MULTIPLE USER INFORMATION THEORY

by

Abbas El Gamal
Thomas M. Cover

TECHNICAL REPORT NO. 42
JUNE 1980

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT ESC 78-23334

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
MULTIPLE USER INFORMATION THEORY†

Abbas El Gamal‡‡
Thomas M. Cover‴‴

Abstract

A unified framework is given for the theory of multiple user information networks. The focus is on broadcast, multiple access, relay, and other channels for which the recent theory is relatively well developed. A discussion of Gaussian version of these channels demonstrates the concreteness of the encoding and decoding necessary to achieve optimal information flow. We also offer speculations about the form of a general theory of information flow in networks.

†Invited paper.
‡‡ Stanford University. This paper was partially supported by the Joint Services Electronics Program through the Air Force Office of Scientific Research (AFSC) under Contract F44620-76-C-0061 and NSF ENG 79-08948.
‴‴ Stanford University. This paper was partially supported by the Joint Services Electronics Program DAAG-29-79-C-0047 and by NSF Grant ECS78-23334.
1. **Introduction**

The Shannon theory of channel capacity has been extended successfully to many interesting communication networks in the past 10 years. We shall attempt to achieve three goals in our exposition of this theory:

1) to make the theory accessible to researchers in communication theory,

2) to provide conditionally novel proofs of the theory for researchers in information theory, and

3) to present an overview of the basic problems in constructing a theory of information flow in networks.

The primary ideas can be obtained by reading the introduction (§1), Gaussian examples (§2), Shannon's theorem (§4), and the summary (§10). The heretofore unpublished information theoretic proofs are those for the multiple access channel (§5), Slepian-Wolf data compression (§6), and the degraded broadcast channel (§8). All proofs, both new and old, are based on the idea of jointly typical sequences.

No claim for comprehensive coverage is given. For that the reader is referred to van der Meulen [1]. Rather, we are concerned with providing a unified approach to the theory. This leads naturally to a discussion of some of the major results. We begin by discussing some of the building blocks for networks.

Suppose \( m \) ground stations are simultaneously communicating to a common satellite as in Figure 1.1. This is known as the multiple access channel. What are the achievable rates of communication? Does the total amount of information flow tend to infinity with the number of stations— or does the interference put an upper limit on the total communication?
Does the signalling strategy change with \( m \)? Here the theory is completely known (Ahlswede [2] and Liao [3]), and all of these questions have quite satisfying answers (Section 5).

In constrast, we can reverse the network, and consider one satellite broadcasting simultaneously to \( m \) stations as shown in Figure 1.2. Here the achievable set of rates is not known except in special cases (Section 7).

Yet another example consists of only one sender and one receiver, but includes extra channels serving as relays. This is the relay channel shown in Figure 1.3.

In general, the underlying goal of work on these special channels is a theory for networks of the general form given in Figure 1.4.

![Figure 1.1](image)

**Figure 1.1**

Multiple Access Network
Figure 1.2
Broadcast Network

Figure 1.3
Relay Network
The interpretation of Figure 1.4 is that at each instant of time the
ith node sends a symbol \( x_i \) that depends on the messages he wishes to send
and (perhaps) his past received \( y_i \) symbols. We assume that the result of
the simultaneous transmission of \( (x_1, x_2, \ldots, x_m) \) is a random collection of
received symbols \( (y_1, y_2, \ldots, y_m) \) drawn according to a conditional prob-
ability \( p(y_1, \ldots, y_m|x_1, \ldots, x_m) \), where \( p(\cdot | \cdot) \) describes all of the effects
of interference and noise in the network.

In a more restricted domain, such as the flow of water in networks of
pipes, the existing theory is very satisfying. For example, in the single
source, single sink network in Figure 1.5, the maximum flow from A to B is
easily computed from the maximum flow minimum cut theorem of Ford and
Fulkerson [4].
\[ \text{Capacity} = \min \{ C_1 + C_2, C_1 + C_3, C_5, C_2 + C_4, C_4 + C_5 \} \]

**Figure 1.5**
The Maximum Flow Minimum Cut Theorem

Assume that the edges have capacities \( C_i \) as shown. Clearly, the maximum flow across any cutset is no greater than the sum of the capacities of the cut edges. Thus minimizing the maximum flow over cutsets yields an upper bound to the total flow from A to B. This flow can actually be achieved, as the Ford and Fulkerson theorem demonstrates.

However, the information flow problem involves "soft" quantities rather than "hard" commodities. A choice of node symbols \( x \) results in a random response \( Y \), and it is difficult to see how to choose as many distinguishable \( x \)'s as possible in this random environment? Consequently, it is gratifying
to find that information problems like the relay channel and cascade channel admit minflow maxcut interpretations. For example, the informally defined cascade network in Figure 1.6,

\[ X_1 \rightarrow Y_1 : X_2 \rightarrow Y_2 \]

\[ C_1 \quad C_2 \]

Figure 1.6

has capacity \( C = \min \{ C_1, C_2 \} \), where \( C_i \) denotes the Shannon capacity of the \( i \)th channel. Also, for the degraded or deterministic relay channel (Section 7) we have a similar maxflow mincut interpretation as shown in Figure 1.7.

\[ I(X_1; Y_1, Y|X_2) \quad Y_1 : X_2 \quad I(X_1, X_2; Y) \]

\[ C \leq \max_{p(x_1, x_2)} \min \left\{ I(X_1; Y_1, Y|X_2), I(X_1, X_2; Y) \right\} \]

Figure 1.7

Degraded and Deterministic Relay Channel Capacity
The structure of this paper is presented in miniature in Section 2. In that section, we use Gaussian channels to run through the major results that will be given in greater generality and detail in the subsequent sections. The physically motivated Gaussian channel lends itself to concrete and easily interpreted answers. Some preliminary technical details on the properties of joint typicality are given in Section 3, followed by a simple proof in Section 4 of Shannon's original capacity theorem. Then treated are the multiple access channel (Section 5), Slepian-Wolf data compression theorem (Section 6), the combination of both (Section 7), the broadcast channel (Section 8), and the relay channel (Section 9).

The final summary (Section 10) is a recapitulation of the paper paralleling Section 2, this time in greater generality.
2. Gaussian Multiple User Channels

We shall begin our treatment of multiple user information theory by investigating Gaussian multiple user channels. This allows us to give concrete descriptions of the coding schemes and associated capacity regions. The proofs of capacity for the discrete memoryless counterparts of these channels will be given in later sections.

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

$$Y_i = x_i + Z_i, \quad i = 1,2,\ldots,$$

where $Z_i$ are independent identically distributed Gaussian random variables with mean zero and variance $N$. The signal $x = (x_1,x_2,\ldots,x_n)$ has a power constraint

$$\frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P.$$

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

$$C = \frac{1}{2} \log \left(1 + \frac{P}{N} \right) \text{ bits/second.} \quad (2.1)$$

The continuous time Gaussian channel capacity is simply related to the discrete time capacity. If the signal $x(t), \quad 0 \leq t \leq T$, has power constraint $P$ and bandwidth constraint $W$, and the white noise $Z(t), \quad 0 \leq t \leq T$, has power spectral density $N$, then the capacity of the channel $Y(t) = x(t) + Z(t)$ is given by

$$C = W \log \left(1 + \frac{P}{NW} \right) \text{ bits/second.} \quad (2.2)$$
The relationship between (2.1) and (2.2) can be seen informally by replacing the continuous time processes by \( n = 2TW \) independent samples from the process and calculating the noise variance per sample. The full theory establishing (2.2) can be found in Wyner [5], Gallager [6], and Pollack, Landau, and Slepian [7].

Having said this, we restrict our treatment to time discrete Gaussian channels.

**Random Codebook:** Shannon observed in 1948 that a randomly selected codebook is usually good when the rate \( R \) of the codebook is less than the channel capacity \( C = \max I(X;Y) \). As mentioned above, for the Gaussian channel the capacity is given by \( C = (1/2) \log(1+P/N) \) bits per transmission.

We now set up a codebook that will be used in all of the multiple user channel models below. The codewords comprising the codebook are vectors of length \( n \) and power \( P \). To generate such a random codebook, simply choose \( 2^{nR} \) independent identically distributed random \( n \)-vectors \( \{x(1), x(2), \ldots, x(2^{nR})\} \), each consisting of \( n \) independent Gaussian random variables with mean zero and variance \( P \). The rate \( R \) will be specified later. Sometimes we will need two or more independently generated codebooks.

In the continuous channel case, one simply lets the white noise generator of power \( P \) and bandwidth \( W \) run for \( T \) seconds. Every \( T \) seconds, a new codeword is generated and we list them until we fill up the codebook.

Now we analyze the Gaussian channels shown in Figure 2.1.

2.1. **The Gaussian Channel:** Here \( Y = X + Z \). Choose an \( R < C = 1/2 \log(1+P/N) \). Choose any index \( i \) in the set \( 2^{nR} \). Send the \( i \)th vector \( x(i) \) from the codebook generated above. The receiver observes
Figure 2.1
Gaussian Multiple User Channels
\[ Y = X(i) + Z, \] then finds the index \( \hat{i} \) of the closest codeword to \( Y \).

If \( n \) is sufficiently large, the probability of error \( P(i \neq \hat{i}) \) will be arbitrarily small. As will be seen in the definitions on joint typicality, this minimum distance decoding scheme for the Gaussian channel is essentially equivalent to finding the codeword in the codebook which is jointly typical with the received vector \( Y \).

2.2. The Multiple Access Channel with \( m \) Users: We consider \( m \) transmitters, each of power \( P \). Let \( Y = \sum_{i=1}^{m} x_i + Z \).

Specializing the results of Section 4 to the Gaussian channel shows that the achievable rate region for the Gaussian channel takes on the simple form given in the following equations:

\[
\begin{align*}
R_i &< C(P/N) \\
R_i + R_j &< C(2P/N) \\
R_i + R_j + R_k &< C(3P/N) \\
&\vdots \\
\sum_{i=1}^{m} R_i &< C(mP/N),
\end{align*}
\]

(2.3)

where

\[
C(x) = \frac{1}{2} \log (1+x)
\]

(2.4)

denotes the capacity of the Gaussian channel with signal to noise ratio \( x \).

When all the rates are the same, the last inequality dominates the others.

Here we need \( m \) codebooks, each with \( 2^{nR_i} \) codewords of power \( P \).

Transmission is simple. Each of the independent transmitters chooses whatever codeword he wishes from his own codebook. The users simultaneously send these vectors. The receiver sees the codewords added together with the Gaussian noise \( Z \).
Optimal decoding consists of looking for the \( m \) codewords, one from each codebook, such that the vector sum is closest to \( Y \) in Euclidean distance. The set of \( m \) codewords achieving the minimum distance to \( Y \) corresponds to the hypothesized collection of messages sent.

If \((R_1, R_2, \ldots, R_m)\) is in the capacity region given above, then the probability of error goes to zero as \( n \) tends to infinity.

Remarks: It is exciting to see in this problem that the sum of the rates \( C(mP/N) \) of the users goes to infinity with \( m \). Thus in a cocktail party with \( m \) celebrants of power \( P \) in the presence of ambient noise \( N \), although the interference grows as the number of speakers increases, the intended listener receives an unbounded amount of information as the number of people goes 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 utilizes the entire time to send his message.

A practical consideration for ground transmission to a satellite involves the possible inability of the ground communicators to synchronize their transmissions. Nonetheless, it can be shown that the capacity is unchanged when there is a lack of synchronization [8].

2.3. The Broadcast Channel: Here we assume that we have a sender of power \( P \) and two distant receivers, one with noise spectral density \( N_1 \) and the other with noise spectral density \( N_2 \). Without loss of generality, assume \( N_1 < N_2 \). Thus in some sense receiver \( Y_1 \) is better 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 which will be discussed in Section 8. Specializing that work, we find the following capacity region for the Gaussian broadcast channel:

\[
R_1 < C(\alpha P/N_1) \\
R_2 < C(\alpha P/(\alpha P + N_2)),
\]

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

To encode the messages, the receiver generates two codebooks, one with power \( \alpha P \) at rate \( R_1 \), and another codebook with power \( \alpha P \) and rate \( R_2 \). He has chosen \( R_1 \) and \( R_2 \) to satisfy the equation 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, he takes codeword \( x(i) \) from the first codebook and codeword \( x(j) \) from the second codebook and computes the sum. He then sends the sum over the channel.

Two receivers must now do the decoding. First consider the bad receiver \( Y_2 \). He merely looks through the second codebook for the closest codeword to his received vector \( \hat{y}_2 \). His effective signal to noise ratio is \( \alpha P / \alpha P + N_2 \) since \( Y_1 \)'s message acts as noise to \( Y_2 \). The good receiver \( Y_1 \) first decodes \( Y_2 \)'s codeword, which he can do because of his lower noise \( N_1 \). He substracts this codeword \( \hat{x}_2 \) from \( Y_1 \). This leaves him with a channel of power \( \alpha P \)
and noise $N_1$. He then looks for the closest codeword in the first code-
book to $Y_1 - \hat{x}_2$. The resulting probability of error can be made as low
as wished.

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 extra information intended for himself.

2.4. The Relay Channel: For the relay channel, we have a sender $X_1$
and an ultimate intended receiver $Y$. Also present, however, is the relay
channel intended solely to help the sender. The channel is given by

$$Y_1 = x_1 + Z_1$$
$$Y_2 = x_1 + Z_1 + x_2 + Z_2,$$

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

$$x_{2i} = f_i(y_{11}, y_{12}, \ldots, y_{1i-1}).$$

The sender $X_1$ has power $P_1$ and the relay $X_2$ has power $P_2$. The re-
results of Section 9 yield the capacity

$$C = \max_{0 < \alpha < 1} \min \left\{ C \left( \frac{P_1 + P_2 + 2\sqrt{\alpha P_2}}{N_1 + N_2} \right), C\left( \frac{\alpha P_1}{N_1} \right) \right\},$$

where $\bar{\alpha} = 1 - \alpha$.

Note that if

$$P_2/N_2 \geq P_1/N_1,$$
it can be seen that $C^* = C(P_1/N_1)$. (This is achieved by $\alpha = 1$.) The channel appears to be noise free after the relay, and the capacity $C(P_1/N_1)$ from $x_1$ to the relay can be achieved. Thus the rate without the relay $C(P_1/(N_1 + N_2))$ is increased by the relay to $C(P_1/N_1)$. For large $N_2$, and for $P_2/N_2 \geq P_1/N_1$, we see that the increment in rate is from $C(P_1/(N_1 + N_2)) \approx 0$ to $C(P_1/N_1)$.

Encoding of Information: Two codebooks are needed. The first codebook has $2^{nR_1}$ words of power $\alpha P_1$. The second has $2^{nR_0}$ codewords of power $\alpha P_1$. We shall use words 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 $R_1 < C(\alpha P_1/N_1)$ but the intended receiver does not. However, the intended receiver has a list of possible codewords of size $2^{n(R_1 - C(\alpha P_1/N_1 + N_2))}$. The list calculation involves a result on list codes.

In the next block the relay and the transmitter would like to cooperate to resolve the receiver's uncertainty about the previously sent codeword on the receiver's list. Unfortunately, they cannot quite be sure what this list is. They do not know what the received signal $Y$ was. 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 what this partition is. 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 $X_1$ and $X_2$ send the same designated codeword. The relay, of course, must scale this codeword so that it meets his power constraint $P_2$. They now simultaneously transmit their codewords. An important point to note 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 \((\sqrt{\alpha P_1} + \sqrt{P_2})^2\).

However, this does not exhaust what the transmitter does in the second block. He also chooses a fresh codeword from his first codebook, adds it "on paper" to the cooperative codeword from his second codebook, and sends this 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 it off, and then calculates a list of indices of size \( n R_0 \) corresponding to all transmitted words from the first book which might have been sent in that second block.

Now it is time for the intended receiver \( Y \) to finish 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. Since the rates and powers have been chosen right, it is highly probable that there will be only one codeword in this 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 adds some fresh information from his first codebook to his 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.
2.5. The Interference Channel: In the interference channel, we have 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. As can be seen, this channel involves interference of each user with the other. It is not quite a broadcast channel because there is only one intended receiver for each sender nor is it quite a multiple access channel, because each receiver is only interested in what is being sent by the corresponding transmitter.

This channel has not been solved in general, even in the Gaussian case. But remarkably, in the case of high interference, Carleial [9] has shown that the solution to 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, since the interference is high, the first receiver can understand perfectly 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 his sender's codebook to find the closest codeword and declares that codeword to be the one sent.
3. The Asymptotic Equipartition Theorem and Shannon Channel Capacity.

Every sequence of $n$ fair coin flips has equal probability $(1/2)^n$. If the coin has bias $p$, all of the sequences having roughly $np$ heads are nearly equally probable and exhaust almost all of the probability. A formalization of this idea for arbitrary random variables is known as the asymptotic equipartition property (AEP) (Shannon [10], MacMillan [11], and Breiman [12]).

Consider a sequence of independent identically distributed random variables $X = (X_1, X_2, \ldots, X_n)$, where $X_i$ is drawn according to probability mass function $p(x)$. We are interested in defining a set $A$ of possible outcomes, each roughly equally probable, such that $P(X \in A) \approx 1$. Toward this end, we shall say that a given sequence $x = (x_1, x_2, \ldots, x_n)$ is $\varepsilon$-typical if

$$|-(1/n) \log p(x) - H(X)| \leq \varepsilon,$$  \hspace{1cm} (3.1)

where

$$H(X) = -\sum p(x) \log p(x)$$  \hspace{1cm} (3.2)

is the Shannon entropy of $p(\cdot)$. Let us define the typical set $A_\varepsilon$ to be the set of all $\varepsilon$-typical $n$-sequences $x$.

By the law of large numbers, for a random i.i.d. sequence $X$

$$-1/n \log p(X) = (1/n) \sum_{i=1}^{n} -\log p(X_i)$$

$$\rightarrow H(X), \text{ with probability one.}$$ \hspace{1cm} (3.3)

Consequently, the following results are true:

1) $P(A_\varepsilon) \rightarrow 1$, as $n \rightarrow \infty$.
2) \( x \in A_{\varepsilon} \) implies \( 2^{-n(H+\varepsilon)} \leq p(x) \leq 2^{-n(H-\varepsilon)} \).

3) The number of elements \( \| A_{\varepsilon} \| \) in \( A_{\varepsilon} \) is \( \leq 2^{n(H+\varepsilon)} \).

The third assertion follows from

\[
1 \geq \sum_{A_{\varepsilon}} p(x) \geq \sum_{A_{\varepsilon}} 2^{-n(H+\varepsilon)} = \| A_{\varepsilon} \| 2^{-n(H+\varepsilon)}. \quad (3.4)
\]

Roughly speaking, a sequence is typical if the proportion of occurrences of each of its symbols is close to the true probability of occurrence.

An immediate application of the AEP yields Shannon's source coding theorem:

**Theorem 1 (Shannon 1948):** Given an independent identically distributed source \( X_1, X_2, \ldots \) with entropy \( H(X) \), there exists, for every \( \varepsilon > 0 \), an integer \( n \) and an encoding \( i: \bar{X}^n \rightarrow \{1, 2, \ldots, 2^{n(H+\varepsilon)}\} \) and decoding rule \( g:\{1, 2, \ldots, 2^{n(H+\varepsilon)}\} \rightarrow \bar{X}^n \), such that

\[
P\{g(i(X)) \neq X\} \leq \varepsilon.
\]

**Proof:** Simply index the elements \( x \) in the typical set \( A_{\varepsilon} \).

**Remark:** Since the index set has \( \leq 2^{n(H+\varepsilon)} \) elements, we see that only \((H+\varepsilon)\) bits/symbol are necessary to describe \( x \).

Now, to make progress with multiple user information theory, the idea of joint typicality is needed. A pair of sequences \( x \) and \( y \) are said to be jointly \( \varepsilon \)-typical if \( x \) is individually \( \varepsilon \)-typical, i.e.,

\[
| -\frac{1}{n} \log p(x) - H(X) | \leq \varepsilon,
\]

\( y \) is individually \( \varepsilon \)-typical, and \((x,y)\) is \( \varepsilon \)-typical, i.e.,
\[- \frac{1}{n} \log p(x, y) - H(X, Y) \leq \epsilon.\]

The picture of jointly typical pairs is given in Figure 3.1.

![Diagram of jointly typical pairs](image)

**Figure 3.1**

Jointly Typical Sequences

The dots in the matrix denote jointly typical pairs. It can be shown by the method of (3.4) that there are \( \leq 2^{nH(Y|X)} \) dots in each row and
\( \leq 2^{nH(X|Y)} \) dots in each column.

The definitions and proofs of the needed results now follow.

Let \( \{x^{(1)}, x^{(2)}, \ldots, x^{(k)}\} \) denote a finite collection of discrete random
variables with some fixed joint distribution \( p(x^{(1)}, x^{(2)}, \ldots, x^{(k)}) \), 
\((x^{(1)}, x^{(2)}, \ldots, x^{(k)}) \in X^{(1)} \times X^{(2)} \times \ldots \times X^{(k)} \). Let \( S \) denote an ordered subset of these r.v.'s, and consider \( n \) independent copies of \( S \). Thus,

\[
\Pr \{ \underline{s} = \underline{s} \} = \prod_{i=1}^{n} \Pr \{ S_i = s_i \} , \quad \underline{s} \in S^n .
\] (3.5)

For example, if \( S = (X^{(j)}, X^{(\ell)}) \), then

\[
\Pr \{ \underline{s} = \underline{s} \} = \Pr \{ (X^{(j)}, X^{(\ell)}) = (\tilde{x}^{(j)}, \tilde{x}^{(\ell)}) \} 
= \prod_{i=1}^{n} p(x^{(j)}, x^{(\ell)}) .
\] (3.6)

By the law of large numbers, for any subset \( S \) of random variables,

\[
- \frac{1}{n} \log p(S_1, S_2, \ldots, S_n) = - \frac{1}{n} \sum_{i=1}^{n} \log p(S_i) \rightarrow H(S) .
\] (3.7)

Convergence in (3.7) take place simultaneously with probability one for all \( 2^k \) subsets

\[
S \subseteq \{ X^{(1)}, X^{(2)}, \ldots, X^{(k)} \} .
\]

**Definition:** The set \( A_\varepsilon \) of \( \varepsilon \)-typical \( n \)-sequences \( (\underline{x}^{(1)}, \underline{x}^{(2)}, \ldots, \underline{x}^{(k)}) \) is defined by

\[
A_\varepsilon(x^{(1)}, x^{(2)}, \ldots, x^{(k)}) = A_\varepsilon = \{ (\underline{x}^{(1)}, \underline{x}^{(2)}, \ldots, \underline{x}^{(k)}) : - \frac{1}{n} \log p(S) - H(S) < \varepsilon , \forall S \subseteq \{ X^{(1)}, X^{(2)}, \ldots, X^{(k)} \} \} .
\] (3.8)
Let $A_\varepsilon(S)$ denote the restriction of $A_\varepsilon$ to the coordinates corresponding to $S$. Thus if $S = (x^{(1)}, x^{(2)})$ we have

$$A_\varepsilon(x^{(1)}, x^{(2)}) = \{ (\underline{x}^{(1)}, \underline{x}^{(2)}) : \left| -\frac{1}{n} \log p(\underline{x}^{(1)}, \underline{x}^{(2)}) - H(x^{(1)}, x^{(2)}) \right| < \varepsilon $$

$$\left| -\frac{1}{n} \log p(\underline{x}^{(1)}) - H(x^{(1)}) \right| < \varepsilon$$

$$\left| -\frac{1}{n} \log p(\underline{x}^{(2)}) - H(x^{(2)}) \right| < \varepsilon \}. \quad (3.9)$$

**Notation:** We shall use the notation

$$a_n = 2^{nb} \quad (3.10)$$

to mean

$$(1/n) \log a_n \to b. \quad (3.11)$$

**Lemma 1:** For any $\varepsilon > 0$, for sufficiently large $n$,

(i) $P\{A_\varepsilon(S)\} \geq 1 - \varepsilon, \quad \forall S \subseteq \{x^{(1)}, \ldots, x^{(k)}\},$

(ii) $\|A_\varepsilon(S)\| \geq 2^n H(S),$

(iii) $s \in A_\varepsilon(S) \Rightarrow p(s) = 2^{-n(H(S) - \varepsilon)},$

(iv) Let $S_1, S_2 \subseteq \{x^{(1)}, x^{(2)}, \ldots, x^{(k)}\}$. If $(s_1, s_2) \in A_\varepsilon(S_1, S_2)$, then

$$p(s_1 | s_2) = 2^{-(H(S_1) - H(S_2) - 2\varepsilon)}.$$ 

**Proof:**

(i) follows from the convergence of the random variables in the definition of $A_\varepsilon(S)$. (iii) follows immediately from the definition of $A_\varepsilon(S)$. (ii) follows from
\[
1 \geq \sum_{s \in A_\varepsilon(S)} p(s) \geq \sum_{A_\varepsilon(S)} 2^{-n(H(S)+\varepsilon)} = \|A_\varepsilon(S)\| 2^{-n(H(S)+\varepsilon)} \tag{3.12}
\]
and
\[
(1-\varepsilon) \leq \sum_{s \in A_\varepsilon(S)} p(s) \leq \|A_\varepsilon(S)\| 2^{-n(H(S)-\varepsilon)} \tag{3.13}
\]

Finally, \((s_1, s_2) \in A_\varepsilon(S_1, S_2)\) implies
\[
p(s_1, s_2) = 2^{-n(H(S_1, S_2)-\varepsilon)}
\]
and
\[
p(s_2) = 2^{-(H(S_2)-\varepsilon)}.
\]

(iv) follows from
\[
p(s_1|s_2) = p(s_1, s_2)/p(s_2).
\]

**Lemma 2:** Let \(S_1, S_2\) be two subsets of \(X^{(1)}, \ldots, X^{(k)}\). For any \(\varepsilon > 0\), define \(\|A_\varepsilon(S_1|S_2)\|\) to be the number of \(s_1\) sequences, jointly \(\varepsilon\)-typical with a particular \(s_2\) sequence. If \(s_2 \in A_\varepsilon(S_2)\), then for arbitrarily large \(n\)
\[
\|A_\varepsilon(S_1|S_2)\| \leq 2^{n(H(S_1|S_2) + 2\varepsilon)}.
\]

Moreover,
\[
(1-\varepsilon) 2^{n(H(S_1|S_2)-2\varepsilon)} \leq \sum_{s_2} p(s_2) \|A_\varepsilon(S_1|S_2)\|.
\]

**Proof:** Similar to part (ii) of Lemma 1.
In achievability proofs, we shall need to know the probability that conditionally independent sequences are jointly typical. Let $S_1, S_2$, and $S_3$ be three subsets of \{ $X^{(1)}, X^{(2)}, \ldots, X^{(k)}$ \}.

**Lemma 3:**

(a) Let

$$P(S_1 = s_1, S_2 = s_2, S_3 = s_3) = \prod_{i=1}^{n} p(s_{1i} | s_{3i}) p(s_{2i} | s_{3i}) p(s_{3i}).$$

Then,

$$P\{ (S_1', S_2', S_3') \in A_\varepsilon \} = 2^{-n I(S_1; S_2 | S_3)}.$$

**Proof:**

$$P\{ (S_1', S_2', S_3') \in A_\varepsilon \} = \sum_{(S_1, S_2, S_3) \in A_\varepsilon} p(s_3) p(s_1 | s_3) p(s_2 | s_3)$$

$$= \|A_\varepsilon(S_1, S_2, S_3)\| 2^{-nH(S_3) - nH(S_1 | S_3) - nH(S_2 | S_3)}$$

$$\leq 2^{-nH(S_1, S_2, S_3) - n(H(S_3) + H(S_1 | S_3) + H(S_2 | S_3))}$$

$$= 2^{-nI(S_1; S_2 | S_3)}.$$
4. **Shannon's Capacity Theorem.**

We begin the theoretical discussion by reproving Shannon's original channel capacity theorem for the discrete memoryless channel (Figure 4.1). When Shannon's work was published in 1948, the proofs were considered to be only heuristic--mere plausibility arguments. The first rigorous proof by Feinstein [13] 8 years later was completely different from Shannon's random coding idea. Further proofs by Wolfowitz, Fano, Gallager, and many others, were also different. The rigorous proof we give here supports the idea of Shannon's outline. It can be found as a problem in Gallager [6] and in Forney's unpublished class notes [14]. This proof technique will come up many times in subsequent sections. The fact that Shannon's original proof is the natural proof for multi-user networks is really quite remarkable.

The basic Shannon model for a communication channel is the discrete memoryless channel \((X, p(y|x), Y)\) consisting of two alphabets--input alphabet \(X\) and output alphabet \(Y\)--and a channel probability matrix \(p(y|x)\). The interpretation is that if \(x\) is sent, then the received symbol \(y\) is drawn according to probability mass function \(p(y|x)\). Discrete memoryless means if a sequence \(x_1, x_2, \ldots, x_n\) is sent, then the received sequence \(y = (y_1, y_2, \ldots, y_n)\) is drawn according to \(\prod_{i=1}^{n} p(y_i|x_i)\).

A \((2^{nR}, n)\) code for a channel consists of a set of integers \([1, 2^{nR}]\) called the message set, an encoding function

\[ x : [1, 2^{nR}] \rightarrow x^n \]

and a decoding function

\[ g : y^n \rightarrow [1, 2^{nR}] \].
If the message $w \in [1, 2^{nR}]$ is sent, let

$$\lambda(w) = P \{ g(Y) \neq w \mid w \text{ sent} \} \quad (4.1)$$

denote the conditional probability of error. Define the average probability of error of the code assuming a uniform distribution over the set of messages $[1, 2^{nR}]$, as

$$\overline{P}_{e}^{n} = \frac{1}{2^{nR}} \sum_{w} \lambda(w). \quad (4.2)$$

The rate $R$ of an $(2^{nR}, n)$ code is said to be achievable by the discrete memoryless channel if, for any $\varepsilon > 0$, there exists a $(2^{nR}, n)$ code for all $n$ sufficiently large such that $\overline{P}_{e}^{n} < \varepsilon$.

The capacity $C$ of the discrete memoryless channel is the supremum of the set of achievable rates. In 1948, Shannon showed that the capacity $C$ of the discrete memoryless finite alphabet channel is given by the following.
Theorem 2 (Shannon): The capacity of the discrete memoryless channel $(X,p(y|x),Y)$ is given by

$$C = \sup_{p(x)} I(X;Y). \tag{4.3}$$

Proof: The proof that $C$ is the capacity of a channel involves proving:

(i) **achievability** of any $R < C$, i.e., that there exists a sequence of $(2^{nR},n)$ codes such that $P_e^n \to 0$, and

(ii) a **converse** showing that given any sequence of codes $(2^{nR},n)$ with $P_e^n \to 0$ then $R < C$.

We shall only prove the achievability part of the theorem. The converses are discussed in the following sections of this paper.

**Achievability**: Assume that the maximization in (4.3) is achieved for a distribution $p^*(x)$ on $X$.

**Random code**: Choose $2^{nR}$ i.i.d. $x$ sequences each with probability

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

**Decoding**: Given $y$, choose the message $w$ such that

$$(y,x(w)) \in A_e(X,Y),$$

if such a $w \in [1,2^{nR}]$ exists and is unique, otherwise declare an error.

By symmetry of the random code construction, the probability of error (averaged over the random code) is independent of the index $w$ sent. Thus without loss of generality assume that 1 is sent. Consider the events
\[ E_w = \{(X(w), Y) \in A_\varepsilon \}, \ w \in [1, 2^n R] . \]

Then by the union of events bound

\[ P_n = P \left( \bigcup_{w \neq 1} E_w \cup E_1^c \right) \]

\[ \leq P(E_1^c) + \sum_{w \neq 1} P(E_w) . \]

From Lemma 1, it follows that

\[ P(E_1^c) \to 0 . \]

From Lemma 3,

\[ P(E_w) \leq 2^{-nC} \text{ for all } w \neq 1 . \]

Therefore, if \( R < C \), then

\[ \sum_{w \neq 1} P(E_w) \leq 2^{n(R-C)} \to 0 . \]

If the error averaged over the random code tends to zero, then there must exist a sequence of codes with \( P_e^n \to 0 \), and achievability is proved.
5. The Multiple Access Channel.

The most important of the completely understood multiple user channels is the multiple access channel shown in Figure 5.1.

![Figure 5.1: The Multiple Access Channel](image)

Many senders each simultaneously attempt to communicate a message to a common receiver. The most common example of this is a satellite receiver with many independent ground stations. After reception, the satellite will broadcast the information back to the ground, but that is the subject of the section on broadcast channels.
We see that the senders must contend not only with the receiver noise but with interference from each other as well.

The discrete memoryless multiple access channel \((X_1 \times X_2, p(y|x_1, x_2), Y)\) consists of three alphabets \(X_1, X_2\), and \(Y\) and a probability transition matrix \(p(y|x_1, x_2)\).

A \((2^{nR_1}, 2^{nR_2}), n\) code for the multiple access channel consists of two sets of integers \(M_1 = [1, 2^{nR_1}], M_2 = [1, 2^{nR_2}]\) called the message sets, two encoding functions

\[
X_1 : M_1 \to X_1^n
\]

\[
X_2 : M_2 \to X_2^n
\]

and a decoding function

\[
g : y^n \to M_1 \times M_2 .
\]

Assuming uniform distribution over the product message sets \(M_1 \times M_2\) i.e., that the messages are independent and equally likely, we define the average probability of error for a \((2^{nR_1}, 2^{nR_2}), n\) code to be
\[
\bar{p}_e^n = \frac{1}{2^{n(R_1 + R_2)}} \sum_{(w_1, w_2) \in M_1 \times M_2} \Pr \{ g(Y) \neq (w_1, w_2) | (w_1, w_2) \text{ sent} \}. \quad (5.1)
\]

A rate pair \((R_1, R_2)\) is said to be achievable for the multiple access channel if there exists a sequence of \(nR_1, nR_2, n\) codes with \(\bar{p}_e^n \to 0\).

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

The capacity region of the multiple access channel was established by Ahlswede [2] and Liao [3]. The following proof is different from theirs.

**Theorem 3: Multiple Access Capacity.**

The capacity of the multiple access channel \((X_1 \times X_2, p(y|x_1, x_2), Y)\) is given by the convex hull of the union of the sets

\[
C(p_1 p_2) = \{(R_1, R_2) : R_1 \leq I(X_1; Y|X_2) \\
R_2 \leq I(X_2; Y|X_1) \\
R_1 + R_2 \leq I(X_1, X_2; Y) \}
\]

where \(p(x_1, x_2) = p_1(x_1)p_2(x_2)\) \(,\) \( \quad (5.2)\)

over all input product probability distributions on \(X_1 \times X_2\).
Proof:

(i) Achievability.

**Random coding:** Fix $p_1(x_1), p_2(x_2)$. Let $p(x_1,x_2) = p_1(x_1)p_2(x_2)$.

Choose a random code of $2^n R_1$'s $\in X_1^n$ i.i.d. $\sim \prod_{i=1}^{n} p_1(x_1)$, and independently choose $2^n R_2$'s $\in X_2^n$ i.i.d. $\sim \prod_{i=1}^{n} p_2(x_2)$. Index these sequences as $x_1(i), x_2(j), i \in [1,2^R_1], j \in [1,2^R_2]$.

**Decoding:** For decoding, given $v$, simply choose the pair $(i,j)$ such that

$$(x_1(i), x_2(j), y) \in A_N^c,$$

if such an $(i,j) \in [1,2^{nR_1}] \times [1,2^{nR_2}]$ exists and is unique—otherwise declare an error.

By symmetry of the random code construction, the probability of error (averaged over the random code) is independent of the index $(i,j)$ sent. Thus without loss of generality assume that $(i,j) = (1,1)$ is sent.

Consider the events

$$E_{ij} = \{(x_1(i), x_2(j), y) \in A_N^c\}.$$  

Then, by the union of events bound

$$P_n = P(E_{11}^c \cup \bigcup_{(i,j) \neq (1,1)} E_{ij})$$  \hspace{1cm} (5.3)

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

From Lemma 1, $P(E_{11}^c) \to 0$.

Next, for $i \neq 1$, 

-32-
\[
P(E_{11}) = \sum_{x_1,x_2,y \in A_{\epsilon}} p(x_1)p(x_2,y) 
= \frac{-nH(X_1)}{2} - \frac{nH(X_2,Y)}{2} 
\approx \frac{-nI(X_1;Y|X_2)}{2},
\]

where the first equality follows from $i \neq 1$, which implies the independence of $X_1$ from $(X_2,Y)$. The second and third inequalities follow from the definition of $A_{\epsilon}$, Lemma 1, and from $I(X_1;X_2,Y) = I(X_1;Y|X_2)$ when $X_1$ and $X_2$ are independent.

Similarly, for $j \neq 1$,
\[
P(E_{1j}) \approx 2^{-nI(X_2;Y|X_1)}
\]
and for $i \neq 1, j \neq 1$,
\[
P(E_{ij}) \approx 2^{-nI(X_1;X_2;Y)}.
\]

It follows that
\[
P_n \leq P(E_{11}^c) + 2^{-nR_1} 2^{-nI(X_1;Y|X_2)} + 2^{-nR_2} 2^{-nI(X_2;Y|X_1)} + 2^{-n(R_1 + R_2)} 2^{-nI(X_1,X_2;Y)}
\]

(5.4)

Thus the conditions of the theorem cause each term to tend to zero as $n \to \infty$.

Time-sharing allows any $(R_1,R_2)$ in the convex hull to be achieved, and the theorem is proved.

We now prove the converse to Theorem 3. This should provide the reader with some of the basic converse proof techniques.
(ii) Converse. Given a $(2, nR_1, nR_2)$ code for the MAC, the empirical probability distribution on $w_1 \times w_2 \times x_1^n \times x_2^n \times y^n$ is of the form

$$p(w_1, w_2, x_1, x_2, y) = \frac{1}{n(R_1 + R_2)} p(x_1 | w_1) p(x_2 | w_2) p(y | x_1, x_2).$$  \hspace{1cm} (5.5)

Fano's inequality \cite{6} requires that

$$H(W_1, W_2 | Y) \leq n(R_1 + R_2) \bar{p}_e^n + h(\bar{p}_e^n) \Delta \varepsilon_n.$$  \hspace{1cm} (5.6)

Consequently,

$$H(W_1 | Y) \leq n \varepsilon_n$$

and

$$H(W_2 | Y) \leq n \varepsilon_n.$$  \hspace{1cm} (5.7)

Now consider,

\hspace{1cm} (i) \hspace{1cm} nR_1 = H(W_1) = I(W_1; Y) + H(W_1 | Y) \leq I(W_1; Y) + n \varepsilon_n.$$  \hspace{1cm} (5.8)

By the data processing inequality it follows that

$$nR_1 \leq I(W_1; Y) + n \varepsilon_n \leq I(X_1; Y) + n \varepsilon_n.$$  \hspace{1cm} (5.9)

But since $X_1$ and $X_2$ are independent, we have

$$nR_1 \leq I(X_1; Y | X_2) + n \varepsilon_n.$$  \hspace{1cm} (5.9)

By symmetry it can be shown that

\hspace{1cm} (ii) \hspace{1cm} nR_2 \leq I(X_2; Y | X_1) + n \varepsilon_n.$$  \hspace{1cm} (5.10)
and

\[ (iii) \quad n(R_1 + R_2) \leq I(W_1;W_2;Y) + n \varepsilon_n. \quad (5.11) \]

The data processing theorem yields

\[ I(W_1;W_2;Y) \leq I(X_1;X_2;Y). \quad (5.12) \]

Now the discrete memorylessness of the channel is easily applied to yield

\[ R_1 \leq 1/n I(X_1;Y|X_2) + \varepsilon_n \leq 1/n \sum_{i=1}^{n} I(X_{1i};Y_{1i}|X_{2i}) + \varepsilon_n, \]

\[ R_2 \leq 1/n I(X_2;Y|X_1) + \varepsilon_n \leq 1/n \sum_{i=1}^{n} I(X_{2i};Y_{1i}|X_{1i}) + \varepsilon_n, \]

and

\[ R_1 + R_2 \leq 1/n I(X_1,X_2;Y) + \varepsilon_n \leq 1/n \sum_{i=1}^{n} I(X_{1i},X_{2i};Y_{1i}) + \varepsilon_n, \quad (5.13) \]

and the converse is proved.
6. The Slepian-Wolf Source Coding Theorem

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 \( R > H(X,Y) \)
is sufficient. But what if the \( X \)-source and the \( Y \)-source must be separately
described for some user who wishes to reconstruct \( X \) and \( Y \)? Clearly, by
separately encoding \( X \) and \( Y \), it is seen that a rate \( R = R_X + R_Y > H(X) + H(Y) \)
is sufficient. However, in the surprising and fundamental paper of
Slepian and Wolf [15], it is shown that a total rate \( R = H(X,Y) \) suffices.

Example: Let \( X \) be a sequence of 1's and 0's generated by flipping a
fair coin. Let \( Z \) be a sequence of 0's and 1's generated by flipping a coin
with \( \Pr(Z=1) = p = .11 \). Let \( Y = X \oplus Z \). Note that \( Y \) is also a fair
coin sequence.

Suppose that \( R_Y = 1 \), so that \( Y \) is perfectly described. Then the
information rate needed to describe \( X \) is not \( R_X = 1 \) but is \( R_X = H(X|Y) = h(p) = 1/2 \). This is true despite the fact that the describer of \( X \) does
not know the \( Y \) on which his description will be conditioned.

We now give a formal description of the Slepian-Wolf problem depicted
in Figure 6.1. The proof is different from that in [15] and [16] in that
no time sharing is required.

![Diagram of Slepian-Wolf Source Coding](image)

**Figure 6.1**

Correlated Source Coding

-36-
A sequence \( \{(X_i,Y_i)\}_{i=1}^{n} \) of independent copies of the pair \((X,Y)\) of discrete random variables is to be encoded by two separate encoders; an \(X\)-encoder that observes the \(\{X_i\}_{i=1}^{n}\) sequence and maps it into an integer \(i \in [1,2^n]^{X}\), and a \(Y\)-encoder that observes the \(\{Y_i\}_{i=1}^{n}\) sequence and maps it into the integer \(j \in [1,2^n]^{Y}\). The integers \((i,j)\) are then communicated to a common decoder that tries to reproduce the \(\{(X_i,Y_i)\}_{i=1}^{n}\) sequence. A \(((2^n)^X,2^n)^Y,n)\) compression scheme for this problem consists of an integer \(n\), two encoding functions

\[
i: x^n \rightarrow [1,2^n]^X
\]
\[
j: y^n \rightarrow [1,2^n]^Y
\]

and a decoding function

\[
g: [1,2^n]^X \times [1,2^n]^Y \rightarrow x^n \times y^n.
\]

The average probability of incorrect reproduction is given by

\[
\bar{p}_e^n \triangleq P\{g(i(X),j(Y)) \neq (X,Y)\}
\]

is said to be achievable if there exist a sequence of compression schemes with \(\bar{p}_e^n \rightarrow 0\). The problem is to find the set \(R\) of achievable compression rates \((R_X,R_Y)\). The rate region \(R\) (shown in Figure 6.2) was established by Slepian and Wolf [\textsuperscript{1}].

**Theorem 4 (Slepian-Wolf):** The set \(R\) of all achievable rates is given by

\[
R = \{(R_X,R_Y) : R_X > H(X|Y) \\
R_Y > H(Y|X) \\
R_X + R_Y > H(X,Y)\}
\]
Figure 6.2
Rate Region for Slepian-Wolf Data Compression

The idea is to divide the $X^n$-space into $2^{nR_x}$ bins and the $Y^n$-space into $2^{nR_y}$ bins as shown in Figure 6.3.

Figure 6.3
Bins for Slepian and Wolf
The dots in this figure correspond to jointly typical \((x,y)\) pairs, and they \(\varepsilon\)-exhaust the probability. If \(R_x\) and \(R_y\) are large enough, there will be no more than one dot per product bin. Thus the name \(i\) of the \(x\) bin and the name \(j\) of the \(y\) bin will uniquely define the sequence \((x,y)\) that falls in the product bin \((i,j)\). We now prove this result.

Proof:

**Encoding:** Randomly assign every \(x \in \mathcal{X}^n\) to one of \(2^{nR_x}\) bins, according to a uniform distribution over the integers \([1,2^{nR_x}]\). More precisely, for every \(x \in \mathcal{X}^n\), let \(p(i(x) = i) = 2^{-nR_x}\) for \(i \in [1,2^{nR_x}]\).

Similarly, randomly assign every \(y \in \mathcal{Y}^n\) to one of \(2^{nR_y}\) bins, such that \(p(j(y) = j) = 2^{-nR_y}\) for \(j \in [1,2^{nR_y}]\).

**Decoding:** Given \((i_0,j_0)\), declare \((\hat{x},\hat{y}) = (x,y)\) was sent if there is one and only one pair of sequences \((x,y)\) such that

\[
i(x) = i_0, \quad j(y) = j_0,
\]

and

\[(x,y) \in A_\varepsilon^n(X,Y).
\]

Otherwise declare an error. To bound \(P_e^n\), define the events

\[
E_0 = \{(x,y) \notin A_\varepsilon^n(X,Y)\},
\]

\[
E_1 = \{ i(x') = i_0, j(y') = j_0 \text{ such that } (x',y') \neq (x,y) \text{ such that } i(x') = i_0, j(y') = j_0 \}.
\]

Using the union of events bound, we obtain

-39-
\[ \bar{P}_n = P(E_0 \cup E_1) \leq P(E_0) + P(E_1). \quad (6.1) \]

The first term in (6.1) \( \to 0 \) as \( n \to \infty \) by typicality (Lemma 1). Notice that the event \( E_1 \) is equal to the union of the events

\[ E_{11} = \{ \exists \text{ some } x' \text{ such that } x' \neq x \text{ and } i(x') = i_0 \text{ and } y = y' \} , \]

\[ E_{12} = \{ \exists \text{ some } y' \text{ such that } y' \neq y \text{ and } j(y') = j_0 \text{ and } x = x' \} , \]

and

\[ E_{13} = \{ \exists \text{ some } (x', y') \text{ such that } x' \neq x \text{ and } y' \neq y \text{ and } i(x') = i_0 \text{ and } j(y') = j_0 \} . \]

First consider the probability of \( E_{11} \). By the union of events bound

\[ P(E_{11}) \leq \sum_{x' \neq x} P\{i(x') = i_0\} \]

\[ = 2^{-nR_X} \| A_\varepsilon(x|y) \| . \]

But from Lemma 2

\[ \| A_\varepsilon(x|y) \| \leq 2^n(H(X|Y) + 2\varepsilon). \]

Therefore,

\[ P(E_{11}) \to 0 \text{ as } n \to \infty , \]

if

\[ R_X > H(X|Y) . \]

Similarly, the probability of the event \( E_{12} \) and \( E_{13} \to 0 \) as \( n \to \infty \) if
\[ R_Y > H(Y|X) \text{ and } R_X + R_Y > H(Y,X) \]

respectively.

Finally, the probability of error is averaged over random partitions. Thus there exists maps \( i^*(\cdot), j^*(\cdot) \) such that \( \bar{p}_e \to 0 \).

The Slepian-Wolf theorem presented in this section is historically the first fundamental result in multiple user source coding theory. For a comprehensive overview of that theory, the reader is referred to Berger [17].
7. The Multiple Access Channels with Correlated Sources

From Slepian-Wolf, we know how to give efficient separate descriptions of correlated sources. From the multiple access channel, we know how to send two separate independent descriptions over a noisy channel. Apparently we can weld these two problems together to obtain a theory of sending correlated sources over a multiple access channel.

We do so in this section and find that the combined theory involves a new ingredient—correlation of the inputs for the multiple access channel. As a byproduct, we shall find a common proof of the Slepian-Wolf theorem and the multiple access channel capacity region [18].

Assume we have two information sources \( U_1, U_2, \ldots \) and \( V_1, V_2, \ldots \) generated by repeated independent drawings of a pair of discrete random variables \( U \) and \( V \) from a given bivariate distribution \( p(u,v) \).

![Diagram](attachment:image.png)

Figure 7.1

The Multiple Access Channel with Arbitrarily Correlated Sources

A block code for the channel consists of an integer \( n \), and two encoding functions

\[
\begin{align*}
    x_1^n & : U^n \to X_1^n \\
    x_2^n & : V^n \to X_2^n
\end{align*}
\]
assigning codewords to the source outputs, and a decoding function

\[ d^n : y^n \to u^n \times v^n. \]

The probability of error is given by

\[ p_n = P \{ (u^n, y^n) \neq d^n (y^n) \} \]

\[ = \sum_{(u, v) \in u^n \times v^n} p(u^n, v^n) p \{ d^n(y^n) \neq (u^n, v^n) | (u^n, v^n) = (u^n, v^n) \}, \]  

(7.1)

where the joint probability mass function is given, for a code assignment \( \{ x_1(u^n), x_2(v^n) \} \), by

\[ p(u^n, v^n) = \prod_{i=1}^{n} p(u_i, v_i) p(y_i | x_1(u^n), x_2(v^n)). \]  

(7.2)

**Definition:** The source \((U, V)\) can be reliably transmitted over the multiple access channel \((X_1 \times X_2, V, p(y|x_1, x_2))\) if there exists a sequence of block codes \((X_1^n(u^n), X_2^n(v^n), d^n(y^n))\) such that

\[ p_n = P \{ d^n(y^n) \neq (u^n, v^n) \} \to 0. \]

**Example:** Consider the transmission of the correlated sources \((U, V)\) with the joint distribution \(p(u, v)\) given by

\[
\begin{array}{c|cc}
  & 0 & 1 \\
  \hline
u & & \\
\end{array}
\]

\[
\begin{array}{c|cc}
  v & 1/3 & 1/3 \\
  \hline
1/3 & 0 & 1/3 \\
\end{array}
\]

over the multiple access channel defined by
\[ X_1 = X_2 = \{0,1\} \]

\[ Y = \{0,1,2\} \]

\[ Y = X_1 + X_2. \]

Here \( H(U,V) = \log 3 = 1.58 \) bits. On the other hand, if \( X_1 \) and \( X_2 \) are independent,

\[
\max \quad I(Y;X_1,X_2) = 1.5 \text{ bits.}
\]

\[
p(x_1)p(x_2)
\]

Thus \( H(U,V) > I(Y;X_1,X_2) \) for all \( p(x_1)p(x_2) \). Consequently there is no way, even with the use of Slepian-Wolf data compression on \( U \) and \( V \), to use the standard multiple access channel capacity region to send \( U \) and \( V \) reliably to \( Y \). However, it is easy to see that with the choice \( X_1 \equiv U \), and \( X_2 \equiv V \), error-free transmission of the source over the channel is possible. This example shows that the separate source and channel coding described above is not optimal—the partial information that each of the random variables \( U \) and \( V \) contains about the other is destroyed in this separation.

To allow partial cooperation between the two transmitters, we allow our codes to depend statistically on the source outputs. This induces dependence between codewords.

We shall outline here a proof of a special case of Theorem 1, [18], in which \( U \) and \( V \) have no common part. In this case, we must show that \( U \) and \( V \) can be reliably sent to \( Y \) if, for

\[
p(u,v)p(x_1|u)p(x_2|v)p(y|x_1,x_2),
\]

\[
H(U|V) < I(X_1;Y|X_2,V)
\]

\[
H(V|U) < I(X_2;Y|X_1,U)
\]

(7.3)
\[ H(U,V) < I(X_1,X_2;Y). \] (7.4)

The proof will employ random coding. We first describe the random code generation and encoding-decoding schemes, and then analyze the probability of error.

**Generating random codes:** Fix \( p(x_1|u) \) and \( p(x_2|v) \); for each \( u \in U^n \) generate one \( x_1 \) sequence drawn according to \( \prod_{i=1}^{n} p(x_{1i}|u_i) \) and for each \( v \in V^n \) generate one \( x_2 \) sequence drawn according to \( \prod_{i=1}^{n} p(x_{2i}|v_i) \). Call these sequences \( x_1(u) \) and \( x_2(v) \) respectively.

**Encoding:** Transmitter 1, upon observing \( u \) at the output of source 1, transmits \( x_1(u) \) and transmitter 2, after observing \( v \) at the output of source 2, transmits \( x_2(v) \). Assume the maps \( x_1(\cdot), x_2(\cdot) \) are known to the receiver.

**Decoding:** Upon receiving \( v \), the decoder finds the only \((u,v)\) pair such that \((u,v,x_1(u),x_2(v),v) \in A_\varepsilon \) where \( A \) is the set of jointly \( \varepsilon \)-typical sequences. If there is no such \((u,v)\) pair, or there exists more than one such pair, the decoder declares an error. A helpful picture is given in Figure 3.

**Error:** Suppose \((U_0,V_0)\) is the source output, and define the events

\[ E(u,v) = \{(u,v,x_1(u),x_2(v),v) \notin A_\varepsilon \} \cdot \]

The probability of error averaged over \((U_0,V_0)\) and the random codes is given by

\[
P_n = P \left\{ E(U_0,V_0) \cup \{(u,v) \neq (U_0,V_0), E(u,v) \} \right\}
\]

\[
\leq P \left\{ E(U_0,V_0) \right\} + P \left\{ \cup_{(u,v) \neq (U_0,V_0)} E(u,v) \right\}.
\]

-45-
Figure 7.2

Picture of Joint Typicality for Multiple Access Channel

The dots correspond to jointly typical \((X_1, X_2)\) pairs.

Note that only \(2^{nH(U,V)}\) \((x_1(u), x_2(v))\) pairs are likely to occur.
The first term $\rightarrow 0$ by Lemma 1. The second term is bounded above by

$$\sum_{(u_0, v_0) \in A_\epsilon} p(u_0, v_0) \sum_{u' \neq u_0 \atop v' = v_0} p\{ (u', v', X_1(u'), X_2(v'), Y) \in A_\epsilon \mid (u_0, v_0) \}$$

$$+ \sum_{(u_0, v_0) \in A_\epsilon} p(u_0, v_0) \sum_{u' = u_0 \atop v' \neq v_0} p\{ \cdot \} + \sum_{(u_0, v_0) \in A_\epsilon} p(u_0, v_0) \sum_{u' \neq u_0 \atop v' \neq v_0} p\{ \cdot \}$$

$$\leq 2^{2n(H(U|V) + \epsilon)} - 2^{2n(I(X_1; Y|X_2, V) - \epsilon)} + 2^{2n(H(V|U) + \epsilon)} - 2^{2n(I(X_2; Y|X_1, U) - \epsilon)} + 2^{2n(H(U, V)} - 2^{2n(I(X_1, X_2; Y) - \epsilon)}$$

$$(7.5)$$

Consequently, $P_n \rightarrow 0$ if the conditions in (7.3) are satisfied.
8. The Broadcast Channel

The broadcast channel [19] is a communication network with one transmitter and many receivers as shown in Figure 8.1.

![Diagram of a broadcast channel with transmitters and receivers]

**Figure 8.1**

Broadcast Channel

The basic problem is to find the set of simultaneously achievable rates \((R_1, R_2, \ldots, R_K)\). To date this problem has not been solved. The special case of sequentially degraded channels has been solved by Bergmans [20] and Gallager [21]. An achievable rate region for the general broadcast channel has been
put forth by Marton [22], but is not known to be the capacity region.

The basic background for broadcast communication are the maximin and time-sharing approaches. Suppose that the transmission channels to the receivers have respective channel capacities $C_1, C_2, \ldots, C_k$ bits per second. If the same information is to be sent to all receivers, then send at rate $C_{\text{min}} = \min \{C_1, C_2, \ldots, C_k\}$. If the channels are "compatible," each receiver will understand the transmitted information perfectly. Here the transmission rate is limited by the worst channel. At the other extreme, information could be sent at rate $R = C_{\text{max}}$, with resulting rates $R_i = 0, i = 1, 2, \ldots, k-1$, for all but the best channel, and rate $R_k = C_{\text{max}}$ for the best channel.

The next idea is that of time sharing. Allocate proportions of time $\lambda_1, \lambda_2, \ldots, \lambda_k, \lambda_i \geq 0, \sum \lambda_i = 1$, to sending at rates $C_1, C_2, \ldots, C_k$. Assuming compatibility of the channels and assuming $C_1 \leq C_2 \leq \ldots \leq C_k$, we find that the rate of transmission of information through the $i$th channel is given by

$$R_i = \lambda_i C_i$$

These regions of achievable rates can be exceeded. Theory demonstrates that one should not transmit simultaneously to several channels at the rate of the worst channel, nor should one attempt to transmit information by time-sharing or time-multiplexing, but rather one should distribute the high-rate information across the low-rate message. This approach is referred to as superposition coding. The following example will help explain this idea.
Figure 8.2

A Degraded Broadcast Channel

Example:

Consider the two-receiver broadcast channel depicted in Figure 8.2. Channel $X + Y$ is noiseless with capacity $C_1 = 1$ and channel $X \rightarrow Z$ is a binary symmetric channel with error probability $p$ and capacity $C_2 = 1 - h(p)$. First, we know from the random coding proof that a good 
\[ 2^{-n(C_2 - \epsilon)} \]
Code can be generated by choosing at random a subset 
\[ S \]
$2^{-n(C_2 - \epsilon)}$ elements from the set of $2^n$ binary $n$-sequences $X^n = \{1, 2\}^n$, and using the decoding rule that assigns the received vector $y$ to the element of $S$ that is jointly typical with it. By time-sharing, i.e., by allocating a proportion of time $\lambda$ to send to $Y$ at rate $C_1 = 1$, and $\bar{\lambda}$ to send to $Z$ at rate $C_2 = 1 - h(p)$, rate pairs $(R_1, R_2)$ satisfying 
\[ R_1 \leq \lambda, \]
(8.1)
\[ R_2 \leq \lambda(1-h(p)) \text{ for some } \lambda \in [0,1], \quad (8.2) \]

can be achieved. We now show how to improve this region by superposition.

First generate a random code for a channel consisting of the cascade of a BSC of parameter \( p \) and a fake BSC of parameter \( \alpha \). This cascade BSC has parameter \( \alpha \bar{p} + \bar{\alpha} p \), where \( \bar{\alpha} = 1-\alpha \). Thus there will be only
\[ 2^n(C(\alpha \bar{p} + \bar{\alpha} p) - \varepsilon) \]
codewords in this set, where \( C(p) = 1-h(p) \). However, a larger noise of size \( n(\alpha \bar{p} + \bar{\alpha} p) \) will be tolerated. This tolerance can then be used to pack in some extra message information intended solely for the good receiver \( Y \).

Pass each codeword \( x \) in \( S \) through the fake BSC with parameter \( \alpha \). Associate all jointly typical sequences at the output of that BSC with \( x \), as suggested by the clouds of points shown in Figure 8.2. There are
\[ \binom{n}{\alpha n} = 2^{nh(\alpha)} \]
such points. This code structure allows the transmission of an arbitrary integer \( r \in [1,2^nC(\alpha \bar{p} + \bar{\alpha} p)] \) to both receivers \( Y \) and \( Z \) and an arbitrary integer \( s \in [1,2^{nh(\alpha)}] \) to receiver \( Y \). If the message pair \((r,s)\) is to be sent, the integer \( r \) refers to the cloud, and \( s \) to the point \( x \) within the cloud. This \( n \)-sequence \( x \) is then transmitted. The independent transmission rate for the \( Y \) channel is
\[ R_1 = h(\alpha). \quad (8.3) \]

Receiver \( Z \) perceives the cloud center as if it had been sent through an additional BSC of parameter \( \alpha \). However since the cloud centers were chosen to be distinguished over a BSC of parameter \( \alpha \bar{p} + \bar{\alpha} p \), \( r \) is correctly decoded by \( Y \). Thus
\[ R_2 = C(\alpha \bar{p} + \bar{\alpha} p). \quad (8.4) \]

It can be easily demonstrated that the set of achievable rates \((R_1, R_2)\) satisfying (8.3) and (8.4) dominate those satisfying (8.2).
We now give the formal definition of the two-receiver broadcast channel problem. A discrete memoryless broadcast channel \((X, p(y, z|x), Y \times Z)\) as depicted in Figure 8.3 consists of three finite alphabets \(X, Y, Z\) and a probability transition matrix \(p(y, z|x)\).

![Diagram of the two-receiver broadcast channel](image)

**Figure 8.3**

The Discrete Memoryless Broadcast Channel

An \((2^{nR_0}, 2^{nR_1}, 2^{nR_2}, n)\) code for a broadcast channel consists of three sets of integers \(M_0 = [1, 2^{nR_0}], M_1 = [1, 2^{nR_1}],\) and \(M_2 = [1, 2^{nR_2}]\), an encoding function

\[X : M_0 \times M_1 \times M_2 \rightarrow X^n\]

and two decoding functions

\[g_1 : Y^n \rightarrow M_0 \times M_1\]

\[g_2 : Z^n \rightarrow M_0 \times M_2\].

-52-
The integer $w_0$ has the interpretation of the common part of the message, while the integers $w_1, w_2$ are called the independent parts of the message. Assuming uniform distribution on the set of messages $M_0 \times M_1 \times M_2$, define

$$p_e^n = \frac{1}{2^{n(R_0 + R_1 + R_2)}} \sum_{(w_0, w_1, w_2) \in M_0 \times M_1 \times M_2} p \left\{ g_1(Y) \neq (w_0, w_1) \text{ or } \right. \\
left. g_2(Z) = (w_0, w_2) \mid (w_0, w_1, w_2) \text{ sent} \right\} \quad (8.5)$$

to be the average probability of error of the code.

The rate triple $(R_0, R_1, R_2)$ is said to be achievable by a broadcast channel if there exists a sequence of $(2^{nR_0}, 2^{nR_1}, 2^{nR_2}, n)$ codes with $p_e^n \to 0$.

The capacity region $C$ for the broadcast channel is the closure of the set of all achievable rate triples $(R_0, R_1, R_2)$. We saw in the example that superposition is natural when one channel can be considered a degraded version of the other. The mathematical definition corresponding to this physical idea is as follows.

**Definition:** The broadcast channel $(X, p(y, z|x), Y \times Z)$ is called degraded if there exists a probability transition matrix $p(z|y)$ such that for all $z \in Z$ and $x \in X$

$$p(z|x) = \sum_{y \in Y} p(y, z|x) = \sum_{y \in Y} p(z|y)p(y|x). \quad (8.6)$$

**Theorem 5:** The capacity region of the degraded broadcast channel $(X, p(y, z|x), Y \times Z)$ is given by
\[ C = \{ (R_0,R_1,R_2) : R_0 + R_2 < I(U;Z) \]  
\[ R_1 < I(X;Y|U) , \]

for some \( p(u)p(x|u) \),

\[ ||u|| \leq \min \{ ||X||, ||Y||, ||Z|| \} \] \hspace{1cm} (8.7)

**Proof:** Fix \( p(u), p(x|u) \).

(a) **Achievability:**

(i) **Random Code:** First generate \( 2^{n(R_1+R_2)} \) i.i.d. sequences \( u \),

each with probability

\[ p(u) = \prod_{i=1}^{n} p(u_k) . \]

Label these \( u(i), i \in [1,2^{n(R_0+R_2)}] \).

For every \( u(i) \), generate \( 2^{nR_1} \) conditionally independent \( x \) sequences,

each with probability

\[ p(x|u(i)) = \prod_{k=1}^{n} p(x_k|u_k(i)) . \]

Label these \( x(i,j), j \in [1,2^{nR_1}] \).

(ii) **Decoding:** If \( y \) is received, declare \( (\hat{x},\hat{u}) = (i,j) \) was sent

if there is one and only one pair \( (i,j) \in [1,2^{n(R_0+R_2)}] \times [1,2^{nR_1}] \), such that \( (u(i),x(i,j),y) \in A(\hat{u},\hat{x},y) \). If \( z \) is received, declare \( \hat{i} = i \) was sent if there is one and only one \( i \) such that \( (u(i),z) \in A(\hat{u},\hat{z},z) \).
Let $I, J$ be independent r.v.'s drawn according to uniform distributions on $[1, 2^{-n(R_1 + R_2)}] \times [1, 2^{-nR_1}]$. Let the code be chosen randomly according to the encoding description. Then the probability of error averaged over $I, J$ and the random code is given by

$$P_n = P \{ \hat{I}, \hat{J} \neq (I, J) \text{ or } \hat{I} \neq I \}.$$  \hfill (8.8)

By the symmetry induced by the random coding, we see that each transmitted message $(i, j)$ yields the same probability of error. Thus, setting $(i, j) = (1, 1)$, we have

$$P_n = P \{ (I, J) \neq (1, 1) \text{ or } \hat{I} \neq 1 | (I, J) = (1, 1) \}.$$

Define the events

$$E_{z_i} = \{(U(i), Z) \in A_e(U, Z)\}$$

$$E_{y_i} = \{(U(i), Y) \in A_e(U, Y)\}$$

$$E_{y_{ij}} = \{(U(i), X(i, j), Y) \in A_e(U, X, Y)\}.$$ Then

$$P_n = P \left\{ E_{z_1}^C \cup E_{y_{11}}^C \cup_{i \neq 1} E_{z_i} \cup_{i \neq 1} E_{y_i} \cup_{j \neq 1} E_{y_{ij}} \right\} \leq P \{ E_{z_1}^C \} + P \{ E_{y_{11}}^C \} + \sum_{i \neq 1} P \{ E_{z_i} \} +$$

$$+ \sum_{i \neq 1} P \{ E_{y_i} \} + \sum_{j \neq 1} P \{ E_{y_{ij}} \}.$$ The first two terms correspond to the event that the correct codeword does not belong to the decoding set. The last terms correspond to the event
that some incorrect codeword belongs to the decoding set.

From Lemma 1, it follows that

\[ P \{ E_{Zi}^C \} \to 0, \quad P \{ E_{Y1}^C \} \to 0. \]

Consider the event \( E_{Zi} \). We observe, for \( i \neq 1 \), that \( U(i) \) and \( Z \) are independent. Thus, by Lemma 3, for \( i \neq 1 \)

\[ P \{ E_{Zi} \} \simeq 2^{-nI(U;Z)}. \]

Consequently,

\[ \sum_{i \neq 1} P \{ E_{Zi} \} \simeq 2^{-n(I(U;Z)-(R_0+R_2))}. \]

This term goes to zero if

\[ R_0 + R_2 < I(U;Z). \]

Similarly, for the event \( E_{Yi} \), \( i \neq 1 \)

\[ P \{ E_{Yi} \} \simeq 2^{-nI(U;Y)}. \]

and

\[ \sum_{i \neq 1} P \{ E_{Yi} \} \to 0, \text{ if } R_0 + R_2 < I(U;Y) \quad (8.9) \]

But since by the degradedness of the broadcast channel \( I(U;Y) \leq I(U;Z) \), the condition (8.9) is redundant.

Next, consider the event \( E_{Y1j} \). From Lemma 3, it follows that

\[ P \{ E_{Y1j} \} \simeq 2^{-nI(X;Y|U)}. \]
Thus the term $\sum_{j \neq 1} P\{E_{Yj}\} \to 0$ if

$$R_1 < I(X;Y|U).$$  \hfill (8.10)

(b) **Converse:** (See Gallager [21] and Bergmans [23])

The capacity region for the general broadcast channel is still an open problem. Several special cases have been recently solved (e.g. degraded message sets [24], more capable class [25], deterministic class [26],[27], and parallel degraded channels [28]). The most general inner bound to the capacity region of the broadcast channel is that given by Marton [22]. In the following theorem, we outline a proof of a special case of Marton's general result where it is assumed that no common part is decoded. However, this special case isolates a new coding idea. This, together with superposition, yields Marton's theorem. The proof of the following theorem is due to El Gamal and Van der Meulen [29].

**Theorem:** Let

$$R_0 = \{(R_1, R_2) : R_1, R_2 \geq 0, \quad
\begin{align*}
R_1 &\leq I(U;Y), \\
R_2 &\leq I(V;Z), \\
R_1 + R_2 &\leq I(U;Y) + I(V;Z) - I(U;V),
\end{align*}$$

for some $p(u,v,x)$ on $U \times V \times X$ \}

Then any rate pair $(R_1, R_2) \in R_0$ is achievable for $(X, p(y,z|x), Y \times Z)$.

**Proof:** (Outline)

Fix $p(u,v), p(x|u,v)$. The channel $p(y,z|x)$ is given. The idea is
to send \( u \) to \( v \) and \( v \) to \( z \).

Random Coding: Generate \( 2^n I(U; Y) \) typical \( u \)'s in \( U^n \). Generate \( 2^n I(V; Z) \) typical \( v \)'s in \( V^n \). Now randomly throw the \( u \)'s into \( 2^{nR_1} \) bins and the \( v \)'s into \( 2^{nR_2} \) bins. For each product bin, find a jointly typical \((u, v)\) pair. This can be done, as shown in Figure , if

\[
R_1 + R_2 < I(U; Y) + I(V; Y) - I(U; V). \tag{8.12}
\]

To see this, we recall from Lemma 3 that independent choices of \( u \) and \( v \) result in a jointly typical \((u, v)\) with probability \( 2^{-nI(U; V)} \). Now there are \( 2^{n(I(U; Y) - R_1)} \) \( u \)'s in any \( U \)-bin, and \( 2^{n(I(V; Z) - R_2)} \) \( v \)'s in any \( V \)-bin. Thus the expected number of jointly typical \((u, v)\) pairs in a given product \( U \times V \) bin is

\[
\frac{n(I(U; Y) - R_1) \cdot n(I(V; Z) - R_2)}{2} - nI(U; V). \tag{8.13}
\]

The desired jointly typical \((u, v)\) pair can be found if this expected number is much greater than 1. This is guaranteed if (8.9) holds.

Continuing with the coding, for each \( U \times V \) bin and its designated jointly typical \( n \) \((u, v)\) pair, generate \( x(u, v) \) according to the conditional distribution

\[
\Pi_{k=1} p(x_k | u_k, v_k).
\]

Encoding: To send \( i \) to \( Y \) and \( j \) to \( Z \), send \( x(u, v) \), where \((u, v)\) is the designated pair in the product bin \((i, j)\).

Decoding: Receiver \( Y \), upon receiving \( y \), finds the \( u \) such that \((u, y)\) is jointly typical. Thus it is necessary that \( R_1 < I(U; Y) \). He then finds the index \( i \) of the bin in which \( u \) lies. Receiver \( Z \) finds the
such that \((v,z)\) is jointly typical. Thus we need \(R_2 < I(V;Z)\). He then finds the index \(j\) of the bin in which \(v\) lies.

A Note on Feedback: In [30], it was shown that feedback cannot increase the capacity of the physically degraded broadcast channel, i.e., broadcast channels for which \(p(y,z|x) = p(y|x)p(z|y)\). Surprisingly, it was later shown by Ozarow [31] and Deuck [32] that feedback can in fact increase the capacity of general broadcast channels.

![Diagram](image-url)

Figure 8.4

Coding for Marton's Theorem

The discrete memoryless relay channel denoted by \((X_1 \times X_2, p(y, y_1 | x_1, x_2), Y \times Y_1)\) consists of four finite sets \(X_1, X_2, Y, Y_1\) and a collection of probability distributions \(p(\cdot, \cdot | x_1, x_2)\) on \(Y \times Y_1\) one for each \((x_1, x_2) \in X_1 \times X_2\). The interpretation is that \(x_1\) is the input to the channel and \(y\) is the output, \(y_1\) is the relay's output and \(x_2\) is the input symbol chosen by the relay as shown in Figure 9.1. The problem is to find the capacity of the channel between the sender \(x_1\) and the receiver \(y\).

The relay channel was introduced by van der Meulen [33]. The following discussion is based on Cover and El Gamal [34].

![Figure 9.1](image)

The Relay Channel
A \((2^{nR}, n)\) code for the relay channel consists of a set of integers 
\([1, 2^{nR}]\), an encoding function

\[ x_1 : [1, 2^{nR}] \rightarrow X_1^n, \]

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

\[ x_{2i} = f_i(Y_{11}, Y_{12}, \ldots, Y_{1i-1}), \quad 1 \leq i \leq n, \]

and a decoding function

\[ g : Y^n \rightarrow [1, 2^{nR}]. \]

Note that the allowed encoding functions actually form part of the
definition of the relay channel because of the nonanticipatory relay condition. The relay channel input \(x_{2i}\) is allowed to depend only on the past
\(y_{11}, y_{12}, \ldots, y_{1i-1}\). This is the definition used by van der Meulen [33].
The channel is memoryless in the sense that \((y_i, y_{1i})\) depends on the past
only through the current transmitted symbols \((x_{1i}, x_{2i})\). Thus,
for any choice \(p(w), w \in M\), and code choice \(x_1 : [1, 2^{nR}] \rightarrow X_1^n\) and relay
functions \(\{f_i\}_{i=1}^\infty\), the joint probability mass function on 
\(M \times X_1^n \times X_2^n \times Y^n \times Y_1^n\) is given by

\[
p(w, x_1, x_2, y, y_1) =
\]

\[
p(w) \prod_{i=1}^n p(x_{1i} | w)p(x_{2i} | y_{11}, y_{12}, \ldots, y_{1i-1})p(y_i, y_{1i} | x_{1i}, x_{2i}). \tag{9.1}
\]

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

\[ \lambda (w) = \Pr \{ g(y) \neq w \mid w \text{ sent} \} \]

-61-
denote the conditional probability of error. We define the average probability of error of the code as

$$\overline{p}_e^n = 2^{-nR} \sum_{w} ^\lambda (w).$$

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

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

**Theorem 9.1:**

For any relay channel $(X_1 \times X_2, p(y,y_1|x_1,x_2), Y \times Y_1)$ the capacity $C$ is bounded above by

$$C \leq \sup_{p(x_1,x_2)} \min \{ I(X_1,X_2;Y), I(X_1;Y,Y_1|X_2) \}$$  \hspace{1cm} (9.2)

This upper bound has a nice max flow min cut interpretation. The first term in (9.2) upper bounds the rate of information flow from senders $X_1$ and $X_2$ to receiver $Y$. The second term bounds the rate of transmission from $X_1$ to $Y$ and $Y_1$.

We now consider a family of relay channels in which the relay receiver $Y_1$ is better than the ultimate receiver $Y$ in the sense defined below.

**Definition:** The relay channel $(X_1 \times X_2, p(y,y_1|x_1,x_2), Y \times Y_1)$ is said to be **degraded** if $p(y,y_1|x_1,x_2)$ can be written in the form

$$p(y,y_1|x_1,x_2) = p(y_1|x_1,x_2)p(y|y_1,x_2).$$  \hspace{1cm} (9.3)
Thus \( Y \) is a random degradation of the relay signal \( Y_1 \).

For the degraded relay channel, the capacity is given by the following

**Theorem 9.2:** The capacity \( C \) of the degraded relay channel is given by

\[
C = \sup_{p(x_1, x_2)} \min \{ I(X_1, X_2; Y), I(X_1; Y_1 | X_2) \} \tag{9.4}
\]

where the supremum is over all joint distributions \( p(x_1, x_2) \) on \( X_1 \times X_2 \).

**Proof:** Converse.

The converse follows from Theorem \( \ldots \) and by degradedness. Thus

\[
I(X_1; Y, Y_1 | X_2) = I(X_1; Y_1 | X_2). \tag{9.5}
\]

This capacity region proof involves 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.

**Achievability (Outline):**

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

In each \( n \)-block \( b = 1, 2, \ldots, B \), we shall use the same doubly indexed set of codewords

\[
C = \{ x_1(w | s), x_2(s) \},
\]

\[
w \in [1, 2^{nR}], \ s \in [1, 2^{nR_0}], \tag{9.6}
\]

\[
x_1(\cdot, \cdot) \in X_1^n, \ x_2(\cdot) \in X_2^n.
\]
We shall also need a partition

\[ S = \left\{ S_1, S_2, \ldots, S_{nR_0} \right\} \quad \text{of} \quad M = \left\{ 1, 2, \ldots, 2^{nR} \right\} \]

into

\[ 2^{nR_0} \] cells, \( S_i \cap S_j = \emptyset, \quad i \neq j, \quad \bigcup S_i = M. \]

The partition \( S \) will allow us to send side information to the receiver in the manner of Slapeian and Wolf [ ].

**Random Coding:** First generate at random \( 2^{nR_0} \) independent identically distributed n-sequences in \( X_2^n \), each drawn according to \( p(x_2) = \prod_{i=1}^{n} p(x_{2i}). \) Index them as \( x_2(s), \ s \in [1, 2^{nR_0}]. \) For each \( x_2(s) \), generate \( 2^{nR} \) conditionally independent n-sequences \( x_1(w|s), \ w \in [1, 2^{nR}] \) drawn according to \( p(x_1|x_2(s)) = \prod_{i=1}^{n} p(x_{1i}|x_{2i}(s)). \) This defines a random codebook \( C = \{ x_1(w|s), x_2(s) \}. \)

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

**Encoding:** Let \( w_i \in [1, 2^{nR}] \) be the new index to be sent in block \( i, \) and assume that \( w_{i-1} \in S_{s_i}. \) Then the encoder sends \( x_1(w_i|s_i). \) 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 \( \hat{w}_{i-1} \in S_{\hat{s}_i}. \) Then the relay encoder sends the codeword \( x_2(s_i) \) 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 relay knows \((w_1, w_2, \ldots, w_{i-1})\)
and consequently \((s_1, \ldots, s_i)\).

The decoding procedures at the end of block \(i\) are as follows.

1) Knowing \(s_i\), and upon receiving \(y_1(i)\), the relay receiver estimates the message of the transmitter \(\hat{w}_i = w\) iff there exists a unique \(w\) such that \((x_{1}(w|s_i), x_{2}(s_i), y_1(i))\) are jointly \(\varepsilon\)-typical. Using Lemma 3 it can be shown that \(\hat{w}_i = w_i\) with arbitrarily small probability of error if

\[
R < I(X_1; Y_1 | X_2).
\]  \hfill (9.7)

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 \((x_2(s), y(i))\) are jointly \(\varepsilon\)-typical. From Lemma 3 we know that \(s_i\) can be decoded with arbitrarily small probability of error if

\[
R_0 < I(X_2; Y),
\]  \hfill (9.8)

and \(n\) is sufficiently large.

3) Assuming that \(s_i\) is decoded successfully at the receiver, then \(\hat{w}_{i-1} = w\) is declared as the index sent in block \((i-1)\) iff there is a unique \(w \in S_{s_i} \cap L(Y(i-1))\) where \(L(Y(i-1))\) is the list of indices \(w\) that the receiver \(y\) considered to be "typical" with \(Y(i-1)\) in the \((i-1)\)th block. If \(n\) is sufficiently large and if

\[
R < I(X_1; Y | X_2) + R_0
\]  \hfill (9.9)

then \(\hat{w}_{i-1} = w_{i-1}\) with arbitrarily small probability of error. Combining the two constraints (8.8) and (8.9) \(R_0\) drops out, leaving

\[
R < I(X_1; Y | X_2) + I(X_2; Y) = I(X_1, X_2; Y).
\]  \hfill (9.10)
For a detailed analysis of the probability of error, the reader is referred to [34].

Theorem 9.1 was also shown to be the capacity for the following classes of relay channels:

(i) Reversely degraded:

\[ p(y, y_1 | x_1, x_2) = p(y_1 | x_1, x_2) p(y | y_1, x_2) . \]

(ii) Relay channel with feedback.

(iii) Deterministic relay channel [35]:

\[ y = f(x_1, x_2), \quad y_1 = g(x_1, x_2) . \]

A general lower bound to the capacity of any relay channel can be found in [34].
10. Summary and Open Problems.

Now that we have presented the Gaussian multiple user channels in Section 2 and proved some of the results in detail in subsequent sections, it is time to abstract the salient points of the theory. We do so by paralleling the discussion of Section 2 for the channels shown in Figure 10.1.

![Diagram of channels](image)

*Figure 10.1*  
Multiple User Channels

1. The Shannon channel: The codewords are $n$-vectors $x(1), x(2), \ldots, x(2^n)$. First suppose $R < I(X;Y)$. The intuitive idea is that $2^n$ sequences, each independent identically distributed according to $\prod_{i=1}^{n} p(x_i)$ will be mutually far apart in the sense that if one sends an $x(i)$ and receives a $Y$, then, looking back from the $Y$, one finds only one $X$ in the codebook that is jointly typical with $Y$. This is the correct $x(i)$.

2. List codes: Suppose that $2^n$ codewords are generated as above for the Shannon channel, but $R > I(X;Y)$. Then there will be exponentially
many codewords jointly typical with \( Y \). In fact, the number of codewords on the list associated with \( Y \) will be \( 2^{n(R-I)} \). Thus the cutoff at capacity is very sharp. One goes from one codeword in the inverse fan for \( R < I \) to an exponential number of codewords for \( R > I \).

3. The multiple access channel: Again, random coding works. Choose \( 2^{nR_1} x_1 \) sequences according to \( \Pi p(x_{1i}) \) and choose \( 2^{nR_2} x_2 \) sequences according to \( \Pi p(x_{2i}) \). Now send one of the codewords from the first codebook, and one of the codewords from the second codebook. These two together generate a random response \( Y \) drawn according to the conditional probability distribution of \( Y \) given the two sequences. With high probability, \( Y \) will be jointly typical with those two sequences. If \( R_1 \) and \( R_2 \) satisfy
\[
R_1 < I(X_1;Y|X_2), \quad R_2 < I(X_2;Y|X_1), \quad R_1 + R_2 < I(X_1,X_2;Y)
\]
for some \( p_1(x_1)p_2(x_2) \), then the receiver \( Y \) will see that, although there are exponentially many \( x_1 \)'s jointly typical with \( Y \) and exponentially many \( x_2 \)'s jointly typical with \( Y \), there will be only one \( (x_1,x_2) \) pair in the codebook that is jointly typical with \( Y \). Thus no error will be made in the decoding.

4. The degraded broadcast channel: We assume that \( X \) sends to two receivers \( Y \) and \( Z \), where \( Z \) is stochastically noisier than \( Y \). Here, the idea of superposition is needed. We first introduce an auxiliary sender \( U \) and an associated channel from \( U \) to \( X \). Thus the overall channel becomes \( U \rightarrow X \rightarrow Y \rightarrow Z \). We now use a random code for the channel \( U \rightarrow Z \) with \( 2^{nR_2} \) codewords \( u \) independently drawn according to \( \Pi p(u_i) \). For each such codeword \( u \), we generate \( 2^{nR_1} \) codewords \( x \) according to \( \Pi p(x_i|u_i) \). If \( R_2 < I(U;Z) \), then \( Z \) perceives \( u \) correctly. And if \( R_1 < I(X;Y|U) \), then receiver \( Y \) perceives \( x \) correctly after decoding \( u \).
In summary, we send crude codewords to the poor receiver \( Z \), but each of these crude codewords is really made up of small satellite codewords distinguishable by \( Y \).

5. **Degraded relay channel**: Random coding works, but here we also need the Slepian-Wolf theorem. The idea is that we randomly choose codewords which are distinguishable by the relay, but not by the ultimate receiver. Nonetheless, by the argument for list codes, the number of codewords on the ultimate receiver's list is given by

\[ 2^{n(R-I(X_1;Y|X_2))} \]

Then the relay, in cooperation with the sender, attempts to send the name of the codeword on the (unknown) list of known size \( 2^nR_0 \). He does this by randomly partitioning all the codewords into \( 2^{nR_0} \) sets each of equal size. This is the Slepian-Wolf step. He then sends the index of that set in cooperation with the transmitter. They do so by superimposing this information in the manner suggested by the degraded broadcast channel. Yet one more ingredient is necessary. The resolution of the index of the sent codeword on \( Y \)'s list does not take place until the next block transmission. Thus, we have a block Markov encoding of the resolution information, while superimposing the fresh information for the relay.

Summing up, we have the following points:

1. We use a random code to send information to the relay.
2. We use the list coding idea to see how many codewords are left to be resolved by the ultimate receiver \( Y \).
3. We have a block Markov encoding scheme in which the resolution of the information is sent in the next block.
4. We use superposition to put cooperative information on top of the fresh infusion of new information to the relay.
When one puts all of these ideas together, one gets the achievable rate region for the degraded relay channel. Moreover, a converse, which we do not prove here, shows that this is in fact the capacity for the degraded relay channel.

We can obviously combine these building blocks to solve networks like that in Figure 10.2 and schematized in Figure 10.3.

Figure 10.2
Communication System
Multiple Access \hspace{1cm} \text{Degraded Relay} \hspace{1cm} \text{Shannon Channel with Feedback} \hspace{1cm} \text{Shannon Channel} \hspace{1cm} \text{Degraded Broadcast Channel}

\begin{align*}
(R_1, R_2) &\in C_{\text{MAC}} \hspace{1cm} R_1 + R_2 < C_{\text{DR}} \hspace{1cm} R_1 + R_2 < C_{\text{FB}} \hspace{1cm} R_1 + R_2 < C \hspace{1cm} (R_1, R_2) \in C_{\text{DB}} \\
C_{\text{FB}} &= C
\end{align*}

\textit{Figure 10.3}

Communication Network with Known Capacity

We have suppressed the input and output labels at the nodes of this graph. In this network, sender 1 wishes to send an index \( W_1 \) at rate \( R_1 \) and sender 2 wishes to send an index \( W_2 \) at rate \( R_2 \) to the receivers respectively designated as \( \hat{W}_1 \) and \( \hat{W}_2 \) with overall probability of error \( P \{ (W_1, W_2) \neq (\hat{W}_1, \hat{W}_2) \} \) tending to zero. This can be accomplished if and only if \( R_1 \) and \( R_2 \) simultaneously satisfy all of the constraints in Figure 10.3. This overall capacity region is the convex subset of the \( (R_1, R_2) \) plane given by the intersection of the individual capacity regions. Thus maxflow mincut holds for the capacity regions in this network.

Optimism is created each time a new capacity theorem is obtained, but a general theory does not yet exist for information flow in networks. Even for the building blocks of networks, some problems remain open. The multiple
access capacity is known, but the capacity of the multiple access channel with feedback is still open. The capacity of the degraded broadcast channel is known, but the capacity of the general broadcast channel remains unknown. The capacity of the general relay channel is still unknown. Progress on Shannon's two-way channel [36] and on the interference channel[37] has just begun. A general theory for this problem is years away. Indeed there may not be a mathematically nice answer. Even if there is a nice answer, it may not yield a practical implementation. But in communication theory, as in physics, one must have some faith. If a theory is found, it will have enormous qualitative theoretical implications. Indeed, even if optimal implementations prove to be too complex, the theory will tell the communication designer when he is close to optimality. Finally, new coding schemes will be suggested from inspection of the proofs. Fortunately, all of the results achieved to date are consistent with an information theoretic description solely in terms of quantities like $I(X_1,X_2;Y_3|X_3)$, where $I(\cdot;\cdot|\cdot)$ denotes conditional mutual information. Whether this is a result of limited technique or is true of the entire theory is yet to be revealed.

Acknowledgement: The authors are thankful to Chris Heegard for many valuable remarks during the preparation of this paper.
REFERENCES


