\hat{i}$ of the closest codeword to $\mathbf{Y}$. If n is sufficiently large, the probability of error $\Pr(i \neq \hat{i})$ will be arbitrarily small. As can be seen from the definition of joint typicality, this minimum distance decoding scheme is essentially equivalent to finding the codeword in the codebook that is jointly typical with the received vector $\mathbf{Y}$.

14.1.2 The Gaussian Multiple Access Channel with m Users

We consider m transmitters, each with a power P. Let

$$Y = \sum_{i=1}^{m} X_i + Z. \tag{14.4}$$

Let

$$C\left(\frac{P}{N}\right) = \frac{1}{2} \log\left(1 + \frac{P}{N}\right) \tag{14.5}$$

denote the capacity of a single user Gaussian channel with signal to noise ratio P/N. The achievable rate region for the Gaussian channel takes on the simple form given in the following equations:

$$R_i < C\left(\frac{P}{N}\right) \tag{14.6}$$

$$R_i + R_j < C\left(\frac{2P}{N}\right) \tag{14.7}$$

$$R_i + R_j + R_k < C\left(\frac{3P}{N}\right) \tag{14.8}$$

$$\vdots \tag{14.9}$$

$$\sum_{i=1}^{m} R_i < C\left(\frac{mP}{N}\right). \tag{14.10}$$

Note that when all the rates are the same, the last inequality dominates the others.

Here we need m codebooks, the ith codebook having 2^{nR_i} codewords of power P. Transmission is simple. Each of the independent transmitters chooses an arbitrary codeword from its own codebook. The users simultaneously send these vectors. The receiver sees these codewords added together with the Gaussian noise $\mathbf{Z}$.

Optimal decoding consists of looking for the m codewords, one from each codebook, such that the vector sum is closest to $\mathbf{Y}$ in Euclidean distance. If $(R_1, R_2, \ldots, R_m)$ is in the capacity region given above, then the probability of error goes to 0 as n tends to infinity.

Remarks: It is exciting to see in this problem that the sum of the rates of the users $C(mP/N)$ goes to infinity with m. Thus in a cocktail party with m celebrants of power P in the presence of ambient noise N, the intended listener receives an unbounded amount of information as the number of people grows to infinity. A similar conclusion holds, of course, for ground communications to a satellite.

It is also interesting to note that the optimal transmission scheme here does not involve time division multiplexing. In fact, each of the transmitters uses all of the bandwidth all of the time.

14.1.3 The Gaussian Broadcast Channel

Here we assume that we have a sender of power P and two distant receivers, one with Gaussian noise power N_1 and the other with Gaussian noise power N_2. Without loss of generality, assume $N_1 < N_2$. Thus receiver Y_1 is less noisy than receiver Y_2. The model for the channel is $Y_1 = X + Z_1$ and $Y_2 = X + Z_2$, where Z_1 and Z_2 are arbitrarily correlated Gaussian random variables with variances N_1 and N_2, respectively. The sender wishes to send independent messages at rates R_1 and R_2 to receivers Y_1 and Y_2, respectively.

Fortunately, all Gaussian broadcast channels belong to the class of degraded broadcast channels discussed in Section 14.6.2. Specializing that work, we find that the capacity region of the Gaussian broadcast channel is

$$R_1 < C\left(\frac{\alpha P}{N_1}\right) \tag{14.11}$$

$$R_2 < C\left(\frac{(1-\alpha)P}{\alpha P + N_2}\right) \tag{14.12}$$

where α may be arbitrarily chosen ($0 \leq \alpha \leq 1$) to trade off rate R_1 for rate R_2 as the transmitter wishes.

To encode the messages, the transmitter generates two codebooks, one with power αP at rate R_1, and another codebook with power $\bar{\alpha}P$ at rate R_2, where R_1 and R_2 lie in the capacity region above. Then to send an index $i \in \{1, 2, \ldots, 2^{nR_1}\}$ and $j \in \{1, 2, \ldots, 2^{nR_2}\}$ to Y_1 and Y_2, respectively, the transmitter takes the codeword $\mathbf{X}(i)$ from the first codebook and codeword $\mathbf{X}(j)$ from the second codebook and computes the sum. He sends the sum over the channel.

The receivers must now decode their messages. First consider the bad receiver Y_2. He merely looks through the second codebook to find the closest codeword to the received vector $\mathbf{Y}_2$. His effective signal to noise ratio is $\bar{\alpha}P/(\alpha P + N_2)$, since Y_1's message acts as noise to Y_2. (This can be proved.)

The good receiver Y_1 first decodes Y_2's codeword, which he can accomplish because of his lower noise N_1. He subtracts this codeword $\hat{\mathbf{X}}_2$ from $\mathbf{Y}_1$. He then looks for the codeword in the first codebook closest to $\mathbf{Y}_1 - \hat{\mathbf{X}}_2$. The resulting probability of error can be made as low as desired.

A nice dividend of optimal encoding for degraded broadcast channels is that the better receiver Y_1 always knows the message intended for receiver Y_2 in addition to the message intended for himself.

14.1.4 The Gaussian Relay Channel

For the relay channel, we have a sender X and an ultimate intended receiver Y. Also present is the relay channel intended solely to help the receiver. The Gaussian relay channel (Figure 14.30) is given by

$$Y_1 = X + Z_1, \tag{14.13}$$

$$Y = X + Z_1 + X_1 + Z_2, \tag{14.14}$$

where Z_1 and Z_2 are independent zero mean Gaussian random variables with variance N_1 and N_2, respectively. The allowed encoding by the relay is the causal sequence

$$X_{1i} = f_i(Y_{11}, Y_{12}, \ldots, Y_{1i-1}). \tag{14.15}$$

The sender X has power P and sender X_1 has power P_1. The capacity is

$$C = \max_{0 \le \alpha \le 1} \min\left\{ C\left(\frac{P + P_1 + 2\sqrt{\bar{\alpha} P P_1}}{N_1 + N_2}\right), C\left(\frac{\alpha P}{N_1}\right)\right\}, \tag{14.16}$$

where $\bar{\alpha} = 1 - \alpha$. Note that if

$$\frac{P_1}{N_2} \ge \frac{P}{N_1}, \tag{14.17}$$

it can be seen that $C = C(P/N_1)$,which is achieved by $\alpha = 1$. The channel appears to be noise-free after the relay, and the capacity $C(P/N_1)$ from X to the relay can be achieved. Thus the rate $C(P/(N_1 + N_2))$ without the relay is increased by the presence of the relay to $C(P/N_1)$. For large N_2, and for $P_1/N_2 \ge P/N_1$, we see that the increment in rate is from $C(P/(N_1 + N_2)) \approx 0$ to $C(P/N_1)$.

Let $R_1 < C(\alpha P/N_1)$. Two codebooks are needed. The first codebook has 2^{nR_1} words of power αP. The second has 2^{nR_0} codewords of power $\bar{\alpha} P$. We shall use codewords from these codebooks successively in order to create the opportunity for cooperation by the relay. We start by sending a codeword from the first codebook. The relay now knows the index of this codeword since $R_1 < C(\alpha P/N_1)$, but the intended receiver has a list of possible codewords of size $2^{n(R_1 - C(\alpha P/(N_1+N_2)))}$. This list calculation involves a result on list codes.

In the next block, the transmitter and the relay wish to cooperate to resolve the receiver's uncertainty about the previously sent codeword on the receiver's list. Unfortunately, they cannot be sure what this list is because they do not know the received signal Y. Thus they randomly partition the first codebook into 2^{nR_0} cells with an equal number of codewords in each cell. The relay, the receiver, and the transmitter agree on this partition. The relay and the transmitter find the cell of the partition in which the codeword from the first codebook lies and cooperatively send the codeword from the second codebook with that index. That is, both X and X_1 send the same designated codeword. The relay, of course, must scale this codeword so that it meets his power constraint P_1. They now simultaneously transmit their codewords. An important point to note here is that the cooperative information sent by the relay and the transmitter is sent coherently. So the power of the sum as seen by the receiver Y is $(\sqrt{\bar{\alpha} P} + \sqrt{P_1})^2$.

However, this does not exhaust what the transmitter does in the second block. He also chooses a fresh codeword from the first codebook, adds it "on paper" to the cooperative codeword from the second codebook, and sends the sum over the channel.

The reception by the ultimate receiver Y in the second block involves first finding the cooperative index from the second codebook by looking for the closest codeword in the second codebook. He subtracts the codeword from the received sequence, and then calculates a list of

indices of size 2^{nR_0} corresponding to all codewords of the first codebook that might have been sent in the second block.

Now it is time for the intended receiver to complete computing the codeword from the first codebook sent in the first block. He takes his list of possible codewords that might have been sent in the first block and intersects it with the cell of the partition that he has learned from the cooperative relay transmission in the second block. The rates and powers have been chosen so that it is highly probable that there is only one codeword in the intersection. This is Y's guess about the information sent in the first block.

We are now in steady state. In each new block, the transmitter and the relay cooperate to resolve the list uncertainty from the previous block. In addition, the transmitter superimposes some fresh information from his first codebook to this transmission from the second codebook and transmits the sum.

The receiver is always one block behind, but for sufficiently many blocks, this does not affect his overall rate of reception.

14.1.5 The Gaussian Interference Channel

The interference channel has two senders and two receivers. Sender 1 wishes to send information to receiver 1. He does not care what receiver 2 receives or understands. Similarly with sender 2 and receiver 2. Each channel interferes with the other. This channel is illustrated in Figure 14.5. It is not quite a broadcast channel since there is only one intended receiver for each sender, nor is it a multiple access channel because each receiver is only interested in what is being sent by the corresponding transmitter. For symmetric interference, we have

$$Y_1 = X_1 + aX_2 + Z_1 \tag{14.18}$$

$$Y_2 = X_2 + aX_1 + Z_2, \tag{14.19}$$

where Z_1, Z_2 are independent $\mathcal{N}(0, N)$ random variables. This channel

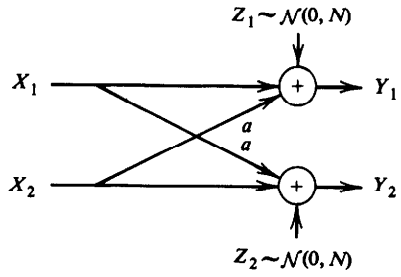


Figure 14.5. The Gaussian interference channel.

has not been solved in general even in the Gaussian case. But remarkably, in the case of high interference, it can be shown that the capacity region of this channel is the same as if there were no interference whatsoever.

To achieve this, generate two codebooks, each with power P and rate $C(P/N)$. Each sender independently chooses a word from his book and sends it. Now, if the interference a satisfies $C(a^2P/(P+N)) > C(P/N)$, the first transmitter perfectly understands the index of the second transmitter. He finds it by the usual technique of looking for the closest codeword to his received signal. Once he finds this signal, he subtracts it from his received waveform. Now there is a clean channel between him and his sender. He then searches the sender's codebook to find the closest codeword and declares that codeword to be the one sent.

14.1.6. The Gaussian Two-Way Channel

The two-way channel is very similar to the interference channel, with the additional provision that sender 1 is attached to receiver 2 and sender 2 is attached to receiver 1 as shown in Figure 14.6. Hence, sender 1 can use information from previous received symbols of receiver 2 to decide what to send next. This channel introduces another fundamental aspect of network information theory, namely, feedback. Feedback enables the senders to use the partial information that each has about the other's message to cooperate with each other.

The capacity region of the two-way channel is not known in general. This channel was first considered by Shannon [246], who derived upper and lower bounds on the region. (See Problem 15 at the end of this chapter.) For Gaussian channels, these two bounds coincide and the capacity region is known; in fact, the Gaussian two-way channel decomposes into two independent channels.

Let P_1 and P_2 be the powers of transmitters 1 and 2 respectively and let N_1 and N_2 be the noise variances of the two channels. Then the rates $R_1 < C(P_1/N_1)$ and $R_2 < C(P_2/N_2)$ can be achieved by the techniques described for the interference channel. In this case, we generate two codebooks of rates R_1 and R_2. Sender 1 sends a codeword from the first codebook. Receiver 2 receives the sum of the codewords sent by the two senders plus some noise. He simply subtracts out the codeword of sender

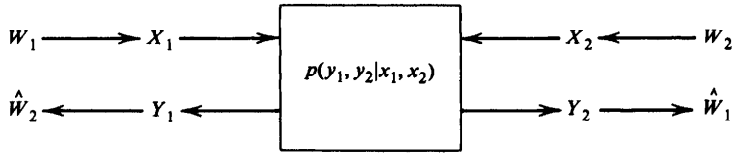


Figure 14.6. The two-way channel.

2 and he has a clean channel from sender 1 (with only the noise of variance N_1). Hence the two-way Gaussian channel decomposes into two independent Gaussian channels. But this is not the case for the general two-way channel; in general there is a trade-off between the two senders so that both of them cannot send at the optimal rate at the same time.

14.2 JOINTLY TYPICAL SEQUENCES

We have previewed the capacity results for networks by considering multi-user Gaussian channels. We will begin a more detailed analysis in this section, where we extend the joint AEP proved in Chapter 8 to a form that we will use to prove the theorems of network information theory. The joint AEP will enable us to calculate the probability of error for jointly typical decoding for the various coding schemes considered in this chapter.

Let $(X_1, X_2, \ldots, X_k)$ denote a finite collection of discrete random variables with some fixed joint distribution, $p(x_1, x_2, \ldots, x_k)$, $(x_1, x_2, \ldots, x_k) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_k$. Let S denote an ordered subset of these random variables and consider n independent copies of S. Thus

$$\Pr\{S = s\} = \prod_{i=1}^{n} \Pr\{S_i = s_i\}, \qquad s \in \mathcal{S}^n. \tag{14.20}$$

For example, if $S = (X_j, X_l)$, then

$$\Pr\{S = s\} = \Pr\{(\mathbf{X}_j, \mathbf{X}_l) = (\mathbf{x}_j, \mathbf{x}_l)\} \tag{14.21}$$

$$= \prod_{i=1}^{n} p(x_{ji}, x_{li}). \tag{14.22}$$

To be explicit, we will sometimes use $X(S)$ for S. By the law of large numbers, for any subset S of random variables,

$$-\frac{1}{n} \log p(S_1, S_2, \ldots, S_n) = -\frac{1}{n} \sum_{i=1}^{n} \log p(S_i) \to H(S), \tag{14.23}$$

where the convergence takes place simultaneously with probability 1 for all 2^k subsets, $S \subseteq \{X_1, X_2, \ldots, X_k\}$.

Definition: The set $A_\epsilon^{(n)}$ of ϵ-typical n-sequences $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$ is defined by

$$A_\epsilon^{(n)}(X_1, X_2, \ldots, X_k)$$

$$= A_\epsilon^{(n)}$$

$$= \left\{(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k): \right.$$

$$\left. \left| -\frac{1}{n} \log p(\mathbf{s}) - H(S) \right| < \epsilon, \quad \forall S \subseteq \{X_1, X_2, \ldots, X_k\} \right\}. \tag{14.24}$$

Let $A_\epsilon^{(n)}(S)$ denote the restriction of $A_\epsilon^{(n)}$ to the coordinates of S. Thus if $S = (X_1, X_2)$, we have

$$A_\epsilon^{(n)}(X_1, X_2) = \left\{(\mathbf{x}_1, \mathbf{x}_2): \right.$$

$$\left| -\frac{1}{n} \log p(\mathbf{x}_1, \mathbf{x}_2) - H(X_1, X_2) \right| < \epsilon,$$

$$\left| -\frac{1}{n} \log p(\mathbf{x}_1) - H(X_1) \right| < \epsilon,$$

$$\left. \left| -\frac{1}{n} \log p(\mathbf{x}_2) - H(X_2) \right| < \epsilon \right\}. \tag{14.25}$$

Definition: We will use the notation $a_n \doteq 2^{n(b \pm \epsilon)}$ to mean

$$\left| \frac{1}{n} \log a_n - b \right| < \epsilon \tag{14.26}$$

for n sufficiently large.

Theorem 14.2.1: *For any $\epsilon > 0$, for sufficiently large n,*

1. $$P(A_\epsilon^{(n)}(S)) \geq 1 - \epsilon, \quad \forall S \subseteq \{X_1, X_2, \ldots, X_k\}. \tag{14.27}$$

2. $$\mathbf{s} \in A_\epsilon^{(n)}(S) \Rightarrow p(\mathbf{s}) \doteq 2^{-n(H(S) \pm \epsilon)}. \tag{14.28}$$

3. $$|A_\epsilon^{(n)}(S)| \doteq 2^{n(H(S) \pm 2\epsilon)}. \tag{14.29}$$

4. *Let $S_1, S_2 \subseteq \{X_1, X_2, \ldots, X_k\}$. If $(\mathbf{s}_1, \mathbf{s}_2) \in A_\epsilon^{(n)}(S_1, S_2)$, then*

$$p(\mathbf{s}_1 | \mathbf{s}_2) \doteq 2^{-n(H(S_1|S_2) \pm 2\epsilon)}. \tag{14.30}$$

Proof:

1. This follows from the law of large numbers for the random variables in the definition of $A_\epsilon^{(n)}(S)$.

2. This follows directly from the definition of $A_\epsilon^{(n)}(S)$.
3. This follows from

$$1 \geq \sum_{\mathbf{s} \in A_\epsilon^{(n)}(S)} p(\mathbf{s}) \tag{14.31}$$

$$\geq \sum_{\mathbf{s} \in A_\epsilon^{(n)}(S)} 2^{-n(H(S)+\epsilon)} \tag{14.32}$$

$$= |A_\epsilon^{(n)}(S)| 2^{-n(H(S)+\epsilon)} . \tag{14.33}$$

If n is sufficiently large, we can argue that

$$1 - \epsilon \leq \sum_{\mathbf{s} \in A_\epsilon^{(n)}(S)} p(\mathbf{s}) \tag{14.34}$$

$$\leq \sum_{\mathbf{s} \in A_\epsilon^{(n)}(S)} 2^{-n(H(S)-\epsilon)} \tag{14.35}$$

$$= |A_\epsilon^{(n)}(S)| 2^{-n(H(S)-\epsilon)} . \tag{14.36}$$

Combining (14.33) and (14.36), we have $|A_\epsilon^{(n)}(S)| \doteq 2^{n(H(S)\pm 2\epsilon)}$ for sufficiently large n.

4. For $(\mathbf{s}_1, \mathbf{s}_2) \in A_\epsilon^{(n)}(S_1, S_2)$, we have $p(\mathbf{s}_1) \doteq 2^{-n(H(S_1)\pm\epsilon)}$ and $p(\mathbf{s}_1, \mathbf{s}_2) \doteq 2^{-n(H(S_1, S_2)\pm\epsilon)}$. Hence

$$p(\mathbf{s}_2|\mathbf{s}_1) = \frac{p(\mathbf{s}_1, \mathbf{s}_2)}{p(\mathbf{s}_1)} \doteq 2^{-n(H(S_2|S_1)\pm 2\epsilon)} . \quad \square \tag{14.37}$$

The next theorem bounds the number of conditionally typical sequences for a given typical sequence.

Theorem 14.2.2: *Let S_1, S_2 be two subsets of $X_1, X_2, \ldots, X_k$. For any $\epsilon > 0$, define $A_\epsilon^{(n)}(S_1|\mathbf{s}_2)$ to be the set of $\mathbf{s}_1$ sequences that are jointly ϵ-typical with a particular $\mathbf{s}_2$ sequence. If $\mathbf{s}_2 \in A_\epsilon^{(n)}(S_2)$, then for sufficiently large n, we have*

$$|A_\epsilon^{(n)}(S_1|\mathbf{s}_2)| \leq 2^{n(H(S_1|S_2)+2\epsilon)} , \tag{14.38}$$

and

$$(1 - \epsilon) 2^{n(H(S_1|S_2)-2\epsilon)} \leq \sum_{\mathbf{s}_2} p(\mathbf{s}_2) |A_\epsilon^{(n)}(S_1|\mathbf{s}_2)| . \tag{14.39}$$

Proof: As in part 3 of the previous theorem, we have

$$1 \geq \sum_{\mathbf{s}_1 \in A_\epsilon^{(n)}(S_1|\mathbf{s}_2)} p(\mathbf{s}_1|\mathbf{s}_2) \tag{14.40}$$

$$\geq \sum_{\mathbf{s}_1 \in A_\epsilon^{(n)}(S_1|\mathbf{s}_2)} 2^{-n(H(S_1|S_2)+2\epsilon)} \tag{14.41}$$

$$= |A_\epsilon^{(n)}(S_1|\mathbf{s}_2)| 2^{-n(H(S_1|S_2)+2\epsilon)} . \tag{14.42}$$

If n is sufficiently large, then we can argue from (14.27) that

$$1 - \epsilon \leq \sum_{\mathbf{s}_2} p(\mathbf{s}_2) \sum_{\mathbf{s}_1 \in A_\epsilon^{(n)}(S_1|\mathbf{s}_2)} p(\mathbf{s}_1|\mathbf{s}_2) \tag{14.43}$$

$$\leq \sum_{\mathbf{s}_2} p(\mathbf{s}_2) \sum_{\mathbf{s}_1 \in A_\epsilon^{(n)}(S_1|\mathbf{s}_2)} 2^{-n(H(S_1|S_2)-2\epsilon)} \tag{14.44}$$

$$= \sum_{\mathbf{s}_2} p(\mathbf{s}_2) |A_\epsilon^{(n)}(S_1|\mathbf{s}_2)| 2^{-n(H(S_1|S_2)-2\epsilon)} . \quad \square \tag{14.45}$$

To calculate the probability of decoding error, we need to know the probability that conditionally independent sequences are jointly typical. Let S_1, S_2 and S_3 be three subsets of $\{X_1, X_2, \ldots, X_k\}$. If S_1' and S_2' are conditionally independent given S_3' but otherwise share the same pairwise marginals of (S_1, S_2, S_3), we have the following probability of joint typicality.

Theorem 14.2.3: *Let $A_\epsilon^{(n)}$ denote the typical set for the probability mass function $p(s_1, s_2, s_3)$, and let*

$$P(\mathbf{S}_1' = \mathbf{s}_1, \mathbf{S}_2' = \mathbf{s}_2, \mathbf{S}_3' = \mathbf{s}_3) = \prod_{i=1}^{n} p(s_{1i}|s_{3i}) p(s_{2i}|s_{3i}) p(s_{3i}) . \tag{14.46}$$

Then

$$P\{(\mathbf{S}_1', \mathbf{S}_2', \mathbf{S}_3') \in A_\epsilon^{(n)}\} \doteq 2^{n(I(S_1; S_2|S_3)\pm 6\epsilon)} . \tag{14.47}$$

Proof: We use the $\doteq$ notation from (14.26) to avoid calculating the upper and lower bounds separately. We have

$$P\{(\mathbf{S}_1', \mathbf{S}_2', \mathbf{S}_3') \in A_\epsilon^{(n)}\}$$

$$= \sum_{(\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3) \in A_\epsilon^{(n)}} p(\mathbf{s}_3) p(\mathbf{s}_1|\mathbf{s}_3) p(\mathbf{s}_2|\mathbf{s}_3) \tag{14.48}$$

$$\doteq |A_\epsilon^{(n)}(S_1, S_2, S_3)| 2^{-n(H(S_3)\pm\epsilon)} 2^{-n(H(S_1|S_3)\pm 2\epsilon)} 2^{-n(H(S_2|S_3)\pm 2\epsilon)} \tag{14.49}$$

$$\doteq 2^{n(H(S_1, S_2, S_3)\pm\epsilon)} 2^{-n(H(S_3)\pm\epsilon)} 2^{-n(H(S_1|S_3)\pm 2\epsilon)} 2^{-n(H(S_2|S_3)\pm 2\epsilon)} \tag{14.50}$$

$$\doteq 2^{-n(I(S_1; S_2|S_3)\pm 6\epsilon)} . \quad \square \tag{14.51}$$

We will specialize this theorem to particular choices of S_1, S_2 and S_3 for the various achievability proofs in this chapter.

14.3 THE MULTIPLE ACCESS CHANNEL

The first channel that we examine in detail is the multiple access channel, in which two (or more) senders send information to a common receiver. The channel is illustrated in Figure 14.7.

A common example of this channel is a satellite receiver with many independent ground stations. We see that the senders must contend not only with the receiver noise but with interference from each other as well.

Definition: A *discrete memoryless multiple access channel* consists of three alphabets, $\mathcal{X}_1$, $\mathcal{X}_2$ and $\mathcal{Y}$, and a probability transition matrix $p(y|x_1, x_2)$.

Definition: A $((2^{nR_1}, 2^{nR_2}), n)$ code for the multiple access channel consists of two sets of integers $\mathcal{W}_1 = \{1, 2, \ldots, 2^{nR_1}\}$ and $\mathcal{W}_2 = \{1, 2, \ldots, 2^{nR_2}\}$ called the *message sets*, two *encoding functions*,

$$X_1 : \mathcal{W}_1 \to \mathcal{X}_1^n , \tag{14.52}$$

$$X_2 : \mathcal{W}_2 \to \mathcal{X}_2^n \tag{14.53}$$

and a *decoding function*

$$g : \mathcal{Y}^n \to \mathcal{W}_1 \times \mathcal{W}_2 . \tag{14.54}$$

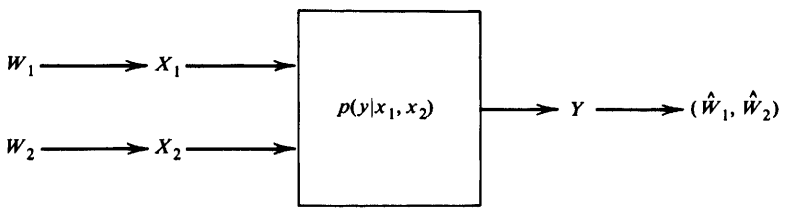


Figure 14.7. The multiple access channel.

There are two senders and one receiver for this channel. Sender 1 chooses an index W_1 uniformly from the set $\{1, 2, \ldots, 2^{nR_1}\}$ and sends the corresponding codeword over the channel. Sender 2 does likewise. Assuming that the distribution of messages over the product set $\mathcal{W}_1 \times \mathcal{W}_2$ is uniform, i.e., the messages are independent and equally likely, we define the *average probability of error* for the $((2^{nR_1}, 2^{nR_2}), n)$ code as follows:

$$P_e^{(n)} = \frac{1}{2^{n(R_1+R_2)}} \sum_{(w_1, w_2) \in \mathcal{W}_1 \times \mathcal{W}_2} \Pr\{g(Y^n) \neq (w_1, w_2) | (w_1, w_2) \text{ sent}\} . \tag{14.55}$$

Definition: A rate pair (R_1, R_2) is said to be *achievable* for the multiple access channel if there exists a sequence of $((2^{nR_1}, 2^{nR_2}), n)$ codes with $P_e^{(n)} \to 0$.

Definition: The *capacity region* of the multiple access channel is the closure of the set of achievable (R_1, R_2) rate pairs.

An example of the capacity region for a multiple access channel is illustrated in Figure 14.8.

We first state the capacity region in the form of a theorem.

Theorem 14.3.1 (*Multiple access channel capacity*): *The capacity of a multiple access channel* $(\mathcal{X}_1 \times \mathcal{X}_2, p(y|x_1, x_2), \mathcal{Y})$ *is the closure of the convex hull of all* (R_1, R_2) *satisfying*

$$R_1 < I(X_1; Y|X_2) , \tag{14.56}$$

$$R_2 < I(X_2; Y|X_1) , \tag{14.57}$$

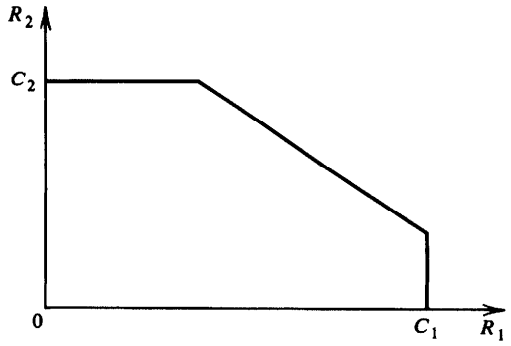


Figure 14.8. Capacity region for a multiple access channel.

$$R_1 + R_2 < I(X_1, X_2; Y) \tag{14.58}$$

for some product distribution $p_1(x_1)p_2(x_2)$ *on* $\mathcal{X}_1 \times \mathcal{X}_2$.

Before we prove that this is the capacity region of the multiple access channel, let us consider a few examples of multiple access channels:

Example 14.3.1 (*Independent binary symmetric channels*): Assume that we have two independent binary symmetric channels, one from sender 1 and the other from sender 2, as shown in Figure 14.9.

In this case, it is obvious from the results of Chapter 8 that we can send at rate $1 - H(p_1)$ over the first channel and at rate $1 - H(p_2)$ over the second channel. Since the channels are independent, there is no interference between the senders. The capacity region in this case is shown in Figure 14.10.

Example 14.3.2 (*Binary multiplier channel*): Consider a multiple access channel with binary inputs and output

$$Y = X_1 X_2 . \tag{14.59}$$

Such a channel is called a binary multiplier channel. It is easy to see that by setting $X_2 = 1$, we can send at a rate of 1 bit per transmission from sender 1 to the receiver. Similarly, setting $X_1 = 1$, we can achieve $R_2 = 1$. Clearly, since the output is binary, the combined rates $R_1 + R_2$ of

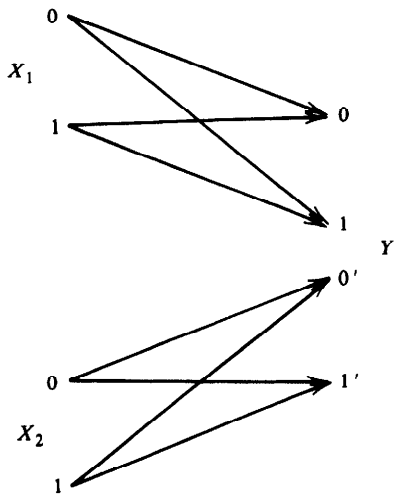


Figure 14.9. Independent binary symmetric channels.

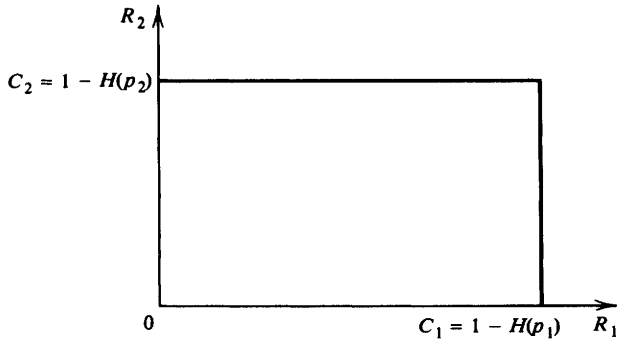


Figure 14.10. Capacity region for independent BSC's.

sender 1 and sender 2 cannot be more than 1 bit. By timesharing, we can achieve any combination of rates such that $R_1 + R_2 = 1$. Hence the capacity region is as shown in Figure 14.11.

Example 14.3.3 (*Binary erasure multiple access channel*): This multiple access channel has binary inputs, $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$ and a ternary output $Y = X_1 + X_2$. There is no ambiguity in (X_1, X_2) if $Y = 0$ or $Y = 2$ is received; but $Y = 1$ can result from either $(0, 1)$ or $(1, 0)$.

We now examine the achievable rates on the axes. Setting $X_2 = 0$, we can send at a rate of 1 bit per transmission from sender 1. Similarly, setting $X_1 = 0$, we can send at a rate $R_2 = 1$. This gives us two extreme points of the capacity region.

Can we do better? Let us assume that $R_1 = 1$, so that the codewords of X_1 must include all possible binary sequences; X_1 would look like a

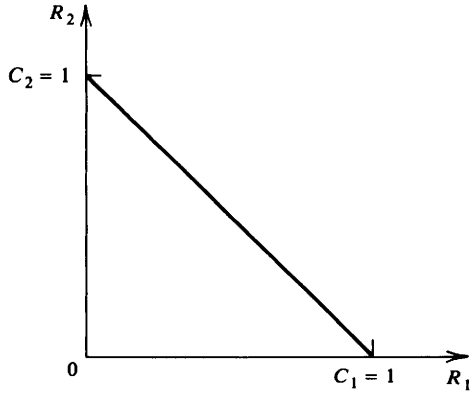


Figure 14.11. Capacity region for binary multiplier channel.

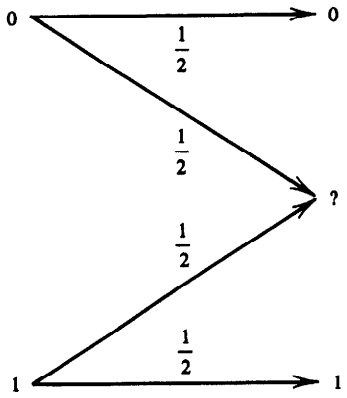


Figure 14.12. Equivalent single user channel for user 2 of a binary erasure multiple access channel.

Bernoulli($\frac{1}{2}$) process. This acts like noise for the transmission from X_2. For X_2, the channel looks like the channel in Figure 14.12.

This is the binary erasure channel of Chapter 8. Recalling the results, the capacity of this channel is $\frac{1}{2}$ bit per transmission.

Hence when sending at maximum rate 1 for sender 1, we can send an additional $\frac{1}{2}$ bit from sender 2. Later on, after deriving the capacity region, we can verify that these rates are the best that can be achieved.

The capacity region for a binary erasure channel is illustrated in Figure 14.13.

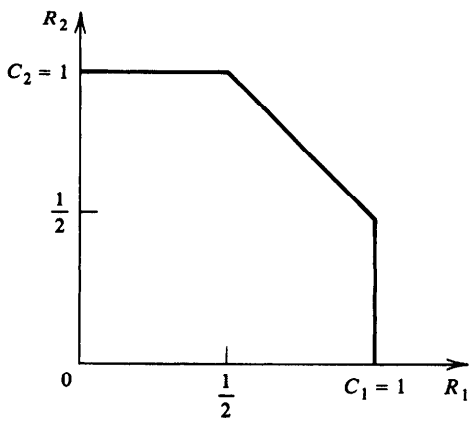


Figure 14.13. Capacity region for binary erasure multiple access channel.

14.3.1 Achievability of the Capacity Region for the Multiple Access Channel

We now prove the achievability of the rate region in Theorem 14.3.1; the proof of the converse will be left until the next section. The proof of achievability is very similar to the proof for the single user channel. We will therefore only emphasize the points at which the proof differs from the single user case. We will begin by proving the achievability of rate pairs that satisfy (14.58) for some fixed product distribution $p(x_1)p(x_2)$. In Section 14.3.3, we will extend this to prove that all points in the convex hull of (14.58) are achievable.

Proof (*Achievability in Theorem 14.3.1*): Fix $p(x_1, x_2) = p_1(x_1)p_2(x_2)$.

Codebook generation. Generate 2^{nR_1} independent codewords $\mathbf{X}_1(i)$, $i \in \{1, 2, \ldots, 2^{nR_1}\}$, of length n, generating each element i.i.d. $\sim \prod_{i=1}^{n} p_1(x_{1i})$. Similarly, generate 2^{nR_2} independent codewords $\mathbf{X}_2(j)$, $j \in \{1, 2, \ldots, 2^{nR_2}\}$, generating each element i.i.d. $\sim \prod_{i=1}^{n} p_2(x_{2i})$. These codewords form the codebook, which is revealed to the senders and the receiver.

Encoding. To send index i, sender 1 sends the codeword $\mathbf{X}_1(i)$. Similarly, to send j, sender 2 sends $\mathbf{X}_2(j)$.

Decoding. Let $A_\epsilon^{(n)}$ denote the set of typical $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y})$ sequences. The receiver Y^n chooses the pair (i, j) such that

$$(\mathbf{x}_1(i), \mathbf{x}_2(j), \mathbf{y}) \in A_\epsilon^{(n)} \tag{14.60}$$

if such a pair (i, j) exists and is unique; otherwise, an error is declared.

Analysis of the probability of error. By the symmetry of the random code construction, the conditional probability of error does not depend on which pair of indices is sent. Thus the conditional probability of error is the same as the unconditional probability of error. So, without loss of generality, we can assume that $(i, j) = (1, 1)$ was sent.

We have an error if either the correct codewords are not typical with the received sequence or there is a pair of incorrect codewords that are typical with the received sequence. Define the events

$$E_{ij} = \{(\mathbf{X}_1(i), \mathbf{X}_2(j), \mathbf{Y}) \in A_\epsilon^{(n)}\} . \tag{14.61}$$

Then by the union of events bound,

$$P_e^{(n)} = P(E_{11}^c \cup \cup_{(i,j)\neq(1,1)} E_{ij}) \tag{14.62}$$

$$\leq P(E_{11}^c) + \sum_{i\neq 1, j=1} P(E_{i1}) + \sum_{i=1, j\neq 1} P(E_{1j}) + \sum_{i\neq 1, j\neq 1} P(E_{ij}), \tag{14.63}$$

where P is the conditional probability given that $(1, 1)$ was sent. From the AEP, $P(E_{11}^c) \to 0$.

By Theorem 14.2.1 and Theorem 14.2.3, for $i \neq 1$, we have

$$P(E_{i1}) = P((\mathbf{X}_1(i), \mathbf{X}_2(1), \mathbf{Y}) \in A_\epsilon^{(n)}) \tag{14.64}$$

$$= \sum_{(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) \in A_\epsilon^{(n)}} p(\mathbf{x}_1) p(\mathbf{x}_2, \mathbf{y}) \tag{14.65}$$

$$\leq |A_\epsilon^{(n)}| 2^{-n(H(X_1)-\epsilon)} 2^{-n(H(X_2, Y)-\epsilon)} \tag{14.66}$$

$$\leq 2^{-n(H(X_1)+H(X_2, Y)-H(X_1, X_2, Y)-3\epsilon)} \tag{14.67}$$

$$= 2^{-n(I(X_1; X_2, Y)-3\epsilon)} \tag{14.68}$$

$$= 2^{-n(I(X_1; Y|X_2)-3\epsilon)}, \tag{14.69}$$

since X_1 and X_2 are independent, and therefore $I(X_1; X_2, Y) = I(X_1; X_2) + I(X_1; Y|X_2) = I(X_1; Y|X_2)$.

Similarly, for $j \neq 1$,

$$P(E_{1j}) \leq 2^{-n(I(X_2; Y|X_1)-3\epsilon)}, \tag{14.70}$$

and for $i \neq 1, j \neq 1$,

$$P(E_{ij}) \leq 2^{-n(I(X_1, X_2; Y)-4\epsilon)}. \tag{14.71}$$

It follows that

$$P_e^{(n)} \leq P(E_{11}^c) + 2^{nR_1} 2^{-n(I(X_1; Y|X_2)-3\epsilon)} + 2^{nR_2} 2^{-n(I(X_2; Y|X_1)-3\epsilon)} + 2^{n(R_1+R_2)} 2^{-n(I(X_1, X_2; Y)-4\epsilon)}. \tag{14.72}$$

Since $\epsilon > 0$ is arbitrary, the conditions of the theorem imply that each term tends to 0 as $n \to \infty$.

The above bound shows that the average probability of error, averaged over all choices of codebooks in the random code construction, is arbitrarily small. Hence there exists at least one code $\mathscr{C}^*$ with arbitrarily small probability of error.

This completes the proof of achievability of the region in (14.58) for a fixed input distribution. Later, in Section 14.3.3, we will show that

timesharing allows any (R_1, R_2) in the convex hull to be achieved, completing the proof of the forward part of the theorem. □

14.3.2 Comments on the Capacity Region for the Multiple Access Channel

We have now proved the achievability of the capacity region of the multiple access channel, which is the closure of the convex hull of the set of points (R_1, R_2) satisfying

$$R_1 < I(X_1; Y|X_2), \tag{14.73}$$

$$R_2 < I(X_2; Y|X_1), \tag{14.74}$$

$$R_1 + R_2 < I(X_1, X_2; Y) \tag{14.75}$$

for some distribution $p_1(x_1)p_2(x_2)$ on $\mathcal{X}_1 \times \mathcal{X}_2$.

For a particular $p_1(x_1)p_2(x_2)$, the region is illustrated in Figure 14.14.

Let us now interpret the corner points in the region. The point A corresponds to the maximum rate achievable from sender 1 to the receiver when sender 2 is not sending any information. This is

$$\max R_1 = \max_{p_1(x_1)p_2(x_2)} I(X_1; Y|X_2). \tag{14.76}$$

Now for any distribution $p_1(x_1)p_2(x_2)$,

$$I(X_1; Y|X_2) = \sum_{x_2} p_2(x_2) I(X_1; Y|X_2 = x_2) \tag{14.77}$$

$$\leq \max_{x_2} I(X_1; Y|X_2 = x_2), \tag{14.78}$$

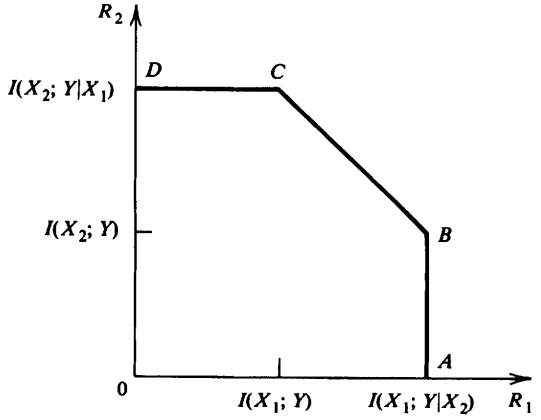


Figure 14.14. Achievable region of multiple access channel for a fixed input distribution.

since the average is less than the maximum. Therefore, the maximum in (14.76) is attained when we set $X_2 = x_2$, where x_2 is the value that maximizes the conditional mutual information between X_1 and Y. The distribution of X_1 is chosen to maximize this mutual information. Thus X_2 must facilitate the transmission of X_1 by setting $X_2 = x_2$.

The point B corresponds to the maximum rate at which sender 2 can send as long as sender 1 sends at his maximum rate. This is the rate that is obtained if X_1 is considered as noise for the channel from X_2 to Y. In this case, using the results from single user channels, X_2 can send at a rate $I(X_2; Y)$. The receiver now knows which X_2 codeword was used and can "subtract" its effect from the channel. We can consider the channel now to be an indexed set of single user channels, where the index is the X_2 symbol used. The X_1 rate achieved in this case is the average mutual information, where the average is over these channels, and each channel occurs as many times as the corresponding X_2 symbol appears in the codewords. Hence the rate achieved is

$$\sum_{x_2} p(x_2) I(X_1; Y|X_2 = x_2) = I(X_1; Y|X_2) . \tag{14.79}$$

The points C and D correspond to B and A respectively with the roles of the senders reversed.

The non-corner points can be achieved by timesharing. Thus, we have given a single user interpretation and justification for the capacity region of a multiple access channel.

The idea of considering other signals as part of the noise, decoding one signal and then "subtracting" it from the received signal is a very useful one. We will come across the same concept again in the capacity calculations for the degraded broadcast channel.

14.3.3 Convexity of the Capacity Region of the Multiple Access Channel

We now recast the capacity region of the multiple access channel in order to take into account the operation of taking the convex hull by introducing a new random variable. We begin by proving that the capacity region is convex.

Theorem 14.3.2: *The capacity region* $\mathscr{C}$ of a multiple access channel is convex, i.e., if $(R_1, R_2) \in \mathscr{C}$ *and* $(R_1', R_2') \in \mathscr{C}$, *then* $(\lambda R_1 + (1-\lambda)R_1', \lambda R_2 + (1-\lambda)R_2') \in \mathscr{C}$ *for* $0 \le \lambda \le 1$.

Proof: The idea is timesharing. Given two sequences of codes at different rates $\mathbf{R} = (R_1, R_2)$ and $\mathbf{R}' = (R_1', R_2')$, we can construct a third codebook at a rate $\lambda\mathbf{R} + (1-\lambda)\mathbf{R}'$ by using the first codebook for the first λn symbols and using the second codebook for the last $(1-\lambda)n$ symbols. The number of X_1 codewords in the new code is

$$2^{n\lambda R_1}2^{n(1-\lambda)R_1'} = 2^{n(\lambda R_1 + (1-\lambda)R_1')} \tag{14.80}$$

and hence the rate of the new code is $\lambda\mathbf{R} + (1-\lambda)\mathbf{R}'$. Since the overall probability of error is less than the sum of the probabilities of error for each of the segments, the probability of error of the new code goes to 0 and the rate is achievable. □

We will now recast the statement of the capacity region for the multiple access channel using a timesharing random variable Q.

Theorem 14.3.3: *The set of achievable rates of a discrete memoryless multiple access channel is given by the closure of the set of all* (R_1, R_2) *pairs satisfying*

$$R_1 < I(X_1; Y|X_2, Q),$$

$$R_2 < I(X_2; Y|X_1, Q),$$

$$R_1 + R_2 < I(X_1, X_2; Y|Q) \tag{14.81}$$

for some choice of the joint distribution $p(q)p(x_1|q)p(x_2|q)p(y|x_1, x_2)$ *with* $|\mathcal{Q}| \leq 4$.

Proof: We will show that every rate pair lying in the region in the theorem is achievable, i.e., it lies in the convex closure of the rate pairs satisfying Theorem 14.3.1. We will also show that every point in the convex closure of the region in Theorem 14.3.1 is also in the region defined in (14.81).

Consider a rate point $\mathbf{R}$ satisfying the inequalities (14.81) of the theorem. We can rewrite the right hand side of the first inequality as

$$I(X_1; Y|X_2, Q) = \sum_{k=1}^{m} p(q)I(X_1; Y|X_2, Q = q) \tag{14.82}$$

$$= \sum_{k=1}^{m} p(q)I(X_1; Y|X_2)_{p_{1q}, p_{2q}}, \tag{14.83}$$

where m is the cardinality of the support set of Q. We can similarly expand the other mutual informations in the same way.

For simplicity in notation, we will consider a rate pair as a vector and denote a pair satisfying the inequalities in (14.58) for a specific input product distribution $p_{1q}(x_1)p_{2q}(x_2)$ as $\mathbf{R}_q$. Specifically, let $\mathbf{R}_q = (R_{1q}, R_{2q})$ be a rate pair satisfying

$$R_{1q} < I(X_1; Y|X_2)_{p_{1q}(x_1)p_{2q}(x_2)}, \tag{14.84}$$

$$R_{2q} < I(X_2; Y|X_1)_{p_{1q}(x_1)p_{2q}(x_2)}, \tag{14.85}$$

$$R_{1q} + R_{2q} < I(X_1, X_2; Y)_{p_{1q}(x_1)p_{2q}(x_2)} . \tag{14.86}$$

Then by Theorem 14.3.1, $\mathbf{R}_q = (R_{1q}, R_{2q})$ is achievable.

Then since $\mathbf{R}$ satisfies (14.81), and we can expand the right hand sides as in (14.83), there exists a set of pairs $\mathbf{R}_q$ satisfying (14.86) such that

$$\mathbf{R} = \sum_{q=1}^{m} p(q)\mathbf{R}_q . \tag{14.87}$$

Since a convex combination of achievable rates is achievable, so is $\mathbf{R}$. Hence we have proved the achievability of the region in the theorem. The same argument can be used to show that every point in the convex closure of the region in (14.58) can be written as the mixture of points satisfying (14.86) and hence can be written in the form (14.81).

The converse will be proved in the next section. The converse shows that all achievable rate pairs are of the form (14.81), and hence establishes that this is the capacity region of the multiple access channel.

The cardinality bound on the time-sharing random variable Q is a consequence of Carathéodory's theorem on convex sets. See the discussion below. □

The proof of the convexity of the capacity region shows that any convex combination of achievable rate pairs is also achievable. We can continue this process, taking convex combinations of more points. Do we need to use an arbitrary number of points ? Will the capacity region be increased? The following theorem says no.

Theorem 14.3.4 (*Carathéodory*): *Any point in the convex closure of a connected compact set A in a d dimensional Euclidean space can be represented as a convex combination of $d+1$ or fewer points in the original set A.*

Proof: The proof can be found in Eggleston [95] and Grünbaum [127], and is omitted here. □

This theorem allows us to restrict attention to a certain finite convex combination when calculating the capacity region. This is an important property because without it we would not be able to compute the capacity region in (14.81), since we would never know whether using a larger alphabet $\mathcal{Q}$ would increase the region.

In the multiple access channel, the bounds define a connected compact set in three dimensions. Therefore all points in its closure can be

defined as the convex combination of four points. Hence, we can restrict the cardinality of Q to at most 4 in the above definition of the capacity region.

14.3.4 Converse for the Multiple Access Channel

We have so far proved the achievability of the capacity region. In this section, we will prove the converse.

Proof (*Converse to Theorem 14.3.1 and Theorem 14.3.3*): We must show that given any sequence of $((2^{nR_1}, 2^{nR_2}), n)$ codes with $P_e^{(n)} \to 0$, that the rates must satisfy

$$R_1 \le I(X_1; Y|X_2, Q),$$

$$R_2 \le I(X_2; Y|X_1, Q),$$

$$R_1 + R_2 \le I(X_1, X_2; Y|Q) \tag{14.88}$$

for some choice of random variable Q defined on $\{1, 2, 3, 4\}$ and joint distribution $p(q)p(x_1|q)p(x_2|q)p(y|x_1, x_2)$.

Fix n. Consider the given code of block length n. The joint distribution on $\mathcal{W}_1 \times \mathcal{W}_2 \times \mathcal{X}_1^n \times \mathcal{X}_2^n \times \mathcal{Y}^n$ is well defined. The only randomness is due to the random uniform choice of indices W_1 and W_2 and the randomness induced by the channel. The joint distribution is

$$p(w_1, w_2, x_1^n, x_2^n, y^n) = \frac{1}{2^{nR_1}} \frac{1}{2^{nR_2}} p(x_1^n|w_1)p(x_2^n|w_2) \prod_{i=1}^{n} p(y_i|x_{1i}, x_{2i}), \tag{14.89}$$

where $p(x_1^n|w_1)$ is either 1 or 0 depending on whether $x_1^n = \mathbf{x}_1(w_1)$, the codeword corresponding to w_1, or not, and similarly, $p(x_2^n|w_2) = 1$ or 0 according to whether $x_2^n = \mathbf{x}_2(w_2)$ or not. The mutual informations that follow are calculated with respect to this distribution.

By the code construction, it is possible to estimate (W_1, W_2) from the received sequence Y^n with a low probability of error. Hence the conditional entropy of (W_1, W_2) given Y^n must be small. By Fano's inequality,

$$H(W_1, W_2|Y^n) \le n(R_1 + R_2)P_e^{(n)} + H(P_e^{(n)}) \triangleq n\epsilon_n. \tag{14.90}$$

It is clear that $\epsilon_n \to 0$ as $P_e^{(n)} \to 0$.

Then we have

$$H(W_1|Y^n) \le H(W_1, W_2|Y^n) \le n\epsilon_n, \tag{14.91}$$

$$H(W_2|Y^n) \le H(W_1, W_2|Y^n) \le n\epsilon_n . \tag{14.92}$$

We can now bound the rate R_1 as

$$nR_1 = H(W_1) \tag{14.93}$$

$$= I(W_1; Y^n) + H(W_1|Y^n) \tag{14.94}$$

$$\overset{(a)}{\le} I(W_1; Y^n) + n\epsilon_n \tag{14.95}$$

$$\overset{(b)}{\le} I(X_1^n(W_1); Y^n) + n\epsilon_n \tag{14.96}$$

$$= H(X_1^n(W_1)) - H(X_1^n(W_1)|Y^n) + n\epsilon_n \tag{14.97}$$

$$\overset{(c)}{\le} H(X_1^n(W_1)|X_2^n(W_2)) - H(X_1^n(W_1)|Y^n, X_2^n(W_2)) + n\epsilon_n \tag{14.98}$$

$$= I(X_1^n(W_1); Y^n|X_2^n(W_2)) + n\epsilon_n \tag{14.99}$$

$$= H(Y^n|X_2^n(W_2)) - H(Y^n|X_1^n(W_1), X_2^n(W_2)) + n\epsilon_n \tag{14.100}$$

$$\overset{(d)}{=} H(Y^n|X_2^n(W_2)) - \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X_1^n(W_1), X_2^n(W_2)) + n\epsilon_n \tag{14.101}$$

$$\overset{(e)}{=} H(Y^n|X_2^n(W_2)) - \sum_{i=1}^{n} H(Y_i|X_{1i}, X_{2i}) + n\epsilon_n \tag{14.102}$$

$$\overset{(f)}{\le} \sum_{i=1}^{n} H(Y_i|X_2^n(W_2)) - \sum_{i=1}^{n} H(Y_i|X_{1i}, X_{2i}) + n\epsilon_n \tag{14.103}$$

$$\overset{(g)}{\le} \sum_{i=1}^{n} H(Y_i|X_{2i}) - \sum_{i=1}^{n} H(Y_i|X_{1i}, X_{2i}) + n\epsilon_n \tag{14.104}$$

$$= \sum_{i=1}^{n} I(X_{1i}; Y_i|X_{2i}) + n\epsilon_n , \tag{14.105}$$

where

(a) follows from Fano's inequality,

(b) from the data processing inequality,

(c) from the fact that since W_1 and W_2 are independent, so are $X_1^n(W_1)$ and $X_2^n(W_2)$, and hence $H(X_1^n(W_1)|X_2^n(W_2)) = H(X_1^n(W_1))$, and $H(X_1^n(W_1)|Y^n, X_2^n(W_2)) \le H(X_1^n(W_1)|Y^n)$ by conditioning,

(d) follows from the chain rule,

(e) from the fact that Y_i depends only on X_{1i} and X_{2i} by the memoryless property of the channel,

(f) from the chain rule and removing conditioning, and

(g) follows from removing conditioning.

Hence, we have

$$R_1 \le \frac{1}{n} \sum_{i=1}^{n} I(X_{1i}; Y_i|X_{2i}) + \epsilon_n . \tag{14.106}$$

Similarly, we have

$$R_2 \le \frac{1}{n} \sum_{i=1}^{n} I(X_{2i}; Y_i|X_{1i}) + \epsilon_n . \tag{14.107}$$

To bound the sum of the rates, we have

$$n(R_1 + R_2) = H(W_1, W_2) \tag{14.108}$$

$$= I(W_1, W_2; Y^n) + H(W_1, W_2|Y^n) \tag{14.109}$$

$$\overset{(a)}{\le} I(W_1, W_2; Y^n) + n\epsilon_n \tag{14.110}$$

$$\overset{(b)}{\le} I(X_1^n(W_1), X_2^n(W_2); Y^n) + n\epsilon_n \tag{14.111}$$

$$= H(Y^n) - H(Y^n|X_1^n(W_1), X_2^n(W_2)) + n\epsilon_n \tag{14.112}$$

$$\overset{(c)}{=} H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X_1^n(W_1), X_2^n(W_2)) + n\epsilon_n \tag{14.113}$$

$$\overset{(d)}{=} H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_{1i}, X_{2i}) + n\epsilon_n \tag{14.114}$$

$$\overset{(e)}{\le} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_{1i}, X_{2i}) + n\epsilon_n \tag{14.115}$$

$$= \sum_{i=1}^{n} I(X_{1i}, X_{2i}; Y_i) + n\epsilon_n , \tag{14.116}$$

where

(a) follows from Fano's inequality,
(b) from the data processing inequality,
(c) from the chain rule,
(d) from the fact that Y_i depends only on X_{1i} and X_{2i} and is conditionally independent of everything else, and
(e) follows from the chain rule and removing conditioning.

Hence we have

$$R_1 + R_2 \le \frac{1}{n} \sum_{i=1}^{n} I(X_{1i}, X_{2i}; Y_i) + \epsilon_n . \tag{14.117}$$

The expressions in (14.106), (14.107) and (14.117) are the averages of the mutual informations calculated at the empirical distributions in column i of the codebook. We can rewrite these equations with the new variable Q, where $Q = i \in \{1, 2, \ldots, n\}$ with probability $\frac{1}{n}$. The equations become

$$R_1 \leq \frac{1}{n} \sum_{i=1}^{n} I(X_{1i}; Y_i | X_{2i}) + \epsilon_n \tag{14.118}$$

$$= \frac{1}{n} \sum_{i=1}^{n} I(X_{1q}; Y_q | X_{2q}, Q = i) + \epsilon_n \tag{14.119}$$

$$= I(X_{1Q}; Y_Q | X_{2Q}, Q) + \epsilon_n \tag{14.120}$$

$$= I(X_1; Y | X_2, Q) + \epsilon_n , \tag{14.121}$$

where $X_1 \triangleq X_{1Q}$, $X_2 \triangleq X_{2Q}$ and $Y \triangleq Y_Q$ are new random variables whose distributions depend on Q in the same way as the distributions of X_{1i}, X_{2i} and Y_i depend on i. Since W_1 and W_2 are independent, so are $X_{1i}(W_1)$ and $X_{2i}(W_2)$, and hence

$$\Pr(X_{1i}(W_1) = x_1, X_{2i}(W_2) = x_2)$$

$$\triangleq \Pr\{X_{1Q} = x_1 | Q = i\} \Pr\{X_{2Q} = x_2 | Q = i\} . \tag{14.122}$$

Hence, taking the limit as $n \to \infty$, $P_e^{(n)} \to 0$, we have the following converse:

$$R_1 \leq I(X_1; Y | X_2, Q) ,$$

$$R_2 \leq I(X_2; Y | X_1, Q) ,$$

$$R_1 + R_2 \leq I(X_1, X_2; Y | Q) \tag{14.123}$$

for some choice of joint distribution $p(q)p(x_1|q)p(x_2|q)p(y|x_1, x_2)$.

As in the previous section, the region is unchanged if we limit the cardinality of $\mathcal{Q}$ to 4.

This completes the proof of the converse. □

Thus the achievability of the region of Theorem 14.3.1 was proved in Section 14.3.1. In Section 14.3.3, we showed that every point in the region defined by (14.88) was also achievable. In the converse, we showed that the region in (14.88) was the best we can do, establishing that this is indeed the capacity region of the channel. Thus the region in

(14.58) cannot be any larger than the region in (14.88), and this is the capacity region of the multiple access channel.

14.3.5 m-User Multiple Access Channels

We will now generalize the result derived for two senders to m senders, $m \geq 2$. The multiple access channel in this case is shown in Figure 14.15.

We send independent indices $w_1, w_2, \ldots, w_m$ over the channel from the senders $1, 2, \ldots, m$ respectively. The codes, rates and achievability are all defined in exactly the same way as the two sender case.

Let $S \subseteq \{1, 2, \ldots, m\}$. Let S^c denote the complement of S. Let $R(S) = \Sigma_{i \in S} R_i$, and let $X(S) = \{X_i : i \in S\}$. Then we have the following theorem.

Theorem 14.3.5: *The capacity region of the m-user multiple access channel is the closure of the convex hull of the rate vectors satisfying*

$$R(S) \leq I(X(S); Y|X(S^c)) \quad \text{for all } S \subseteq \{1, 2, \ldots, m\} \tag{14.124}$$

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

Proof: The proof contains no new ideas. There are now $2^m - 1$ terms in the probability of error in the achievability proof and an equal number of inequalities in the proof of the converse. Details are left to the reader. □

In general, the region in (14.124) is a beveled box.

14.3.6 Gaussian Multiple Access Channels

We now discuss the Gaussian multiple access channel of Section 14.1.2 in somewhat more detail.

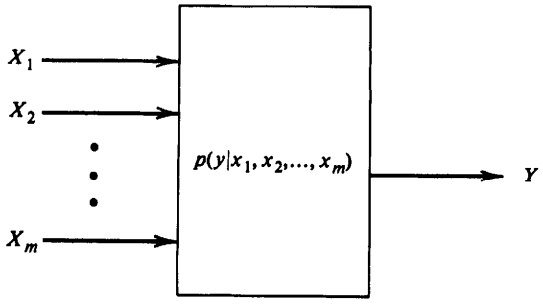


Figure 14.15. m-user multiple access channel.

There are two senders, X_1 and X_2, sending to the single receiver Y. The received signal at time i is

$$Y_i = X_{1i} + X_{2i} + Z_i, \tag{14.125}$$

where $\{Z_i\}$ is a sequence of independent, identically distributed, zero mean Gaussian random variables with variance N (Figure 14.16). We will assume that there is a power constraint P_j on sender j, i.e., for each sender, for all messages, we must have

$$\frac{1}{n}\sum_{i=1}^{n} x_{ji}^2(w_j) \le P_j, \quad w_j \in \{1, 2, \ldots, 2^{nR_j}\}, j = 1, 2. \tag{14.126}$$

Just as the proof of achievability of channel capacity for the discrete case (Chapter 8) was extended to the Gaussian channel (Chapter 10), we can extend the proof the discrete multiple access channel to the Gaussian multiple access channel. The converse can also be extended similarly, so we expect the capacity region to be the convex hull of the set of rate pairs satisfying

$$R_1 \le I(X_1; Y|X_2), \tag{14.127}$$

$$R_2 \le I(X_2; Y|X_1), \tag{14.128}$$

$$R_1 + R_2 \le I(X_1, X_2; Y) \tag{14.129}$$

for some input distribution $f_1(x_1)f_2(x_2)$ satisfying $EX_1^2 \le P_1$ and $EX_2^2 \le P_2$.

Now, we can expand the mutual information in terms of relative entropy, and thus

$$I(X_1; Y|X_2) = h(Y|X_2) - h(Y|X_1, X_2) \tag{14.130}$$

$$= h(X_1 + X_2 + Z|X_2) - h(X_1 + X_2 + Z|X_1, X_2) \tag{14.131}$$

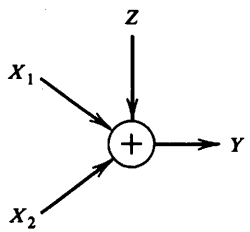


Figure 14.16. Gaussian multiple access channel.

$$= h(X_1 + Z|X_2) - h(Z|X_1, X_2) \tag{14.132}$$

$$= h(X_1 + Z|X_2) - h(Z) \tag{14.133}$$

$$= h(X_1 + Z) - h(Z) \tag{14.134}$$

$$= h(X_1 + Z) - \frac{1}{2}\log(2\pi e)N \tag{14.135}$$

$$\leq \frac{1}{2}\log(2\pi e)(P_1 + N) - \frac{1}{2}\log(2\pi e)N \tag{14.136}$$

$$= \frac{1}{2}\log\left(1 + \frac{P_1}{N}\right), \tag{14.137}$$

where (14.133) follows from the fact that Z is independent of X_1 and X_2, (14.134) from the independence of X_1 and X_2, and (14.136) from the fact that the normal maximizes entropy for a given second moment. Thus the maximizing distribution is $X_1 \sim \mathcal{N}(0, P_1)$ and $X_2 \sim \mathcal{N}(0, P_2)$ with X_1 and X_2 independent. This distribution simultaneously maximizes the mutual information bounds in (14.127)–(14.129).

Definition: We define the channel capacity function

$$C(x) \triangleq \frac{1}{2}\log(1 + x), \tag{14.138}$$

corresponding to the channel capacity of a Gaussian white noise channel with signal to noise ratio x.

Then we write the bound on R_1 as

$$R_1 \leq C\left(\frac{P_1}{N}\right). \tag{14.139}$$

Similarly,

$$R_2 \leq C\left(\frac{P_2}{N}\right), \tag{14.140}$$

and

$$R_1 + R_2 \leq C\left(\frac{P_1 + P_2}{N}\right). \tag{14.141}$$

These upper bounds are achieved when $X_1 \sim \mathcal{N}(0, P_1)$ and $X_2 = \mathcal{N}(0, P_2)$ and define the capacity region.

The surprising fact about these inequalities is that the sum of the rates can be as large as $C(\frac{P_1 + P_2}{N})$, which is that rate achieved by a single transmitter sending with a power equal to the sum of the powers.

The interpretation of the corner points is very similar to the interpretation of the achievable rate pairs for a discrete multiple access channel for a fixed input distribution. In the case of the Gaussian channel, we can consider decoding as a two-stage process: in the first stage, the receiver decodes the second sender, considering the first sender as part of the noise. This decoding will have low probability of error if $R_2 < C(\frac{P_2}{P_1+N})$. After the second sender has been successfully decoded, it can be subtracted out and the first sender can be decoded correctly if $R_1 < C(\frac{P_1}{N})$. Hence, this argument shows that we can achieve the rate pairs at the corner points of the capacity region.

If we generalize this to m senders with equal power, the total rate is $C(\frac{mP}{N})$, which goes to ∞ as $m \to \infty$. The average rate per sender, $\frac{1}{m}C(\frac{mP}{N})$ goes to 0. Thus when the total number of senders is very large, so that there is a lot of interference, we can still send a total amount of information which is arbitrarily large even though the rate per individual sender goes to 0.

The capacity region described above corresponds to Code Division Multiple Access (CDMA), where orthogonal codes are used for the different senders, and the receiver decodes them one by one. In many practical situations, though, simpler schemes like time division multiplexing or frequency division multiplexing are used.

With frequency division multiplexing, the rates depend on the bandwidth allotted to each sender. Consider the case of two senders with powers P_1 and P_2 and using bandwidths non-intersecting frequency bands W_1 and W_2, where $W_1 + W_2 = W$ (the total bandwidth). Using the formula for the capacity of a single user bandlimited channel, the following rate pair is achievable:

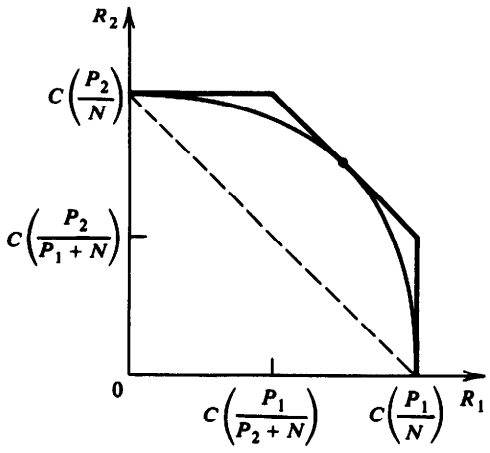


Figure 14.17. Gaussian multiple access channel capacity.

$$R_1 = \frac{W_1}{2}\log\left(1 + \frac{P_1}{NW_1}\right), \tag{14.142}$$

$$R_2 = \frac{W_2}{2}\log\left(1 + \frac{P_2}{NW_2}\right). \tag{14.143}$$

As we vary W_1 and W_2, we trace out the curve as shown in Figure 14.17. This curve touches the boundary of the capacity region at one point, which corresponds to allotting bandwidth to each channel proportional to the power in that channel. We conclude that no allocation of frequency bands to radio stations can be optimal unless the allocated powers are proportional to the bandwidths.

As Figure 14.17 illustrates, in general the capacity region is larger than that achieved by time division or frequency division multiplexing. But note that the multiple access capacity region derived above is achieved by use of a common decoder for all the senders. However in many practical systems, simplicity of design is an important consideration, and the improvement in capacity due to the multiple access ideas presented earlier may not be sufficient to warrant the increased complexity.

For a Gaussian multiple access system with m sources with powers $P_1, P_2, \ldots, P_m$ and ambient noise of power N, we can state the equivalent of Gauss's law for any set S in the form

$$\sum_{i \in S} R_i = \text{Total rate of information flow across boundary of } S \tag{14.144}$$

$$\leq C\left(\frac{\Sigma_{i \in S} P_i}{N}\right). \tag{14.145}$$

14.4 ENCODING OF CORRELATED SOURCES

We now turn to distributed data compression. This problem is in many ways the data compression dual to the multiple access channel problem.

We know how to encode a source X. A rate $R > H(X)$ is sufficient. Now suppose that there are two sources $(X, Y) \sim p(x, y)$. A rate $H(X, Y)$ is sufficient if we are encoding them together. But what if the X-source and the Y-source must be separately described for some user who wishes to reconstruct both X and Y? Clearly, by separate encoding X and Y, it is seen that a rate $R = R_x + R_y > H(X) + H(Y)$ is sufficient. However, in a surprising and fundamental paper by Slepian and Wolf [255], it is shown that a total rate $R = H(X, Y)$ is sufficient even for separate encoding of correlated sources.

Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be a sequence of jointly distributed random variables i.i.d. $\sim p(x, y)$. Assume that the X sequence is available at a

location A and the Y sequence is available at a location B. The situation is illustrated in Figure 14.18.

Before we proceed to the proof of this result, we will give a few definitions.

Definition: A $((2^{nR_1}, 2^{nR_2}), n)$ *distributed source code* for the joint source (X, Y) consists of two encoder maps,

$$f_1 : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR_1}\}, \tag{14.146}$$

$$f_2 : \mathcal{Y}^n \to \{1, 2, \ldots, 2^{nR_2}\} \tag{14.147}$$

and a decoder map,

$$g : \{1, 2, \ldots, 2^{nR_1}\} \times \{1, 2, \ldots, 2^{nR_2}\} \to \mathcal{X}^n \times \mathcal{Y}^n . \tag{14.148}$$

Here $f_1(X^n)$ is the index corresponding to X^n, $f_2(Y^n)$ is the index corresponding to Y^n and (R_1, R_2) is the rate pair of the code.

Definition: The *probability of error* for a distributed source code is defined as

$$P_e^{(n)} = P(g(f_1(X^n), f_2(Y^n)) \neq (X^n, Y^n)) . \tag{14.149}$$

Definition: A rate pair (R_1, R_2) is said to be *achievable* for a distributed source if there exists a sequence of $((2^{nR_1}, 2^{nR_2}), n)$ distributed source codes with probability of error $P_e^{(n)} \to 0$. The *achievable rate region* is the closure of the set of achievable rates.

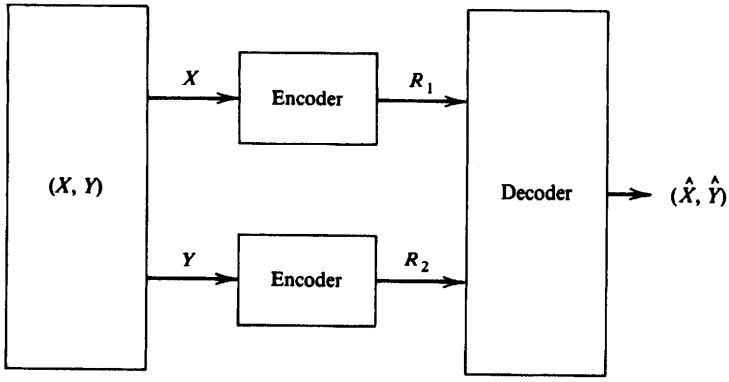


Figure 14.18. Slepian-Wolf coding.

Theorem 14.4.1 (*Slepian-Wolf*): *For the distributed source coding problem for the source* (X, Y) *drawn i.i.d* $\sim p(x, y)$, *the achievable rate region is given by*

$$R_1 \geq H(X|Y), \tag{14.150}$$

$$R_2 \geq H(Y|X), \tag{14.151}$$

$$R_1 + R_2 \geq H(X, Y). \tag{14.152}$$

Let us illustrate the result with some examples.

Example 14.4.1: Consider the weather in Gotham and Metropolis. For the purposes of our example, we will assume that Gotham is sunny with probability 0.5 and that the weather in Metropolis is the same as in Gotham with probability 0.89. The joint distribution of the weather is given as follows:

$p(x, y)$	Metropolis Rain	Metropolis Shine
Gotham Rain	0.445	0.055
Gotham Shine	0.055	0.445

Assume that we wish to transmit 100 days of weather information to the National Weather Service Headquarters in Washington. We could send all the 100 bits of the weather in both places, making 200 bits in all. If we decided to compress the information independently, then we would still need $100H(0.5) = 100$ bits of information from each place for a total of 200 bits.

If instead we use Slepian-Wolf encoding, we need only $H(X) + H(Y|X) = 100H(0.5) + 100H(0.89) = 100 + 50 = 150$ bits total.

Example 14.4.2: Consider the following joint distribution:

$p(u, v)$	0	1
0	1/3	1/3
1	0	1/3

In this case, the total rate required for the transmission of this source is $H(U) + H(V|U) = \log 3 = 1.58$ bits, rather than the 2 bits which would

be needed if the sources were transmitted independently without Slepian-Wolf encoding.

14.4.1 Achievability of the Slepian-Wolf Theorem

We now prove the achievability of the rates in the Slepian-Wolf theorem. Before we proceed to the proof, we will first introduce a new coding procedure using random bins.

The essential idea of random bins is very similar to hash functions: we choose a large random index for each source sequence. If the set of typical source sequences is small enough (or equivalently, the range of the hash function is large enough), then with high probability, different source sequences have different indices, and we can recover the source sequence from the index.

Let us consider the application of this idea to the encoding of a single source. In Chapter 3, the method that we considered was to index all elements of the typical set and not bother about elements outside the typical set. We will now describe the random binning procedure, which indexes all sequences, but rejects untypical sequences at a later stage.

Consider the following procedure: For each sequence X^n, draw an index at random from $\{1, 2, \ldots, 2^{nR}\}$. The set of sequences X^n which have the same index are said to form a bin, since this can be viewed as first laying down a row of bins and then throwing the X^n's at random into the bins. For decoding the source from the bin index, we look for a typical X^n sequence in the bin. If there is one and only one typical X^n sequence in the bin, we declare it to be the estimate $\hat{X}^n$ of the source sequence; otherwise, an error is declared.

The above procedure defines a source code. To analyze the probability of error for this code, we will now divide the X^n sequences into two types, the typical sequences and the non-typical sequences.

If the source sequence is typical, then the bin corresponding to this source sequence will contain at least one typical sequence (the source sequence itself). Hence there will be an error only if there is more than one typical sequence in this bin. If the source sequence is non-typical, then there will always be an error. But if the number of bins is much larger than the number of typical sequences, the probability that there is more than one typical sequence in a bin is very small, and hence the probability that a typical sequence will result in an error is very small.

Formally, let $f(X^n)$ be the bin index corresponding to X^n. Call the decoding function g. The probability of error (averaged over the random choice of codes f) is

$$\begin{aligned} P(g(f(\mathbf{X})) \neq \mathbf{X}) &\leq P(\mathbf{X} \notin A_\epsilon^{(n)}) + \sum_{\mathbf{x}} P(\exists \mathbf{x}' \neq \mathbf{x} : \mathbf{x}' \in A_\epsilon^{(n)}, f(\mathbf{x}') = f(\mathbf{x}))p(\mathbf{x}) \\ &\leq \epsilon + \sum_{\mathbf{x}} \sum_{\substack{\mathbf{x}' \in A_\epsilon^{(n)} \\ \mathbf{x}' \neq \mathbf{x}}} P(f(\mathbf{x}') = f(\mathbf{x}))p(\mathbf{x}) \end{aligned} \tag{14.153}$$

$$\leq \epsilon + \sum_{\mathbf{x}} \sum_{\mathbf{x}' \in A_\epsilon^{(n)}} 2^{-nR} p(\mathbf{x}) \tag{14.154}$$

$$= \epsilon + \sum_{\mathbf{x}' \in A_\epsilon^{(n)}} 2^{-nR} \sum_{\mathbf{x}} p(\mathbf{x}) \tag{14.155}$$

$$\leq \epsilon + \sum_{\mathbf{x}' \in A_\epsilon^{(n)}} 2^{-nR} \tag{14.156}$$

$$\leq \epsilon + 2^{n(H(X)+\epsilon)} 2^{-nR} \tag{14.157}$$

$$\leq 2\epsilon \tag{14.158}$$

if $R > H(X) + \epsilon$ and n is sufficiently large. Hence if the rate of the code is greater than the entropy, the probability of error is arbitrarily small and the code achieves the same results as the code described in Chapter 3.

The above example illustrates the fact that there are many ways to construct codes with low probabilities of error at rates above the entropy of the source; the universal source code is another example of such a code. Note that the binning scheme does not require an explicit characterization of the typical set at the encoder; it is only needed at the decoder. It is this property that enables this code to continue to work in the case of a distributed source, as will be illustrated in the proof of the theorem.

We now return to the consideration of the distributed source coding and prove the achievability of the rate region in the Slepian-Wolf theorem.

Proof (*Achievability in Theorem 14.4.1*): The basic idea of the proof is to partition the space of $\mathcal{X}^n$ into 2^{nR_1} bins and the space of $\mathcal{Y}^n$ into 2^{nR_2} bins.

Random code generation. Independently assign every $\mathbf{x} \in \mathcal{X}^n$ to one of 2^{nR_1} bins according to a uniform distribution on $\{1, 2, \ldots, 2^{nR_1}\}$. Similarly, randomly assign every $\mathbf{y} \in \mathcal{Y}^n$ to one of 2^{nR_2} bins. Reveal the assignments f_1 and f_2 to both the encoder and decoder.

Encoding. Sender 1 sends the index of the bin to which $\mathbf{X}$ belongs. Sender 2 sends the index of the bin to which $\mathbf{Y}$ belongs.

Decoding. Given the received index pair (i_0, j_0), declare $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = (\mathbf{x}, \mathbf{y})$, if there is one and only one pair of sequences $(\mathbf{x}, \mathbf{y})$ such that $f_1(\mathbf{x}) = i_0$, $f_2(\mathbf{y}) = j_0$ and $(\mathbf{x}, \mathbf{y}) \in A_\epsilon^{(n)}$. Otherwise declare an error.

The scheme is illustrated in Figure 14.19. The set of X sequences and the set of Y sequences are divided into bins in such a way that the pair of indices specifies a product bin.

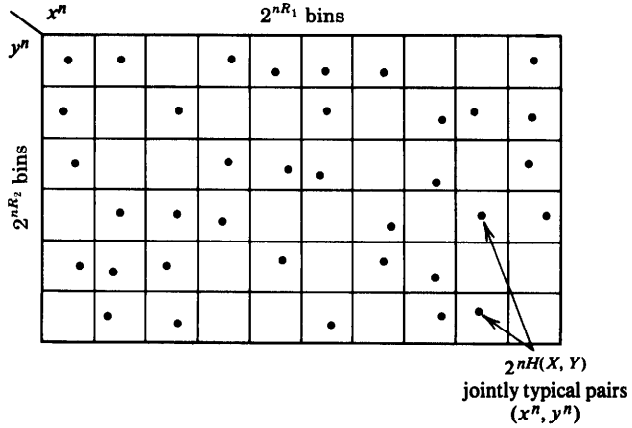


Figure 14.19. Slepian-Wolf encoding: the jointly typical pairs are isolated by the product bins.

Probability of error. Let $(X_i, Y_i) \sim p(x, y)$. Define the events

$$E_0 = \{(\mathbf{X}, \mathbf{Y}) \notin A_\epsilon^{(n)}\}, \tag{14.159}$$

$$E_1 = \{\exists \mathbf{x}' \neq \mathbf{X}: f_1(\mathbf{x}') = f_1(\mathbf{X}) \text{ and } (\mathbf{x}', \mathbf{Y}) \in A_\epsilon^{(n)}\}, \tag{14.160}$$

$$E_2 = \{\exists \mathbf{y}' \neq \mathbf{Y}: f_2(\mathbf{y}') = f_2(\mathbf{Y}) \text{ and } (\mathbf{X}, \mathbf{y}') \in A_\epsilon^{(n)}\}, \tag{14.161}$$

and

$$E_{12} = \{\exists (\mathbf{x}', \mathbf{y}'): \mathbf{x}' \neq \mathbf{X}, \mathbf{y}' \neq \mathbf{Y}, f_1(\mathbf{x}') = f_1(\mathbf{X}), f_2(\mathbf{y}') \neq f_2(\mathbf{Y}) \text{ and } (\mathbf{x}', \mathbf{y}') \in A_\epsilon^{(n)}\}. \tag{14.162}$$

Here $\mathbf{X}, \mathbf{Y}, f_1$ and f_2 are random. We have an error if $(\mathbf{X}, \mathbf{Y})$ is not in $A_\epsilon^{(n)}$ or if there is another typical pair in the same bin. Hence by the union of events bound,

$$P_e^{(n)} = P(E_0 \cup E_1 \cup E_2 \cup E_{12}) \tag{14.163}$$

$$\leq P(E_0) + P(E_1) + P(E_2) + P(E_{12}). \tag{14.164}$$

First consider E_0. By the AEP, $P(E_0) \to 0$ and hence for n sufficiently large, $P(E_0) < \epsilon$.

To bound $P(E_1)$, we have

$$P(E_1) = P\{\exists \mathbf{x}' \neq \mathbf{X}: f_1(\mathbf{x}') = f_1(\mathbf{X}), \text{ and } (\mathbf{x}', \mathbf{Y}) \in A_\epsilon^{(n)}\} \tag{14.165}$$

$$= \sum_{(\mathbf{x}, \mathbf{y})} p(\mathbf{x}, \mathbf{y}) P\{\exists \mathbf{x}' \neq \mathbf{x}: f_1(\mathbf{x}') = f_1(\mathbf{x}), (\mathbf{x}', \mathbf{y}) \in A_\epsilon^{(n)}\} \quad (14.166)$$

$$\leq \sum_{(\mathbf{x}, \mathbf{y})} p(\mathbf{x}, \mathbf{y}) \sum_{\substack{\mathbf{x}' \neq \mathbf{x} \\ (\mathbf{x}', \mathbf{y}) \in A_\epsilon^{(n)}}} P(f_1(\mathbf{x}') = f_1(\mathbf{x})) \quad (14.167)$$

$$= \sum_{(\mathbf{x}, \mathbf{y})} p(\mathbf{x}, \mathbf{y}) 2^{-nR_1} |A_\epsilon(X|\mathbf{y})| \quad (14.168)$$

$$\leq 2^{-nR_1} 2^{n(H(X|Y)+\epsilon)} \text{ (by Theorem 14.2.2)}, \quad (14.169)$$

which goes to 0 if $R_1 > H(X|Y)$. Hence for sufficiently large n, $P(E_1) < \epsilon$. Similarly, for sufficiently large n, $P(E_2) < \epsilon$ if $R_2 > H(Y|X)$ and $P(E_{12}) < \epsilon$ if $R_1 + R_2 > H(X, Y)$.

Since the average probability of error is $< 4\epsilon$, there exists at least one code (f_1^*, f_2^*, g^*) with probability of error $< 4\epsilon$. Thus, we can construct a sequence of codes with $P_e^{(n)} \to 0$ and the proof of achievability is complete. □

14.4.2 Converse for the Slepian-Wolf Theorem

The converse for the Slepian-Wolf theorem follows obviously from from the results for single source, but we will provide it for completeness.

Proof (*Converse to Theorem 14.4.1*): As usual, we begin with Fano's inequality. Let f_1, f_2, g be fixed. Let $I_0 = f_1(X^n)$ and $J_0 = f_2(Y^n)$. Then

$$H(X^n, Y^n | I_0, J_0) \leq P_e^{(n)} n(\log|\mathcal{X}| + \log|\mathcal{Y}|) + 1 = n\epsilon_n, \quad (14.170)$$

where $\epsilon_n \to 0$ as $n \to \infty$. Now adding conditioning, we also have

$$H(X^n | Y^n, I_0, J_0) \leq P_e^{(n)} n\epsilon_n, \quad (14.171)$$

and

$$H(Y^n | X^n, I_0, J_0) \leq P_e^{(n)} n\epsilon_n. \quad (14.172)$$

We can write a chain of inequalities

$$n(R_1 + R_2) \overset{(a)}{\geq} H(I_0, J_0) \quad (14.173)$$

$$= I(X^n, Y^n; I_0, J_0) + H(I_0, J_0 | X^n, Y^n) \quad (14.174)$$

$$\overset{(b)}{=} I(X^n, Y^n; I_0, J_0) \quad (14.175)$$

$$= H(X^n, Y^n) - H(X^n, Y^n | I_0, J_0) \tag{14.176}$$

$$\overset{(c)}{\geq} H(X^n, Y^n) - n\epsilon_n \tag{14.177}$$

$$\overset{(d)}{=} nH(X, Y) - n\epsilon_n , \tag{14.178}$$

where

(a) follows from the fact that $I_0 \in \{1, 2, \ldots, 2^{nR_1}\}$ and $J_0 \in \{1, 2, \ldots, 2^{nR_2}\}$,
(b) from the fact the I_0 is a function of X^n and J_0 is a function of Y^n,
(c) from Fano's inequality (14.170), and
(d) from the chain rule and the fact that (X_i, Y_i) are i.i.d.

Similarly, using (14.171), we have

$$nR_1 \overset{(a)}{\geq} H(I_0) \tag{14.179}$$

$$\geq H(I_0 | Y^n) \tag{14.180}$$

$$= I(X^n; I_0 | Y^n) + H(I_0 | X^n, Y^n) \tag{14.181}$$

$$\overset{(b)}{=} I(X^n; I_0 | Y^n) \tag{14.182}$$

$$= H(X^n | Y^n) - H(X^n | I_0, J_0, Y^n) \tag{14.183}$$

$$\overset{(c)}{\geq} H(X^n | Y^n) - n\epsilon_n \tag{14.184}$$

$$\overset{(d)}{=} nH(X|Y) - n\epsilon_n , \tag{14.185}$$

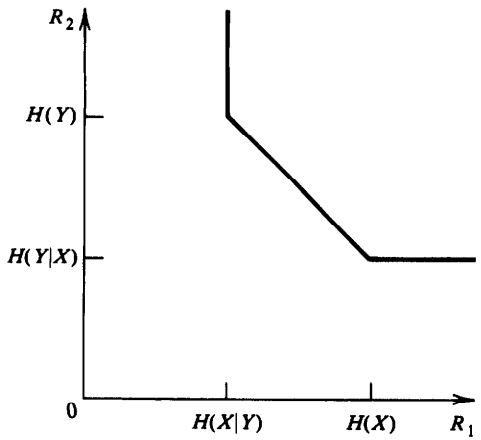


Figure 14.20. Rate region for Slepian-Wolf encoding.

where the reasons are the same as for the equations above. Similarly, we can show that

$$nR_2 \geq nH(Y|X) - n\epsilon_n . \tag{14.186}$$

Dividing these inequalities by n and taking the limit as $n \to \infty$, we have the desired converse. □

The region described in the Slepian-Wolf theorem is illustrated in Figure 14.20.

14.4.3 Slepian-Wolf Theorem for Many Sources

The results of the previous section can easily be generalized to many sources. The proof follows exactly the same lines.

Theorem 14.4.2: *Let* $(X_{1i}, X_{2i}, \ldots, X_{mi})$ *be i.i.d.* $\sim p(x_1, x_2, \ldots, x_m)$. *Then the set of rate vectors achievable for distributed source coding with separate encoders and a common decoder is defined by*

$$R(S) > H(X(S)|X(S^c)) \tag{14.187}$$

for all $S \subseteq \{1, 2, \ldots, m\}$ *where*

$$R(S) = \sum_{i \in S} R_i , \tag{14.188}$$

and $X(S) = \{X_j : j \in S\}$.

Proof: The proof is identical to the case of two variables and is omitted. □

The achievability of Slepian-Wolf encoding has been proved for an i.i.d. correlated source, but the proof can easily be extended to the case of an arbitrary joint source that satisfies the AEP; in particular, it can be extended to the case of any jointly ergodic source [63]. In these cases the entropies in the definition of the rate region are replaced by the corresponding entropy rates.

14.4.4 Interpretation of Slepian-Wolf Coding

We will consider an interpretation of the corner points of the rate region in Slepian-Wolf encoding in terms of graph coloring. Consider the point with rate $R_1 = H(X), R_2 = H(Y|X)$. Using $nH(X)$ bits, we can encode X^n efficiently, so that the decoder can reconstruct X^n with arbitrarily low probability of error. But how do we code Y^n with $nH(Y|X)$ bits?

Looking at the picture in terms of typical sets, we see that associated with every X^n is a typical "fan" of Y^n sequences that are jointly typical with the given X^n as shown in Figure 14.21.

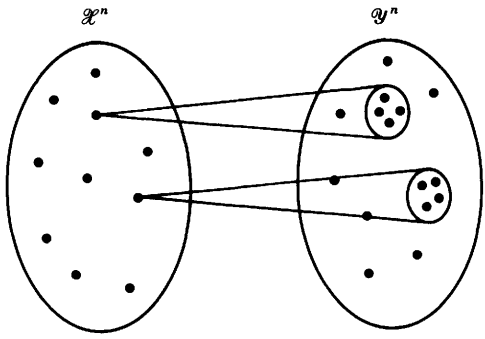


Figure 14.21. Jointly typical fans.

If the Y encoder knows X^n, the encoder can send the index of the Y^n within this typical fan. The decoder, also knowing X^n, can then construct this typical fan and hence reconstruct Y^n. But the Y encoder does not know X^n. So instead of trying to determine the typical fan, he randomly colors all Y^n sequences with 2^{nR_2} colors. If the number of colors is high enough, then with high probability, all the colors in a particular fan will be different and the color of the Y^n sequence will uniquely define the Y^n sequence within the X^n fan. If the rate $R_2 > H(Y|X)$, the number of colors is exponentially larger than the number of elements in the fan and we can show that the scheme will have exponentially small probability of error.

14.5 DUALITY BETWEEN SLEPIAN-WOLF ENCODING AND MULTIPLE ACCESS CHANNELS

With multiple access channels, we considered the problem of sending independent messages over a channel with two inputs and only one output. With Slepian-Wolf encoding, we considered the problem of sending a correlated source over a noiseless channel, with a common decoder for recovery of both sources. In this section, we will explore the duality between the two systems.

In Figure 14.22, two independent messages are to be sent over the channel as X_1^n and X_2^n sequences. The receiver estimates the messages from the received sequence. In Figure 14.23, the correlated sources are encoded as "independent" messages i and j. The receiver tries to estimate the source sequences from knowledge of i and j.

In the proof of the achievability of the capacity region for the multiple access channel, we used a random map from the set of messages to the sequences X_1^n and X_2^n. In the proof for Slepian-Wolf coding, we used a random map from the set of sequences X^n and Y^n to a set of messages.

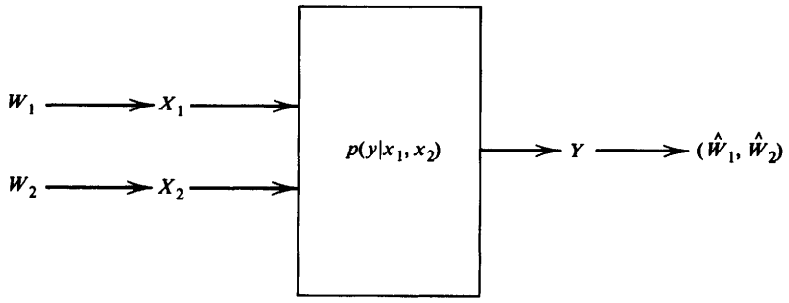


Figure 14.22. Multiple access channels.

In the proof of the coding theorem for the multiple access channel, the probability of error was bounded by

$$P_e^{(n)} \leq \epsilon + \sum_{\text{codewords}} \Pr(\text{codeword jointly typical with received sequence}) \tag{14.189}$$

$$= \epsilon + \sum_{2^{nR_1} \text{ terms}} 2^{-nI_1} + \sum_{2^{nR_2} \text{ terms}} 2^{-nI_2} + \sum_{2^{n(R_1+R_2)} \text{ terms}} 2^{-nI_3}, \tag{14.190}$$

where ϵ is the probability the sequences are not typical, R_i are the rates corresponding to the number of codewords that can contribute to the probability of error, and I_i is the corresponding mutual information that corresponds to the probability that the codeword is jointly typical with the received sequence.

In the case of Slepian-Wolf encoding, the corresponding expression for the probability of error is

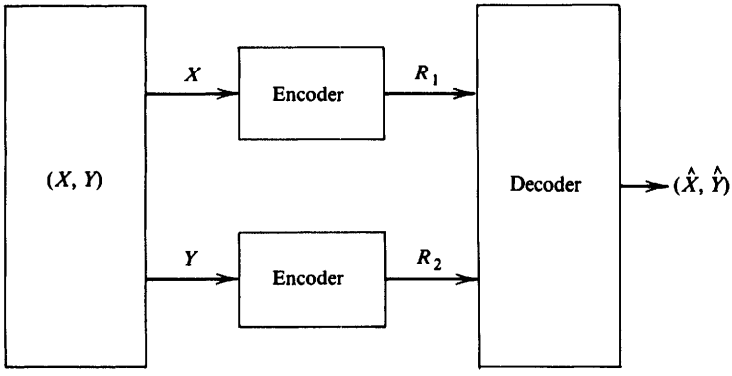


Figure 14.23. Correlated source encoding.

$$P_e^{(n)} \leq \epsilon + \sum_{\text{Jointly typical sequences}} \Pr(\text{have same codeword}) \tag{14.191}$$

$$= \epsilon + \sum_{2^{nH_1} \text{ terms}} 2^{-nR_1} + \sum_{2^{nH_2} \text{ terms}} 2^{-nR_2} + \sum_{2^{nH_3} \text{ terms}} 2^{-n(R_1+R_2)} \tag{14.192}$$

where again the probability that the constraints of the AEP are not satisfied is bounded by ϵ, and the other terms refer to the various ways in which another pair of sequences could be jointly typical and in the same bin as the given source pair.

The duality of the multiple access channel and correlated source encoding is now obvious. It is rather surprising that these two systems are duals of each other; one would have expected a duality between the broadcast channel and the multiple access channel.

14.6 THE BROADCAST CHANNEL

The broadcast channel is a communication channel in which there is one sender and two or more receivers. It is illustrated in Figure 14.24. The basic problem is to find the set of simultaneously achievable rates for communication in a broadcast channel.

Before we begin the analysis, let us consider some examples:

Example 14.6.1 (*TV station*): The simplest example of the broadcast channel is a radio or TV station. But this example is slightly degenerate in the sense that normally the station wants to send the same information to everybody who is tuned in; the capacity is essentially $\max_{p(x)} \min_i I(X; Y_i)$, which may be less than the capacity of the worst receiver.

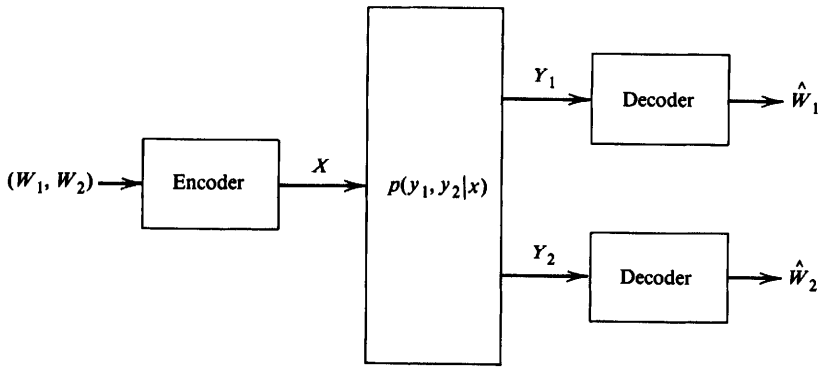


Figure 14.24. Broadcast channel.

But we may wish to arrange the information in such a way that the better receivers receive extra information, which produces a better picture or sound, while the worst receivers continue to receive more basic information. As TV stations introduce High Definition TV (HDTV), it may be necessary to encode the information so that bad receivers will receive the regular TV signal, while good receivers will receive the extra information for the high definition signal. The methods to accomplish this will be explained in the discussion of the broadcast channel.

Example 14.6.2 (*Lecturer in classroom*): A lecturer in a classroom is communicating information to the students in the class. Due to differences among the students, they receive various amounts of information. Some of the students receive most of the information; others receive only a little. In the ideal situation, the lecturer would be able to tailor his or her lecture in such a way that the good students receive more information and the poor students receive at least the minimum amount of information. However, a poorly prepared lecture proceeds at the pace of the weakest student. This situation is another example of a broadcast channel.

Example 14.6.3 (*Orthogonal broadcast channels*): The simplest broadcast channel consists of two independent channels to the two receivers. Here we can send independent information over both channels, and we can achieve rate R_1 to receiver 1 and rate R_2 to receiver 2, if $R_1 < C_1$ and $R_2 < C_2$. The capacity region is the rectangle shown in Figure 14.25.

Example 14.6.4 (*Spanish and Dutch speaker*): To illustrate the idea of superposition, we will consider a simplified example of a speaker who can speak both Spanish and Dutch. There are two listeners: one understands only Spanish and the other understands only Dutch. Assume for simplicity that the vocabulary of each language is 2^{20} words

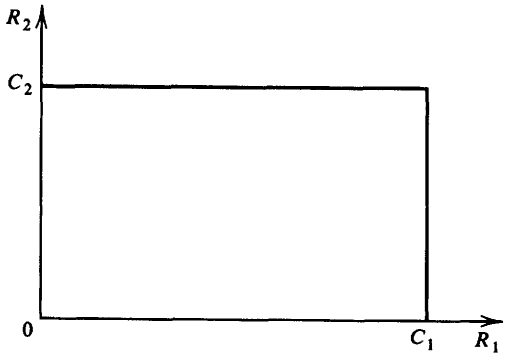


Figure 14.25. Capacity region for two orthogonal broadcast channels.

and that the speaker speaks at the rate of 1 word per second in either language. Then he can transmit 20 bits of information per second to receiver 1 by speaking to him all the time; in this case, he sends no information to receiver 2. Similarly, he can send 20 bits per second to receiver 2 without sending any information to receiver 1. Thus he can achieve any rate pair with $R_1 + R_2 = 20$ by simple timesharing. But can he do better?

Recall that the Dutch listener, even though he does not understand Spanish, can recognize when the word is Spanish. Similarly, the Spanish listener can recognize when Dutch occurs. The speaker can use this to convey information; for example, if the proportion of time he uses each language is 50%, then of a sequence of 100 words, about 50 will be Dutch and about 50 will be Spanish. But there are many ways to order the Spanish and Dutch words; in fact, there are about $\binom{100}{50} \approx 2^{100H(\frac{1}{2})}$ ways to order the words. Choosing one of these orderings conveys information to both listeners. This method enables the speaker to send information at a rate of 10 bits per second to the Dutch receiver, 10 bits per second to the Spanish receiver, and 1 bit per second of common information to both receivers, for a total rate of 21 bits per second, which is more than that achievable by simple time sharing. This is an example of superposition of information.

The results of the broadcast channel can also be applied to the case of a single user channel with an unknown distribution. In this case, the objective is to get at least the minimum information through when the channel is bad and to get some extra information through when the channel is good. We can use the same superposition arguments as in the case of the broadcast channel to find the rates at which we can send information.

14.6.1 Definitions for a Broadcast Channel

Definition: A *broadcast channel* consists of an input alphabet $\mathcal{X}$ and two output alphabets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ and a probability transition function $p(y_1, y_2|x)$. The broadcast channel will be said to be *memoryless* if $p(y_1^n, y_2^n|x^n) = \prod_{i=1}^n p(y_{1i}, y_{2i}|x_i)$.

We define codes, probability of error, achievability and capacity regions for the broadcast channel as we did for the multiple access channel.

A $((2^{nR_1}, 2^{nR_2}), n)$ code for a broadcast channel with independent information consists of an encoder,

$$X:(\{1, 2, \ldots, 2^{nR_1}\} \times \{1, 2, \ldots, 2^{nR_2}\}) \rightarrow \mathcal{X}^n, \tag{14.193}$$

and two decoders,

$$g_1 : \mathcal{Y}_1^n \to \{1, 2, \ldots, 2^{nR_1}\} \tag{14.194}$$

and

$$g_2 : \mathcal{Y}_2^n \to \{1, 2, \ldots, 2^{nR_2}\} . \tag{14.195}$$

We define the average probability of error as the probability the decoded message is not equal to the transmitted message, i.e.,

$$P_e^{(n)} = P(g_1(Y_1^n) \neq W_1 \text{ or } g_2(Y_2^n) \neq W_2) , \tag{14.196}$$

where (W_1, W_2) are assumed to be uniformly distributed over $2^{nR_1} \times 2^{nR_2}$.

Definition: A rate pair (R_1, R_2) is said to be *achievable* for the broadcast channel if there exists a sequence of $((2^{nR_1}, 2^{nR_2}), n)$ codes with $P_e^{(n)} \to 0$.

We will now define the rates for the case where we have common information to be sent to both receivers.

A $((2^{nR_0}, 2^{nR_1}, 2^{nR_2}), n)$ code for a broadcast channel with common information consists of an encoder,

$$X : (\{1, 2, \ldots, 2^{nR_0}\} \times \{1, 2, \ldots, 2^{nR_1}\} \times \{1, 2, \ldots, 2^{nR_2}\}) \to \mathcal{X}^n , \tag{14.197}$$

and two decoders,

$$g_1 : \mathcal{Y}_1^n \to \{1, 2, \ldots, 2^{nR_0}\} \times \{1, 2, \ldots, 2^{nR_1}\} \tag{14.198}$$

and

$$g_2 : \mathcal{Y}_2^n \to \{1, 2, \ldots, 2^{nR_0}\} \times \{1, 2, \ldots, 2^{nR_2}\} . \tag{14.199}$$

Assuming that the distribution on (W_0, W_1, W_2) is uniform, we can define the probability of error as the probability the decoded message is not equal to the transmitted message, i.e.,

$$P_e^{(n)} = P(g_1(Y_1^n) \neq (W_0, W_1) \text{ or } g_2(Z^n) \neq (W_0, W_2)) . \tag{14.200}$$

Definition: A rate triple (R_0, R_1, R_2) is said to be *achievable* for the broadcast channel with common information if there exists a sequence of $((2^{nR_0}, 2^{nR_1}, 2^{nR_2}), n)$ codes with $P_e^{(n)} \to 0$.

Definition: The *capacity region* of the broadcast channel is the closure of the set of achievable rates.

Theorem 14.6.1: *The capacity region of a broadcast channel depends only on the conditional marginal distributions $p(y_1|x)$ and $p(y_2|x)$.*

Proof: See exercises. □

14.6.2 Degraded Broadcast Channels

Definition: A broadcast channel is said to be *physically degraded* if $p(y_1, y_2|x) = p(y_1|x)p(y_2|y_1)$.

Definition: A broadcast channel is said to be *stochastically degraded* if its conditional marginal distributions are the same as that of a physically degraded broadcast channel, i.e., if there exists a distribution $p'(y_2|y_1)$ such that

$$p(y_2|x) = \sum_{y_1} p(y_1|x)p'(y_2|y_1) . \tag{14.201}$$

Note that since the capacity of a broadcast channel depends only on the conditional marginals, the capacity region of the stochastically degraded broadcast channel is the same as that of the corresponding physically degraded channel. In much of the following, we will therefore assume that the channel is physically degraded.

14.6.3 Capacity Region for the Degraded Broadcast Channel

We now consider sending independent information over a degraded broadcast channel at rate R_1 to Y_1 and rate R_2 to Y_2.

Theorem 14.6.2: *The capacity region for sending independent information over the degraded broadcast channel $X \rightarrow Y_1 \rightarrow Y_2$ is the convex hull of the closure of all (R_1, R_2) satisfying*

$$R_2 \leq I(U; Y_2) , \tag{14.202}$$

$$R_1 \leq I(X; Y_1|U) \tag{14.203}$$

for some joint distribution $p(u)p(x|u)p(y, z|x)$, where the auxiliary random variable U has cardinality bounded by $|\mathcal{U}| \leq \min\{|\mathcal{X}|, |\mathcal{Y}_1|, |\mathcal{Y}_2|\}$.

Proof: The cardinality bounds for the auxiliary random variable U are derived using standard methods from convex set theory and will not be dealt with here.

We first give an outline of the basic idea of superposition coding for the broadcast channel. The auxiliary random variable U will serve as a cloud center that can be distinguished by both receivers Y_1 and Y_2. Each

cloud consists of 2^{nR_1} codewords X^n distinguishable by the receiver Y_1. The worst receiver can only see the clouds, while the better receiver can see the individual codewords within the clouds.

The formal proof of the achievability of this region uses a random coding argument: Fix $p(u)$ and $p(x|u)$.

Random codebook generation. Generate 2^{nR_2} independent codewords of length n, $\mathbf{U}(w_2)$, $w_2 \in \{1, 2, \ldots, 2^{nR_2}\}$, according to $\prod_{i=1}^{n} p(u_i)$.

For each codeword $\mathbf{U}(w_2)$, generate 2^{nR_1} independent codewords $\mathbf{X}(w_1, w_2)$ according to $\prod_{i=1}^{n} p(x_i|u_i(w_2))$.

Here $\mathbf{u}(i)$ plays the role of the cloud center understandable to both Y_1 and Y_2, while $\mathbf{x}(i, j)$ is the jth satellite codeword in the ith cloud.

Encoding. To send the pair (W_1, W_2), send the corresponding codeword $\mathbf{X}(W_1, W_2)$.

Decoding. Receiver 2 determines the unique $\hat{\hat{W}}_2$ such that $(\mathbf{U}(\hat{\hat{W}}_2), \mathbf{Y}_2) \in A_\epsilon^{(n)}$. If there are none such or more than one such, an error is declared.

Receiver 1 looks for the unique $(\hat{W}_1, \hat{W}_2)$ such that $(\mathbf{U}(\hat{W}_2), \mathbf{X}(\hat{W}_1, \hat{W}_2), \mathbf{Y}_1) \in A_\epsilon^{(n)}$. If there are none such or more than one such, an error is declared.

Analysis of the probability of error. By the symmetry of the code generation, the probability of error does not depend on which codeword was sent. Hence, without loss of generality, we can assume that the message pair $(W_1, W_2) = (1, 1)$ was sent. Let $P(\cdot)$ denote the conditional probability of an event given that $(1, 1)$ was sent.

Since we have essentially a single user channel from U to Y_2, we will be able to decode the U codewords with low probability of error if $R_2 < I(U; Y_2)$. To prove this, we define the events

$$E_{Yi} = \{(\mathbf{U}(i), \mathbf{Y}_2) \in A_\epsilon^{(n)}\} . \tag{14.204}$$

Then the probability of error at receiver 2 is

$$P_e^{(n)}(2) = P(E_{Y1}^c \bigcup \bigcup_{i \neq 1} E_{Yi}) \tag{14.205}$$

$$\leq P(E_{Y1}^c) + \sum_{i \neq 1} P(E_{Yi}) \tag{14.206}$$

$$\leq \epsilon + 2^{nR_2} 2^{-n(I(U; Y_2) - 2\epsilon)} \tag{14.207}$$

$$\leq 2\epsilon \tag{14.208}$$

if n is large enough and $R_2 < I(U; Y_2)$, where (14.207) follows from the AEP.

Similarly, for decoding for receiver 1, we define the following events

$$\tilde{E}_{Yi} = \{(\mathbf{U}(i), \mathbf{Y}_1) \in A_\epsilon^{(n)}\}, \tag{14.209}$$

$$\tilde{E}_{Yij} = \{(\mathbf{U}(i), \mathbf{X}(i, j), \mathbf{Y}_1) \in A_\epsilon^{(n)}\}, \tag{14.210}$$

where the tilde refers to events defined at receiver 1. Then, we can bound the probability of error as

$$P_e^{(n)}(1) = P\left(\tilde{E}_{Y1}^c \cup \bigcup_{i \neq 1} \tilde{E}_{Yi} \cup \bigcup_{j \neq 1} \tilde{E}_{Y1j}\right) \tag{14.211}$$

$$\leq P(\tilde{E}_{Y1}^c) + \sum_{i \neq 1} P(\tilde{E}_{Yi}) + \sum_{j \neq 1} P(\tilde{E}_{Y1j}). \tag{14.212}$$

By the same arguments as for receiver 2, we can bound $P(\tilde{E}_{Yi}) \leq 2^{-n(I(U; Y_1)-3\epsilon)}$. Hence the second term goes to 0 if $R_2 < I(U; Y_1)$. But by the data processing inequality and the degraded nature of the channel, $I(U; Y_1) \geq I(U; Y_2)$, and hence the conditions of the theorem imply that the second term goes to 0. We can also bound the third term in the probability of error as

$$P(\tilde{E}_{Y1j}) = P((\mathbf{U}(1), \mathbf{X}(1, j), \mathbf{Y}_1) \in A_\epsilon^{(n)}) \tag{14.213}$$

$$= \sum_{(\mathbf{U}, \mathbf{X}, \mathbf{Y}_1) \in A_\epsilon^{(n)}} P((\mathbf{U}(1), \mathbf{X}(1, j), \mathbf{Y}_1)) \tag{14.214}$$

$$= \sum_{(\mathbf{U}, \mathbf{X}, \mathbf{Y}_1) \in A_\epsilon^{(n)}} P(\mathbf{U}(1)) P(\mathbf{X}(1, j)|\mathbf{U}(1)) P(\mathbf{Y}_1|\mathbf{U}(1)) \tag{14.215}$$

$$\leq \sum_{(\mathbf{U}, \mathbf{X}, \mathbf{Y}_1) \in A_\epsilon^{(n)}} 2^{-n(H(U)-\epsilon)} 2^{-n(H(X|U)-\epsilon)} 2^{-n(H(Y_1|U)-\epsilon)} \tag{14.216}$$

$$\leq 2^{n(H(U, X, Y_1)+\epsilon)} 2^{-n(H(U)-\epsilon)} 2^{-n(H(X|U)-\epsilon)} 2^{-n(H(Y_1|U)-\epsilon)} \tag{14.217}$$

$$= 2^{-n(I(X; Y_1|U)-4\epsilon)} \tag{14.218}$$

Hence, if $R_1 < I(X; Y_1|U)$, the third term in the probability of error goes to 0. Thus we can bound the probability of error

$$P_e^{(n)}(1) \leq \epsilon + 2^{nR_2} 2^{-n(I(U; Y_1)-3\epsilon)} + 2^{nR_1} 2^{-n(I(X; Y_1|U)-4\epsilon)} \tag{14.219}$$

$$\leq 3\epsilon \tag{14.220}$$

if n is large enough and $R_2 < I(U; Y_1)$ and $R_1 < I(X; Y_1|U)$. The above bounds show that we can decode the messages with total probability of error that goes to 0. Hence theres exists a sequence of good $((2^{nR_1}, 2^{nR_2}), n)$ codes $\mathcal{C}_n^*$ with probability of error going to 0.

With this, we complete the proof of the achievability of the capacity region for the degraded broadcast channel. The proof of the converse is outlined in the exercises. □

So far we have considered sending independent information to both receivers. But in certain situations, we wish to send common information to both the receivers. Let the rate at which we send common information be R_0. Then we have the following obvious theorem:

Theorem 14.6.3: *If the rate pair (R_1, R_2) is achievable for a broadcast channel with independent information, then the rate triple $(R_0, R_1 - R_0, R_2 - R_0)$ with a common rate R_0 is achievable, provided that $R_0 \leq \min(R_1, R_2)$.*

In the case of a degraded broadcast channel, we can do even better. Since by our coding scheme the better receiver always decodes all the information that is sent to the worst receiver, one need not reduce the amount of information sent to the better receiver when we have common information. Hence we have the following theorem:

Theorem 14.6.4: *If the rate pair (R_1, R_2) is achievable for a degraded broadcast channel, the rate triple $(R_0, R_1, R_2 - R_0)$ is achievable for the channel with common information, provided that $R_0 < R_2$.*

We will end this section by considering the example of the binary symmetric broadcast channel.

Example 14.6.5: Consider a pair of binary symmetric channels with parameters p_1 and p_2 that form a broadcast channel as shown in Figure 14.26.

Without loss of generality in the capacity calculation, we can recast this channel as a physically degraded channel. We will assume that $p_1 < p_2 < \frac{1}{2}$. Then we can express a binary symmetric channel with parameter p_2 as a cascade of a binary symmetric channel with parameter p_1 with another binary symmetric channel. Let the crossover probability of the new channel be α. Then we must have

$$p_1(1 - \alpha) + (1 - p_1)\alpha = p_2 , \tag{14.221}$$

or

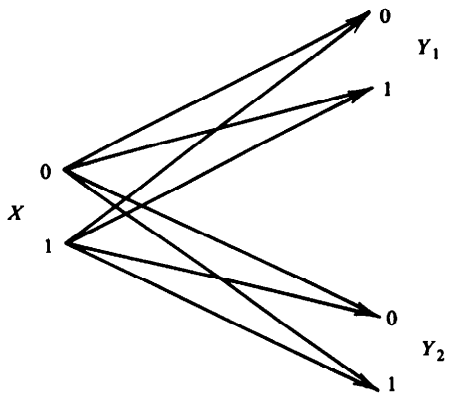


Figure 14.26. Binary symmetric broadcast channel.

$$\alpha = \frac{p_2 - p_1}{1 - 2p_1} . \tag{14.222}$$

We now consider the auxiliary random variable in the definition of the capacity region. In this case, the cardinality of U is binary from the bound of the theorem. By symmetry, we connect U to X by another binary symmetric channel with parameter β, as illustrated in Figure 14.27.

We can now calculate the rates in the capacity region. It is clear by symmetry that the distribution on U that maximizes the rates is the uniform distribution on $\{0, 1\}$, so that

$$I(U; Y_2) = H(Y_2) - H(Y_2|U) \tag{14.223}$$

$$= 1 - H(\beta * p_2), \tag{14.224}$$

where

$$\beta * p_2 = \beta(1 - p_2) + (1 - \beta)p_2 . \tag{14.225}$$

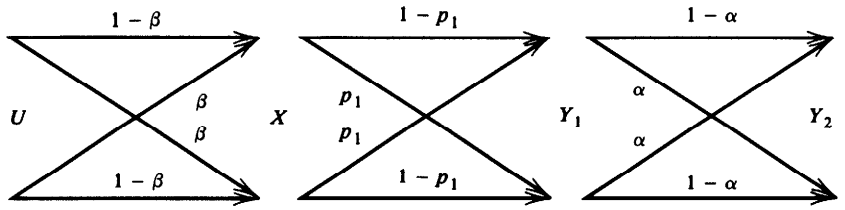


Figure 14.27. Physically degraded binary symmetric broadcast channel.

Similarly,

$$I(X; Y_1|U) = H(Y_1|U) - H(Y_1|X, U) \tag{14.226}$$

$$= H(Y_1|U) - H(Y_1|X) \tag{14.227}$$

$$= H(\beta * p_1) - H(p_1), \tag{14.228}$$

where

$$\beta * p_1 = \beta(1 - p_1) + (1 - \beta)p_1 . \tag{14.229}$$

Plotting these points as a function of β, we obtain the capacity region in Figure 14.28.

When $\beta = 0$, we have maximum information transfer to Y_2, i.e., $R_2 = 1 - H(p_2)$ and $R_1 = 0$. When $\beta = \frac{1}{2}$, we have maximum information transfer to Y_1, i.e., $R_1 = 1 - H(p_1)$, and no information transfer to Y_2. These values of β give us the corner points of the rate region.

Example 14.6.6 (*Gaussian broadcast channel*): The Gaussian broadcast channel is illustrated in Figure 14.29. We have shown it in the case where one output is a degraded version of the other output. Later, we will show that all Gaussian broadcast channels are equivalent to this type of degraded channel.

$$Y_1 = X + Z_1 , \tag{14.230}$$

$$Y_2 = X + Z_2 = Y_1 + Z_2' , \tag{14.231}$$

where $Z_1 \sim \mathcal{N}(0, N_1)$ and $Z_2' \sim \mathcal{N}(0, N_2 - N_1)$.

Extending the results of this section to the Gaussian case, we can show that the capacity region of this channel is given by

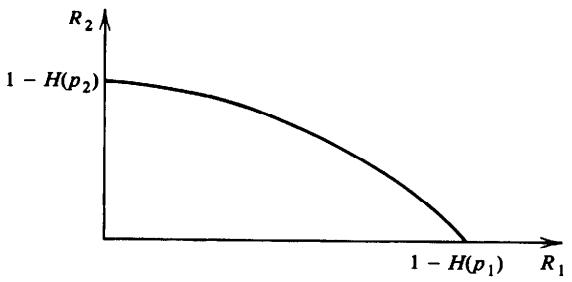


Figure 14.28. Capacity region of binary symmetric broadcast channel.

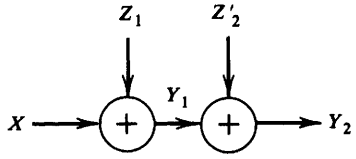


Figure 14.29. Gaussian broadcast channel.

$$R_1 < C\left(\frac{\alpha P}{N_1}\right) \tag{14.232}$$

$$R_2 < C\left(\frac{(1-\alpha)P}{\alpha P + N_2}\right) \tag{14.233}$$

where α may be arbitrarily chosen ($0 \leq \alpha \leq 1$). The coding scheme that achieves this capacity region is outlined in Section 14.1.3.

14.7 THE RELAY CHANNEL

The relay channel is a channel in which there is one sender and one receiver with a number of intermediate nodes which act as relays to help the communication from the sender to the receiver. The simplest relay channel has only one intermediate or relay node. In this case the channel consists of four finite sets $\mathcal{X}$, $\mathcal{X}_1$, $\mathcal{Y}$ and $\mathcal{Y}_1$ and a collection of probability mass functions $p(\cdot, \cdot | x, x_1)$ on $\mathcal{Y} \times \mathcal{Y}_1$, one for each $(x, x_1) \in \mathcal{X} \times \mathcal{X}_1$. The interpretation is that x is the input to the channel and y is the output of the channel, y_1 is the relay's observation and x_1 is the input symbol chosen by the relay, as shown in Figure 14.30. The problem is to find the capacity of the channel between the sender X and the receiver Y.

The relay channel combines a broadcast channel (X to Y and Y_1) and a multiple access channel (X and X_1 to Y). The capacity is known for the special case of the physically degraded relay channel. We will first prove an outer bound on the capacity of a general relay channel and later establish an achievable region for the degraded relay channel.

Definition: A $(2^{nR}, n)$ code for a relay channel consists of a set of integers $\mathcal{W} = \{1, 2, \ldots, 2^{nR}\}$, an encoding function

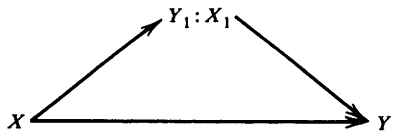


Figure 14.30. The relay channel.

$$X: \{1, 2, \ldots, 2^{nR}\} \to \mathcal{X}^n , \tag{14.234}$$

a set of relay functions $\{f_i\}_{i=1}^n$ such that

$$x_{1i} = f_i(Y_{11}, Y_{12}, \ldots, Y_{1i-1}), \qquad 1 \le i \le n , \tag{14.235}$$

and a decoding function,

$$g: \mathcal{Y}^n \to \{1, 2, \ldots, 2^{nR}\} . \tag{14.236}$$

Note that the definition of the encoding functions includes the nonanticipatory condition on the relay. The relay channel input is allowed to depend only on the past observations $y_{11}, y_{12}, \ldots, y_{1i-1}$. The channel is memoryless in the sense that (Y_i, Y_{1i}) depends on the past only through the current transmitted symbols (X_i, X_{1i}). Thus for any choice $p(w)$, $w \in \mathcal{W}$, and code choice $X: \{1, 2, \ldots, 2^{nR}\} \to \mathcal{X}_i^n$ and relay functions $\{f_i\}_{i=1}^n$, the joint probability mass function on $\mathcal{W} \times \mathcal{X}^n \times \mathcal{X}_1^n \times \mathcal{Y}^n \times \mathcal{Y}_1^n$ is given by

$$p(w, \mathbf{x}, \mathbf{x}_1, \mathbf{y}, \mathbf{y}_1)$$

$$= p(w) \prod_{i=1}^{n} p(x_i|w) p(x_{1i}|y_{11}, y_{12}, \ldots, y_{1i-1}) p(y_i, y_{1i}|x_i, x_{1i}) . \tag{14.237}$$

If the message $w \in [1, 2^{nR}]$ is sent, let

$$\lambda(w) = \Pr\{g(\mathbf{Y}) \neq w | w \text{ sent}\} \tag{14.238}$$

denote the conditional probability of error. We define the average probability of error of the code as

$$P_e^{(n)} = \frac{1}{2^{nR}} \sum_w \lambda(w) . \tag{14.239}$$

The probability of error is calculated under the uniform distribution over the codewords $w \in \{1, 2^{nR}\}$. The rate R is said to be achievable by the relay channel if there exists a sequence of $(2^{nR}, n)$ codes with $P_e^{(n)} \to 0$. The *capacity* C of a relay channel is the supremum of the set of achievable rates.

We first give an upper bound on the capacity of the relay channel.

Theorem 14.7.1: *For any relay channel* $(\mathcal{X} \times \mathcal{X}_1, p(y, y_1|x, x_1), \mathcal{Y} \times \mathcal{Y}_1)$ *the capacity* C *is bounded above by*

$$C \le \sup_{p(x, x_1)} \min\{I(X, X_1; Y), I(X; Y, Y_1|X_1)\} . \tag{14.240}$$

Proof: The proof is a direct consequence of a more general max flow min cut theorem to be given in Section 14.10. □

This upper bound has a nice max flow min cut interpretation. The first term in (14.240) upper bounds the maximum rate of information transfer from senders X and X_1 to receiver Y. The second terms bound the rate from X to Y and Y_1.

We now consider a family of relay channels in which the relay receiver is better than the ultimate receiver Y in the sense defined below. Here the max flow min cut upper bound in the (14.240) is achieved.

Definition: The relay channel $(\mathcal{X} \times \mathcal{X}_1, p(y, y_1|x, x_1), \mathcal{Y} \times \mathcal{Y}_1)$ is said to be *physically degraded* if $p(y, y_1|x, x_1)$ can be written in the form

$$p(y, y_1|x, x_1) = p(y_1|x, x_1)p(y|y_1, x_1) . \tag{14.241}$$

Thus Y is a random degradation of the relay signal Y_1.

For the physically degraded relay channel, the capacity is given by the following theorem.

Theorem 14.7.2: *The capacity C of a physically degraded relay channel is given by*

$$C = \sup_{p(x, x_1)} \min\{I(X, X_1; Y), I(X; Y_1|X_1)\} , \tag{14.242}$$

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

Proof (*Converse*): The proof follows from Theorem 14.7.1 and by degradedness, since for the degraded relay channel, $I(X; Y, Y_1|X_1) = I(X; Y_1|X_1)$.

Achievability. The proof of achievability involves a combination of the following basic techniques: (1) random coding, (2) list codes, (3) Slepian-Wolf partitioning, (4) coding for the cooperative multiple access channel, (5) superposition coding, and (6) block Markov encoding at the relay and transmitter.

We provide only an outline of the proof.

Outline of achievability. We consider B blocks of transmission, each of n symbols. A sequence of $B-1$ indices, $w_i \in \{1, \ldots, 2^{nR}\}$, $i = 1, 2, \ldots, B-1$ will be sent over the channel in nB transmissions. (Note that as $B \to \infty$, for a fixed n, the rate $R(B-1)/B$ is arbitrarily close to R.)

We define a doubly-indexed set of codewords:

$$\mathscr{C} = \{\mathbf{x}(w|s), \mathbf{x}_1(s)\} : w \in \{1, 2^{nR}\}, s \in \{1, 2^{nR_0}\}, \mathbf{x} \in \mathscr{X}^n, \mathbf{x}_1 \in \mathscr{X}_1^n . \tag{14.243}$$

We will also need a partition

$$\mathscr{S} = \{S_1, S_2, \ldots, S_{2^{nR_0}}\} \text{ of } \mathscr{W} = \{1, 2, \ldots, 2^{nR}\} \tag{14.244}$$

into 2^{nR_0} cells, with $S_i \cap S_j = \phi$, $i \neq j$, and $\cup S_i = \mathscr{W}$. The partition will enable us to send side information to the receiver in the manner of Slepian and Wolf [255].

Generation of random code. Fix $p(x_1)p(x|x_1)$.

First generate at random 2^{nR_0} i.i.d. n-sequences in $\mathscr{X}_1^n$, each drawn according to $p(\mathbf{x}_1) = \prod_{i=1}^n p(x_{1i})$. Index them as $\mathbf{x}_1(s)$, $s \in \{1, 2, \ldots, 2^{nR_0}\}$. For each $\mathbf{x}_1(s)$, generate 2^{nR} conditionally independent n-sequences $\mathbf{x}(w|s)$, $w \in \{1, 2, \ldots, 2^{nR}\}$, drawn independently according to $p(\mathbf{x}|\mathbf{x}_1(s)) = \prod_{i=1}^n p(x_i|x_{1i}(s))$. This defines the random codebook $\mathscr{C} = \{\mathbf{x}(w|s), \mathbf{x}_1(s)\}$.

The random partition $\mathscr{S} = \{S_1, S_2, \ldots, S_{2^{nR_0}}\}$ of $\{1, 2, \ldots, 2^{nR}\}$ is defined as follows. Let each integer $w \in \{1, 2, \ldots, 2^{nR}\}$ be assigned independently, according to a uniform distribution over the indices $s = 1, 2, \ldots, 2^{nR_0}$, to cells S_s.

Encoding. Let $w_i \in \{1, 2, \ldots, 2^{nR}\}$ be the new index to be sent in block i, and let s_i be defined as the partition corresponding to w_{i-1}, i.e., $w_{i-1} \in S_{s_i}$. The encoder sends $\mathbf{x}(w_i|s_i)$. The relay has an estimate $\hat{w}_{i-1}$ of the previous index w_{i-1}. (This will be made precise in the decoding section.) Assume that $\hat{w}_{i-1} \in S_{\hat{s}_i}$. The relay encoder sends $\mathbf{x}_1(\hat{s}_i)$ in block i.

Decoding. We assume that at the end of block $i-1$, the receiver knows $(w_1, w_2, \ldots, w_{i-2})$ and $(s_1, s_2, \ldots, s_{i-1})$ and the relay knows $(w_1, w_2, \ldots, w_{i-1})$ and consequently $(s_1, s_2, \ldots, s_i)$.

The decoding procedures at the end of block i are as follows:

1. Knowing s_i and upon receiving $\mathbf{y}_1(i)$, the *relay* receiver estimates the message of the transmitter $\hat{w}_i = w$ if and only if there exists a unique w such that $(\mathbf{x}(w|s_i), \mathbf{x}_1(s_i), \mathbf{y}_1(i))$ are jointly ϵ-typical. Using Theorem 14.2.3, it can be shown that $\hat{w}_i = w_i$ with an arbitrarily small probability of error if

$$R < I(X; Y_1|X_1) \tag{14.245}$$

and n is sufficiently large.

2. The *receiver* declares that $\hat{s}_i = s$ was sent iff there exists one and only one s such that $(\mathbf{x}_1(s), \mathbf{y}(i))$ are jointly ϵ-typical. From Theorem 14.2.1, we know that s_i can be decoded with arbitrarily small probability of error if

$$R_0 < I(X_1; Y) \tag{14.246}$$

and n is sufficiently large.

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

$$R < I(X; Y|X_1) + R_0\,, \tag{14.247}$$

then $\hat{w}_{i-1} = w_{i-1}$ with arbitrarily small probability of error. Combining the two constraints (14.246) and (14.247), R_0 drops out, leaving

$$R < I(X; Y|X_1) + I(X_1; Y) = I(X, X_1; Y)\,. \tag{14.248}$$

For a detailed analysis of the probability of error, the reader is referred to Cover and El Gamal [67]. □

Theorem 14.7.2 can also shown to be the capacity for the following classes of relay channels.

(i) Reversely degraded relay channel, i.e.,

$$p(y, y_1|x, x_1) = p(y|x, x_1)p(y_1|y, x_1)\,. \tag{14.249}$$

(ii) Relay channel with feedback.

(iii) Deterministic relay channel,

$$y_1 = f(x, x_1), \qquad y = g(x, x_1)\,. \tag{14.250}$$

14.8 SOURCE CODING WITH SIDE INFORMATION

We now consider the distributed source coding problem where two random variables X and Y are encoded separately but only X is to be recovered. We now ask how many bits R_1 are required to describe X if we are allowed R_2 bits to describe Y.

If $R_2 > H(Y)$, then Y can be described perfectly, and by the results of Slepian-Wolf coding, $R_1 = H(X|Y)$ bits suffice to describe X. At the other extreme, if $R_2 = 0$, we must describe X without any help, and $R_1 = H(X)$ bits are then necessary to describe X. In general, we will use $R_2 = I(Y; \hat{Y})$ bits to describe an approximate version of Y. This will allow us to describe X using $H(X|\hat{Y})$ bits in the presence of side information $\hat{Y}$. The following theorem is consistent with this intuition.

Theorem 14.8.1: *Let $(X, Y) \sim p(x, y)$. If Y is encoded at rate R_2 and X is encoded at rate R_1, we can recover X with an arbitrarily small probability of error if and only if*

$$R_1 \geq H(X|U), \tag{14.251}$$

$$R_2 \geq I(Y; U) \tag{14.252}$$

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

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

Proof (*Converse*): Consider any source code for Figure 14.31. The source code consists of mappings $f_n(X^n)$ and $g_n(Y^n)$ such that the rates of f_n and g_n are less than R_1 and R_2, respectively, and a decoding mapping h_n such that

$$P_e^{(n)} = \Pr\{h_n(f_n(X^n), g_n(Y^n)) \neq X^n\} < \epsilon . \tag{14.253}$$

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

$$H(X^n|S, T) \leq n\epsilon_n . \tag{14.254}$$

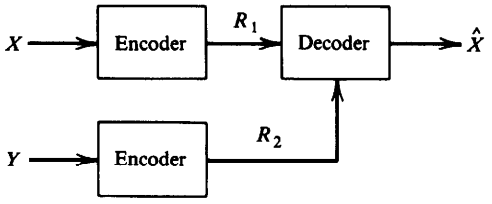


Figure 14.31. Encoding with side information.

Then

$$nR_2 \overset{(a)}{\geq} H(T) \tag{14.255}$$

$$\overset{(b)}{\geq} I(Y^n; T) \tag{14.256}$$

$$= \sum_{i=1}^{n} I(Y_i; T|Y_1, \ldots, Y_{i-1}) \tag{14.257}$$

$$\overset{(c)}{=} \sum_{i=1}^{n} I(Y_i; T, Y_1, \ldots, Y_{i-1}) \tag{14.258}$$

$$\overset{(d)}{=} \sum_{i=1}^{n} I(Y_i; U_i) \tag{14.259}$$

where

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

We also have another chain for R_1,

$$nR_1 \overset{(a)}{\geq} H(S) \tag{14.260}$$

$$\overset{(b)}{\geq} H(S|T) \tag{14.261}$$

$$= H(S|T) + H(X^n|S, T) - H(X^n|S, T) \tag{14.262}$$

$$\overset{(c)}{\geq} H(X^n, S|T) - n\epsilon_n \tag{14.263}$$

$$\overset{(d)}{=} H(X^n|T) - n\epsilon_n \tag{14.264}$$

$$\overset{(e)}{=} \sum_{i=1}^{n} H(X_i|T, X_1, \ldots, X_{i-1}) - n\epsilon_n \tag{14.265}$$

$$\overset{(f)}{\geq} \sum_{i=1}^{n} H(X_i|T, X^{i-1}, Y^{i-1}) - n\epsilon_n \tag{14.266}$$

$$\overset{(g)}{=} \sum_{i=1}^{n} H(X_i|T, Y^{i-1}) - n\epsilon_n \tag{14.267}$$

$$\overset{(h)}{=} \sum_{i=1}^{n} H(X_i|U_i) - n\epsilon_n \tag{14.268}$$

where

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

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

$$R_1 \geq \frac{1}{n} \sum_{i=1}^{n} H(X_i|U_i) \tag{14.269}$$

$$R_2 \geq \frac{1}{n} \sum_{i=1}^{n} I(Y_i; U_i) . \tag{14.270}$$

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

$$R_1 \geq \frac{1}{n} \sum_{i=1}^{n} H(X_i|U_i, Q = i) = H(X_Q|U_Q, Q) \tag{14.271}$$

$$R_2 \geq \frac{1}{n} \sum_{i=1}^{n} I(Y_i; U_i|Q = i) = I(Y_Q; U_Q|Q) \tag{14.272}$$

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

$$I(Y_Q; U_Q|Q) = I(Y_Q; U_Q, Q) - I(Y_Q; Q) = I(Y_Q; U_Q, Q) . \tag{14.273}$$

Now X_Q and Y_Q have the joint distribution $p(x, y)$ in the theorem.

Defining $U = (U_Q, Q)$, $X = X_Q$, and $Y = Y_Q$, we have shown the existence of a random variable U such that

$$R_1 \geq H(X|U), \tag{14.274}$$

$$R_2 \geq I(Y; U) \tag{14.275}$$

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

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

$$\left| \frac{1}{n} N(a, b, c|x^n, y^n, z^n) - p(a, b, c) \right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|}. \tag{14.276}$$

In particular, this implies that (x^n, y^n) are jointly strongly typical and that (y^n, z^n) are also jointly strongly typical. But the converse is not true: the fact that $(x^n, y^n) \in A_\epsilon^{*(n)}(X, Y)$ and $(y^n, z^n) \in A_\epsilon^{*(n)}(Y, Z)$ does not in general imply that $(x^n, y^n, z^n) \in A_\epsilon^{*(n)}(X, Y, Z)$. But if $X \to Y \to Z$ forms a Markov chain, this implication is true. We state this as a lemma without proof [28, 83].

Lemma 14.8.1: *Let (X, Y, Z) form a Markov chain $X \to Y \to Z$, i.e., $p(x, y, z) = p(x, y)p(z|y)$. If for a given $(y^n, z^n) \in A_\epsilon^{*(n)}(Y, Z)$, X^n is drawn $\sim \prod_{i=1}^n p(x_i|y_i)$, then $\Pr\{(X^n, y^n, z^n) \in A_\epsilon^{*(n)}(X, Y, Z)\} > 1 - \epsilon$ for n sufficiently large.*

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

We now outline the proof of achievability in Theorem 14.8.1.

Proof (*Achievability in Theorem 14.8.1*): Fix $p(u|y)$. Calculate $p(u) = \sum_y p(y)p(u|y)$.

Generation of codebooks. Generate 2^{nR_2} independent codewords of length n, $\mathbf{U}(w_2)$, $w_2 \in \{1, 2, \ldots, 2^{nR_2}\}$ according to $\prod_{i=1}^n p(u_i)$.

Randomly bin all the X^n sequences into 2^{nR_1} bins by independently generating an index b uniformly distributed on $\{1, 2, \ldots, 2^{nR_1}\}$ for each X^n. Let $B(i)$ denote the set of X^n sequences allotted to bin i.

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

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

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

1. The pair (X^n, Y^n) generated by the source is not typical. The probability of this is small if n is large. Hence, without loss of generality, we can condition on the event that the source produces a particular typical sequence $(x^n, y^n) \in A_\epsilon^{*(n)}$.
2. The sequence Y^n is typical, but there does not exist a $U^n(s)$ in the codebook which is jointly typical with it. The probability of this is small from the arguments of Section 13.6, where we showed that if there are enough codewords, i.e., if

$$R_2 > I(Y; U), \tag{14.277}$$

   then we are very likely to find a codeword that is jointly strongly typical with the given source sequence.
3. The codeword $U^n(s)$ is jointly typical with y^n but not with x^n. But by Lemma 14.8.1, the probability of this is small since $X \to Y \to U$ forms a Markov chain.
4. We also have an error if there exists another typical $X^n \in B(i)$ which is jointly typical with $U^n(s)$. The probability that any other X^n is jointly typical with $U^n(s)$ is less than $2^{-n(I(U;X)-3\epsilon)}$, and therefore the probability of this kind of error is bounded above by

$$|B(i) \cap A_\epsilon^{*(n)}(X)| 2^{-n(I(X;U)-3\epsilon)} \leq 2^{n(H(X)+\epsilon)} 2^{-nR_1} 2^{-n(I(X;U)-3\epsilon)}, \tag{14.278}$$

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

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

14.9 RATE DISTORTION WITH SIDE INFORMATION

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

We will begin with a few definitions. Let (X_i, Y_i) be i.i.d. $\sim p(x, y)$ and encoded as shown in Figure 14.32.

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

$$\limsup_{n\to\infty} Ed(X^n, g_n(Y^n, i_n(X^n))) \leq D . \tag{14.279}$$

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

Theorem 14.9.1 (*Rate distortion with side information*): *Let* (X, Y) *be drawn i.i.d.* $\sim p(x, y)$ *and let* $d(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^n d(x_i, \hat{x}_i)$ *be given. The rate distortion function with side information is*

$$R_Y(D) = \min_{p(w|x)} \min_f (I(X; W) - I(Y; W)) \tag{14.280}$$

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

$$\sum_x \sum_w \sum_y p(x, y)p(w|x)\, d(x, f(y, w)) \leq D . \tag{14.281}$$

The function f in the theorem corresponds to the decoding map that maps the encoded version of the X symbols and the side information Y to

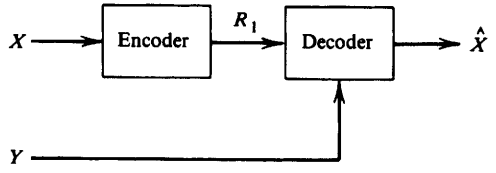


Figure 14.32. Rate distortion with side information.

the output alphabet. We minimize over all conditional distributions on W and functions f such that the expected distortion for the joint distribution is less than D.

We first prove the converse after considering some of the properties of the function $R_Y(D)$ defined in (14.280).

Lemma 14.9.1: *The rate distortion function with side information $R_Y(D)$ defined in (14.280) is a non-increasing convex function of D.*

Proof: The monotonicity of $R_Y(D)$ follows immediately from the fact that the domain of minimization in the definition of $R_Y(D)$ increases with D.

As in the case of rate distortion without side information, we expect $R_Y(D)$ to be convex. However, the proof of convexity is more involved because of the double rather than single minimization in the definition of $R_Y(D)$ in (14.280). We outline the proof here.

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

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

$$D = Ed(X, \hat{X}) \tag{14.282}$$

$$= \lambda Ed(X, f_1(W_1, Y)) + (1-\lambda)Ed(X, f_2(W_2, Y)) \tag{14.283}$$

$$= \lambda D_1 + (1-\lambda)D_2, \tag{14.284}$$

and (14.280) becomes

$$I(W; X) - I(W; Y) = H(X) - H(X|W) - H(Y) + H(Y|W) \tag{14.285}$$

$$= H(X) - H(X|W_Q, Q) - H(Y) + H(Y|W_Q, Q) \tag{14.286}$$

$$= H(X) - \lambda H(X|W_1) - (1-\lambda)H(X|W_2)$$

$$- H(Y) + \lambda H(Y|W_1) + (1-\lambda)H(Y|W_2) \tag{14.287}$$

$$= \lambda(I(W_1, X) - I(W_1; Y))$$

$$+ (1-\lambda)\Big(I(W_2, X) - I(W_2; Y)\Big), \tag{14.288}$$

and hence

$$R_Y(D) = \min_{U: Ed \le D} (I(U; X) - I(U; Y)) \tag{14.289}$$

$$\le I(W; X) - I(W; Y) \tag{14.290}$$

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

$$= \lambda R_Y(D_1) + (1 - \lambda) R_Y(D_2), \tag{14.291}$$

proving the convexity of $R_Y(D)$. □

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

Proof (*Converse to Theorem 14.9.1*): Consider any rate distortion code with side information. Let the encoding function be $f_n: \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\}$. Let the decoding function be $g_n: \mathcal{Y}^n \times \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n$ and let $g_{ni}: \mathcal{Y}^n \times \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}$ denote the ith symbol produced by the decoding function. Let $T = f_n(X^n)$ denote the encoded version of X^n. We must show that if $Ed(X^n, g_n(Y^n, f_n(X^n))) \le D$, then $R \ge R_Y(D)$.

We have the following chain of inequalities:

$$nR \overset{(a)}{\ge} H(T) \tag{14.292}$$

$$\overset{(b)}{\ge} H(T|Y^n) \tag{14.293}$$

$$\ge I(X^n; T|Y^n) \tag{14.294}$$

$$\overset{(c)}{=} \sum_{i=1}^{n} I(X_i; T|Y^n, X^{i-1}) \tag{14.295}$$

$$= \sum_{i=1}^{n} H(X_i|Y^n, X^{i-1}) - H(X_i|T, Y^n, X^{i-1}) \tag{14.296}$$

$$\overset{(d)}{=} \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|T, Y^{i-1}, Y_i, Y_{i+1}^n, X^{i-1}) \tag{14.297}$$

$$\overset{(e)}{\ge} \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|T, Y^{i-1}, Y_i, Y_{i+1}^n) \tag{14.298}$$

$$\overset{(f)}{=} \sum_{i=1}^{n} H(X_i|Y_i) - H(X_i|W_i, Y_i) \tag{14.299}$$

$$\overset{(g)}{=} \sum_{i=1}^{n} I(X_i; W_i|Y_i) \tag{14.300}$$

$$= \sum_{i=1}^{n} H(W_i|Y_i) - H(W_i|X_i, Y_i) \tag{14.301}$$

$$\overset{(h)}{=} \sum_{i=1}^{n} H(W_i|Y_i) - H(W_i|X_i) \tag{14.302}$$

$$= \sum_{i=1}^{n} H(W_i) - H(W_i|X_i) - H(W_i) + H(W_i|Y_i) \tag{14.303}$$

$$= \sum_{i=1}^{n} I(W_i; X_i) - I(W_i; Y_i) \tag{14.304}$$

$$\overset{(i)}{\geq} \sum_{i=1}^{n} R_Y(Ed(X_i, g'_{ni}(W_i, Y_i))) \tag{14.305}$$

$$= n\frac{1}{n}\sum_{i=1}^{n} R_Y(Ed(X_i, g'_{ni}(W_i, Y_i))) \tag{14.306}$$

$$\overset{(j)}{\geq} nR_Y\left(E\,\frac{1}{n}\sum_{i=1}^{n} d(X_i, g'_{ni}(W_i, Y_i))\right) \tag{14.307}$$

$$\overset{(k)}{\geq} nR_Y(D), \tag{14.308}$$

where

(a) follows from the fact that the range of T is $\{1, 2, \ldots, 2^{nR}\}$,
(b) from the fact that conditioning reduces entropy,
(c) from the chain rule for mutual information,
(d) from the fact that X_i is independent of the past and future Y's and X's given Y_i,
(e) from the fact that conditioning reduces entropy,
(f) follows by defining $W_i = (T, Y^{i-1}, Y_{i+1}^n)$,
(g) follows from the defintion of mutual information,
(h) follows from the fact that since Y_i depends only on X_i and is conditionally independent of T and the past and future Y's, and therefore $W_i \to X_i \to Y_i$ forms a Markov chain,
(i) follows from the definition of the (information) conditional rate distortion function, since $\hat{X}_i = g_{ni}(T, Y^n) \overset{\Delta}{=} g'_{ni}(W_i, Y_i)$, and hence $I(W_i; X_i) - I(W_i; Y_i) \geq \min_{W:Ed(X,\hat{X})\leq D_i} I(W; X) - I(W; Y) = R_Y(D_i)$,
(j) follows from Jensen's inequality and the convexity of the conditional rate distortion function (Lemma 14.9.1), and
(k) follows from the definition of $D = E\frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i)$. □

It is easy to see the parallels between this converse and the converse for rate distortion without side information (Section 13.4). The proof of

achievability is also parallel to the proof of the rate distortion theorem using strong typicality. However, instead of sending the index of the codeword that is jointly typical with the source, we divide these codewords into bins and send the bin index instead. If the number of codewords in each bin is small enough, then the side information can be used to isolate the particular codeword in the bin at the receiver. Hence again we are combining random binning with rate distortion encoding to find a jointly typical reproduction codeword. We outline the details of the proof below.

Proof (*Achievability of Theorem 14.9.1*): Fix $p(w|x)$ and the function $f(w, y)$. Calculate $p(w) = \sum_x p(x)p(w|x)$.

Generation of codebook. Let $R_1 = I(X; W) + \epsilon$. Generate 2^{nR} i.i.d. codewords $W^n(s) \sim \prod_{i=1}^n p(w_i)$, and index them by $s \in \{1, 2, \ldots, 2^{nR_1}\}$.

Let $R_2 = I(X; W) - I(Y; W) + 5\epsilon$. Randomly assign the indices $s \in \{1, 2, \ldots, 2^{nR_1}\}$ to one of 2^{nR_2} bins using a uniform distribution over the bins. Let $B(i)$ denote the indices assigned to bin i. There are approximately $2^{n(R_1 - R_2)}$ indices in each bin.

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

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

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

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

$$R_1 > I(W; X). \tag{14.309}$$

3. The pair of sequences $(X^n, W^n(s)) \in A_\epsilon^{*(n)}$ but $(W^n(s), Y^n) \notin A_\epsilon^{*(n)}$,

i.e., the codeword is not jointly typical with the Y^n sequence. By the Markov lemma (Lemma 14.8.1), the probability of this event is small if n is large enough.

4. There exists another s' with the same bin index such that $(W^n(s'), Y^n) \in A_\epsilon^{*(n)}$. Since the probability that a randomly chosen W^n is jointly typical with Y^n is $\approx 2^{-nI(Y;W)}$, the probability that there is another W^n in the same bin that is typical with Y^n is bounded by the number of codewords in the bin times the probability of joint typicality, i.e.,

$$\Pr(\exists s' \in B(i): (W^n(s'), Y^n) \in A_\epsilon^{*(n)}) \leq 2^{n(R_1 - R_2)} 2^{-n(I(W;Y)-3\epsilon)}, \tag{14.310}$$

which goes to zero since $R_1 - R_2 < I(Y; W) - 3\epsilon$.

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

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

The reader is referred to Wyner and Ziv [284] for the details of the proof.

After the discussion of the various situations of compressing distributed data, it might be expected that the problem is almost completely solved. But unfortunately this is not true. An immediate generalization

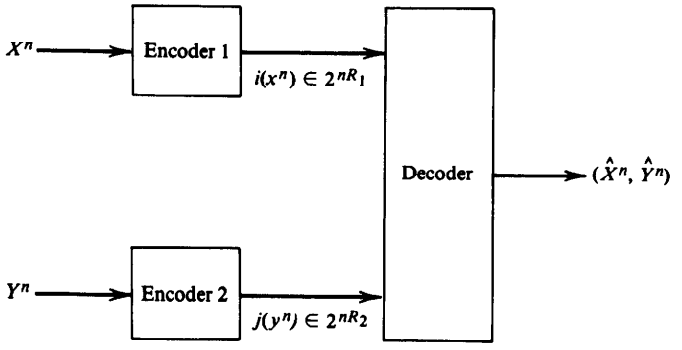


Figure 14.33. Rate distortion for two correlated sources.

of all the above problems is the rate distortion problem for correlated sources, illustrated in Figure 14.33. This is essentially the Slepian-Wolf problem with distortion in both X and Y. It is easy to see that the three distributed source coding problems considered above are all special cases of this setup. Unlike the earlier problems, though, this problem has not yet been solved and the general rate distortion region remains unknown.

14.10 GENERAL MULTITERMINAL NETWORKS

We conclude this chapter by considering a general multiterminal network of senders and receivers and deriving some bounds on the rates achievable for communication in such a network.

A general multiterminal network is illustrated in Figure 14.34. In this section, superscripts denote node indices and subscripts denote time indices. There are m nodes, and node i has an associated transmitted variable $X^{(i)}$ and a received variable $Y^{(i)}$. The node i sends information at rate $R^{(ij)}$ to node j. We assume that all the messages $W^{(ij)}$ being sent from node i to node j are independent and uniformly distributed over their respective ranges $\{1, 2, \ldots, 2^{nR^{(ij)}}\}$.

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

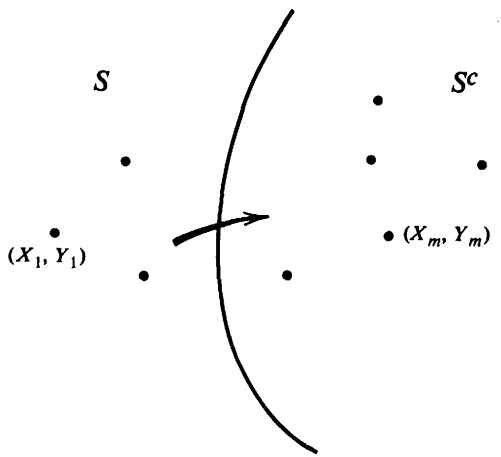


Figure 14.34. A general multiterminal network.

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

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

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

$$P_e^{(n)(ij)} = \Pr(\hat{W}^{(ij)}(\mathbf{Y}^{(j)}, W^{(j1)}, \ldots, W^{(jm)}) \neq W^{(ij)}), \tag{14.311}$$

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

A set of rates $\{R^{(ij)}\}$ is said to be achievable if there exist encoders and decoders with block length n with $P_e^{(n)(ij)} \to 0$ as $n \to \infty$ for all $i, j \in \{1, 2, \ldots, m\}$.

We use this formulation to derive an upper bound on the flow of information in any multiterminal network. We divide the nodes into two sets, S and the complement S^c. We now bound the rate of flow of information from nodes in S to nodes in S^c.

Theorem 14.10.1: *If the information rates $\{R^{(ij)}\}$ are achievable, then there exists some joint probability distribution $p(x^{(1)}, x^{(2)}, \ldots, x^{(m)})$, such that*

$$\sum_{i \in S,\, j \in S^c} R^{(ij)} \leq I(X^{(S)}; Y^{(S^c)} | X^{(S^c)}), \tag{14.312}$$

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

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

$$n \sum_{i \in S,\, j \in S^c} R^{(ij)} \tag{14.313}$$

$$\stackrel{(a)}{=} \sum_{i \in S,\, j \in S^c} H(W^{(ij)}) \tag{14.314}$$

$$\stackrel{(b)}{=} H(W^{(T)}) \tag{14.315}$$

$$\stackrel{(c)}{=} H(W^{(T)}|W^{(T^c)}) \tag{14.316}$$

$$= I(W^{(T)}; Y_1^{(S^c)}, \ldots, Y_n^{(S^c)}|W^{(T^c)}) \tag{14.317}$$

$$+ H(W^{(T)}|Y_1^{(S^c)}, \ldots, Y_n^{(S^c)}, W^{(T^c)}) \tag{14.318}$$

$$\stackrel{(d)}{\leq} I(W^{(T)}; Y_1^{(S^c)}, \ldots, Y_n^{(S^c)}|W^{(T^c)}) + n\epsilon_n \tag{14.319}$$

$$\stackrel{(e)}{=} \sum_{k=1}^{n} I(W^{(T)}; Y_k^{(S^c)}|Y_1^{(S^c)}, \ldots, Y_{k-1}^{(S^c)}, W^{(T^c)}) + n\epsilon_n \tag{14.320}$$

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

$$- H(Y_k^{(S^c)}|Y_1^{(S^c)}, \ldots, Y_{k-1}^{(S^c)}, W^{(T^c)}, W^{(T)}) + n\epsilon_n \tag{14.321}$$

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

$$- H(Y_k^{(S^c)}|Y_1^{(S^c)}, \ldots, Y_{k-1}^{(S^c)}, W^{(T^c)}, W^{(T)}, X_k^{(S)}, X_k^{(S^c)}) + n\epsilon_n \tag{14.322}$$

$$\stackrel{(h)}{\leq} \sum_{k=1}^{n} H(Y_k^{(S^c)}|X_k^{(S^c)}) - H(Y_k^{(S^c)}|X_k^{(S^c)}, X_k^{(S)}) + n\epsilon_n \tag{14.323}$$

$$= \sum_{k=1}^{n} I(X_k^{(S)}; Y_k^{(S^c)}|X_k^{(S^c)}) + n\epsilon_n \tag{14.324}$$

$$\stackrel{(i)}{=} n \frac{1}{n} \sum_{k=1}^{n} I(X_Q^{(S)}; Y_Q^{(S^c)}|X_Q^{(S^c)}, Q = k) + n\epsilon_n \tag{14.325}$$

$$\stackrel{(j)}{=} nI(X_Q^{(S)}; Y_Q^{(S^c)}|X_Q^{(S^c)}, Q) + n\epsilon_n \tag{14.326}$$

$$= n(H(Y_Q^{(S^c)}|X_Q^{(S^c)}, Q) - H(Y_Q^{(S^c)}|X_Q^{(S)}, X_Q^{(S^c)}, Q)) + n\epsilon_n \tag{14.327}$$

$$\stackrel{(k)}{\leq} n(H(Y_Q^{(S^c)}|X_Q^{(S^c)}) - H(Y_Q^{(S^c)}|X_Q^{(S)}, X_Q^{(S^c)}, Q)) + n\epsilon_n \tag{14.328}$$

$$\stackrel{(l)}{\leq} n(H(Y_Q^{(S^c)}|X_Q^{(S^c)}) - H(Y_Q^{(S^c)}|X_Q^{(S)}, X_Q^{(S^c)})) + n\epsilon_n \tag{14.329}$$

$$= nI(X_Q^{(S)}; Y_Q^{(S^c)}|X_Q^{(S^c)}) + n\epsilon_n\,, \tag{14.330}$$

where

(a) follows from the fact that the messages $W^{(ij)}$ are uniformly distributed over their respective ranges $\{1, 2, \ldots, 2^{nR^{(ij)}}\}$,
(b) follows from the definition of $W^{(T)} = \{W^{(ij)}: i \in S, j \in S^c\}$ and the fact that the messages are independent,
(c) follows from the independence of the messages for T and T^c,
(d) follows from Fano's inequality since the messages $W^{(T)}$ can be decoded from $Y^{(S)}$ and $W^{(T^c)}$,
(e) is the chain rule for mutual information,
(f) follows from the definition of mutual information,
(g) follows from the fact that $X_k^{(S^c)}$ is a function of the past received symbols $Y^{(S^c)}$ and the messages $W^{(T^c)}$ and the fact that adding conditioning reduces the second term,
(h) from the fact that $Y_k^{(S^c)}$ depends only on the current input symbols $X_k^{(S)}$ and $X_k^{(S^c)}$,
(i) follows after we introduce a new timesharing random variable Q uniformly distributed on $\{1, 2, \ldots, n\}$,
(j) follows from the definition of mutual information,
(k) follows from the fact that conditioning reduces entropy, and
(l) follows from the fact that $Y_Q^{(S^c)}$ depends only the inputs $X_Q^{(S)}$ and $X_Q^{(S^c)}$ and is conditionally independent of Q.

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

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

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

- *Multiple access channel.* The multiple access channel is a network with many input nodes and one output node. For the case of a two-user multiple access channel, the bounds of Theorem 14.10.1 reduce to

$$R_1 \le I(X_1; Y|X_2), \tag{14.331}$$

$$R_2 \le I(X_2; Y|X_1), \tag{14.332}$$

$$R_1 + R_2 \le I(X_1, X_2; Y) \tag{14.333}$$

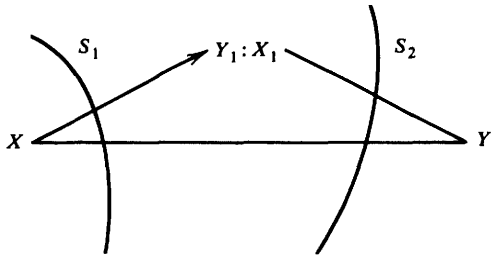


Figure 14.35. The relay channel.

for some joint distribution $p(x_1, x_2)p(y|x_1, x_2)$. These bounds coincide with the capacity region if we restrict the input distribution to be a product distribution and take the convex hull (Theorem 14.3.1).

- *Relay channel.* For the relay channel, these bounds give the upper bound of Theorem 14.7.1 with different choices of subsets as shown in Figure 14.35. Thus

$$C \leq \sup_{p(x, x_1)} \min\{I(X, X_1; Y), I(X; Y, Y_1|X_1)\} . \quad (14.334)$$

This upper bound is the capacity of a physically degraded relay channel, and for the relay channel with feedback [67].

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

- *The source channel separation theorem.* In Section 8.13, we discussed the source channel separation theorem, which proves that we can transmit the source noiselessly over the channel if and only if the entropy rate is less than the channel capacity. This allows us to characterize a source by a single number (the entropy rate) and the channel by a single number (the capacity).

  What about the multi-user case? We would expect that a distributed source could be transmitted over a channel if and only if the rate region for the noiseless coding of the source lay within the capacity region of the channel. To be specific, consider the transmission of a distributed source over a multiple access channel, as shown in Figure 14.36. Combining the results of Slepian-Wolf encoding with the capacity results for the multiple access channel, we can show that we can transmit the source over the channel and recover it with a low probability of error if

$$H(U|V) \leq I(X_1; Y|X_2, Q) , \quad (14.335)$$

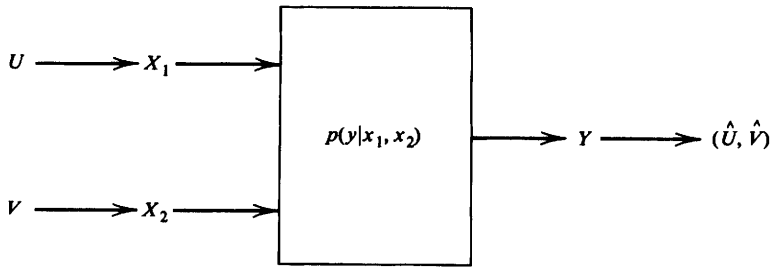


Figure 14.36. Transmission of correlated sources over a multiple access channel.

$$H(V|U) \leq I(X_2; Y|X_1, Q), \tag{14.336}$$

$$H(U, V) \leq I(X_1, X_2; Y, Q) \tag{14.337}$$

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

But is this condition also necessary? No, as a simple example illustrates. Consider the transmission of the source of Example 14.4 over the binary erasure multiple access channel (Example 14.3). The Slepian-Wolf region does not intersect the capacity region, yet it is simple to devise a scheme that allows the source to be transmitted over the channel. We just let $X_1 = U$, and $X_2 = V$, and the value of Y will tell us the pair (U, V) with no error. Thus the conditions (14.337) are not necessary.

The reason for the failure of the source channel separation theorem lies in the fact that the capacity of the multiple access channel increases with the correlation between the inputs of the channel. Therefore, to maximize the capacity, one should preserve the correlation between the inputs of the channel. Slepian-Wolf encoding, on the other hand, gets rid of the correlation. Cover, El Gamal and Salehi [69] proposed an achievable region for transmission of a correlated source over a multiple access channel based on the idea of preserving the correlation. Han and Costa [131] have proposed a similar region for the transmission of a correlated source over a broadcast channel.

Capacity regions with feedback. Theorem 8.12.1 shows that feedback does not increase the capacity of a single user discrete memoryless channel. For channels with memory, on the other hand, feedback enables the sender to predict something about the noise and to combat it more effectively, thus increasing capacity.

What about multi-user channels? Rather surprisingly, feedback does increase the capacity region of multi-user channels, even when the channels are memoryless. This was first shown by Gaarder and Wolf [117], who showed how feedback helps increase the capacity of the binary erasure multiple access channel. In essence, feedback from the receiver to the two senders acts as a separate channel between the two senders. The senders can decode each other's transmissions before the receiver does. They then cooperate to resolve the uncertainty at the receiver, sending information at the higher cooperative capacity rather than the non-cooperative capacity. Using this scheme, Cover and Leung [73] established an achievable region for multiple access channel with feedback. Willems [273] showed that this region was the capacity for a class of multiple access channels that included the binary erasure multiple access channel. Ozarow [204] established the capacity region for the two user Gaussian multiple access channel. The problem of finding the capacity region for the multiple access channel with feedback is closely related to the capacity of a two-way channel with a common output.

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

SUMMARY OF CHAPTER 14

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

$$R_1 < I(X_1; Y|X_2), \tag{14.338}$$

$$R_2 < I(X_2; Y|X_1), \tag{14.339}$$

$$R_1 + R_2 < I(X_1, X_2; Y) \tag{14.340}$$

for some distribution $p_1(x_1)p_2(x_2)$ on $\mathcal{X}_1 \times \mathcal{X}_2$.

The capacity region of the m-user multiple access channel is the closure of the convex hull of the rate vectors satisfying

$$R(S) \leq I(X(S); Y|X(S^c)) \quad \text{for all } S \subseteq \{1, 2, \ldots, m\} \tag{14.341}$$

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

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

$$R_1 \leq C\left(\frac{P_1}{N}\right), \tag{14.342}$$

$$R_2 \leq C\left(\frac{P_2}{N}\right), \tag{14.343}$$

$$R_1 + R_2 \leq C\left(\frac{P_1 + P_2}{N}\right), \tag{14.344}$$

where

$$C(x) = \frac{1}{2}\log(1+x). \tag{14.345}$$

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

$$R_1 > H(X|Y), \tag{14.346}$$

$$R_2 > H(Y|X), \tag{14.347}$$

$$R_1 + R_2 > H(X, Y). \tag{14.348}$$

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

$$R_2 \leq I(U; Y_2), \tag{14.349}$$

$$R_1 \leq I(X; Y_1|U) \tag{14.350}$$

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

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

$$C = \sup_{p(x, x_1)} \min\{I(X, X_1; Y), I(X; Y_1|X_1)\}, \tag{14.351}$$

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

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

$$R_1 \geq H(X|U), \tag{14.352}$$

$$R_2 \geq I(Y; U) \tag{14.353}$$

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

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

$$R_Y(D) = \min_{p(w|x)} \min_{f: \mathcal{Y} \times \mathcal{W} \rightarrow \hat{\mathcal{X}}} I(X; W) - I(Y; W), \tag{14.354}$$

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

$$\sum_x \sum_w \sum_y p(x, y)p(w|x)d(x, f(y, w)) \leq D. \tag{14.355}$$

PROBLEMS FOR CHAPTER 14

1. *The cooperative capacity of a multiple access channel.* (See Figure 14.37.)

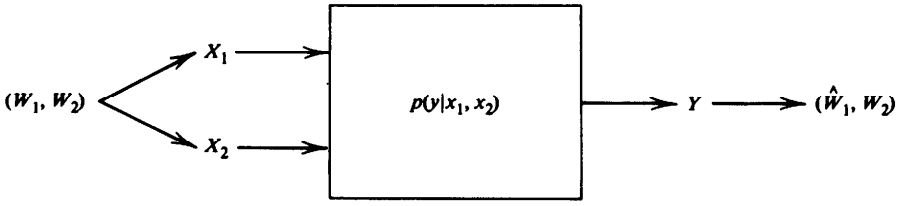


Figure 14.37. Multiple access channel with cooperating senders.

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

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

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

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

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

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

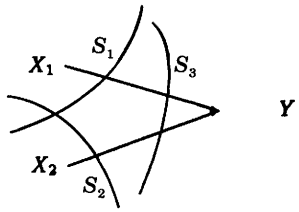
$$R_1 < I(X_1; Y|X_2), \tag{14.356}$$

$$R_2 < I(X_2; Y|X_1), \tag{14.357}$$

$$R_1 + R_2 < I(X_1, X_2; Y), \tag{14.358}$$

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

$$I(X_1; Y|X_2) = I(X_1; Y, X_2).$$



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

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

$$R_1 < \frac{1}{2}\log\left(1 + \frac{P_1}{N}\right), \tag{14.359}$$

$$R_2 < \frac{1}{2}\log\left(1 + \frac{P_2}{N}\right), \tag{14.360}$$

$$R_1 + R_2 < \frac{1}{2}\log\left(1 + \frac{P_1 + P_2}{N}\right). \tag{14.361}$$

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

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

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

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

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

   (c) Argue that there are points in the capacity region of this multiple access channel that can only be achieved by timesharing, i.e., there exist achievable rate pairs (R_1, R_2) which lie in the capacity region for the channel but not in the region defined by

$$R_1 \leq I(X_1; Y|X_2), \tag{14.362}$$

$$R_2 \le I(X_2; Y|X_1), \tag{14.363}$$

$$R_1 + R_2 \le I(X_1, X_2; Y) \tag{14.364}$$

for any product distribution $p(x_1)p(x_2)$. Hence the operation of convexification strictly enlarges the capacity region. This channel was introduced independently by Csiszar and Körner [83] and Bierbaum and Wallmeier [33].

7. *Convexity of capacity region of broadcast channel.* Let $\mathbf{C} \subseteq \mathbf{R}^2$ be the capacity region of all achievable rate pairs $\mathbf{R} = (R_1, R_2)$ for the broadcast channel. Show that $\mathbf{C}$ is a convex set by using a timesharing argument.

   Specifically, show that if $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ are achievable, then $\lambda\mathbf{R}^{(1)} + (1 - \lambda)\mathbf{R}^{(2)}$ is achievable for $0 \le \lambda \le 1$.

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

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

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

$$P_1^{(n)} = P\{\hat{W}_1(\mathbf{Y}_1) \ne W_1\}, \tag{14.365}$$

$$P_2^{(n)} = P\{\hat{W}_2(\mathbf{Y}_2) \ne W_2\}, \tag{14.366}$$

$$P^{(n)} = P\{(\hat{W}_1, \hat{W}_2) \ne (W_1, W_2)\}. \tag{14.367}$$

Then show

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

The result now follows by a simple argument.

Remark: The probability of error $P^{(n)}$ *does* depend on the conditional joint distribution $p(y_1, y_2|x)$. But whether or not $P^{(n)}$ can be driven to zero (at rates (R_1, R_2)) *does not* (except through the conditional marginals $p(y_1|x)$, $p(y_2|x)$).

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

Setup for converse for degraded broadcast channel capacity

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

Encoding

$$f_n : 2^{nR_1} \times 2^{nR_2} \to \mathcal{X}^n$$

Decoding

$$g_n : \mathcal{Y}^n \to 2^{nR_1}, \qquad h_n : \mathcal{Z}^n \to 2^{nR_2}$$

Let $U_i = (W_2, Y^{i-1})$. Then

$$nR_2 \overset{\cdot}{\leq}_{Fano} I(W_2; Z^n) \tag{14.368}$$

$$\overset{(a)}{=} \sum_{i=1}^{n} I(W_2; Z_i \,|Z^{i-1}) \tag{14.369}$$

$$\overset{(b)}{=} \sum_{i} (H(Z_i|Z^{i-1}) - H(Z_i \,|W_2, Z^{i-1})) \tag{14.370}$$

$$\overset{(c)}{\leq} \sum_{i} (H(Z_i) - H(Z_i|W_2, Z^{i-1}, Y^{i-1})) \tag{14.371}$$

$$\overset{(d)}{\leq} \sum_{i} (H(Z_i) - H(Z_i \,|W_2, Y^{i-1})) \tag{14.372}$$

$$\overset{(e)}{=} \sum_{i=1}^{n} I(U_i; Z_i)\,. \tag{14.373}$$

Continuation of converse. Give reasons for the labeled inequalities:

$$nR_1 \overset{\cdot}{\leq}_{Fano} I(W_1; Y^n) \tag{14.374}$$

$$\overset{(f)}{\leq} I(W_1; Y^n, W_2) \tag{14.375}$$

$$\overset{(g)}{\leq} I(W_1; Y^n \,|W_2) \tag{14.376}$$

$$\overset{(h)}{=} \sum_{i-1}^{n} I(W_1; Y_i|Y^{i-1}, W_2) \tag{14.377}$$

$$\overset{(i)}{\leq} \sum_{i=1}^{n} I(X_i; Y_i|U_i)\,. \tag{14.378}$$

12. *Capacity points.*
   (a) For the degraded broadcast channel $X \to Y_1 \to Y_2$, find the points a and b where the capacity region hits the R_1 and R_2 axes (Figure 14.38).
   (b) Show that $b \leq a$.

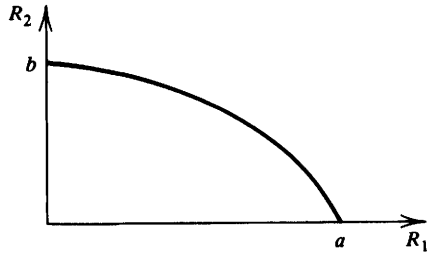


Figure 14.38. Capacity region of a broadcast channel.

13. *Degraded broadcast channel.* Find the capacity region for the degraded broadcast channel in Figure 14.39.

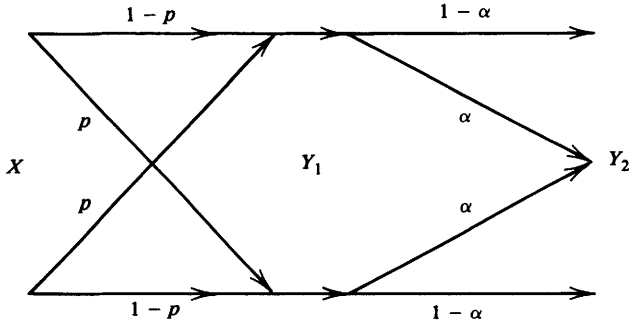


Figure 14.39. Broadcast channel–BSC and erasure channel.

14. *Channels with unknown parameters.* We are given a binary symmetric channel with parameter p. The capacity is $C = 1 - H(p)$.

    Now we change the problem slightly. The receiver knows only that $p \in \{p_1, p_2\}$, i.e., $p = p_1$ or $p = p_2$, where p_1 and p_2 are given real numbers. The transmitter knows the actual value of p. Devise two codes for use by the transmitter, one to be used if $p = p_1$, the other to be used if $p = p_2$, such that transmission to the receiver can take place at rate $\approx C(p_1)$ if $p = p_1$ and at rate $\approx C(p_2)$ if $p = p_2$.

    Hint: Devise a method for revealing p to the receiver without affecting the asymptotic rate. Prefixing the codeword by a sequence of 1's of appropriate length should work.

15. *Two-way channel.* Consider the two-way channel shown in Figure 14.6. The outputs Y_1 and Y_2 depend only on the current inputs X_1 and X_2.

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

$$R_1 < I(X_1; Y_2|X_2), \tag{14.379}$$

$$R_2 < I(X_2; Y_1|X_1) \tag{14.380}$$

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

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

$$R_1 \leq I(X_1; Y_2|X_2), \tag{14.381}$$

$$R_2 \leq I(X_2; Y_1|X_1) \tag{14.382}$$

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

The inner and outer bounds on the capacity of the two-way channel are due to Shannon [246]. He also showed that the inner bound and the outer bound do not coincide in the case of the binary multiplying channel $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y}_1 = \mathcal{Y}_2 = \{0, 1\}$, $Y_1 = Y_2 = X_1X_2$. The capacity of the two-way channel is still an open problem.

HISTORICAL NOTES

This chapter is based on the review in El Gamal and Cover [98]. The two-way channel was studied by Shannon [246] in 1961. He derived inner and outer bounds on the capacity region. Dueck [90] and Schalkwijk [232, 233] suggested coding schemes for two-way channels which achieve rates exceeding Shannon's inner bound; outer bounds for this channel were derived by Zhang, Berger and Schalkwijk [287] and Willems and Hekstra [274].

The multiple access channel capacity region was found by Ahlswede [3] and Liao [178] and was extended to the case of the multiple access channel with common information by Slepian and Wolf [254]. Gaarder and Wolf [117] were the first to show that feedback increases the capacity of a discrete memoryless multiple access channel. Cover and Leung [73] proposed an achievable region for the multiple access channel with feedback, which was shown to be optimal for a class of multiple access channels by Willems [273]. Ozarow [204] has determined the capacity region for a two user Gaussian multiple access channel with feedback. Cover, El Gamal and Salehi [69] and Ahlswede and Han [6] have considered the problem of transmission of a correlated source over a multiple access channel.

The Slepian-Wolf theorem was proved by Slepian and Wolf [255], and was extended to jointly ergodic sources by a binning argument in Cover [63].

Broadcast channels were studied by Cover in 1972 [60]; the capacity region for the degraded broadcast channel was determined by Bergmans [31] and Gallager [119]. The superposition codes used for the degraded broadcast channel are also optimal for the less noisy broadcast channel (Körner and Marton [160]) and the more capable broadcast channel (El Gamal [97]) and the broadcast channel with degraded message sets (Körner and Marton [161]). Van der Meulen [26] and Cover [62] proposed achievable regions for the general broadcast channel. The best known achievable region for broadcast channel is due to Marton [189]; a simpler proof of Marton's region was given by El Gamal and Van der Meulen [100]. The deterministic broadcast channel capacity was determined by Pinsker [211] and Marton [189]. El Gamal [96] showed that feedback does not increase the capacity of a physically degraded broadcast channel. Dueck [91] introduced an example to illustrate that feedback could increase the capacity of a memoryless

broadcast channel; Ozarow and Leung [205] described a coding procedure for the Gaussian broadcast channel with feedback which increased the capacity region.

The relay channel was introduced by Van der Meulen [262]; the capacity region for the degraded relay channel was found by Cover and El Gamal [67]. The interference channel was introduced by Shannon [246]. It was studied by Ahlswede [4], who gave an example to show that the region conjectured by Shannon was not the capacity region of the interference channel. Carleial [49] introduced the Gaussian interference channel with power constraints, and showed that very strong interference is equivalent to no interference at all. Sato and Tanabe [231] extended the work of Carleial to discrete interference channels with strong interference. Sato [229] and Benzel [26] dealt with degraded interference channels. The best known achievable region for the general interference channel is due to Han and Kobayashi [132]. This region gives the capacity for Gaussian interference channels with interference parameters greater than 1, as was shown in Han and Kobayashi [132] and Sato [230]. Carleial [48] proved new bounds on the capacity region for interference channels.

The problem of coding with side information was introduced by Wyner and Ziv [283] and Wyner [280]; the achievable region for this problem was described in Ahlswede and Körner [7] and in a series of papers by Gray and Wyner [125] and Wyner [281, 282]. The problem of finding the rate distortion function with side information was solved by Wyner and Ziv [284]. The problem of multiple descriptions is treated in El Gamal and Cover [99].

The special problem of encoding a function of two random variables was discussed by Körner and Marton [162], who described a simple method to encode the modulo two sum of two binary random variables. A general framework for the description of source networks can be found in Csiszár and Körner [82], [83]. A common model which includes Slepian-Wolf encoding, coding with side information, and rate distortion with side information as special cases was described by Berger and Yeung [30].

Comprehensive surveys of network information theory can be found in El Gamal and Cover [98], Van der Meulen [262, 263, 264], Berger [28] and Csiszár and Körner [83].