# SEQUENTIAL DECODING FOR MULTIPLE ACCESS CHANNELS

by

ERDAL ARIKAN

B.S.E.E., California Institute of Technology
(1981)

S.M.E.E., Massachusetts Institute of Technology
(1982)

SUBMITTED TO THE DEPARTMENT OF
ELECTRICAL ENGINEEERING AND COMPUTER SCIENCE
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

November 1985

Signature of Author ________________________________________
Department of Electrical Engineering and Computer Science
November 18, 1985

Certified by ____________________________________________
Professor Robert G. Gallager
Thesis Supervisor

Accepted by ____________________________________________
Arthur C. Smith
Chairman, Departmental Committe on Graduate Students

SEQUENTIAL DECODING FOR MULTIPLE ACCESS CHANNELS

by

ERDAL ARIKAN

Submitted to the Department of Electrical Engineering and
Computer Science on November 18, 1985
in partial fulfillment of the requirements for the
Degree of Doctor of Philosophy

## ABSTRACT

Sequential decoding is a decoding algorithm for tree codes originally developed for single-user channels (i.e., channels with one transmitter and one receiver). Sequential decoding relies on what is called a metric to direct its search and find the path in the tree that corresponds to the encoded message. The decoding complexity in sequential decoding, that is, the number of computations to decode a source digit, is a random variable. A rate is said to be achievable by sequential decoding if it is possible to select a code with that rate and a metric such that the expected value of the decoding complexity is finite. In the single-user case, the largest achievable rate is called the cut-off rate of sequential decoding.

Multiple access channels are models of communication systems where there are a number of users all sharing the same transmission medium to communicate their messages to a common receiver. This thesis explores the possibility of using sequential decoding on multiple access channels. Immediate generalizations of the metrics, in particular of the Fano metric, that have been used in the past for single-user sequential decoding, do not work satisfactorily in the multi-user case. A new metric is introduced which works quite satisfactorily not only for multiple access channels but also for single-user ones. The achievable rate region of sequential decoding under this new metric is evaluated. It is shown by examples that sequential decoding has the potential of achieving rates (throughputs) beyond those achievable by conventional ways of using multiple access channels, such as time-division multiplexing, frequency division multiplexing, and Aloha-like schemes.

Outer bounds to the achievable rate region of sequential decoding are considered. The cut-off rate of sequential decoding (in the single-user case) is determined, thus settling a long-standing open question. Also, the achievable rate region of sequential decoding is determined in the case of multiple access channels that have a property known as pairwise-reversibility. The achievable rate region of sequential decoding for arbitrary multiple access channels remains undetermined.

An alternative approach to sequential decoding, in which there is a separate sequential decoder for each user in the system, is considered and an inner bound to its achievable rate region is given. Non-joint sequential decoding, as this approach is called, has the advantage of being simple: each sequential decoder is responsible for decoding the message of a single user, so it does not have to know the tree codes of the other users. An example is given for which non-joint sequential decoding, in addition to being simpler, also achieves rates that are unachievable by ordinary sequential decoding.

Name and Title of Thesis Supervisor:
Robert G. Gallager
Professor of Electrical Engineering and Computer Science

## ACKNOWLEDGMENTS

I wish to express my deepest gratitude to Professor Robert Gallager for his guidance and support throughout my graduate study. This work could not have been possible without his supervision. Working with him was a truly exciting and enlightening experience.

I am thankful to Professors Peter Elias and Pierre Humblet, my thesis readers, for their helpful comments at the final stages of this work.

Life as a graduate student has been endurable in part because I was fortunate to share the same office with Isidro Castineyra, Julio Escobar, Jeannine Mosely and Edward Tiedemann. Their camaraderie will be missed.

This research was conducted at M.I.T. Laboratory for Information and Decision Systems with support provided by Defence Advanced Research Projects Agency under Contract N000 14-84-K-0357, for which I am grateful.

# CONTENTS

# Chapter 1

# INTRODUCTION

Multiple access channels are models of communication systems in which there are a number of uncoordinated users sharing a transmission medium to transmit messages to a common destination. Some examples of multiple access channels are a satellite transponder shared by several ground stations, a radio network in which users transmit over the same frequency band to exchange messages, and a computer network where several computers send messages over a common bus.

One common approach to multiple access communications is to employ time-sharing (time-division multiplexing), in which at any given time only one user is allowed to transmit a message. This idea of splitting a given channel into non-interfering subchannels and giving the use of each subchannel exclusively to a single user also underlies frequency-division multiplexing and other techniques that aim at elimination of multi-user interference.

Another approach, which is much less common than time-sharing, is to let all users transmit simultaneously, thus allowing them to interfere with each other. In this approach, a sufficient amount of redundancy is embedded into what is transmitted by each user so that, with high probability, the receiver can reconstruct the messages correctly. This is the coding approach to multiple access communications. Theoretically, coding affords a channel utilization (throughput) always as high as, and often significantly higher than, what is possible by time-sharing. The reason for being interested in coding for multiple access channels is thus the desire to communicate at higher rates, or more reliably at a given rate.

While coding is potentially superior to time-sharing in terms of throughput, it requires more complexity in the form of encoders and decoders. In addition, there is the problem of finding an encoder-decoder pair achieving a given desired rate. This thesis examines a particular approach to coding for multi-access channels, namely, tree coding and sequential decoding, and establishes it as a practically applicable method for achieving rates beyond those achievable by time-sharing.

## 1.1. The Multiple Access Channel Model

The multiple access channel model used in this thesis has, as its central element, a channel (in the information theoretic sense of the word), which has one input for each user and a single output to the common destination (Figure 1.1.1).

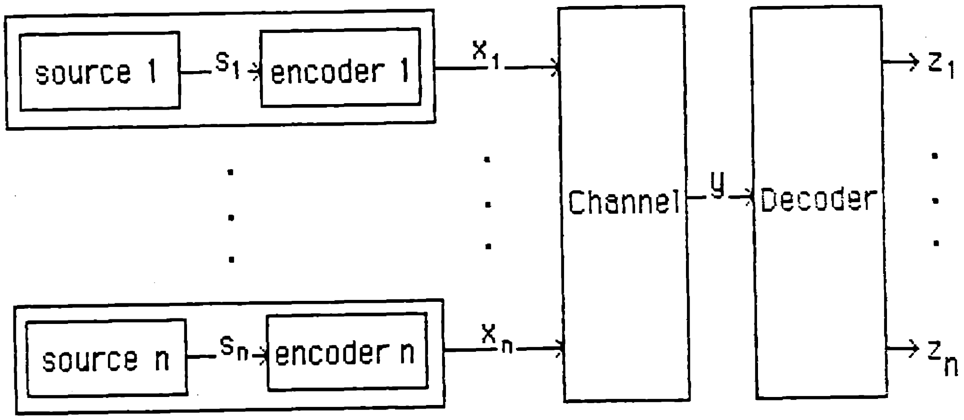


Figure 1.1.1. Multi-user communication system model.

Our study is restricted to the class of channels which have the following properties.

1) The channel operates in discrete time; it can be used only once a second, say.

2) The channel is discrete; that is, the channel input and output alphabets are finite sets.

3) The channel is memoryless and stationary. Memorylessness is the property that the statistics of the output at any given time depends only on the inputs at that time, and possibly on the time itself. Stationarity rules out the dependence of channel statistics on time.

A channel in this class with n users can be identified by its input alphabets $X_1,...,X_n$, its output alphabet Y, and its transition probabilities $P=\{P(\eta|\xi):\eta\epsilon Y,\ \xi\epsilon X_1\times\cdots\times X_n\}$. $P(\eta|\xi)$ is the probability of receiving $\eta$ given that $\xi$ is transmitted. If $\xi=(\xi_1,...,\xi_n)$, an alternative notation for $P(\eta|\xi)$ is $P(\eta|\xi_1,...,\xi_n)$; $P(\eta|\xi_1,...,\xi_n)$ is thus the probability that $\eta$ is received given that user i transmits $\xi_i$, i=1,...,n. A channel with these parameters will be denoted by $(P;X_1,...,X_n;Y)$.

The encoders in this model are what we call (M,k) encoders, where M and k are arbitrary positive integers. An (M,k) encoder is a device which sends k symbols to the channel for each digit it receives from the source; M designates the size of the source alphabet.

In general, each user may have encoders with arbitrary parameters, say, $(M_i,k_i)$ for user i, i=1,...,n. We shall, however, consider only those cases where $k_i$ is the same for all i, and denote the parameter of such a collection of encoders by $(M_1,...,M_n,k)$.

A source for an $(M,\cdot)$ encoder is viewed as an infinite shift-register holding digits from a set with M elements. It is assumed that each digit in each source register is a random variable, uniformly distributed, and independent of all other source digits in the same or in other registers. Viewing the sources in this way eliminates the source coding problem, and thus, enables us to focus on the problem of channel coding, which is the problem of main interest here.

At this point, we view the decoder quite generally as any device that generates an estimate for each source digit.

Notice that, as a result of the statistical independence at the source level and the lack of cooperation among the users in the encoding of their messages, the inputs to the channel by different users are statistically independent. This is the essential difference between a multi-user channel, say, $(P;X_1,\dots,X_n;Y)$ and its single-user counterpart $(P;X_1\times\cdots\times X_n;Y)$.

The two main performance criteria for the analysis of this model will be the expected system delay and the probability of decoding error. System delay for a source digit is defined as the time lag from the time that digit is accepted by its encoder to the time the decoder delivers its estimate about that digit. System delay is permitted to be a random variable; but clearly, a system can not be used in practice unless the expected system delay is uniformly bounded over all source digits.

Probability of decoding error for a source digit is the probability that the decoder estimate for that digit is in error. We are interested in finding ways of reducing the probability of decoding error to arbitrarily low levels for each source digit, while keeping the expected system delay bounded.

In order to describe the model precisely, and also for future reference, we now list the notation that will be used throughout this thesis.

Notation, Concepts, and Conventions

Transmissions start at time 1, and take place at times 1,2,3,...

As a convention, in the following notation, subscripts refer to user identity, arguments refer to time.

Generically, $\mathbf{e}_i$ stands for the encoder (the device) and the encoding operation for user i; the parameter of $\mathbf{e}_i$ is denoted by $(M_i,k)$; and the number of users is denoted by n. $\mathbf{e}$ denotes the collection of encoders $\mathbf{e}_1,\dots,\mathbf{e}_n$, and also the joint encoding operation.

Source Outputs, Encoder Inputs

$s_i(m)$ is the $m^{th}$ input to $\mathbf{e}_i$, or equivalently, the $m^{th}$ output of source i.
$s_i(..m)=(s_i(1),....,s_i(m))$ is the first m inputs to $\mathbf{e}_i$.
$s_i=s_i(1),s_i(2),...$ is the input sequence to $\mathbf{e}_i$.

It is important to note that $s_i$ denotes the actual output of source i. Throughout what follows, the letter s is reserved for denoting actual source outputs. When there is need to mention a possible but arbitrary output sequence for source i, we write $u_i$ or $\tilde{u}_i$ or $v_i$, but never $s_i$. Thus, $u_i$ denotes an arbitrary sequence of letters from $\{1,...,M_i\}$. We denote the $m^{th}$ letter of $u_i$ by $u_i(m)$, and the first m letters of $u_i$ by $u_i(..m)$.

Encoder Outputs, Channel Inputs

$x_i(m)$ is the $m^{th}$ output block of $\mathbf{e}_i$, m=1,2,...
$x_i(m,j)$ is the $j^{th}$ digit of $x_i(m)$, j=1,...,k.
$x_i(..m)=(x_i(1),....,x_i(m))$ is the first m output blocks of $\mathbf{e}_i$.
$x_i=x_i(1,1),...,x_i(1,k),x_i(2,1),...$ is the output sequence of $\mathbf{e}_i$.

$x_i$ is the actual output of encoder $\mathbf{e}_i$; in other words, it is the sequence of channel symbols transmitted by user i. $x_i$ and $s_i$ are related through the equation $x_i(m)=\mathbf{e}_i(s_i(..m))$. As stated earlier, $\mathbf{e}_i$ is regarded not only as a device (the encoder) but also as the encoding operation itself. In this second sense, $\mathbf{e}_i$ is a causal operator mapping source sequences into channel input sequences.

Our model allows $x_i(m)$ to depend on all of $s_i(1),...,s_i(m)$, no matter how large m is. If $x_i(m)$ does not depend on $s_i(m-b-1)$ for any $b \geq b_0$ and $b_0$ is the smallest integer with this property, then $b_0$ is said to be the memory of $\mathbf{e}_i$.

Encoders with zero memory are called <u>block encoders</u>, and they will be discussed in the next section. The discussion of block codes aims at introducing certain theorems that are useful in understanding the coding problem in multi-access channels. Our focus in this thesis is on tree codes, which are generated by encoders that may have arbitrarily large memories.

We often use the following notation for the actual channel inputs:

$$\mathbf{e}_i s_i = x_i \ , \qquad \mathbf{e}_i s_i(m) = x_i(m) \ , \quad \mathbf{e}_i s_i(..m) = x_i(..m) \ .$$

We use the following notation in relation to what would be observed as the output of $\mathbf{e}_i$ if $u_i$ were the input to $\mathbf{e}_i$.

$\mathbf{e}_i u_i(m)=\mathbf{e}_i(u_i(..m))$, the $m^{th}$ output block of $\mathbf{e}_i$ in response to $u_i$.
$\mathbf{e}_i u_i(..m)=(\mathbf{e}_i u_i(1),...,\mathbf{e}_i u_i(m))$, the first m blocks in response to $u_i$.
$\mathbf{e}_i u_i=\mathbf{e}_i u_i(1),\mathbf{e}_i u_i(2),....$ , the output sequence in response to $u_i$.

<u>Inputs and Outputs for the Joint Encoder</u>

$s(m)=(s_1(m),...,s_n(m))$ is the $m^{th}$ input to $\mathbf{e}$.
$s(..m)=(s(1),...,s(m))$ is the first m inputs to $\mathbf{e}$.
$s=s(1),s(2),...$ is the input sequence to $\mathbf{e}$.

$x(m,j)=(x_1(m,j),...,x_n(m,j))$ is the $j^{th}$ digit in the $m^{th}$ output block of $\mathbf{e}$.
$x(m)=(x(m,1),.....,x(m,k))$ is the $m^{th}$ output block of $\mathbf{e}$.
$x(..m)=(x(1),...,x(m))$ is the first m output blocks of $\mathbf{e}$.
$x=x(1,1),...,x(1,k),x(2,1),...$ is the output sequence of $\mathbf{e}$.

The functional relationship between the <u>joint</u> source output s and the <u>joint</u> channel input x will be expressed by writing $x(m)=\mathbf{e}(s(..m))$. Thus, $\mathbf{e}$ is regarded both as a label for the collection of all encoders and as an operator mapping sequences of letters from $\{1,..,M_1\}\times\cdots\times\{1,...,M_n\}$ into sequences of letters from $X_1\times\cdots\times X_n$. In this second sense, **e** is an encoder with parameter $(M_1\cdots M_n,k)$, input alphabet $\{1,..,M_1\}\times\cdots\times\{1,...,M_n\}$, and output alphabet $X_1\times\cdots\times X_n$.

We often use the following notation for joint channel inputs.

$$\mathbf{e}s=x\ ,\qquad \mathbf{e}s(m)=x(m)\ ,\qquad \mathbf{e}s(..m)=x(..m)\ .$$

As in the case of individual source sequences, the letter s is reserved for denoting the actual <u>joint</u> source outputs. Arbitrary joint source sequences are denoted by u or $\tilde{u}$ or v, etc. Thus, u denotes a sequence of elements from $\{1,...,M_1\}\times\cdots\times\{1,...,M_n\}$; u(m) denotes the m$^{th}$ letter of u; and u(..m) denotes the first m letters of u.

We use the following notation in relation to what would be observed as the output of **e** if u were the input to **e**.

**e**u(m)=**e**(u(..m)), the m$^{th}$ output block of **e** in response to u.
**e**u(..m)=(**e**u(1),...,**e**u(m)), the first m blocks in response to u.
**e**u=**e**u(1),**e**u(2),.... , the output sequence in response to u.

<u>Channel and Decoder Outputs</u>

y(m,j) is the channel output in response to x(m,j).
y(m)=(y(m,1),.....,y(m,k)) is the m$^{th}$ channel output block.
y(..m)=(y(1),...,y(m)) is the first m channel output blocks.
y=y(1,1),...,y(1,k),y(2,1),... is the channel output sequence.

$z_i(m)$ is the decoder estimate for $s_i(m)$.
$z(m)=(z_1(m),...,z_n(m))$ is the decoder estimate for s(m).
An error in the decoding of $s_i(m)$ is the event that $z_i(m)\neq s_i(m)$.

This completes the basic list of notation.

We now introduce an operation to simplify the notation.

For any collection of sets $A_1,\dots,A_h$, any integer t, and any collection of $\xi_i=(\xi_{i,1},\dots,\xi_{i,t})\epsilon A_i{}^t$, $i=1,\dots,n$, we define

$$\xi_1\times\xi_2\times\cdots\times\xi_n = ((\xi_{1,1},\xi_{2,1},\dots,\xi_{n,1}),(\xi_{1,2},\xi_{2,2},\dots,\xi_{n,2}),\dots\dots,(\xi_{1,t},\dots\dots,\xi_{n,t})).$$

If $\xi_i=\xi_{i,1},\xi_{i,2},\xi_{i,3},\dots.$ with $\xi_{i,j}\epsilon A_i$, then we define

$$\xi_1\times\xi_2\times\cdots\times\xi_n=(\xi_{1,1},\xi_{2,1},\dots,\xi_{n,1}),(\xi_{1,2},\xi_{2,2},\dots,\xi_{n,2}),\dots.$$

Some of the preceding relations can now be restated as follows.

$$s(m)=s_1(m)\times\cdots\times s_n(m)\ ,\qquad s(..m)=s_1(..m)\times\cdots\times s_n(..m)\ ,\qquad s=s_1\times\cdots\times s_n\ .$$

$$x(m)=x_1(m)\times\cdots\times x_n(m)\ ,\qquad x(..m)=x_1(..m)\times\cdots\times x_n(..m)\ ,\qquad x=x_1\times\cdots\times x_n\ .$$

$\mathbf{e}s=\mathbf{e}_1s_1\times\cdots\times\mathbf{e}_ns_n$,
$\mathbf{e}s(m)=\mathbf{e}_1s_1(m)\times\cdots\times\mathbf{e}_ns_n(m)$,
$\mathbf{e}s(..m)=\mathbf{e}_1s_1(..m)\times\cdots\times\mathbf{e}_ns_n(..m)$.

If $u_i$ is an arbitrary input sequence for $\mathbf{e}_i$, $i=1,\dots,n$, and $u=u_1\times\cdots\times u_n$, then
$\mathbf{e}u=\mathbf{e}_1u_1\times\cdots\times\mathbf{e}_nu_n$,
$\mathbf{e}u(m)=\mathbf{e}_1u_1(m)\times\cdots\times\mathbf{e}_nu_n(m)$,
$\mathbf{e}u(..m)=\mathbf{e}_1u_1(..m)\times\cdots\times\mathbf{e}_nu_n(..m)$.

## 1.2. Capacity and Coding for Multiple Access Channels

Interest in multiple access channels (and other types of multi-user channels) goes back to Shannon's 1961 paper [1]. Since the publication of that paper considerable theoretical work has been done about such channels. This section presents two well-known results about multiple access channels which provide the motivation and the framework for the work reported in this thesis. To keep the notation simple, the discussion is limited to the two-user case.

Two-User Block Coding

A $(M_1,M_2,k)$ block code for a two-user channel with input alphabets $X_1$ and $X_2$ is a mapping

$$f:\{1,\ldots,M_1\}\times\{1,\ldots,M_2\} \longrightarrow (X_1\times X_2)^k$$

which has the property that, for each $(i,j)\epsilon\{1,\ldots,M_1\}\times\{1,\ldots,M_2\}$,

$$f(i,j)=f_1(i)\times f_2(j),$$

for some pair of functions $f_1$ and $f_2$ such that

$$f_1:\{1,\ldots,M_1\} \longrightarrow X_1{}^k,$$

$$f_2:\{1,\ldots,M_2\} \longrightarrow X_2{}^k.$$

The operation × is as defined in §1.1.

The above definition forces a two-user block code f to be decomposable into two component block codes $f_1$ and $f_2$. This reflects the requirement that in a two-user channel the channel inputs must be independently encoded.

The implementation of a block code f, with component codes $f_1$ and $f_2$, on a channel $K=(P;X_1,X_2;Y)$ results in the following functional relationships.

$x_1(m)=f_1(s_1(m))$, $\qquad$ $x_2(m)=f_2(s_2(m))$, $\qquad$ $x(m)=f(s(m))$.

Any function g, $g:Y^k \longrightarrow \{1,\ldots,M_1\}\times\{1,..,M_2\}$, can be used as a decoder for the above block code by simply letting $z(m)=g(y(m))$.

An error is said to occur in the decoding of $s(m)$ if $s(m)\neq z(m)$. Under our assumption that the source output letters are independent and uniformly distributed, the probability of $s(m)\neq z(m)$ is independent of m; it equals

$$P_e(f,g) = \sum_{i=1}^{M_1} (1/M_1) \sum_{j=1}^{M_2} (1/M_2) \sum_{\eta\epsilon Y^k:g(\eta)\neq(i,j)} P(\eta\,|\,f(i,j)).$$

$P_e(f,g)$ is minimized if g has the property that, for each $\eta\epsilon Y^k$, $g(\eta)=(i,j)$ only if $P(\eta\,|\,f(i,j))\geq P(\eta\,|\,f(h,m))$ for all $(h,m)\epsilon\{1,\ldots,M_1\}\times\{1,\ldots,M_2\}$. Such a decoder is called a <u>maximum-likelihood</u> (ML) decoder. The way ties are broken in ML decoding does not affect the probability of decoding error; so, we denote the probability of error for ML decoders by $P_e(f)$.

## <u>Capacity Region</u>

The capacity region **C**(K) of a two-user channel $K=(P;X_1,X_2;Y)$ is defined as the closure of the following region.

$$C(K)=\text{convex-hull} \bigcup_{Q_1,Q_2} C(Q_1,Q_2)$$

where the union is over all $Q_1$ and $Q_2$ such that $Q_1$ is a probability distribution (p.d.) on $X_1$ and $Q_2$ is a p.d. on $X_2$; and $C(Q_1,Q_2)$ is defined as the set of points $(R_1,R_2)$ such that

$$0 \leq R_1 < \sum_{\xi_1\epsilon X_1} Q_1(\xi_1) \sum_{\xi_2\epsilon X_2} Q_2(\xi_2) \sum_{\eta\epsilon Y} P(\eta\,|\,\xi_1,\xi_2) \ln \frac{P(\eta\,|\,\xi_1,\xi_2)}{\sum_{\zeta\epsilon X_2} Q_2(\zeta)P(\eta\,|\,\xi_1,\zeta)},$$

$$0 \le R_2 < \sum_{\xi_1\in X_1} Q_1(\xi_1) \sum_{\xi_2\in X_2} Q_2(\xi_2) \sum_{\eta\in Y} P(\eta|\xi_1,\xi_2)\ln \frac{P(\eta|\xi_1,\xi_2)}{\sum_{\zeta\in X_1} Q_1(\zeta)P(\eta|\zeta,\xi_2)},$$

$$R_1+R_2 < \sum_{\xi_1\in X_1} Q_1(\xi_1) \sum_{\xi_2\in X_2} Q_2(\xi_2) \sum_{\eta\in Y} P(\eta|\xi_1,\xi_2)\ln \frac{P(\eta|\xi_1,\xi_2)}{\sum_{\zeta_1\in X_1} Q_1(\zeta_1)\sum_{\zeta_2\in X_2} Q_2(\zeta)P(\eta|\zeta_1,\zeta_2)}.$$

**Theorem 1.2.1.** (Ahlswede [2], Liao [3])
For any two-user channel $K=(P;X_1,X_2;Y)$ and any pair of real numbers $R_1$ and $R_2$, we have:

i) If $(R_1,R_2)\epsilon C(K)$, then, for any $\epsilon>0$, there exists a $(M_1,M_2,k)$ block code f such that $P_e(f)<\epsilon$ and $(1/k)\ln M_i \ge R_i$, $i=1,2$.

ii) If $(R_1,R_2)$ lies outside $\mathbf{C}(K)$, then $P_e(f,g)$ is bounded away from zero for all f and g, so long as $(M_1,M_2,k)$, the parameter of f, is such that $(1/k)\ln M_i \ge R_i$, $i=1,2$. □

In words, Theorem 1.2.1 states that, for any channel K, i) communication with arbitrarily low probability of error is possible if the source rates lie in C(K), and ii) probability of error can not be made arbitrarily small (i.e., reliable communication is not possible) if the source rates lie outside **C**(K). The theorem does not assert anything about points which belong to **C**(K) but not to C(K).

**Example 1.2.1.**
To illustrate the capacity theorem and to explain certain approaches to multi-access communications, we now discuss the two-user erasure channel (TEC) of Figure 1.2.1. We observe from the figure and by the channel capacity theorem that sum rates, $R_1+R_2$, of up to 1.5 bits are achievable (with arbitrarily small probability of error) by using block codes.

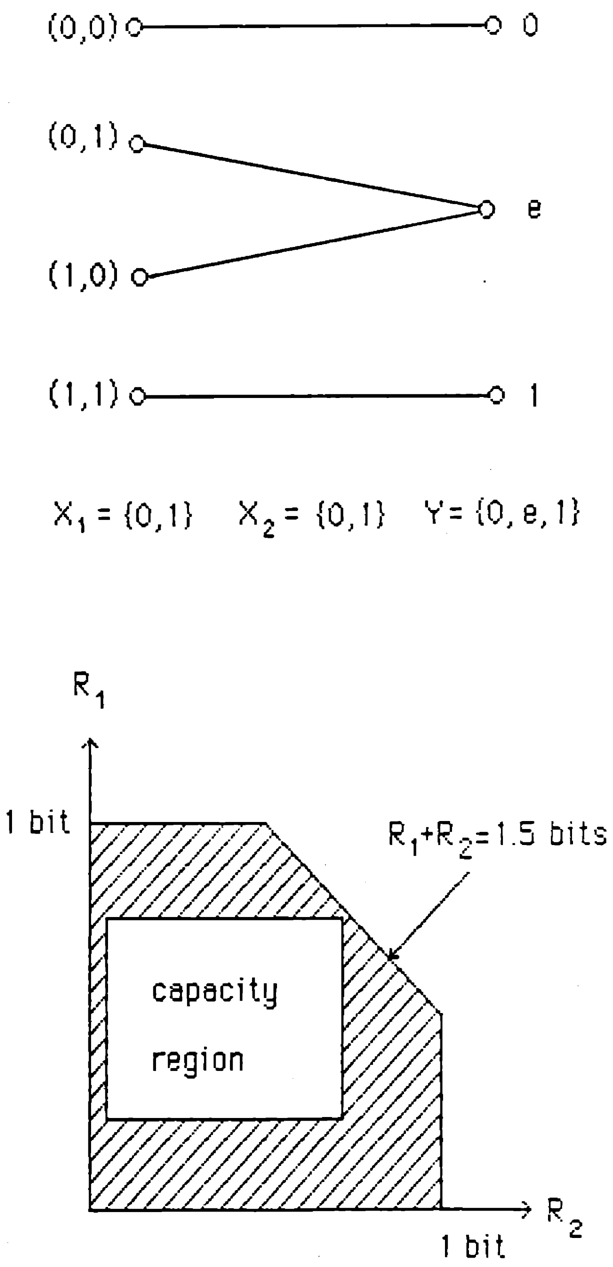


Figure 1.2.1. Two-user erasure channel and its capacity region.

Let us look at some simple block codes for this channel. It is easy to see that the following code achieves the rate pair (0.5 bits, 0.5 bits) with zero probability of error.

Code 1.

| User 1 | | User 2 | |
|---|---|---|---|
| Message | Codeword | Message | Codeword |
| 1 | 00 | 1 | 00 |
| 2 | 01 | 2 | 10 |

In this code, the first user sends no information in the first digit of a codeword (it always transmits a 0); similarly, the second user is "quiet" in the second digit of each codeword. For this reason, this code is said to have no multi-user interference: user 1's message can be estimated independently of user 2's message without any loss of optimality. Thus, elimination of multi-user interference simplifies decoding, but codes without multi-user interference are limited to sum rates of at most 1 bit in the case of the TEC, which is significantly below the theoretically possible 1.5 bits.

Code 1 is typical of a class of straightforward approaches to multiple access communications, such as time division multiplexing, frequency division multiplexing, and the like, which are based on the idea of splitting the channel into non-interfering subchannels and giving the use of each subchannel exclusively to a single user. The main advantage of these approaches is the ease of decoding, but as here, their operation is often restricted to a small portion of the capacity region. Coding for multiple access channels aims, at the very least, at finding practical techniques for achieving rates beyond what is achievable by such simple schemes.

One can easily improve upon Code 1; for example, Kasami and Lin [4] give the following code, which achieves a sum rate of $0.5+(1/2)\log_2 3 \approx 1.3$ bits.

Code 2.

| User 1 | | User 2 | |
|---|---|---|---|
| Message | Codeword | Message | Codeword |
| 1 | 00 | 1 | 01 |
| 2 | 11 | 2 | 10 |
| | | 3 | 11 |

In this code, unlike the previous one, both users transmit information in both digits of each codeword; as a result, each received digit is corrupted by multi-user interference. Hence, if optimality is desired, the decoder must deal with the codes of both users simultaneously. So, an increase in the rates comes at the cost of increased decoding complexity. As a general rule, allowing the users to interfere with each other requires untangling a more complicated set of possibilities at the decoder, hence, an increased decoding complexity.

If we wish to communicate at still higher sum rates, and at the same time keep the probability of error below a given level, we find out that codes with longer block lengths must be considered. The channel capacity theorem does not tell us how large the block length has to be before we can be sure that there exists a block code with that block length which satisfies our rate and reliability requirements; the following theorem provides an answer to this question.

**Theorem 1.2.2.** (Slepian and Wolf [5])
For any two-user channel K, there exists a function $E_K(R_1,R_2)$ which has the following properties. 1) $E_K(R_1,R_2)$ is positive if $(R_1,R_2)\epsilon C(K)$ and zero otherwise. 2) For any $(R_1,R_2)$, there exists a block code f with parameter $(M_1,M_2,k)$ such that a) $(1/k)\ln M_i \geq R_i$ for $i=1,2$ and b) $P_e(f) \leq \exp -kE_K(R_1,R_2)$.

For the purposes of our discussion, the explicit form of $E_K(R_1,R_2)$ is not important. The important point is that, for any given rate in C(K), this theorem establishes the possibility of making the probability of decoding error at that rate approach zero exponentially by increasing the block length. This suggests a favorable trade-off between reliability and system complexity, as long as the desired rate is in C(K). A more complete discussion of this issue lies outside the scope of this thesis. For that the interested reader is referred to [6], which covers all the material given up to here in greater detail and from a broader perspective, and also gives an overview of several approaches to coding for multiple access channels, which we will not discuss at all.

## 1.3. Multi-User Tree Codes

A multi-user tree code is simply another name for the joint encoding operation described in §1.1. The name derives from the fact that the mapping generated by causal encoders with long memory is most easily visualized as a tree. This section starts by considering a single-user tree code to introduce the basic terminology and concepts; then a two-user tree code is considered; next the form of the concepts and the notation for an arbitrary number of users is indicated; and, finally, random tree code ensembles are introduced.

### Single-User Tree Codes

As in the case of encoders, a single-user tree code with parameter (M,k) has an input alphabet of size M and, for each source digit accepted, it generates k channel digits. The rate of such a tree code is defined as $(1/k)\ln M$ (nats) or, equivalently, as $(1/k)\log_2 M$ (bits).

As an example, consider a (2,2) tree code for which the source and the channel alphabets are both equal to {0,1} and the encoding operation **e** is defined as follows.

$$\mathbf{e}(u(..m)) = \begin{cases} (u(1),u(1)) & \text{for } m=1; \\ (u(m-1)+u(m),u(m)) & \text{for } m=2,3,... \end{cases}$$

Here, + denotes modulo 2 addition, and u denotes an arbitrary source sequence.

The first three levels of the code tree for **e** are shown in Figure 1.3.1. The tree representation is based on establishing a one-to-one mapping from source sequences to paths in the tree. In the present example, the mapping is indicated by the arrows at the left side of the diagram. In order to generate the encoded sequence, the encoder uses the source output as a sequence of instructions and follows the "upper" or the "lower" branch going out from the current node depending on whether the next source digit is, respectively, a 0 or a 1.

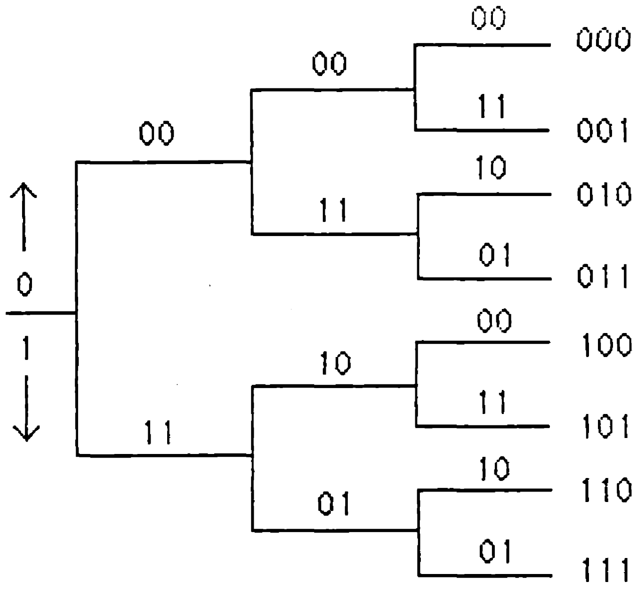


Figure 1.3.1. Example of a single-user tree code.

For example, if the first three digits of the source output are 0,1,0, then the first three blocks (branches) of the encoded sequence are 00,11,10. Thus, each source sequence is mapped to a unique path. Hence, we refer to source sequences as paths and to initial segments of source sequences as nodes. For any path u, and any m=1,2,..., the branch connecting node u(..m-1) (for m=1, take u(..m-1) as the origin) to node u(..m) is labelled by **e**(u(..m)).

In the tree representation of a (M,k) tree code, each node at each level is connected to M nodes at the next higher level; each branch is labelled by a block of k channel input digits; M is referred to as the degree of the tree.

The path corresponding to s, the actual source sequence, is called the correct path. Nodes on the correct path are called the correct nodes. The branch labels on the correct path are thus the channel symbols that get transmitted over the channel.

Two-User Tree Codes

We illustrate the relationship between a pair of single-user tree codes, $\mathbf{e}_1$ and $\mathbf{e}_2$, and the corresponding joint two-user tree code, $\mathbf{e}$, by using the example shown in Figure 1.3.2. We observe that the parameters of $\mathbf{e}_1$ and $\mathbf{e}_2$ are both equal to (2,2). In general, if $(M_1,k)$ and $(M_2,k)$ are the parameters of $\mathbf{e}_1$ and $\mathbf{e}_2$, then $(M_1,M_2,k)$ is the parameter of $\mathbf{e}$. So, here, the parameter of $\mathbf{e}$ is (2,2,2).

With reference to Figure 1.3.2, observe that, for each pair of nodes, $u_1(..m)$ in $\mathbf{e}_1$ and $u_2(..m)$ in $\mathbf{e}_2$, $u_1(..m) \times u_2(..m)$ is a node in $\mathbf{e}$. Likewise, for each pair of paths, $u_1$ in $\mathbf{e}_1$ and $u_2$ in $\mathbf{e}_2$, $u_1 \times u_2$ is a path in $\mathbf{e}$.

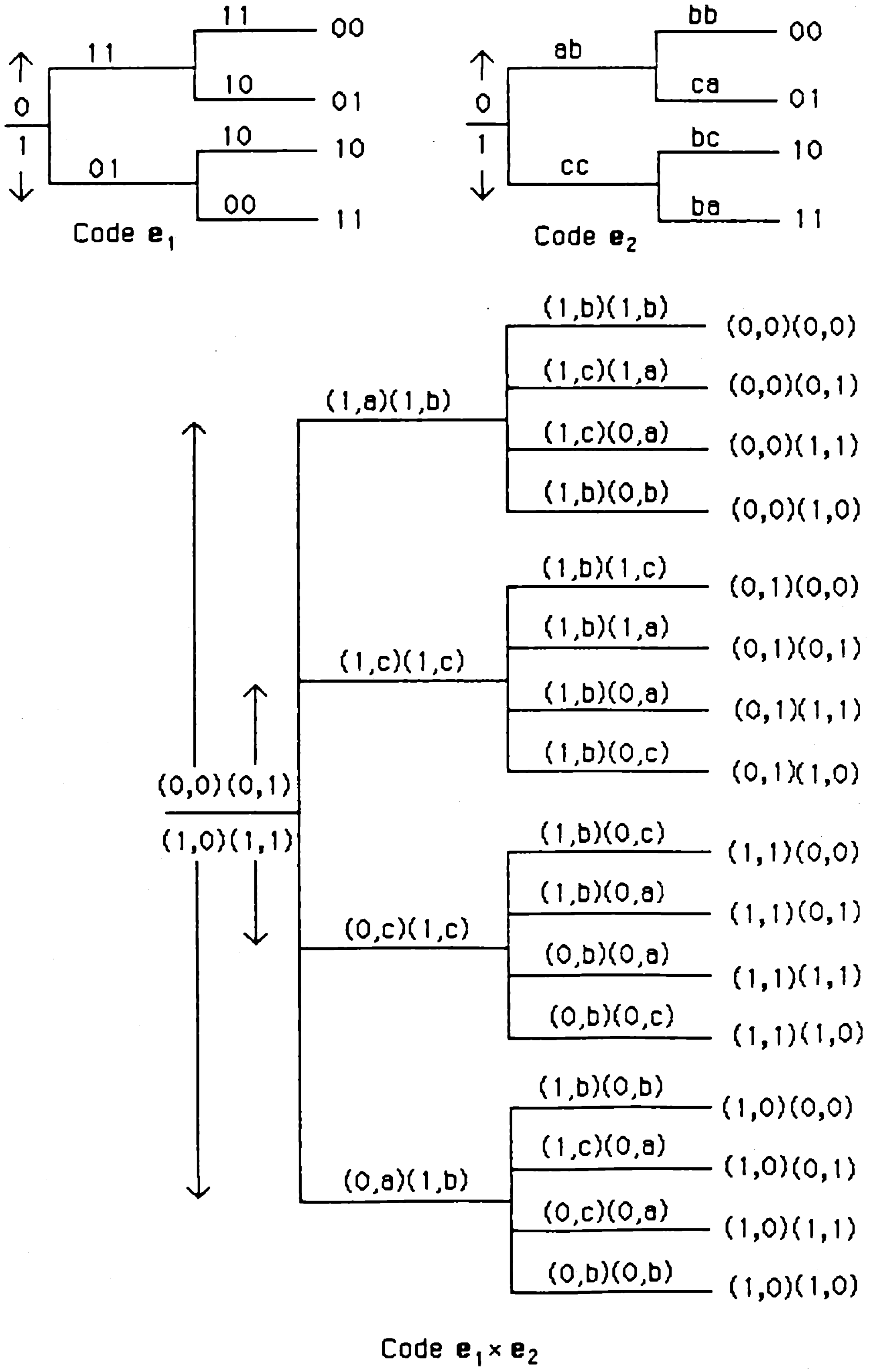
The path $s=s_1 \times s_2$, where $s_1$ is the correct path in $\mathbf{e}_1$ and $s_2$ is the correct path in $\mathbf{e}_2$, is called the joint correct path, or the correct path in $\mathbf{e}$.

Basic Concepts and Notation for Multi-User Tree Codes

Generically, $\mathbf{e}_i$ denotes the tree code for user i, and $\mathbf{e}$ denotes the joint tree code. $(M_i,k)$ denotes the parameter of $\mathbf{e}_i$; n denotes the number of users; and $(M_1,\ldots,M_n,k)$ denotes the parameter of $\mathbf{e}$. The rate of $\mathbf{e}_i$ is defined as $R_i=(1/k)\ln M_i$, and that of $\mathbf{e}$ as $(R_1,\ldots,R_n)$.

If $u_i$ is a path in $\mathbf{e}_i$ for each $i \epsilon \{1,\ldots,n\}$, then $u_1 \times \cdots \times u_n$ is a path in $\mathbf{e}$. It is called the product path or the joint path corresponding to $u_1,\ldots,u_n$. $u_i$ is said to be a component path of $u_1 \times \cdots \times u_n$.

The path in $\mathbf{e}_i$ corresponding to $s_i$, the actual source output, is called the correct path in $\mathbf{e}_i$; $s_1 \times \cdots \times s_n$ is called the correct path in $\mathbf{e}$, or the joint correct path. Nodes on $s_1 \times \cdots \times s_n$ are called correct nodes.



Code $\mathbf{e}_1$

Code $\mathbf{e}_2$

Code $\mathbf{e}_1 \times \mathbf{e}_2$

Figure 1.3.2. Example of a two-user tree code.

If $u_i(..m)$ is a node in $\mathbf{e}_i$ for each $i \in \{1,...,n\}$, then $u_1(..m) \times \cdots \times u_n(..m)$ is a node in $\mathbf{e}$. It is called the joint node or the product node corresponding to $u_1(..m),....,u_n(..m)$. $u_i(..m)$ is said to be a component node of $u_1(..m) \times \cdots \times u_n(..m)$.

For any pair of nodes in $\mathbf{e}$, $u(..m)=u_1(..m) \times \cdots \times u_n(..m)$ and $\tilde{u}(..m)=\tilde{u}_1(..m) \times \cdots \times \tilde{u}_n(..m)$, the type of $u(..m)$ with respect to $\tilde{u}(..m)$ is defined as the vector $(T_1,...,T_m)$ where $T_j$, $1 \le j \le m$, is the set of $i$ such that $u_i(..j) \ne \tilde{u}_i(..j)$. (For example, in Figure 1.3.2, the type of node $((1,1),(1,0))$ with respect to $((1,1),(0,0))$ is $(\phi,\{1\})$.)

For any node $u(..m)$ and any path $\tilde{u}$ in $\mathbf{e}$, the type of $u(..m)$ with respect to $\tilde{u}$ is defined as the type of $u(..m)$ with respect to $\tilde{u}(..m)$.

For any path $u$ in $\mathbf{e}$, the $m^{th}$ $(m \ge 1)$ incorrect subtree of $u$, denoted by $I_m(u)$, is defined as the set of nodes $\tilde{u}(..j)$ in $\mathbf{e}$ such that a) $j \ge m$, b) $\tilde{u}(..m) \ne u(..m)$, and c) if $m \ge 2$, $\tilde{u}(..m-1)=u(..m-1)$.

The number of types of nodes at level $m$ equals $(m+1)^n$. This can be seen by observing that, if $(T_1,...,T_m)$ is the type of a node, $T_j$ must be a subset of $T_h$ for all $h>j$. Thus, for each user, there are $m+1$ ways that that user first appears (one possibility is that it never appears) in the sequence of sets $T_1,...,T_m$.

## Ensembles of Tree Codes

We end this section by introducing a certain type of tree code ensembles, which will be used mainly for proving theorems.

For any parameter $(M,k)$, any channel input alphabet $X$, and any p.d. $Q$ on $X$, the single-user tree code ensemble $Ens(M;k;X;Q)$ is a set of tree codes $\Omega(M,k,X)$ with a probability measure $\mu$ on it. $\Omega(M,k,X)$ is the set of all

(M,k) tree codes with channel input alphabet X. $\mu$ is a measure defined on the class of events that are expressable as countable unions and intersections (the $\sigma$-algebra) of elementary events of the form

$$E(u(..i),\xi)=\left\{\mathbf{e}\epsilon\Omega(M,k,X):\ \mathbf{e}(u(..i))=\xi\epsilon X^k\right\}.$$

$E(u(..i),\xi)$ is the set of tree codes in $\Omega(M,k,X)$ for which $\xi$ is the label of the branch immediately preceding node u(..i). $\mu$ is the extension measure corresponding to the following probability assignment: For any collection of distinct nodes $u_1(..m_1),...,u_r(..m_r)$ and any $\xi_1,...,\xi_r\epsilon X^k$,

$$Pr\left\{E(u_1(..m_1),\xi_1),....,E(u_r(..m_r),\xi_r)\right\} = Q(\xi_1)\cdots Q(\xi_r).$$

Thus, the statistical properties of a code chosen at random according to $\mu$ coincides with those of a (M,k) tree code each of whose branches gets a label $\xi$, $\xi\epsilon X^k$, with probability $Q(\xi)$, independently of what is assigned to other branches.

For any n-user parameter $(M_1,...,M_n,k)$, any collection of channel input alphabets $X_1,....,X_n$, and any collection of $Q_1,...,Q_n$, where $Q_i$ is a p.d. on $X_i$, the <u>n-user tree code ensemble</u> $Ens(M_1,...,M_n;k;X_1,...,X_n;Q_1,...,Q_n)$ is defined as the set of all $(M_1,...,M_n,k)$ tree codes for which $X_i$ is user i's channel input alhabet, with the following probability measure $\mu$ on this set. $\mu$ is best described by saying that it is the measure that would exist on the joint tree code **e** corresponding to a collection of random, mutually independent tree codes $\mathbf{e}_1,...,\mathbf{e}_n$, where $\mathbf{e}_i$ is selected according to the probability measure associated with $Ens(M_i;k;X_i;Q_i)$. In other words, the statistical properties of a code chosen at random according to $\mu$ are identical to those of a joint tree code in the situation where each branch of each user's tree code is labelled independently of each other branch, in such a way that $Q_i$ is the p.d. for branch labels in user i's tree code, i=1,...,n.

## 1.4. Sequential Decoding for Multi-User Tree Codes

Sequential decoding is a decoding algorithm for tree codes invented by Wozencraft [7], and later developed by Fano [8]. This section describes the stack algorithm, a version of sequential decoding due to Zigangirov [9] and Jelinek [10], and defines the concept of achievability for sequential decoding. Familiarity with sequential decoding, to the extent that it is given in any one of the references [11], [12], and [13], is assumed.

### Sequential Decoding and Its Metric

Sequential decoding is a tree search algorithm for finding the correct path in a code tree based on the information available from the received sequence. The algorithm relies on what is called a metric for directing its search. The metric in sequential decoding is not a metric in the usual mathematical sense of the word. Ordinarily, the metric is intended to be a function that measures the statistical correlation between the received sequence and the hypothesized transmitted sequence.

Formally, a metric for a channel $K=(P;X_1,\ldots,X_n;Y)$ and a $(M_1,\ldots,M_n,k)$ tree code $\mathbf{e}$ is any function of the form

$$\Gamma : \bigcup_{h=1}^{\infty} (X_1\times\cdots\times X_n)^{hk}\times Y^{hk} \longrightarrow [-\infty,+\infty).$$

The value of the metric at a node $u(..m)$ for a received sequence $y$ is given by $\Gamma(\mathbf{e}u(..m),y(..m))$, where the notation is as given in §1.1.

It is important to note that $\Gamma(\mathbf{e}u(..m),y(..m))$ does not depend on $y(m+1)$, $y(m+2),....$, the portion of the received sequence beyond level m. This restriction is an integral part of sequential decoding; and without it, some results of this thesis would not hold.

Also notice that the metric is allowed to take on the value $-\infty$. As will be clear soon, this makes it possible to rule out a node permanently from further consideration when there is no doubt that it is incorrect.

**Example 1.4.1.** The Fano Metric

The most well-known metric for sequential decoding is the Fano metric, which was originally introduced by Fano for single-user channels [8]. In the case of an n-user channel $K=(P;X_1,...,X_n;Y)$ and a $(M_1,...,M_n,k)$ tree code **e**, the Fano metric takes the following form.

$$\Gamma(\mathbf{e}u(..m),y(..m)) = \sum_{h=1}^{m} \{\ln \frac{P(y(h)|\mathbf{e}u(h))}{\omega(y(h))} - kR\},$$

where $\omega$ is a p.d. on $Y^k$ and $R=(1/k)\sum_{i=1}^{n} \ln M_i$.

In practice, one might pick **e** at random according to the probability measure associated with an ensemble $Ens(M_1,...,M_n;k;X_1,...,X_n;Q_1,...,Q_n)$ and set

$$\omega(\eta) = \sum_{\xi_1 \epsilon X_1{}^k} Q_1(\xi_1) \cdots \sum_{\xi_n \epsilon X_n{}^k} Q_n(\xi_n) P(\eta | \xi_1,...,\xi_n)$$

for each $\eta \epsilon Y^k$.

The Fano metric is branchwise additive; that is,

$$\Gamma(\mathbf{e}u(..m),y(..m)) = \Gamma(\mathbf{e}u(..m-1),y(..m-1)) + \gamma(\mathbf{e}u(m),y(m)),$$

where $\gamma(\mathbf{e}u(m),y(m)) = \ln \frac{P(y(m)|\mathbf{e}u(m))}{\omega(y(m))} - kR.$

Branchwise additive metrics are simpler to implement and easier to analyze; but these are not compelling reasons to restrict our discussion to this class of metrics, and we do not do so.

The Stack Algorithm

There are two well-known versions of sequential decoding, namely, the

Fano algorithm and the stack algorithm. For practical purposes, the Fano algorithm is preferable since it requires almost no storage. However, in this thesis, we shall consider only the stack algorithm, mainly because it is much simpler to describe and analyze. Let us point out that the results of our analyses hold for the Fano algorithm without any essential changes.

In the stack algorithm, there is a list of nodes in which nodes are ordered with respect to their metric values. This list is referred to as the stack. The metric values of the nodes in the stack increase towards the top of the stack. Ties between the metric values in the ordering of nodes are broken by some fixed but arbitrary rule. Each step of the stack algorithm consists of deleting the node at the stack-top and inserting its immediate descendants into the stack. At the start of the algorithm, the origin is the only node in the stack, and it has a metric value of zero.

In practice, all tree codes are truncated at some finite level, and the stack algorithm stops as soon as a node at the last level of the code tree reaches the stack-top. The stack-top-node is then taken as the output of the sequential decoder. If the rate is sufficiently small, reliability of the decoder output can be improved by increasing the length of the finite tree code. The remarkable point about sequential decoding is the possibility of making the average decoding complexity independent of the length of the tree code, and thus, of the desired level of reliability.

The following definitions formalize the concept of decoding complexity.

**Definition 1.4.1.** A Measure of Decoding Complexity

If the stack algorithm is used, with $\Gamma$ as its metric, in decoding a tree code $\mathbf{e}$ over a channel K, then $C_j(K,\mathbf{e},\Gamma,s,y)$ denotes the number of nodes in $I_j(s)$, the $j^{th}$ incorrect subset of the correct path, which reach the stack-top, conditional on s being the correct path and y being the received sequence.

$C_j(K,\mathbf{e},\Gamma)$ denotes the expected value of $C_j(K,\mathbf{e},\Gamma,s,y)$ with respect to the joint p.d. on s and y. That is, $C_j(K,\mathbf{e},\Gamma)=E_s E_{y|\mathbf{es}} C_j(K,\mathbf{e},\Gamma,s,y)$ where $E_s$ denotes expectation with respect to the p.d. on s and $E_{y|\mathbf{es}}$ denotes expectation with respect to the p.d. on y conditional on **es** being the transmitted sequence.

For each L, $D_L(K,\mathbf{e},\Gamma)$ is defined to be $(C_1(K,\mathbf{e},\Gamma)+\cdots+C_L(K,\mathbf{e},\Gamma))/L$. □

Observe that $LD_L(K,\mathbf{e},\Gamma)$ is an upper bound on the expected number of nodes which reach the stack-top before the algorithm reaches level L on the correct path for the first time. Hence, for large L, $D_L$ can be taken as an approximate measure of the average number of computations for the algorithm to move one step along the correct path. These considerations motivate the following definition.

**Definition 1.4.2.** A Criterion of Applicability

A point $R=(R_1,\ldots,R_n)$ is said to be an achievable rate for sequential decoding on a channel $K=(P;X_1,\ldots,X_n;Y)$ if

1) $R_i \geq 0$ for each $i=1,\ldots,n$, and

2) there exists a finite constant A, $A=A(K,R)$, such that, for every L, there exist

i) a code **e** with rate at least as large as R

and ii) a metric $\Gamma$

such that $D_L(K,\mathbf{e},\Gamma)<A$.

(Condition i) above means that, if $(M_1,\ldots,M_n,k)$ is the parameter of **e**, then $(1/k)\ln M_i \geq R_i$ for each $i=1,\ldots,n$.)

The closure of the set of all such R is called the achievable rate region of sequential decoding and is denoted by $\mathbf{R}(K)$.□

The above definition of achievability allows **e** and $\Gamma$ to depend on L. Now, one may ask, quite justifiably, why the definition of achievability does not read as follows.

**Definition 1.4.3.** An Alternative Criterion of Applicability
A point $R=(R_1,\ldots,R_n)$ is said to be a strongly achievable rate for sequential decoding on a channel $K=(P;X_1,\ldots,X_n;Y)$ if

1) $R_i \geq 0$ for each $i=1,\ldots,n$, and

2) there exists a finite constant $A=A(K,R)$ such that there exist
   i) a code **e** with rate at least as large as R
   and ii) a metric $\Gamma$
   such that $D_L(K,\mathbf{e},\Gamma)<A$ for all L.□

Unlike Definition 1.4.2, Definition 1.4.3 requires that **e** and $\Gamma$ be chosen independently of L. Clearly, if R is achievable in the sense of Def. 1.4.3, then R is also achievable in the sense of Def. 1.4.2.

The concept of achievability used in the literature on sequential decoding coincides with that of Def. 1.4.2. It is not known if strong achievability and achievability are equivalent, even for the single-user case. (Resolving this question might contribute greatly to our understanding of sequential decoding.) Strong achievability is not used anywhere in this thesis for the following reasons. First, despite some efforts, we have not been able to prove that any non-trivial rate is strongly achievable. Second, for finite tree codes, which are the only type of tree codes of practical interest, strong achievability is unnecessarily restrictive.

To illustrate that achievability in the sense of Def. 1.4.2 is sufficient for practical purposes, consider a situation where the desired rate and the desired level of reliability are given. Suppose that the desired rate is achievable. Then, given any L, there exists an infinite tree code **e** with the desired rate and a metric $\Gamma$ such that $D_L(K,\mathbf{e},\Gamma)<A$, where A is a finite constant, independent of **e**, L, and $\Gamma$. The idea is to pick L large enough so that, among those code-metric pairs satisfying $D_L(K,\mathbf{e},\Gamma)<A$, there exist **e** and $\Gamma$ such that: When the stack algorithm is applied, with $\Gamma$ as its metric, to the finite tree code that is obtained by truncating **e** at level L, the desired reliability is also satisfied. A tail, i.e. a part where no branching occurs, may be appended to the truncated code in order to increase the reliability of the final digits of the decoded sequence.

## 1.5. Summary of Results

The research reported in this thesis has been aimed mainly at finding a characterization of **R**, the achievable rate region of sequential decoding. This goal has not been achieved and no general characterization of **R** is known at present; there are, however, some partial results, which we now summarize.

The Result on Achievability

The following theorem is the main result of this thesis on achievability. For notational simplicity, it is stated here for the two-user case. In Chapter 2, it is restated and proved for an arbitrary number of users.

**Theorem 2.2.1.**
For any two-user channel $K=(P;X_1,X_2;Y)$, $\mathbf{R}(K)$ is inner-bounded by $\mathbf{R}_0(K)$, which is defined as follows.

$$\mathbf{R}_0(K) = \bigcup_{\mathbf{Q}} \mathbf{R}_0(K,\mathbf{Q})$$

where the union is over all $\mathbf{Q}=(\mathbf{Q}_1,\mathbf{Q}_2)$ such that $\mathbf{Q}_1$ is a p.d. on $X_1^k$ and $\mathbf{Q}_2$ is a p.d. on $X_2^k$ for some arbitrary integer k (same k for both $\mathbf{Q}_1$ and $\mathbf{Q}_2$); and for any such $\mathbf{Q}$, $\mathbf{R}_0(K,\mathbf{Q})$ is defined as the set of all $(R_1,R_2)$ such that

$$0 \le R_1 \le -(1/k)\ln \sum_{\boldsymbol{\xi}_2\epsilon X_2^k} \mathbf{Q}_2(\boldsymbol{\xi}_2) \sum_{\boldsymbol{\eta}\epsilon Y^k} \left\{ \sum_{\boldsymbol{\xi}_1\epsilon X_1^k} \mathbf{Q}_1(\boldsymbol{\xi}_1)\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)} \right\}^2,$$

$$0 \le R_2 \le -(1/k)\ln \sum_{\boldsymbol{\xi}_1\epsilon X_1^k} \mathbf{Q}_1(\boldsymbol{\xi}_1) \sum_{\boldsymbol{\eta}\epsilon Y^k} \left\{ \sum_{\boldsymbol{\xi}_2\epsilon X_2^k} \mathbf{Q}_2(\boldsymbol{\xi}_2)\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)} \right\}^2,$$

$$R_1+R_2 \le -(1/k)\ln \sum_{\boldsymbol{\eta}\epsilon Y^k} \left\{ \sum_{\boldsymbol{\xi}_2\epsilon X_2^k} \mathbf{Q}_2(\boldsymbol{\xi}_2) \sum_{\boldsymbol{\xi}_1\epsilon X_1^k} \mathbf{Q}_1(\boldsymbol{\xi}_1)\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)} \right\}^2.$$

This theorem is proved by showing that $\mathbf{R}_0$ is achievable by the following class of metrics: Members of the class are identified by a parameter $(K,k,\mathbf{Q},B)$ where K is a channel, say $K=(P;X_1,X_2;Y)$; k is a positive integer; $\mathbf{Q}=(\mathbf{Q}_1,\mathbf{Q}_2)$ where $\mathbf{Q}_1$ is a p.d. on $X_1{}^k$ and $\mathbf{Q}_2$ is a p.d. on $X_2{}^k$; and $B=(B_1,B_2,B_3)$ is what is called the bias function. The member of the class with parameter $(K,k,\mathbf{Q},B)$ is based on a branch metric

$$\gamma : (X_1 \times X_2)^k \times Y^k \longrightarrow [-\infty,+\infty),$$

such that, for each $\boldsymbol{\eta} \epsilon Y^k$ and $\boldsymbol{\xi}=\boldsymbol{\xi}_1 \times \boldsymbol{\xi}_2$, where $\boldsymbol{\xi}_1 \epsilon X_1{}^k$, $\boldsymbol{\xi}_2 \epsilon X_2{}^k$,

$$\gamma(\boldsymbol{\xi},\boldsymbol{\eta})=\min\{\gamma_1(\boldsymbol{\xi},\boldsymbol{\eta}),\ \gamma_2(\boldsymbol{\xi},\boldsymbol{\eta}),\ \gamma_3(\boldsymbol{\xi},\boldsymbol{\eta})\},$$

where

$$\gamma_1(\boldsymbol{\xi},\boldsymbol{\eta}) = \ln \frac{\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}}{\sum\limits_{\boldsymbol{\zeta} \epsilon X_2{}^k} \mathbf{Q}_2(\boldsymbol{\zeta})\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_1,\boldsymbol{\zeta})}} - kB_1,$$

$$\gamma_2(\boldsymbol{\xi},\boldsymbol{\eta}) = \ln \frac{\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}}{\sum\limits_{\boldsymbol{\zeta} \epsilon X_1{}^k} \mathbf{Q}_1(\boldsymbol{\zeta})\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\zeta},\boldsymbol{\xi}_2)}} - kB_2, \text{ and}$$

$$\gamma_3(\boldsymbol{\xi},\boldsymbol{\eta}) = \ln \frac{\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}}{\sum\limits_{\boldsymbol{\zeta}_1 \epsilon X_1{}^k} \mathbf{Q}_1(\boldsymbol{\zeta}_1) \sum\limits_{\boldsymbol{\zeta}_2 \epsilon X_2{}^k} \mathbf{Q}_2(\boldsymbol{\zeta}_2)\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)}} - kB_3.$$

Here, $\mathbf{P}$ is the transition probability of K over blocks of length k. (We use boldface characters to indicate quantities relating to blocks.) $\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_1,\boldsymbol{\zeta})$ is the probability that $\boldsymbol{\eta}$ is received given that user 1 transmits $\boldsymbol{\xi}_1$ and user 2 transmits $\boldsymbol{\zeta}$.

A full intuitive account of the above metric cannot be given at this point, because the form of the metric itself is closely related to the method we use in §2.1 to prove that a given rate is achievable.

This metric is the only metric known to achieve $\mathbf{R}_0(K)$ for all K. Our efforts to show that the Fano metric (or simple modifications of it) achieves $\mathbf{R}_0$ have not been successful. In view of this, we regard the introduction of the above metric as a major contribution of this thesis.

Converse Results

Converse arguments aim at finding outer bounds to the achievable rate region of sequential decoding. The main converse results of this thesis are as follows.

**Theorem 3.2.1.** For any single-user channel K, $\mathbf{R}(K)=\mathbf{R}_0(K)$. □

For single-user channels, $\mathbf{R}_0(K)=[0,R_0(K)]$ (see §2.3 or pp.149-50 of [12]), where

$$R_0(K) = \max_{Q} -\ln \sum_{\eta\epsilon Y} \left\{ \sum_{\xi\epsilon X} Q(\xi)\sqrt{P(\eta\,|\,\xi)}\right\}^2 ,$$

where the maximum is taken over all p.d.'s Q on X.

The achievability of all R, for $R\epsilon[0,R_0(K))$, is a special case of Theorem 2.2.1 and it has been well-known, see, e.g., [11], [12], or [13]. But the converse statement, that rates greater than $R_0(K)$ are not achievable, is new and will be proved in §3.2.

The strongest converse prior to this was due to Jacobs and Berlekamp [14], which stated that rates in excess of $\hat{E}_0(K,1)$ are not achievable. Here, $\hat{E}_0(K,1)$ is the value, at $\rho=1$, of $\hat{E}_0(K,\rho)$, which Jacobs and Berlekamp defined as the smallest concave function greater than or equal to

$$E_0(K,\rho) = \max_{Q} -\ln \sum_{\eta\epsilon Y} \left\{ \sum_{\xi\epsilon X} Q(\xi)\, P(\eta\,|\,\xi)^{1/(1+\rho)}\right\}^{(1+\rho)} ,$$

where the maximization is over all p.d.'s Q on X.

Note that $E_0(K,1)=R_0(K)$; hence, our result is an improvement over that of Jacobs and Berlekamp only for channels for which $E_0(K,1)<\hat{E}_0(K,1)$. We do not have an example for which $E_0(K,1)<\hat{E}_0(K,1)$, but we believe that such channels exist. It is known, for example, that there exists K for which $E_0(K,\rho)$ is not a concave function of $\rho$ [14]; for any such K, $E_0(K,\rho)<\hat{E}_0(K,\rho)$ at some $\rho\geq 0$.

$R_0(K)$ has been called the <u>cut-off rate</u> of channel K with the understanding that at rates above $R_0(K)$ the average complexity of sequential decoding is infinite. The above theorem justifies the use of this term.

**Theorem 3.3.1.** $\mathbf{R}(K)=\mathbf{R}_0(K)$ for any channel $K=(P;X_1,\dots,X_n;Y)$ which has the property that

$$\sum_{\eta\in Y}\sqrt{P(\eta\,|\,\xi_1,\dots,\xi_n)P(\eta\,|\,\zeta_1,\dots,\zeta_n)}\log\left\{P(\eta\,|\,\xi_1,\dots,\xi_n)/P(\eta\,|\,\zeta_1,\dots,\zeta_n)\right\}=0$$

for every $\xi_i,\zeta_i\in X_i$, $i=1,\dots,n$. □

Channels with the above property are called pairwise reversible channels [16]; an example is the TEC of Figure 1.2.1.

The above converses determine $\mathbf{R}$ for two special classes of channels. However, $\mathbf{R}$ remains undetermined in the general case. It might be that $\mathbf{R}(K)$ equals $\mathbf{R}_0(K)$ for all K, but this has not been proved yet, except in an ensemble average sense (see Theorem 3.4.1). No examples have been found for which $\mathbf{R}$ is strictly larger than $\mathbf{R}_0$, either.

<u>Non-Joint Sequential Decoding</u>

Chapter 4 considers an alternative approach to sequential decoding and finds an inner bound to its (appropriately defined) achievable rate region. Non-joint sequential decoding, as this approach is called, uses a separate sequential decoder for each user; the decoder for a given user decodes that user's message without any knowledge of the tree codes of the remaining users.

In exchange for the increase in the number of decoders, non-joint decoding allows each decoder to be much simpler than a joint decoder. It is demonstrated by an example in Chapter 4 that non-joint sequential decoding, in addition to being simpler, sometimes achieves rates that are unachievable by ordinary sequential decoding. This seemingly paradoxical result is then explained, and conclusions are drawn about the nature of achievability in sequential decoding.

This completes the summary of the main results. In the remaining part of this section, we shall consider some examples and try to answer some specific questions about sequential decoding.

**Example 1.5.1.**

a) Two-User OR Channel (Figure 1.5.1)
For this channel, it is known that $\mathbf{R}=\mathbf{R}_0=\mathbf{C}$; in other words, the achievable rate region of sequential decoding coincides with the capacity region.

One particular feature of the OR channel, which we wish to discuss, is that it is noiseless; that is, the channel output is completely determined by the channel inputs. Noiseless channels are pairwise reversible. Hence, by Theorem 3.3.1, $\mathbf{R}(K)=\mathbf{R}_0(K)$ for all noiseless K. Furthermore, for any noiseless K, one can achieve $\mathbf{R}_0(K)$ by simply using a metric that has only two values, namely 0 and $-\infty$. This metric assigns 0 to consistent nodes and $-\infty$ to inconsistent ones. A node u(..j) is said to be consistent if its correctness can not be ruled out on the basis of y(..j), the first j blocks of the received sequence.

b) Two-User Erasure Channel (TEC) (Figure 1.5.2)
This is another noiseless channel, so we know that $\mathbf{R}_0$(TEC)=$\mathbf{R}$(TEC). The shaded region in Figure 1.5.2 is an inner bound to $\mathbf{R}_0$(TEC), obtained by computing $\mathbf{R}_0$(TEC,$\mathbf{Q}$) for $\mathbf{Q}=(\mathbf{Q}_1,\mathbf{Q}_2)$ with $\mathbf{Q}_1=\mathbf{Q}_2$=the uniform distribution on {0,1}. $\mathbf{R}_0$(TEC,$\mathbf{Q}$) is not equal to $\mathbf{R}_0$(TEC), because clearly, the points (0,1) and (1,0) belong to $\mathbf{R}_0$(TEC). So, a larger inner bound to $\mathbf{R}_0$(TEC) can be obtained by taking the convex-hull of the union of the shaded region with the points (0,1) and (1,0).

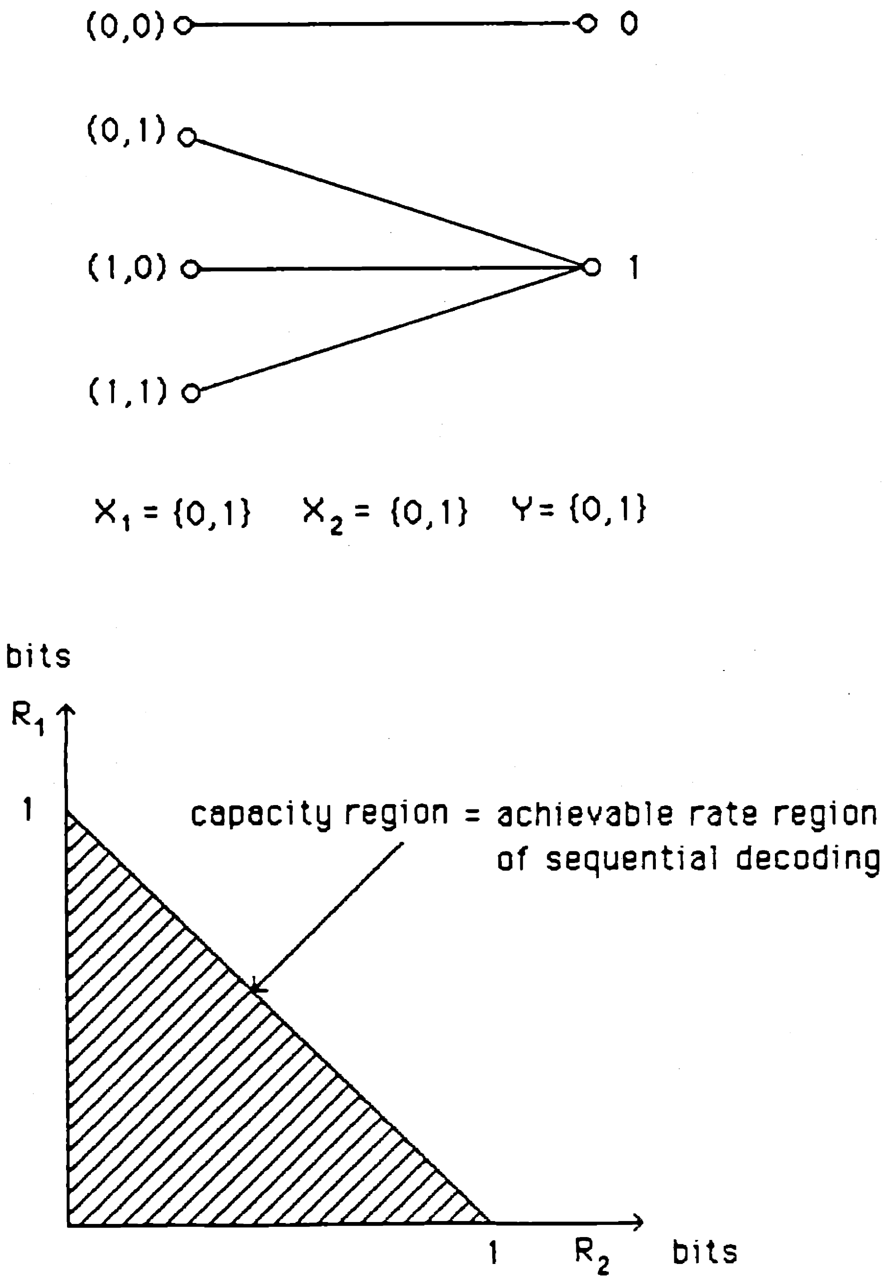

(0,0)
0
(0,1)
(1,0)
1
(1,1)
X1 = {0,1}   X2 = {0,1}   Y = {0,1}
bits
R1
1
capacity region = achievable rate region
of sequential decoding
1
R2
bits

Figure 1.5.1. Two-user OR channel.

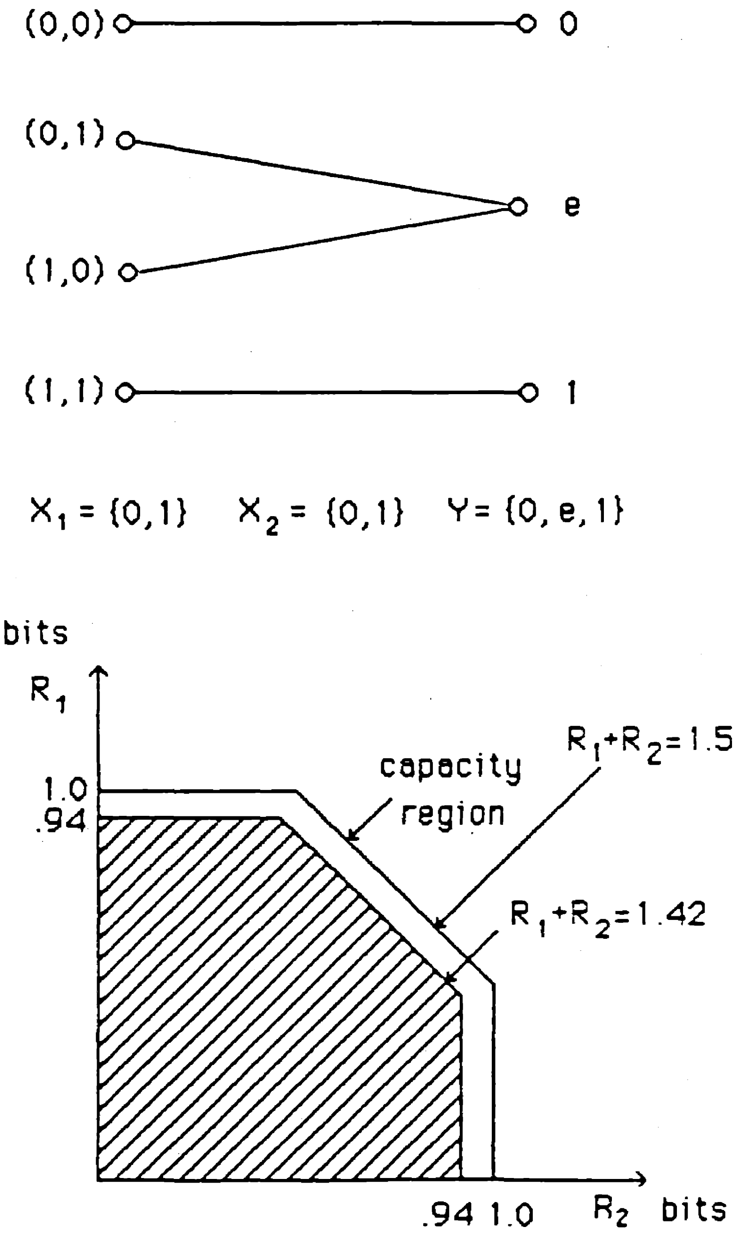


Figure 1.5.2. Two-user erasure channel.

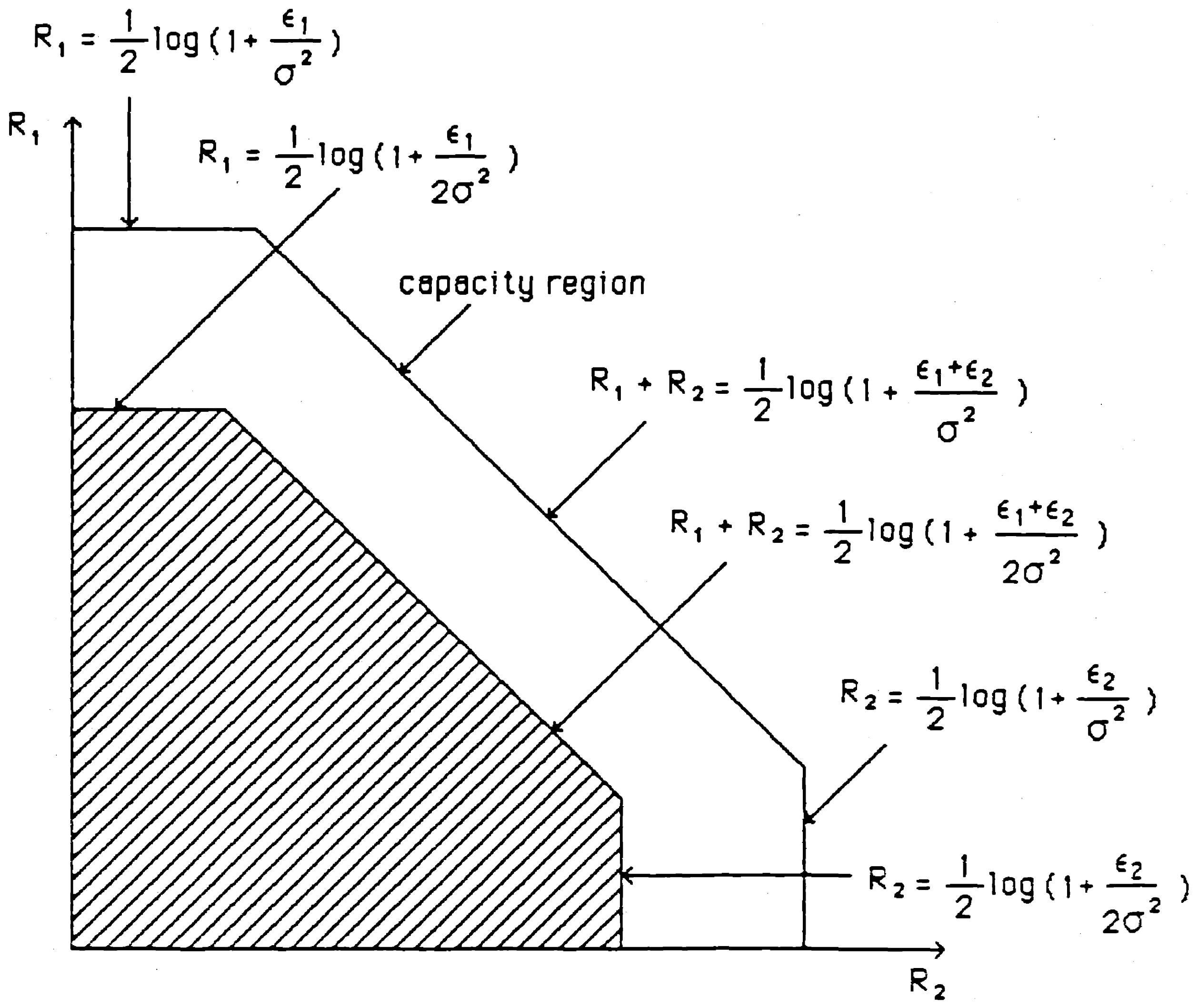
Figure 1.5.2 shows that sum rates of up to 1.42 bits are achievable by sequential decoding. In Example 1.2.1, a simple block code achieving a sum rate of approximately 1.3 bits was given. We do not know, however, of any comparably simple block codes which achieve sum rates as high as 1.42 bits, while maintaining an arbitrarily small probability of error.

c) Two-User Additive Gaussian Noise Channel (AGNC) (Figure 1.5.3)
This is a channel with non-discrete input and output alphabets. Our results do not directly apply to such channels since we are considering only discrete channels. Nevertheless, the AGNC is of special interest because of its practical relevance. The treatment here is brief, however; and we refer to [6] for more about this channel.

The channel input and output alphabets for AGNC's are the set of real numbers. If $\eta$, $\xi_1$, and $\xi_2$ denote, respectively, the received number, the number transmitted by user 1, and the number transmitted by user 2, then $\eta-\xi_1-\xi_2$ (the noise) is a random variable with distribution $N(0,\sigma^2)$. Here, $N(0,\sigma^2)$ is the Gaussian density function with mean 0 and variance $\sigma^2$. There are energy constraints on the inputs of the form: $E(\xi_1^2)\le\epsilon_1$ and $E(\xi_2^2)\le\epsilon_2$, where E denotes expected value in a time and code average sense. (In the absence of energy constraints, the capacity region and the achievable rate region of sequential decoding are unbounded.)

Figure 1.5.3 shows **C**(AGNC), the capacity region, and an inner bound to $\mathbf{R}_0$(AGNC). The inner bound is obtained by computing $\mathbf{R}_0$(AGNC,**q**) for $\mathbf{q}=(N(0,\epsilon_1),N(0,\epsilon_2))$. The computation of $\mathbf{R}_0$(AGNC,**q**) is carried out in the same way as for discrete channels, except that sums are replaced by integrals and probability distributions by densities.

Notice that, if $\sigma^2$ is fixed, the achievable rate region of sequential decoding for an AGNC with constraints $E(\xi_1^2)\le 2\epsilon_1$ and $E(\xi_2^2)\le 2\epsilon_2$ is at least as large as the capacity region of an AGNC with constraints $E(\xi_1^2)\le\epsilon_1$ and $E(\xi_2^2)\le\epsilon_2$. So, at the expense of at most doubling the "energy", we can achieve all points in the capacity region of a given AGNC by sequential decoding.


$R_1 = \frac{1}{2}\log(1+\frac{\epsilon_1}{\sigma^2})$
$R_1$
$R_1 = \frac{1}{2}\log(1+\frac{\epsilon_1}{2\sigma^2})$
capacity region
$R_1 + R_2 = \frac{1}{2}\log(1+\frac{\epsilon_1+\epsilon_2}{\sigma^2})$
$R_1 + R_2 = \frac{1}{2}\log(1+\frac{\epsilon_1+\epsilon_2}{2\sigma^2})$
$R_2 = \frac{1}{2}\log(1+\frac{\epsilon_2}{\sigma^2})$
$R_2 = \frac{1}{2}\log(1+\frac{\epsilon_2}{2\sigma^2})$
$R_2$

$\eta = \xi_1 + \xi_2 + n$ $\quad$ $E(\xi_1^2) \leq 2\epsilon_1$

$n \sim N(0,\sigma^2)$ $\quad$ $E(\xi_2^2) \leq 2\epsilon_2$

Figure 1.5.3. Additive Gaussian noise channel.

## Complementary Remarks

Here we wish to discuss informally some questions that may have arisen up to this point.

**Q.** What makes sequential decoding of multi-user tree codes a different, if not a more difficult, problem than sequential decoding of one-user tree codes?

**A.** The complication in multi-user sequential decoding is due to the presence of different types of incorrect paths which have markedly different statistical properties in relation to the correct path. Despite this, one has to design a metric that distinguishes the correct path from these various types of incorrect paths. While the design of such a metric may not seem to be a problem (because the correct path has a higher correlation with the channel output sequence than any other path), it is not at all clear whether such additional constraints on the metric do not force the achievable rate region of sequential decoding to be much too small to make it attractive.

To discuss the above ideas in more concrete terms, consider a two-user tree code. Let $s_1$ and $s_2$ be the correct paths for users 1 and 2. Sequential decoding aims at finding $s_1 \times s_2$ based on the information available from the received sequence $y$. For simplicity, let us consider only the incorrect paths in $I_1(s_1 \times s_2)$, the first incorrect subtree of the correct path. There are three types of paths in $I_1(s_1 \times s_2)$: 1) Totally incorrect paths of the form $u_1 \times u_2$ where $u_1 \neq s_1$ and $u_2 \neq s_2$. 2) Half incorrect paths of the form $u_1 \times s_2$ where $u_1 \neq s_1$. 3) Half incorrect paths of the form $s_1 \times u_2$ where $u_2 \neq s_2$.

Paths of type 1 have no correlation with $y$; hence, they are relatively easy to detect and eliminate from further search. But paths of types 2 and 3 are correlated with $y$. This is precisely the point where multi-user sequential decoding differs from and becomes more difficult than single-user sequential decoding.

**Q.** Do we know of simpler characterizations of the regions $\mathbf{R}$ and $\mathbf{R}_0$?

**A.** In general, there are no known characterizations of the regions $\mathbf{R}$ and $\mathbf{R}_0$ which are simpler than their definitions. Clearly, the definitions of these regions do not immediately suggest any algorithms for determining whether a given point belongs to these regions.

While so little is known in terms of computing $\mathbf{R}$ and $\mathbf{R}_0$ in general, the situation is completely solved in the one-user case. For any one-user channel K=(P;X;Y), we have

$$\mathbf{R}(K)=\mathbf{R}_0(K)=[0,\sup_{\mathbf{Q}} R_0(K,\mathbf{Q})]\ ,$$

where the supremum is over all p.d.'s $\mathbf{Q}$ on $X^k$ for some arbitrary integer k, and for any such $\mathbf{Q}$,

$$R_0(K,\mathbf{Q}) = -(1/k)\ln \sum_{\boldsymbol{\eta}\epsilon Y^k} \left\{ \sum_{\boldsymbol{\xi}\epsilon X^k} \mathbf{Q}(\boldsymbol{\xi})\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}\right\}^2 .$$

The computation of $\mathbf{R}(K)$ is made possible by Gallager's parallel channels theorem (see pp. 149-50 of [12]), which states that in order to maximize $R_0(K,\mathbf{Q})$ over $\mathbf{Q}$, one needs to consider only p.d.'s over X, i.e.,

$$\sup\left\{ R_0(K,\mathbf{Q}):\mathbf{Q} \text{ is a p.d. on } X^k \text{ for some integer } k\right\}$$

$$= \sup\left\{ R_0(K,Q):Q \text{ is a p.d. on } X\right\}.$$

The computation of $R_0(K):=\sup\{R_0(K,Q):Q$ is a p.d. on $X\}$ is facilitated by the following necessary and sufficient conditions for a p.d. Q on X to maximize $R_0(K,Q)$ (see Theorem 5.6.5 in [12]):

$$\sum_{\eta\epsilon Y}\sqrt{P(\eta\,|\,\xi)}\sum_{\zeta\epsilon X} Q(\zeta)\sqrt{P(\eta\,|\,\zeta)} \geq \sum_{\eta\epsilon Y}\left\{\sum_{\zeta\epsilon X} Q(\zeta)\sqrt{P(\eta\,|\,\zeta)}\right\}^2 \quad \text{all } \xi\epsilon X,$$

with equality if $Q(\xi)>0$.

These conditions are extremely useful in verifying whether a given Q, which may have been guessed on the basis of intuition, does indeed maximize $R_0(K,Q)$. It is unfortunate that there is no analogue of the parallel channels theorem in the multi-user case.

**Q.** Are **R**(K) and $\mathbf{R}_0$(K) convex regions for all K?

**A.** It is not known if **R**(K) is convex for all K. (Note that one may not need to have an explicit characterization of **R**(K) to prove that it is convex.)

It is known that $\mathbf{R}_0$(K) is convex for all K. The convexity of $\mathbf{R}_0$ should not be attributed to the possibility of time-sharing between a number of tree codes and decoding each code by a separate sequential decoder. That argument overlooks the fact that a collection of sequential decoders working on different codes is not equivalent to any single sequential decoder.

The convexity of $\mathbf{R}_0$ can still be explained by the idea of time-sharing, however; but we must consider time-sharing <u>within</u> a code as opposed to between a number of different codes. Time-sharing within a code is achieved by taking the branches of the tree code long enough so that conventional time-sharing can in effect be used within the duration of a branch. The proof of convexity of $\mathbf{R}_0$, along with several other of its properties, is given in §2.3.

**Q.** How well does the metric proposed for multi-user sequential decoding work in the one-user case?

**A.** The achievable rate region of the proposed metric coincides with **R**(K) for every one-user channel K. For K=(P;X;Y), the metric with parameter (K,k,**Q**,B) is given as follows.
For each $\boldsymbol{\xi}\in X^k$, $\boldsymbol{\eta}\in Y^k$,

$$\gamma(\boldsymbol{\xi},\boldsymbol{\eta}) = \ln \frac{\sqrt{P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}}{\sum\limits_{\boldsymbol{\zeta}\in X^k} \mathbf{Q}(\boldsymbol{\zeta})\sqrt{P(\boldsymbol{\eta}\,|\,\boldsymbol{\zeta})}} - kB\ .$$

For any code parameter (M,k) satisfying (1/k)lnM<$R_0$(K), the appropriate parameter to be used is found as follows: **Q** is taken as a p.d. on $X^k$ such that (1/k)lnM<$R_0$(K,**Q**), and B is then set equal to {(1/k)lnM+$R_0$(K,**Q**)}/2.

In Chapter 2, it is proven, as a special case of Theorem 2.2.1, that the above metric achieves the rate (1/k)lnM. It thus follows that all rates up to $R_0(K)$ are achievable.

Now, compare the above metric with the Fano metric, which is given by

$$\gamma_F(\xi,\eta) = \ln \frac{P(\eta\,|\,\xi)}{\omega(\eta)} - kB_F\ ,$$

and which also achieves all rates up to $R_0(K)$ for any single-user channel K, provided that $\omega$ and $B_F$ are chosen appropriately.

Note that these two metrics are not reducible to one another; that is, it is not possible, in general, to choose the parameters of these metrics so that their ratio is fixed.

We conjecture that the following metric, which contains the above two as special cases, also achieves all rates up to $R_0(K)$ for each single-user channel K and for each r, $0.5 \le r \le 1$.

$$\gamma(\xi,\eta) = \ln \frac{P(\eta\,|\,\xi)^r}{\sum\limits_{\zeta \in X^k} Q(\zeta)\, P(\eta\,|\,\zeta)^r} - kB$$

# Chapter 2

# AN INNER BOUND TO THE ACHIEVABLE RATE REGION OF SEQUENTIAL DECODING

The main result of this chapter is the proof that $\mathbf{R}_0(K)$ (to be defined in §2.2) is an inner bound to $\mathbf{R}(K)$ for any multiple access channel K.

## 2.1. Sufficient Conditions on Achievability

Let $K=(P;X_1,\ldots,X_n;Y)$ be an n-user channel; let $\Gamma$ be a branchwise additive metric for $(M_1,\ldots,M_n,k)$ codes for K; let $\gamma$, $\gamma:(X_1\times\cdots\times X_n)^k \longrightarrow [-\infty,\infty)$, be the branch metric for $\Gamma$. The value of $\Gamma$ for a channel input $x(..m)$ and a channel output $y(..m)$ is thus given by

$$\Gamma(x(..m),y(..m))=\sum_{i=1}^{m}\gamma(x(i),y(i)).$$

In this section, we wish to find conditions on K, $(M_1,..,M_n,k)$, and $\gamma$ which, if satisfied, guarantee that the point $R=(R_1,\ldots,R_n)$, where $R_i=(1/k)\ln M_i$, is achievable in the sense of Definition 1.4.2. We fix K, $(M_1,\ldots,M_n,k)$, and $\Gamma$ throughout the following discussion, and suppress them in the notation.

Proving that R is achievable requires exhibiting the existence of a code $\mathbf{e}$, with rate at least as large as R, for which $D_L(\mathbf{e})$ is uniformly bounded. A direct approach to this problem is not feasible, because the computation of $D_L(\mathbf{e})$ is hopelessly complicated for any non-degenerate code $\mathbf{e}$. We try therefore an indirect approach, known as random-coding, which is based on the fact that the expected value of a random variable upper-bounds the value of that random variable at at least one sample point. Thus, instead of a fixed code, we consider an ensemble of codes, and evaluate the expected value of $D_L(\mathbf{e})$ over this ensemble.

The ensemble we use here is $\mathbf{E}=\mathrm{Ens}(M_1,\ldots,M_n;k;X_1,\ldots,X_n;\mathbf{Q}_1,\ldots,\mathbf{Q}_n)$. $\mathbf{E}$ will be fixed throughout the following analysis, and $E_{\mathbf{e}}$ will denote expectation with respect to the probability measure associated with $\mathbf{E}$.

Now, $E_{\mathbf{e}}D_L(\mathbf{e}) = E_{\mathbf{e}}\{C_1(\mathbf{e})+\cdots+C_L(\mathbf{e})\}/L$

$$= \{E_{\mathbf{e}}C_1(\mathbf{e})+\cdots+E_{\mathbf{e}}C_L(\mathbf{e})\}/L. \qquad (1)$$

So, $E_{\mathbf{e}}D_L(\mathbf{e})$ can be upper-bounded by upper-bounding $E_{\mathbf{e}}C_i(\mathbf{e})$ for each i.

$$E_{\mathbf{e}}C_i(\mathbf{e}) = E_{\mathbf{e}}E_sE_{y|\mathbf{e}s}C_i(\mathbf{e},s,y)$$

$$= E_sE_{\mathbf{e}}E_{y|\mathbf{e}s}C_i(\mathbf{e},s,y). \qquad (2)$$

Here, s represents the source sequence; $E_s$ stands for expectation with respect to the source statistics; $E_{y|\mathbf{e}s}$ stands for expectation with respect to the probability measure on the channel output sequence y conditional on **e**s being the channel input sequence.

Changing the order of expectations in (2) is justified by the non-negativity of the terms involved. (See, e.g., page 147 of [19].)

(One can see at this point that $E_{\mathbf{e}}E_{y|\mathbf{e}s}C_i(\mathbf{e},s,y)$ does not depend on s; hence, in (2), $E_s$ can be dropped, and s can be replaced by any fixed source output. But we shall carry along $E_s$ in the following argument.)

$E_{\mathbf{e}}C_i(\mathbf{e})$ will be upper-bounded with the help of the following inequality.

**Lemma 2.1.1.** For any non-negative t,

$$C_i(\mathbf{e},s,y) \le \sum_{u(..j)\in I_i(s)} \sum_{m\ge i} \exp t\Big\{\Gamma(\mathbf{e}u(..j),y(..j)) - \Gamma(\mathbf{e}s(..m),y(..m))\Big\}. \qquad (3)$$

Proof. A node $u(..j)\in I_i(s)$ reaches the stack-top only if

$$\Gamma(\mathbf{e}u(..j),y(..j)) \geq \Gamma(\mathbf{e}s(..m),y(..m)) \qquad \text{for some } m\geq i. \qquad (4)$$

If (4) is not satisfied, $s(..m)$ has precedence over $u(..j)$ in reaching the stack-top for each $m$, $m\geq i$. So, $u(..j)\in I_i(s)$ reaches the stack-top only if

$$1 \leq \sum_{m\geq i} \exp t\Big\{\Gamma(\mathbf{e}u(..j),y(..j)) - \Gamma(\mathbf{e}s(..m),y(..m))\Big\} \quad \text{for all } t\geq 0. \qquad (5)$$

Note that the right hand side of (5) is positive whether or not $u(..j)$ reaches the stack-top; hence, it upper-bounds the indicator function of the event that $u(..j)$ reaches the stack-top. So, by summing the right hand side of (5) over all nodes in $I_i(s)$, we obtain the claimed upper bound on $C_i(\mathbf{e},s,y)$. □

Hereafter, suppose that $t$ is a fixed positive number. Now, from (2) and (3),

$$E_{\mathbf{e}}C_i(\mathbf{e}) \leq E_s \sum_{u(..j)\in I_i(s)} \sum_{m\geq i} \Lambda(s,m,u(..j)), \qquad (6)$$

where, by definition,

$$\Lambda(s,m,u(..j))=E_{\mathbf{e}}E_{y\,|\,\mathbf{e}s}\exp t\Big\{\Gamma(\mathbf{e}u(..j),y(..j)) - \Gamma(\mathbf{e}s(..m),y(..m))\Big\}. \qquad (7)$$

For any $u(..j)\in I_i(s)$,

$$\Lambda(s,m,u(..j)) = E_{\mathbf{e}}E_{y\,|\,\mathbf{e}s} \exp t\{ \sum_{i\leq h\leq j}\gamma(\mathbf{e}u(h),y(h)) - \sum_{i\leq h\leq m}\gamma(\mathbf{e}s(h),y(h)) \};$$

thus, if $j>m$,

$\Lambda(s,m,u(..j)) =$

$$= E_{\mathbf{e}}E_{y\,|\,\mathbf{e}s}\exp t\{\sum_{i\leq h\leq m}[\gamma(\mathbf{e}u(h),y(h)) - \gamma(\mathbf{e}s(h),y(h))]+\sum_{m<h\leq j}\gamma(\mathbf{e}u(h),y(h))\}; \qquad (8)$$

and, if $m \geq j$,

$\Lambda(s,m,u(..j)) =$

$$= E_{\mathbf{e}} E_{y|\mathbf{es}} \exp t\{\sum_{i \leq h \leq j} [\gamma(\mathbf{eu}(h),y(h)) - \gamma(\mathbf{es}(h),y(h))] - \sum_{j<h \leq m} \gamma(\mathbf{es}(h),y(h))\}. \qquad (9)$$

Since the labels on branches at different levels are independent random variables over the ensemble under consideration, (8) and (9) can be rewritten as follows.

For any $u(..j) \epsilon I_i(s)$, if $i>m$,

$\Lambda(s,m,u(..j)) =$

$$\prod_{i \leq h \leq m} E\exp t\{\gamma(\mathbf{eu}(h),y(h)) - \gamma(\mathbf{es}(h),y(h))\} \prod_{m<h \leq j} E\exp t\gamma(\mathbf{eu}(h),y(h)); \qquad (10)$$

and, if $m \geq j$,

$\Lambda(s,m,u(..j)) =$

$$\prod_{i \leq h \leq j} E\exp t\{\gamma(\mathbf{eu}(h),y(h)) - \gamma(\mathbf{es}(h),y(h))\} \prod_{j<h \leq m} E\exp t\gamma(\mathbf{es}(h),y(h)), \qquad (11)$$

where the symbol E has been used as an abbreviation for $E_{\mathbf{e}} E_{y|\mathbf{es}}$.

We now wish to find an explicit expression for $\Lambda(s,m,u(..j))$. Let $u(..j)$ be a fixed node in $I_i(s)$, and $(T_1,\ldots,T_j)$ be the type of $u(..j)$ with respect to s. Now, for any $h \epsilon \{1,\ldots,j\}$, $\boldsymbol{\eta} \epsilon Y^k$, $\boldsymbol{\xi}=\boldsymbol{\xi}_1 \times \cdots \times \boldsymbol{\xi}_n$, and $\boldsymbol{\zeta}=\boldsymbol{\zeta}_1 \times \cdots \times \boldsymbol{\zeta}_n$, where $\boldsymbol{\xi}_r \epsilon X_r^k$, $\boldsymbol{\zeta}_r \epsilon X_r^k$, $r=1,\ldots,n$, the probability that $\mathbf{es}(h)=\boldsymbol{\xi}$ and $\mathbf{eu}(h)=\boldsymbol{\zeta}$ and $y(h)=\boldsymbol{\eta}$ is given as follows.

$\Pr\{es(h)=\boldsymbol{\xi},\ eu(h)=\boldsymbol{\zeta},\ y(h)=\boldsymbol{\eta}\}=$

$$=\prod_{1\le r\le n}\mathbf{Q}_r(\boldsymbol{\xi}_r)\prod_{r\in T_h}\mathbf{Q}_r(\boldsymbol{\zeta}_r)\prod_{r\in T_h{}^c}\chi\{\boldsymbol{\zeta}_r=\boldsymbol{\xi}_r\}\,P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}), \tag{12}$$

where $\chi$ is the indicator function.

To simplify the notation, we shall write

$\mathbf{Q}(\boldsymbol{\xi})$ in place of $\prod_{1\le r\le n}\mathbf{Q}_r(\boldsymbol{\xi}_r)$ ,

$\mathbf{Q}(\boldsymbol{\xi}_T)$ in place of $\prod_{r\in T}\mathbf{Q}_r(\boldsymbol{\xi}_r)$, and

$\chi\{\boldsymbol{\zeta}_T=\boldsymbol{\xi}_T\}$ in place of $\prod_{r\in T}\chi\{\boldsymbol{\zeta}_r=\boldsymbol{\xi}_r\}$.

In this notation, (12) can be rewritten as follows.

$$\Pr\{es(h)=\boldsymbol{\xi},\ eu(h)=\boldsymbol{\zeta},\ y(h)=\boldsymbol{\eta}\} = \mathbf{Q}(\boldsymbol{\xi})\,\mathbf{Q}(\boldsymbol{\zeta}_{T_h})\,\chi\{\boldsymbol{\zeta}_{T_h^c}=\boldsymbol{\xi}_{T_h^c}\}P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}) \tag{13}$$

Now, $E_{\mathbf{e}}E_{y|es}\exp -t\gamma(es(h),y(h))$

$$=\sum_{\boldsymbol{\eta}}\sum_{\boldsymbol{\xi}}\mathbf{Q}(\boldsymbol{\xi})P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})\exp -t\gamma(\boldsymbol{\xi},\boldsymbol{\eta}), \tag{14}$$

$E_{\mathbf{e}}E_{y|es}\exp t\{\gamma(eu(h),y(h))-\gamma(es(h),y(h))\}$

$$=\sum_{\boldsymbol{\eta}}\sum_{\boldsymbol{\xi}}\sum_{\boldsymbol{\zeta}}\mathbf{Q}(\boldsymbol{\xi})\,\mathbf{Q}(\boldsymbol{\zeta}_{T_h})\,\chi\{\boldsymbol{\zeta}_{T_h^c}=\boldsymbol{\xi}_{T_h^c}\}P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})\exp t\{\gamma(\boldsymbol{\zeta},\boldsymbol{\eta})-\gamma(\boldsymbol{\xi},\boldsymbol{\eta})\}, \tag{15}$$

and $E_{\mathbf{e}}E_{y|es}\exp t\gamma(eu(h),y(h))$

$$=\sum_{\boldsymbol{\eta}}\sum_{\boldsymbol{\xi}}\sum_{\boldsymbol{\zeta}}\mathbf{Q}(\boldsymbol{\xi})\,\mathbf{Q}(\boldsymbol{\zeta}_{T_h})\,\chi\{\boldsymbol{\xi}_{T_h^c}=\boldsymbol{\zeta}_{T_h^c}\}P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})\exp t\gamma(\boldsymbol{\zeta},\boldsymbol{\eta}). \tag{16}$$

We see that the left hand side of (14) does not depend on h; and the left hand sides of (15) and (16) depend on h only through $T_h$. So, we define

$$\eta = E_{\mathbf{e}} E_{y|\mathbf{es}} \exp -t\gamma(\mathbf{e}s(h),y(h)),$$

$$\sigma(T_h) = E_{\mathbf{e}} E_{y|\mathbf{es}} \exp t\{\gamma(\mathbf{e}u(h),y(h))-\gamma(\mathbf{e}s(h),y(h))\}, \text{ and}$$

$$\beta(T_h) = E_{\mathbf{e}} E_{y|\mathbf{es}} \exp t\gamma(\mathbf{e}u(h),y(h))).$$

Now, for any node $u(..j)\epsilon I_i(s)$ with type $(T_1,...,T_j)$ wrt s, (10) and (11) can be rewritten as follows.

$$\Lambda(s,m,u(..j)) = \begin{cases} \sigma(T_i)\cdots\sigma(T_m)\beta(T_{m+1})\cdots\beta(T_j) & , j>m; \quad (17) \\ \sigma(T_i)\cdots\sigma(T_m)\eta^{m-j} & , m\geq j. \quad (18) \end{cases}$$

Observe that $\Lambda(s,m,u(..j))$ depends on $u(..j)$ only through the type of $u(..j)$ wrt s. So, let $\Lambda(s,m,\mathbf{T})$ denote $\Lambda(s,m,u(..j))$ whenever $u(..j)$ is a node of type $\mathbf{T}$ wrt s. Letting $\mathbf{T}(i,j)$ be the set of types for level-j nodes in $I_i(s)$, (6) can be rewritten as follows.

$$E_{\mathbf{e}} C_i(\mathbf{e}) \leq E_s \sum_{j=i}^{\infty} \sum_{\mathbf{T}\epsilon\mathbf{T}(i,j)} \sum_{\substack{u(..j): \\ \text{type of } u(..j)=\mathbf{T}}} \sum_{m=i}^{\infty} \Lambda(s,m,u(..j))$$

$$= E_s \sum_{j=i}^{\infty} \sum_{\mathbf{T}\epsilon\mathbf{T}(i,j)} N(\mathbf{T}) \sum_{m=i}^{\infty} \Lambda(s,m,\mathbf{T}), \quad (19)$$

where $N(\mathbf{T})$ denotes the number of nodes of type $\mathbf{T}$.

Define $\Omega(T)=\max\{\sigma(T), \beta(T)\}$. Now, for any $\mathbf{T}=(T_1,...,T_j)\epsilon\mathbf{T}(i,j)$,

$$\Lambda(s,m,\mathbf{T}) \leq \begin{cases} \Omega(T_i)\cdots\Omega(T_j) & , j>m; \quad (20) \\ \Omega(T_i)\cdots\Omega(T_j)\eta^{m-j} & , m\geq j. \quad (21) \end{cases}$$

Thus,

$$\sum_{m=i}^{\infty} \Lambda(s,m,\mathbf{T}) \le \sum_{m=i}^{j-1} \Omega(T_i)\cdots\Omega(T_j) + \sum_{m=j}^{\infty} \Omega(T_i)\cdots\Omega(T_j)\eta^{m-j}$$

$$= \Omega(T_i)\cdots\Omega(T_j)\ (\ j-i+\sum_{h=0}^{\infty} \eta^{h}\ )\ . \tag{22}$$

For any non-empty subset T of $\{1,....,n\}$, let M(T) be the product of $M_i$ for $i\epsilon T$; if $T=\phi$, let $M(T)=1$. For any node type $\mathbf{T}=(T_1,...,T_j)$, let $M(\mathbf{T})=M(T_1)\cdot\cdot$ $\cdot\cdot M(T_j)$. Note that $M(\mathbf{T})$ is an upper bound on $N(\mathbf{T})$, the number of nodes of type $\mathbf{T}$. Also note that, if $\mathbf{T}=(T_1,...,T_j)\epsilon\mathbf{T}(i,j)$, then $M(\mathbf{T})=M(T_i)\cdots M(T_j)$; because $T_h=\phi$ for $1\le h\le i-1$. Define

$$\Psi=\max\Big\{\Omega(T)M(T) : T \text{ is a non-empty subset of } \{1,..,n\}\Big\}.$$

Now, by (22),

$$N(\mathbf{T}) \sum_{m=i}^{\infty} \Lambda(s,m,\mathbf{T}) \le M(\mathbf{T}) \sum_{m=i}^{\infty} \Lambda(s,m,\mathbf{T}) \tag{23}$$

$$\le\ \Psi^{j-i}(\ j-i+\sum_{h=0}^{\infty} \eta^{h}\ ). \tag{24}$$

By (19) and (23)-(24),

$$E_{\mathbf{e}}C_i(\mathbf{e})\ \le\ E_s \sum_{j=i}^{\infty}\ \sum_{\mathbf{T}\epsilon\mathbf{T}(i,j)} \Psi^{j-i}(\ j-i+\sum_{h=0}^{\infty} \eta^{h}\ ) \tag{25}$$

Noting that the number of elements in $\mathbf{T}(i,j)$ is upper-bounded by $(j-i+2)^n$ (see §1.3 for this upper bound), it follows from (25) that

$$E_{\mathbf{e}}C_i(\mathbf{e})\ \le E_s \sum_{j=i}^{\infty}\ (j-i+2)^n\ \Psi^{j-i}(\ j-i+\sum_{h=0}^{\infty} \eta^{h}\ )$$

$$= \sum_{j=0}^{\infty} (j+2)^n \Psi^j ( j+ \sum_{h=0}^{\infty} \eta^h ) \qquad (26)$$

The right side of (26) is independent of i; and, it converges if $\Psi<1$ and $\eta<1$. The conclusion of this discussion can now be stated as follows.

**Theorem 2.1.1.** Sufficient Conditions on Achievability.
Let $K=(P;X_1,\ldots,X_n;Y)$ be a multiple access channel; suppose that there exist a branch metric $\gamma:(X_1\times\cdots\times X_n)^k \longrightarrow [-\infty,\infty)$, an ensemble

$$\mathbf{E}=\mathrm{Ens}(M_1,\ldots,M_n;k;X_1,\ldots,X_n;\mathbf{Q}_1,\ldots,\mathbf{Q}_n),$$

and a positive real number t such that

i) $\eta(t,K,\gamma,\mathbf{E})<1$,

ii) $M(T)\sigma(T,t,K,\gamma,\mathbf{E})<1$ for each non-empty subset T of $\{1,\ldots,n\}$, and

iii) $M(T)\beta(T,t,K,\gamma,\mathbf{E})<1$ for each non-empty subset T of $\{1,\ldots,n\}$.

Then, for all L,

$$E_{\mathbf{e}}D_L(K,\mathbf{e},\Gamma)\leq \sum_{j=0}^{\infty} (j+2)^n\ \Psi(t,K,\gamma,\mathbf{E})^j\ (\ j+1/(1-\eta(t,K,\gamma,\mathbf{E})) < \infty\ ,$$

where $\Gamma$ denotes the metric based on $\gamma$. * □

Thus, if K, $(M_1,\ldots,M_n,k)$, and $\gamma$ satisfy the conditions of the above theorem for some ensemble $\mathbf{E}$, then $(R_1,\ldots,R_n)$, where $R_i=(1/k)\ln M_i$, belongs to $\mathbf{R}(K)$, the achievable rate region of sequential decoding.

* It is possible to prove this theorem with $\eta(t,K,\gamma,\mathbf{E})<1$ relaxed to $\eta(t,K,\gamma,\mathbf{E})\leq 1$ by following Gallager’s proof for n=1 (see App. 6B of [12]).

## 2.2. The Proposed Metric and An Inner Bound to Its Achievable Rate Region

This section considers a class of metrics and finds an inner bound to its achievable rate region by using Theorem 2.1.1. Metrics in this class are parametrized by a four-tuple $(K,k,\mathbf{Q},B)$ where $K$ is a multiple access channel, say $K=(P;X_1,\ldots,X_n;Y)$; $k$ is a positive integer; $\mathbf{Q}=(\mathbf{Q}_1,\ldots,\mathbf{Q}_n)$ where $\mathbf{Q}_i$ is a p.d. on $X_i^k$, $i=1,\ldots,n$; and $B$ is a real-valued function of non-empty subsets of $\{1,\ldots,n\}$. $B(T)$ is called the bias term for subset $T$.

The metric with parameter $(K,k,\mathbf{Q},B)$, denoted by $\mathrm{met}(K,k,\mathbf{Q},B)$, is a branchwise additive metric based on the following branch metric $\gamma$.
For each $\boldsymbol{\eta}\in Y^k$ and $\boldsymbol{\xi}=\boldsymbol{\xi}_1\times\cdots\times\boldsymbol{\xi}_n$ where $\boldsymbol{\xi}_i\in X_i^k$, $i=1,\ldots,n$,

$$\gamma(\boldsymbol{\xi},\boldsymbol{\eta}) = \min_{T}\{\gamma_T(\boldsymbol{\xi},\boldsymbol{\eta})\}, \tag{1}$$

where the minimum is over all non-empty subsets of $\{1,\ldots,n\}$ and

$$\gamma_T(\boldsymbol{\xi},\boldsymbol{\eta})= \ln \frac{\sqrt{P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}}{\displaystyle\sum_{\{\boldsymbol{\zeta}_i\}_{i\in T}} \prod_{i\in T} \mathbf{Q}_i(\boldsymbol{\zeta}_i)\sqrt{P\left\{\boldsymbol{\eta}\,|\,\{\boldsymbol{\xi}_i\}_{i\in T^c}\,,\,\{\boldsymbol{\zeta}_i\}_{i\in T}\right\}}} - kB(T)\,. \tag{2}$$

In (2), the summation is over the cartesian product, over all $i\in T$, of $X_i^k$; $P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})$ is the transition probability of the channel over blocks of length $k$; $P\left\{\boldsymbol{\eta}\,|\,\{\boldsymbol{\xi}_i\}_{i\in T^c}\,,\,\{\boldsymbol{\zeta}_i\}_{i\in T}\right\}$ is the probability that $\boldsymbol{\eta}$ is received given that the transmitted block at input $i$ equals $\boldsymbol{\zeta}_i$ if $i\in T$ and $\boldsymbol{\xi}_i$ if $i\in T^c$.

To simplify the notation, as in the previous section, we shall denote

$$\{\boldsymbol{\xi}_i\}_{i\in T} \quad \text{by} \quad \boldsymbol{\xi}_T,$$

$$\prod_{i\in T} \mathbf{Q}_i(\boldsymbol{\xi}_i) \quad \text{by} \quad \mathbf{Q}(\boldsymbol{\xi}_T), \text{ and}$$

$$\mathbf{P}\left\{\boldsymbol{\eta} \,|\, \{\boldsymbol{\xi}_i\}_{i\in T^c}, \{\boldsymbol{\zeta}_i\}_{i\in T}\right\} \quad \text{by} \quad \mathbf{P}(\boldsymbol{\eta} \,|\, \boldsymbol{\xi}_{T^c}, \boldsymbol{\zeta}_T).$$

In this notation, (2) can be rewritten as follows.

$$\gamma_T(\boldsymbol{\xi},\boldsymbol{\eta}) = \ln \frac{\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}}{\sum\limits_{\boldsymbol{\zeta}_T} \mathbf{Q}(\boldsymbol{\zeta}_T)\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_{T^c},\boldsymbol{\zeta}_T)}} - kB(T) \tag{3}$$

The remainder of this section is devoted to showing that $\mathbf{R}_0(K)$, which we shall define next, is an inner bound to the achievable rate region of the above class of metrics (hence, an inner bound to $\mathbf{R}(K)$) for all K.

**Definition 2.2.1.**
For any channel $K=(P;X_1,\ldots,X_n;Y)$, any $\mathbf{Q}=(\mathbf{Q}_1,\ldots,\mathbf{Q}_n)$ where $\mathbf{Q}_i$ is a p.d. on $X_i^k$, and any subset T of $\{1,\ldots,n\}$, we define

$$R_0(K,\mathbf{Q},T) = -(1/k)\ln \sum_{\boldsymbol{\eta},\boldsymbol{\xi}_{T^c}} \mathbf{Q}(\boldsymbol{\xi}_{T^c})\left\{ \sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})} \right\}^2 \text{ and}$$

$$\mathbf{R}_0(K,\mathbf{Q}) = \left\{(R_1,\ldots,R_n): 0 \le R(T) \le R_0(K,\mathbf{Q},T) \text{ for each subset T of } \{1,\ldots,n\}\right\}.$$

We also define

$$\mathbf{R}_0(K,k) = \bigcup_{\mathbf{Q}} \mathbf{R}_0(K,\mathbf{Q}),$$

where the union is over all $\mathbf{Q}=(\mathbf{Q}_1,\ldots,\mathbf{Q}_n)$ such that $\mathbf{Q}_i$ is a p.d. on $X_i^k$,

and $$\mathbf{R}_0(K) = \bigcup_{k=1}^{\infty} \mathbf{R}_0(K,k). \ \Box$$

$\mathbf{R}_0(K)$ will be shown to be an inner bound to the achievable rate region of met($K,k,\mathbf{Q},B$) with the help of the following fact, which is just a special case of Theorem 2.1.1 at $t=1$ and in the particular way $\mathbf{E}$ is selected.

**Lemma 2.2.1.** Sufficient Conditions on Achievability for met($K,k,\mathbf{Q},B$). For any channel $K=(P;X_1,\ldots X_n;Y)$ and any $(M_1,\ldots,M_n,k)$, the point $(R_1,\ldots,R_n)$, where $R_i=(1/k)\ln M_i$, belongs to the achievable rate region of met($K,k,\mathbf{Q},B$) if the following conditions are satisfied by $\gamma$, the branch metric for met($K,k,\mathbf{Q},B$), and the ensemble $\mathbf{E}=\mathrm{Ens}(M_1,\ldots,M_n;k;X_1,\ldots,X_n;\mathbf{Q}_1,\ldots,\mathbf{Q}_n)$, where $\mathbf{Q}_1,\ldots,\mathbf{Q}_n$ are such that $(\mathbf{Q}_1,\ldots,\mathbf{Q}_n)=\mathbf{Q}$.

1) $\eta(1,K,\gamma,\mathbf{E})<1$,
2) $M(T)\sigma(T,1,K,\gamma,\mathbf{E})<1$ for each non-empty subset $T$ of $\{1,\ldots,n\}$, and
3) $M(T)\beta(T,1,K,\gamma,\mathbf{E})<1$ for each non-empty subset $T$ of $\{1,\ldots,n\}$. □

Note that in the above lemma the distributions parametrizing $\mathbf{E}$ and the metric are identical. Of course, the statement of the lemma would still hold if this were not so, but this less general form is sufficient for our purposes.

In order to restate Lemma 2.2.1 in a simpler, more useful way, we now find upper bounds on $\eta(1,K,\gamma,\mathbf{E})$, $\sigma(T,1,K,\gamma,\mathbf{E})$, and $\beta(T,1,K,\gamma,\mathbf{E})$ for a fixed collection of $K$, $\gamma$, and $\mathbf{E}$, where $\mathbf{E}$ and $\gamma$ are parametrized by the same $\mathbf{Q}=(\mathbf{Q}_1,\ldots,\mathbf{Q}_n)$.

$$\eta(1,K,\gamma,\mathbf{E}) = \sum_{\boldsymbol{\eta},\boldsymbol{\xi}} \mathbf{Q}(\boldsymbol{\xi})\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})\exp-\gamma(\boldsymbol{\xi},\boldsymbol{\eta})$$

$$\leq \sum_{T\neq\phi}\ \sum_{\boldsymbol{\eta},\boldsymbol{\xi}} \mathbf{Q}(\boldsymbol{\xi})\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})\exp-\gamma_T(\boldsymbol{\xi},\boldsymbol{\eta})$$

$$=\sum_{T\neq\phi}\ \sum_{\boldsymbol{\eta},\boldsymbol{\xi}} \mathbf{Q}(\boldsymbol{\xi})\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})\Big\{\sum_{\boldsymbol{\zeta}_T}\mathbf{Q}(\boldsymbol{\zeta}_T)\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_{T^c},\boldsymbol{\zeta}_T)}\Big/\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}\Big\}\exp kB(T)$$

$$= \sum_{T\neq\phi} \sum_{\eta,\xi} Q(\xi) \sqrt{P(\eta\,|\,\xi)} \sum_{\zeta_T} Q(\zeta_T)\sqrt{P(\eta\,|\,\xi_{T^c},\zeta_T)} \exp kB(T)$$

$$= \sum_{T\neq\phi} \sum_{\eta,\xi_{T^c}} Q(\xi_{T^c}) \sum_{\xi_T} Q(\xi_T)\sqrt{P(\eta\,|\,\xi_{T^c},\xi_T)} \sum_{\zeta_T} Q(\zeta_T)\sqrt{P(\eta\,|\,\xi_{T^c},\zeta_T)} \exp kB(T)$$

$$= \sum_{T\neq\phi} \sum_{\eta,\xi_{T^c}} Q(\xi_{T^c})\left\{ \sum_{\xi_T} Q(\xi_T)\sqrt{P(\eta\,|\,\xi)} \right\}^2 \exp kB(T)$$

$$= \sum_{T\neq\phi} \exp -k\left\{R_0(K,Q,T)-B(T)\right\}. \tag{4}$$

For notational convenience, in the following $\gamma(\xi_T,\xi_{T^c},\eta)$ and $\gamma(\xi,\eta)$ will be used interchangeably.

$$\sigma(T,1,K,\gamma,E) = \sum_{\xi,\zeta_T,\eta} Q(\xi)Q(\zeta_T)P(\eta\,|\,\xi)\exp(\gamma(\zeta_T,\xi_{T^c},\eta)-\gamma(\xi,\eta))$$

$$\leq \sum_{\xi,\zeta_T,\eta} Q(\xi)Q(\zeta_T)P(\eta\,|\,\xi)\exp(\gamma_T(\zeta_T,\xi_{T^c},\eta)-\gamma(\xi,\eta))$$

$$= \sum_{\xi,\zeta_T,\eta} Q(\xi)Q(\zeta_T)P(\eta\,|\,\xi)\,\frac{\sqrt{P(\eta\,|\,\zeta_T,\xi_{T^c})}\,\exp-\gamma(\xi,\eta)\exp-kB(T)}{\sum_{\psi_T} Q(\psi_T)\sqrt{P(\eta\,|\,\psi_T,\xi_{T^c})}}$$

$$= \exp -kB(T)\sum_{\xi,\eta} Q(\xi)P(\eta\,|\,\xi)\,\frac{\sum_{\zeta_T} Q(\zeta_T)\sqrt{P(\eta\,|\,\zeta_T,\xi_{T^c})}}{\sum_{\psi_T} Q(\psi_T)\sqrt{P(\eta\,|\,\psi_T,\xi_{T^c})}}\exp-\gamma(\xi,\eta)$$

$$= \exp -kB(T) \sum_{\xi,\eta} Q(\xi)P(\eta \mid \xi) \exp -\gamma(\xi,\eta)$$

$$= \eta(1,K,\gamma,E) \exp -kB(T). \qquad (5)$$

Finally,

$$\beta(T,1,K,\gamma,E) = \sum_{\xi,\zeta_T,\eta} Q(\xi)Q(\zeta_T)P(\eta \mid \xi) \exp \gamma(\zeta_T,\xi_{T^c},\eta)$$

$$\leq \sum_{\xi,\zeta_T,\eta} Q(\xi)Q(\zeta_T)P(\eta \mid \xi) \exp \gamma_T(\zeta_T,\xi_{T^c},\eta)$$

$$= \sum_{\xi,\zeta_T,\eta} Q(\xi)Q(\zeta_T)P(\eta \mid \xi) \frac{\sqrt{P(\eta \mid \zeta_T,\xi_{T^c})}}{\sum_{\psi_T} Q(\psi_T)\sqrt{P(\eta \mid \psi_T,\xi_{T^c})}} \exp -kB(T)$$

$$= \sum_{\xi,\eta} Q(\xi)P(\eta \mid \xi) \frac{\sum_{\zeta_T} Q(\zeta_T)\sqrt{P(\eta \mid \zeta_T,\xi_{T^c})}}{\sum_{\psi_T} Q(\psi_T)\sqrt{P(\eta \mid \psi_T,\xi_{T^c})}} \exp -kB(T)$$

$$= \exp -kB(T) \qquad (6)$$

It follows from (4)-(6) that conditions of Lemma 2.2.1 are satisfied if

$$\sum_{T \neq \phi} \exp\text{-}k\left\{R_0(K,\mathbf{Q},T)-B(T)\right\} < 1, \qquad (7)$$

$$M(T) \exp -kB(T) \sum_{S \neq \phi} \exp\text{-}k\left\{R_0(K,\mathbf{Q},S)-B(S)\right\} < 1 \text{ for each non-empty } T, \qquad (8)$$

and $M(T)\exp -kB(T) < 1$ for each non-empty T. (9)

We notice that (8) is redundant as a condition, because (8) is satisfied whenever (7) and (9) are satisfied.

We can therefore express Lemma 2.2.1 in the following weaker but more readily applicable form.

**Lemma 2.2.2.** For any channel $K=(P;X_1,...X_n;Y)$ and any $(M_1,...,M_n,k)$, the point $R=(R_1,...,R_n)$, where $R_i=(1/k)\ln M_i$, belongs to the achievable rate region of met$(K,k,\mathbf{Q},B)$ if

1) $\sum_{T\neq\phi} \exp -k\left\{R_0(K,\mathbf{Q},T)-B(T)\right\} < 1$ and

2) $M(T)\exp -kB(T) < 1$ for each non-empty T. □

Using this lemma and the following definition, we are now in a position to give an inner bound to the achievable rate region of met$(K,k,\mathbf{Q},B)$.

**Definition 2.2.2.**
For any channel $K=(P;X_1,...,X_n;Y)$, any $M=(M_1,...,M_n,k)$, and any $\mathbf{Q}=(\mathbf{Q}_1,...,\mathbf{Q}_n)$ where $\mathbf{Q}_i$ is a p.d. on $X_i^k$, we define

$$\delta(K,M,\mathbf{Q}) = \min_T\{R_0(K,\mathbf{Q},T)-R(T)\},$$

where the minimum is taken over all non-empty subsets of $\{1,...,n\}$, and $R(T)$ is defined as $(1/k)\ln M(T)$ for any subset T of $\{1,..,n\}$. □

**Lemma 2.2.3.** Inner Bound to Achievable Rate Region of met$(K,k,\mathbf{Q},B)$.
For any $K=(P;X_1,...X_n;Y)$ and $M=(M_1,...,M_n,k)$, the point $R=(R_1,...,R_n)$, where $R_i=(1/k)\ln M_i$, belongs to the achievable rate region of met$(K,k,\mathbf{Q},B)$ if $\delta(K,M,\mathbf{Q}) > (2/k)\ln(2^n-1)$, provided that the bias terms are selected such that $B(T)=(R_0(K,\mathbf{Q},T)+R(T))/2$ for each T.

Proof. Suppose that $\delta(K,M,\mathbf{Q}) > (2/k)\ln(2^n-1)$. It suffices to verify that conditions 1) and 2) of Lemma 2.2.2 are satisfied.

1) $$\sum_{T\neq\phi} \exp\text{-}k\left\{R_0(K,\mathbf{Q},T)-B(T)\right\}$$

$$= \sum_{T\neq\phi} \exp\text{-}k\left\{R_0(K,\mathbf{Q},T)-(R_0(K,\mathbf{Q},T)+R(T))/2\right\}$$

$$= \sum_{T\neq\phi} \exp\text{-}k\left\{(R_0(K,\mathbf{Q},T)-R(T))/2\right\}$$

$$\leq \sum_{T\neq\phi} \exp\text{-}k\left\{\delta(K,M,\mathbf{Q})/2\right\}$$

$$= (2^n-1)\exp\text{-}k\left\{\delta(K,M,\mathbf{Q})/2\right\} < 1.$$

The last two steps follow by noting that $2^n-1$ is the number of non-empty subsets of $\{1,..,n\}$, and that $\delta(K,M,\mathbf{Q}) > (2/k)\ln(2^n-1)$.

2) $M(T)\exp\text{-}kB(T)$

$$= M(T)\exp\text{-}k\left\{(R_0(K,\mathbf{Q},T)+R(T))/2\right\}$$

$$= \exp\text{-}k\left\{(R_0(K,\mathbf{Q},T)-R(T))/2\right\} < 1 \text{ for all non-empty } T,$$

since $\delta(K,M,\mathbf{Q})>0$. □

**Lemma 2.2.4.** For all K, $\mathbf{R}_0(K)$ is an inner bound to the achievable rate region of the proposed class of metrics.

Proof. In view of Lemma 2.2.4, it suffices to prove the following statement: For any channel $K=(P;X_1,...,X_n;Y)$ and any point $R=(R_1,...,R_n)$, suppose that there exist $M=(M_1,...,M_n,k)$ and $\mathbf{Q}=(\mathbf{Q}_1,...,\mathbf{Q}_n)$, where $\mathbf{Q}_i$ is a p.d. on $X_i^k$, such that $(1/k)\ln M_i \geq R_i$, $i=1,...,n$, and $\delta(K,M,\mathbf{Q})>0$. Then, there exist $H=(H_1,\dots,H_n,h)$ and $\mathbf{U}=(\mathbf{U}_1,\dots,\mathbf{U}_n)$, where $\mathbf{U}_i$ is a p.d. on $X_i^h$, such that $(1/h)\ln H_i \geq R_i$, $i=1,\dots,n$, and $\delta(K,H,\mathbf{U})>(2/h)\ln(2^n-1)$.

Suppose that M and **Q** satisfy the hypothesis of the above statement. Let **U** be such that $\mathbf{U}_i$ is the $m^{th}$ product of $\mathbf{Q}_i$, $i=1,\dots,n$; i.e, $\mathbf{U}_i$ is a p.d. on $X_i^{mk}$ such that, for each $(\xi_1,\dots,\xi_{mk})\epsilon X_i^{mk}$,
$\mathbf{U}_i((\xi_1,\dots,\xi_{mk}))=\mathbf{Q}_i((\xi_1,\dots,\xi_k))\mathbf{Q}_i((\xi_{k+1},\dots,\xi_{2k}))\cdots\mathbf{Q}_i((\xi_{(m-1)k+1},\dots,\xi_{mk})$.
Let H be such that $H_i=M_i^m$ and $h=mk$.

It is easy to verify that $R_0(K,\mathbf{Q},T)=R_0(K,\mathbf{U},T)$ for all T, and that $\delta(K,M,\mathbf{Q})=\delta(K,H,\mathbf{U})$. So, by simply taking m large enough, we can satisfy $\delta(K,H,\mathbf{U})>(2/h)\ln(2^n-1)$. □

As a corollary to Lemma 2.2.4, we have the main result of this chapter.

**Theorem 2.2.1.** $\mathbf{R}_0(K)$ is an inner bound to $\mathbf{R}(K)$ for all K. □

Complementary Remarks

1) No examples are known for which **R** is strictly larger than $\mathbf{R}_0$. On the other hand, it is not known if $\mathbf{R}_0(K)=\mathbf{R}(K)$ for all K. In the next chapter, it will be shown that $\mathbf{R}_0=\mathbf{R}$ for single-user channels (see §3.2) and also for pairwise reversible channels (see §3.3).

2) At this point, it is natural to ask whether there exists a class of metrics which satisfies the conditions of Theorem 2.1.1 over a set of points larger than $\mathbf{R}_0$. §2.4 will prove that there is no such class.

3) One might also ask whether the metric of Example 1.4.1 (the Fano metric) satisfies the conditions of Theorem 2.1.1 over all (interior) points of $\mathbf{R}_0(K)$ for all K. Assuming that the parameters of the metric are set in the way suggested in Example 1.4.1, the answer is no. A simple counter-example is a (pseudo) two-user channel which is the parallel combination of two independent binary symmetric channels. By choosing

the crossover probabilities of the binary symmetric channels appropriately (one close to 1/2, the other close to 0), one can obtain a situation where the Fano metric has a positive drift (in an ensemble average sense) on each path whose component path for the less noisy subchannel is correct.

4) The proof of Lemma 2.2.4 suggests a method for finding an appropriate metric in any given situation. Suppose, for example, that $K=(P;X_1,\ldots,X_n;Y)$ is the channel and $R=(R_1,\ldots,R_n)$ is the desired rate. We first try to find $M=(M_1,\ldots,M_n,k)$ and $\mathbf{Q}=(\mathbf{Q}_1,\ldots,\mathbf{Q}_n)$, where $\mathbf{Q}_i$ is a p.d. on $X_i{}^k$, such that $(1/k)\ln M_i \geq R_i$ and $\delta(K,M,\mathbf{Q})>(2/k)\ln(2^n-1)$. Supposing that such a pair is found, then the metric $\mathrm{met}(K,k,\mathbf{Q},B)$, with bias $B(T)=(R_0(K,\mathbf{Q},T)+R(T))/2$ for each T, is an appropriate metric for this situation. If we decide to use this metric, then we may select the tree code at random according to the probability measure associated with the ensemble $\mathrm{Ens}(M_1,\ldots,M_n;k;X_1,\ldots,X_n;\mathbf{Q}_1,\ldots,\mathbf{Q}_n)$. There is no guarantee that such a randomly selected code will perform satisfactorily; but the probability that its performance is much worse than average is small.

5) If the stack algorithm is applied to a tree code with parameter $M=(M_1,\ldots,M_n,k)$, each step of the algorithm requires the evaluation of the metric values of $M_1\cdots M_n$ nodes. Ordinarily, one is given a desired rate $R=(R_1,\ldots,R_n)$ and the code parameter $M=(M_1,\ldots,M_n,k)$ is chosen so that $(1/k)\ln M_i \geq R_i$ is satisfied for each $i\in\{1,\ldots,n\}$. From the viewpoint of computational complexity, it is thus preferable to select M so that k is the minimum possible subject to the rate constraints.

If we wish to use $\mathrm{met}(K,k,\mathbf{Q},B)$, with bias $B(T)=(R_0(K,\mathbf{Q},T)+R(T))/2$ for each T, there is an additional constraint that M has to meet, namely, $\delta(K,M,\mathbf{Q})>(2/k)\ln(2^n-1)$. This constraint is unpleasant because it forces k to get large as the desired rate approaches the boundary of $\mathbf{R}_0(K)$. It is not known at present whether a constraint of this type is inherent in multi-user sequential decoding or whether one can find metrics which do not suffer from this problem.

## 2.3. Some Properties of $\mathbf{R}_0$

This section summarizes some of what is known about the $\mathbf{R}_0$ region.

In §1.5, it was shown that $\mathbf{R}_0(K)=\mathbf{R}_0(K,1)$ for any single-user channel K. In the case of multi-user channels, however, this is no longer true; there are channels for which $\mathbf{R}_0(K)\neq\mathbf{R}_0(K,1)$. An example is the two-user M-ary collision channel $K=(P;X_1,X_2;Y)$, where M is an integer greater than 2, $X_1=X_2=\{0,1,\ldots,M-1\}$, $Y=\{e,0,1,\ldots,M-1\}$, and the transition probabilities are as follows. $P(x_1|x_1,0)=P(x_2|0,x_2)=1$ for each $x_1\in X_1$ and $x_2\in X_2$; $P(e|x_1,x_2)=1$ if $x_1\in\{1,\ldots,M-1\}$ and $x_2\in\{1,\ldots,M-1\}$; and, all other transitions have zero probability. We leave it to the reader to verify that the point $((1/2)\ln M$ nats, $(1/2)\ln M$ nats$)$ belongs to $\mathbf{R}_0(K,2)$ but not to $\mathbf{R}_0(K,1)$.

By considering collision channels with larger numbers of users, it can be seen that, for any fixed m, there exists a channel K for which $\mathbf{R}_0(K)\neq \mathbf{R}_0(K,1)\cup\cdots\cup\mathbf{R}_0(K,m)$.

$\mathbf{R}_0$ is convex. This is a simple result of admitting probability distributions over blocks of arbitrary length in the definition of $\mathbf{R}_0$. The convexity of $\mathbf{R}_0$ can be proved by observing that, for any pair, $\mathbf{Q}_1$ and $\mathbf{Q}_2$, of vectors of p.d.'s over block-lengths $k_1$ and $k_2$, and for any pair of integers $m_1$ and $m_2$, the vector of p.d.'s $\mathbf{Q}$, defined as $\mathbf{Q}=\mathbf{Q}_1^{k_2m_1}\mathbf{Q}_2^{k_1m_2}$, satisfies $(m_1+m_2)R_0(\mathbf{Q},T)=m_1R_0(\mathbf{Q}_1,T)+m_2R_0(\mathbf{Q}_2,T)$ for all T. Here, the components of $\mathbf{Q}_1^{k_2m_1}$ are $k_2m_1$-order product forms of the corresponding components of $\mathbf{Q}_1$, and similarly for $\mathbf{Q}_2^{k_1m_2}$. The components of $\mathbf{Q}$ are product forms of the corresponding components of $\mathbf{Q}_1^{k_2m_1}$ and $\mathbf{Q}_2^{k_1m_2}$. The components of $\mathbf{Q}_1^{k_2m_1}$ and $\mathbf{Q}_2^{k_1m_2}$ are thus p.d.'s over block-lengths of $k_1k_2m_1$ and $k_1k_2m_2$, respectively; and the components of $\mathbf{Q}$ are p.d.'s over a block-length of $k_1k_2(m_1+m_2)$.

For any given m, there exists a channel K (e.g., a collision channel) for which $\mathbf{R}_0(K,m)$ is not convex. It is not known, however, if there exists K such that $\mathbf{R}_0(K)\neq$convex-hull$\,\mathbf{R}_0(K,1)$. If there were no such channel, then we would have a characterization of $\mathbf{R}_0$ similar to that for the capacity region.

By using the parallel channels theorem (pp.149-150, [12]), it can be proved that, for any K, i, and m,

$$\max\{R_i : (0,..,R_i,..,0) \in \text{convex-hull}\, \mathbf{R}_0(K,1)\} =$$
$$\max\{R_i : (0,..,R_i,..,0) \in \text{convex-hull}\, \mathbf{R}_0(K,m)\}.$$

This can be seen directly by noting that, if all users, except for user i, are constrained to transmit at rate zero (which means that each such user transmits a fixed sequence), then the situation reduces to the single-user case, for which we know that the stated result holds. This result is useful in that it provides some information about the relative sizes of the regions $\mathbf{R}_0(K,m)$, m=1,2,...

We now prove some inequalities about the $\mathbf{R}_0$ region.

For any K, **Q**, S, and T, if T is a subset of S, then

$$R_0(K,\mathbf{Q},T) \leq R_0(K,\mathbf{Q},S). \tag{1}$$

Proof. Let m be the block-length for **Q**. Now,

$$mR_0(K,\mathbf{Q},S) = -\ln \sum_{\boldsymbol{\eta},\boldsymbol{\xi}_{S^c}} \mathbf{Q}(\boldsymbol{\xi}_{S^c}) \left\{ \sum_{\boldsymbol{\xi}_S} \mathbf{Q}(\boldsymbol{\xi}_S)\sqrt{P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})} \right\}^2$$

$$= -\ln \sum_{\boldsymbol{\eta},\boldsymbol{\xi}_{S^c}} \mathbf{Q}(\boldsymbol{\xi}_{S^c}) \left\{ \sum_{\boldsymbol{\xi}_{S\setminus T}} \mathbf{Q}(\boldsymbol{\xi}_{S\setminus T}) \sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\sqrt{P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})} \right\}^2 \tag{2}$$

$$\geq -\ln \sum_{\boldsymbol{\eta},\boldsymbol{\xi}_{S^c}} \mathbf{Q}(\boldsymbol{\xi}_{S^c}) \sum_{\boldsymbol{\xi}_{S\setminus T}} \mathbf{Q}(\boldsymbol{\xi}_{S\setminus T}) \left\{ \sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\sqrt{P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})} \right\}^2 \tag{3}$$

$$= -\ln \sum_{\boldsymbol{\eta},\boldsymbol{\xi}_{T^c}} \mathbf{Q}(\boldsymbol{\xi}_{T^c}) \left\{ \sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\sqrt{P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})} \right\}^2$$

$$= mR_0(K,\mathbf{Q},T),$$

where (3) follows from (2) by Jensen's inequality:

$$\Big\{\sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}\Big\}^2 \le \sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}).$$

In the proof of (1), if we replace T by the empty set, we obtain the proof of another basic fact, namely, $R_0(K,\mathbf{Q},S)\ge 0$ for all K, $\mathbf{Q}$, and non-empty S.

For any subset of users T, let $\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_T) = \sum_{\boldsymbol{\xi}_{T^c}} \mathbf{Q}(\boldsymbol{\xi}_{T^c})\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})$.

$\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_T)$ is the transition probability that would be observed between the users in set T and the receiver if the users in set $T^c$ collectively transmitted a given symbol $\boldsymbol{\xi}_{T^c}$ with probability $\mathbf{Q}(\boldsymbol{\xi}_{T^c})$. If one is only interested in decoding the messages of the users in a set T, then one may model the remaining users as noise sources and thus obtain a reduced channel. Such schemes will be the subject of Chapter 4. The following inequality is of interest in comparing the achievable rates for the reduced channel with those for the original one.

For any K, $\mathbf{Q}$, and T,

$$-\ln \sum_{\boldsymbol{\eta}} \Big\{\sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_T)}\Big\}^2 \le mR_0(K,\mathbf{Q},T), \tag{4}$$

where m is the block-length for $\mathbf{Q}$.

Proof.

$$mR_0(K,\mathbf{Q},T) = -\ln \sum_{\boldsymbol{\eta},\boldsymbol{\xi}_{T^c}} \mathbf{Q}(\boldsymbol{\xi}_{T^c})\Big\{\sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\sqrt{\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_T)}\ \Big\}^2$$

$$= -\ln \sum_{\boldsymbol{\eta}} \sum_{\boldsymbol{\xi}_{T^c}} \Big\{ \sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\sqrt{\mathbf{Q}(\boldsymbol{\xi}_{T^c})\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}\Big\}^2 \tag{5}$$

$$\ge -\ln \sum_{\boldsymbol{\eta}} \Big\{ \sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\sqrt{\sum_{\boldsymbol{\xi}_{T^c}} \mathbf{Q}(\boldsymbol{\xi}_{T^c})\mathbf{P}(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}\Big\}^2 \tag{6}$$

$$= -\ln \sum_{\eta}\Big\{\sum_{\xi_T} Q(\xi_T)\sqrt{P(\eta\,|\,\xi_T)}\Big\}^2,$$

where (6) follows from (5) by the following inequality.

$$\sum_{\xi_{T^c}}\Big\{\sum_{\xi_T} Q(\xi_T)\sqrt{Q(\xi_{T^c})P(\eta\,|\,\xi)}\Big\}^2 \le \Big\{\sum_{\xi_T} Q(\xi_T)\sqrt{\sum_{\xi_{T^c}} Q(\xi_{T^c})P(\eta\,|\,\xi)}\Big\}^2 \qquad (7)$$

(7) is proved by using Minkowsky's inequality (see inequality h on p.524 in [12]), which states that, for any collection of non-negative real numbers $\{a_{jk}\}$ and any p.d. $\{Q_j\}$,

$$\sum_{k}\Big\{\sum_{j} Q_j\sqrt{a_{jk}}\Big\}^2 \le \Big\{\sum_{j} Q_j\sqrt{\sum_{k} a_{jk}}\Big\}^2 . \ \square$$

In a sense, this inequality confirms the obvious fact that codebook knowledge of all users can be used to improve the achievable rate region in sequential decoding.

## 2.4. A Result on the Method of §2.1

In this section we prove that there is no branchwise additive metric which satisfies the sufficient conditions on achievability of Theorem 2.1.1 at any given point outside $\mathbf{R}_0$. This means that, if there is an achievable point outside $\mathbf{R}_0$, the achievability of that point cannot be shown by using Theorem 2.1.1. This, of course, does not mean that $\mathbf{R}_0$ equals $\mathbf{R}$, the achievable rate region of sequential decoding. Thus, the results of this section are not directly related to sequential decoding, but rather to the limitations of the particular method of §2.1 in terms of proving achievability.

The above result is proved in two steps. First, Theorem 2.4.1 gives an outer bound, for any given metric, to the rate region where the ensemble average of decoding complexity is finite. Then, Lemma 2.4.1 shows that $\mathbf{R}_0$ outer-bounds the outer bound of Theorem 2.4.1 for any given branchwise additive metric.

Let the following be fixed but otherwise completely arbitrary throughout this section: A channel $K=(P;X_1,\ldots,X_n;Y)$, a code parameter $M=(M_1,\ldots,M_n,k)$, a branchwise additive metric $\Gamma$ which can be used in decoding codes over $K$ with parameter $M$, and an ensemble $\mathbf{E}=\text{Ens}(M_1,\ldots,M_n;k;X_1,\ldots,X_n;\mathbf{Q}_1,\ldots,\mathbf{Q}_n)$.

We define $D_L$ to be $E_{\mathbf{e}}D_L(K,\mathbf{e},\Gamma)$ for each $L$, where $E_{\mathbf{e}}$ denotes expectation with respect to the probability measure associated with $\mathbf{E}$. We also define, as usual, $R_i=(1/k)\ln M_i$, $i=1,\ldots,n$; and we let $\gamma$ denote the branch metric for $\Gamma$.

**Theorem 2.4.1.** If $\inf\{\sigma(T,t,K,\gamma,\mathbf{E}):t\geq 0\}>\exp -kR(T)$ for some non-empty subset $T$ of $\{1,\ldots,n\}$, then $D_L$ increases without bound as $L$ increases.

**Lemma 2.4.1.** If $t\geq 0$ and $T$ is a non-empty subset of $\{1,\ldots,n\}$, then

$$-\ln\sigma(T,t,K,\gamma,\mathbf{E})\geq kR_0(K,\mathbf{Q},T).$$

Proof of Lemma 2.4.1. By definition,

$$\sigma(T,t,K,\gamma,E) = \sum_{\boldsymbol{\xi},\boldsymbol{\zeta}_T,\boldsymbol{\eta}} Q(\boldsymbol{\zeta}_T)Q(\boldsymbol{\xi})P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})\,\mathrm{expt}(\gamma(\boldsymbol{\zeta}_T,\boldsymbol{\xi}_{T^c},\boldsymbol{\eta})-\gamma(\boldsymbol{\xi},\boldsymbol{\eta}))$$

$$=\sum_{\boldsymbol{\xi}_{T^c}} Q(\boldsymbol{\xi}_{T^c})\sum_{\boldsymbol{\xi}_T,\boldsymbol{\zeta}_T,\boldsymbol{\eta}} Q(\boldsymbol{\zeta}_T)Q(\boldsymbol{\xi}_T)P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})\,\mathrm{expt}(\gamma(\boldsymbol{\zeta}_T,\boldsymbol{\xi}_{T^c},\boldsymbol{\eta})-\gamma(\boldsymbol{\xi}_T,\boldsymbol{\xi}_{T^c},\boldsymbol{\eta})). \qquad (1)$$

Now,

$$\sum_{\boldsymbol{\xi}_T,\boldsymbol{\zeta}_T,\boldsymbol{\eta}} Q(\boldsymbol{\zeta}_T)Q(\boldsymbol{\xi}_T)P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_T,\boldsymbol{\xi}_{T^c})\,\mathrm{expt}(\gamma(\boldsymbol{\zeta}_T,\boldsymbol{\xi}_{T^c},\boldsymbol{\eta})-\gamma(\boldsymbol{\xi}_T,\boldsymbol{\xi}_{T^c},\boldsymbol{\eta}))$$

$$=\sqrt{\sum_{\boldsymbol{\xi}_T,\boldsymbol{\zeta}_T,\boldsymbol{\eta}} Q(\boldsymbol{\zeta}_T)Q(\boldsymbol{\xi}_T)P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_T,\boldsymbol{\xi}_{T^c})\,\mathrm{expt}(\gamma(\boldsymbol{\zeta}_T,\boldsymbol{\xi}_{T^c},\boldsymbol{\eta})-\gamma(\boldsymbol{\xi}_T,\boldsymbol{\xi}_{T^c},\boldsymbol{\eta}))}$$

$$\sqrt{\sum_{\boldsymbol{\xi}_T,\boldsymbol{\zeta}_T,\boldsymbol{\eta}} Q(\boldsymbol{\zeta}_T)Q(\boldsymbol{\xi}_T)P(\boldsymbol{\eta}\,|\,\boldsymbol{\zeta}_T,\boldsymbol{\xi}_{T^c})\,\mathrm{expt}(\gamma(\boldsymbol{\xi}_T,\boldsymbol{\xi}_{T^c},\boldsymbol{\eta})-\gamma(\boldsymbol{\zeta}_T,\boldsymbol{\xi}_{T^c},\boldsymbol{\eta}))} \qquad (2)$$

$$\geq \sum_{\boldsymbol{\xi}_T,\boldsymbol{\zeta}_T,\boldsymbol{\eta}} Q(\boldsymbol{\zeta}_T)Q(\boldsymbol{\xi}_T)\sqrt{P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_T,\boldsymbol{\xi}_{T^c})P(\boldsymbol{\eta}\,|\,\boldsymbol{\zeta}_T,\boldsymbol{\xi}_{T^c})}, \qquad (3)$$

where (2) follows by reversing the roles of $\boldsymbol{\xi}_T$ and $\boldsymbol{\zeta}_T$, and (3) follows by Cauchy's inequality. (For arbitrary non-negative reals $a_i$, $b_i$, $i=1,\dots,m$, Cauchy's inequality states that $(\Sigma a_i \Sigma b_i)^{1/2} \geq \Sigma\sqrt{a_i b_i}$, with equality iff, for some constant c, $a_i = cb_i$ for all i.)

Substituting (3) into (1), we get

$$\sigma(T,t,K,\gamma,E) \geq \sum_{\boldsymbol{\xi}_{T^c}} Q(\boldsymbol{\xi}_{T^c}) \sum_{\boldsymbol{\xi}_T,\boldsymbol{\zeta}_T,\boldsymbol{\eta}} Q(\boldsymbol{\zeta}_T)Q(\boldsymbol{\xi}_T)\sqrt{P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_T,\boldsymbol{\xi}_{T^c})P(\boldsymbol{\eta}\,|\,\boldsymbol{\zeta}_T,\boldsymbol{\xi}_{T^c})}$$

$= \exp -kR_0(K,\mathbf{Q},T)$, which is the desired result.

Proof of Theorem 2.4.1. Let the nodes at level L be labelled by integers $1,\ldots,M_L$, where $M_L$ denotes the total number of nodes at level L. Let $\Gamma_k{}^i$ denote the value of the metric at the $k^{th}$ node on the path to level-L node i. $\Gamma_k{}^i$ is thus a random variable whose distribution is determined by the source, channel, and ensemble statistics.

For any pair of nodes i and j at level L, let us define
$A(i,j)$ = the event that $\min\{\Gamma_k{}^j : 1 \le k \le L\} > \min\{\Gamma_k{}^i : 1 \le k \le L\}$,
$B(i,j)$ = the event that $\Gamma_k{}^j > \Gamma_k{}^i$ for each k, $1 \le k \le L$, and
$C(i,j)$ = the event that $\Gamma_L{}^j > \Gamma_L{}^i$.

Let $P_i$ denote probabilities conditional on node i being the correct node at level L.

Theorem 2.4.1 follows by the following sequence of inequalities, each of which is justified subsequently.

$$LD_L \ge \sum_{i=1}^{M_L} (1/M_L) \sum_{j=1}^{M_L} P_i(A(i,j)) \tag{4}$$

$$\ge \sum_{i=1}^{M_L} (1/M_L) \sum_{j=1}^{M_L} P_i(B(i,j)) \tag{5}$$

$$\ge \sum_{i=1}^{M_L} (1/(LM_L)) \sum_{j\,:\,\text{type of } j \text{ wrt } i=(T,\ldots,T)} P_i(C(i,j)) \qquad \text{(for any non-empty T)} \tag{6}$$

$$\ge \sum_{i=1}^{M_L} (1/(LM_L)) \sum_{j\,:\,\text{type of } j \text{ wrt } i=(T,\ldots,T)} (c/\sqrt{L})(\inf\{\sigma(T,t,K,\gamma,E) : t \ge 0\})^L \tag{7}$$

$$\geq \exp\{k(L-1)R(T)\}(c/L^{3/2})(\inf\{\sigma(T,t,K,\gamma,E):t\geq 0\})^{L}. \qquad (8)$$

Supposing for a moment that (4)-(8) hold, it immediately follows that, if $\exp kR(T) > \inf\{\sigma(T,t,K,\gamma,E):t\geq 0\}$ for some T, then $D_L$ goes to infinity as L increases. So, the proof will be complete if we prove (4)-(8).

Proof of (4).
If there exists a node i at level L such that $P_i$(i never reaches the stack-top)>0, then $mD_m=\infty$ for all $m\geq L$. So, without loss of generality, we may assume that $P_i$(i never reaches the stack-top)=0 for each node i at level L and each level L.

Let i be the correct node at level L. If A(i,j) occurs, then, by the properties of the stack algorithm, i cannot reach the stack-top before j. But, by assumption, i reaches the stack-top with probability one; it follows that $P_i(A(i,j))$ is a lower bound to the probability that j reaches the stack-top before i, conditional on i being correct. Summing over j, we obtain a lower bound to the expected number of nodes which reach the stack-top before i, conditional on i being correct; averaging over i, we obtain (4).

Proof of (5).
This follows by the fact that B(i,j) is a subset of A(i,j). To see this, suppose that B(i,j) occurs; in other words, suppose that $\Gamma_k{}^j > \Gamma_k{}^i$ for each k, $1\leq k\leq L$. Now, by taking the minimum of the right side, we obtain $\Gamma_k{}^j > \min\{\Gamma_m{}^i : 1\leq m\leq L\}$, which holds for each k. Taking the minimum of both sides of $\Gamma_k{}^j > \min\{\Gamma_m{}^i : 1\leq m\leq L\}$ over k, we see that whenever B(i,j) occurs so does A(i,j); hence, B(i,j) is a subset of A(i,j).

Proof of (6).
We wish to prove that, for any two nodes i and j, if the type of i with respect to j is uniform, i.e., if it equals (T,...,T) for some non-empty subset T of {1,....,n}, then $P_i(B(i,j))\geq(1/L)P_i(C(i,j))$. We do this with the help of the following fact.

Claim. Let $Z_1,\ldots,Z_L$ be iid (independent, identically-distributed) random variables. Let C be the event that $Z_1+\cdots+Z_L>0$. Let B be the event that

$$\sum_{i=1}^{m} Z_i > 0 \qquad \text{for each } m,\ 1\leq m\leq L.$$

Then, $P(B)\geq(1/L)P(C)$.

Proof of the Claim. Suppose that C occurs; that is, suppose that a sample point $\omega$ occurs such that $Z_1(\omega)+\cdots+Z_L(\omega)>0$. Let h be the maximum index such that $Z_1(\omega)+\cdots+Z_h(\omega)=\min\{Z_1(\omega)+\cdots+Z_k(\omega):1\leq k\leq L\}$. Consider the cyclic permutation $Z_{h+1}(\omega),\ldots,Z_L(\omega),Z_1(\omega),\ldots,Z_h(\omega)$; observe that all partial sums for this permutation, namely $Z_{h+1}(\omega)$, $Z_{h+1}(\omega)+Z_{h+2}(\omega)$, and so on, are positive.

So, if $Z_1(\omega)+\cdots+Z_L(\omega)>0$, then there exists a cyclic permutation for which all partial sums are positive. Since there are L cyclic permutations and since each permutation (cyclic or non-cyclic) of a given realization is equally likely to occur, the claim follows.

The proof follows by substituting $(\Gamma_k{}^j - \Gamma_{k-1}{}^j) - (\Gamma_k{}^i - \Gamma_{k-1}{}^i)$ in place of $Z_k$ in the above claim. Notice that the condition that j be of type (T,...,T) with respect to i ensures that the random variables $(\Gamma_k{}^j-\Gamma_{k-1}{}^j)-(\Gamma_k{}^i-\Gamma_{k-1}{}^i)$, $k=1,\ldots,L$, are identically-distributed.

Proof of (7).
We want to prove that, for any L, any non-empty T, and any pair of nodes i and j at level L, if the type of j wrt i is (T,...,T), then

$$P_i(C(i,j))\geq(c/\sqrt{L})(\inf\{\sigma(T,t,K,\gamma,E):t\geq0\})^L, \qquad (9)$$

where c is a constant.

Let $Z_k = (\Gamma_k^j - \Gamma_{k-1}^j) - (\Gamma_k^i - \Gamma_{k-1}^i)$ for each $k=1,\ldots,L$. Note that $Z_1,\ldots,Z_L$ are iid random variables with a moment generating function $\sigma(T,t,K,\gamma,\mathbf{E})$. Now, we have $P_i(C(i,j)) = P_i(Z_1+\cdots+Z_L>0)$; so, $C(i,j)$ is the event that the sum of L iid random variables exceeds zero.

If $Z_k$ has a non-negative expected value (this corresponds to the situation where the metric tends to increase on a branch of type T at least as fast as it does on a correct branch), then $P_i(C(i,j)) \geq 1/2$ and $\inf\{\sigma(T,t,K,\gamma,\mathbf{E}): t \geq 0\} = 1$; so, in this case, (9) is easily satisfied by taking, say, $c=1/2$.

So, without loss of generality, we may assume that the expected value of $Z_k$ is negative, in which case, (9) follows directly from the asymptotic form of the Chernoff bound, as given by equations 5.4.23 and 5.4.24 of [12].

Proof of (8).
(8) follows from (7) by noting that $\exp k(L-1)R(T)$ is a lower bound on the number of nodes at level L which are of type $(T,\ldots,T)$ wrt (any given) level-L node i. (Also note that $\exp kLR(T)$ is larger than the number of nodes in question.) This completes the proof of Theorem 2.4.1.

# Chapter 3

# OUTER BOUNDS TO THE ACHIEVABLE RATE REGION OF SEQUENTIAL DECODING

## 3.1. A Basic Lemma

**Definition 3.1.1.**
For any channel $K=(P;X_1,\ldots,X_n;Y)$ and any block code $\mathbf{f}$ over K with block length N and codewords $\mathbf{f}(1),\ldots,\mathbf{f}(M)$, define

$$\lambda(K,\mathbf{f})=(1/M)\sum_{i=1}^{M}\sum_{j=1}^{M} P(B(i,j)\,|\,\mathbf{f}(i)),$$

where, for each i and j,

$$B(i,j) = \begin{cases} \{\eta\epsilon Y^N : P(\eta\,|\,\mathbf{f}(i)) \le P(\eta\,|\,\mathbf{f}(j))\} & \text{if } i\neq j, \\ \phi & \text{if } i=j. \;\square \end{cases}$$

$\lambda(K,\mathbf{f})$ is the expected number of incorrect codewords which are at least as likely as the correct codeword conditional on the received word $\eta$, assuming that each codeword is a priori equally likely. $\lambda(K,\mathbf{f})$ will be used in lower-bounding the expected computation in sequential decoding. The link between block codes and sequential decoding is established by Lemma 3.1.1, which will be given after developing some concepts.

**Definition 3.1.2.**
For any channel K, any tree code $\mathbf{e}$ over K, and any positive integer t, define $\Lambda(K,\mathbf{e},\Gamma,t)$ as the expected number of nodes which reach the stack-top before the correct node at level t, assuming that the stack algorithm is used with $\Gamma$ as its metric, and that a priori each path is equally likely to be the correct one.

For any tree code $\mathbf{e}$ and any positive integer t, let $\mathbf{e}(t)$ denote the block code obtained by truncating $\mathbf{e}$ at level t. □

For the purposes of this chapter, it is necessary to state explicitly the tie-breaking rule for ordering those nodes in the stack which have equal metric values. The rule that we shall use is based on the following lexicographical order on the set of nodes.

In our notation, a node u(..j) is associated with a vector (u(1),....,u(j)), where each u(h), $1 \le h \le j$, belongs to a common set, say S. Any ordering relation on the elements of S induces a lexicographical order on the nodes: For any pair of nodes u(..j) and v(..h), u(..j) preceeds v(..h) iff, for some i, $0 \le i \le j-1$, u(..i)=v(..i) and u(i+1) preceeds v(i+1) with respect to the order on S.

We shall assume throughout this chapter that nodes in the stack with equal metric values are ordered in the above lexicographical order. Our interest in the details of the tie-breaking rule is for purposes of precision (and correctness) in the following proofs. For practical purposes, any tie-breaking rule should be as good as any other.

**Lemma 3.1.1.** $\Lambda(K,\mathbf{e},\Gamma,t) \ge (1/2)\lambda(K,\mathbf{e}(t))$. (1)

Remark. Observe that $\lambda(K,\mathbf{e}(t))$ is the expected number of level-t nodes which, conditional on the first t blocks of the received sequence, appear at least as likely as the correct node at level t. Lemma 3.1.1 thus implies that the average decoding complexity in sequential decoding would be minimized if the stack algorithm were able to explore the nodes at any given level t in the same order as they are ordered with respect to their a posteriori likelihoods conditional on the first t blocks of the received sequence. Of course, no sequential decoder can actually do this. So the analysis in this chapter can be seen as an attempt to lower-bound the average decoding complexity of sequential decoding by that of an optimum, but unrealizable, sequential decoder.

Proof. Let $K=(P;X_1,...,X_n;Y)$ be a channel and **e** be a tree code for K with parameter $(M_1,...,M_n,k)$. Consider the situation where the stack algorithm is used in decoding **e** with a metric $\Gamma$.

Let the level-t nodes in $\mathbf{e}$ be labelled by integers 1,...,M(t), where M(t) is the total number of nodes at level t, namely $M(t)=(M_1 \cdots M_n)^t$. Let $\mathbf{e}(t,i)$ denote the encoded sequence for the $i^{th}$ level-t node in $\mathbf{e}$. We shall regard $\mathbf{e}(t,i)$ also as the $i^{th}$ codeword of $\mathbf{e}(t)$.

Claim.

$$\Lambda(K,\mathbf{e},\Gamma,t) \geq (1/M(t)) \sum_{i=1}^{M(t)} \sum_{j=1}^{M(t)} P(A(i,j) \,|\, \mathbf{e}(t,i)). \tag{2}$$

where, by definition, for each pair of distinct level-t nodes i and j,

$A(i,j)=\{\eta\epsilon Y^{kt}$: i cannot reach the stack-top before j given that $\eta$ is the first t blocks of the received sequence$\}$;

and for each level-t node i, $A(i,i)=\phi$.

The definition of A(i,j) would not be meaningful if the stack algorithm (equipped with the lexicographical order discussed above) did not have the property that, given any two nodes at level t, in order to determine which of them reaches the stack-top first, if any reaches it at all, we need to know only the first t blocks of the received sequence. In other words, given a node, the first t blocks of the received sequence, in general, do not tell us if that node reaches the stack-top; but given any two nodes, they tell which of the nodes cannot reach the stack-top before the other.

An explicit characterization of A(i,j) can be given as follows. For any level-t node i, let $\min\Gamma(i,\eta)$ be the minimum of the metric values of the nodes on the path to node i, given that $\eta\epsilon Y^{kt}$ is received. Now, for any two distinct level-t nodes i and j, and any $\eta\epsilon Y^{kt}$,

$\eta\epsilon A(i,j)$ if $\min\Gamma(j,\eta)>\min\Gamma(i,\eta)$ or if $\min\Gamma(j,\eta)=\min\Gamma(i,\eta)$ and j preceeds i with respect to the lexicographical order;

$\eta\epsilon A(j,i)$ otherwise.

Thus, $A(i,j)$ and $A(j,i)$ are complementary sets (in $Y^{kt}$), a fact which will be used in what follows.

Proof of the Claim.
If the probability that the correct node at level t never reaches the stack-top is positive, then $\Lambda(K,\mathbf{e},\Gamma,t)$ is infinite. So, without loss of generality, we may assume that the code and the metric are such that the correct node at level t reaches the stack-top with probability one.

Suppose that node i is the correct node at level t. Let j be some other level-t node. Since i, being the correct node, reaches the stack-top with certainty, the probability that j reaches the stack-top before i equals $P(A(i,j)\,|\,\mathbf{e}(t,i))$. Thus,

$$\sum_{j=1}^{M(t)} P(A(i,j)\,|\,\mathbf{e}(t,i)) \tag{3}$$

is the expected number of level-t nodes which reach the stack-top before node i, conditional on i being correct. Averaging (3) over i, we obtain (2), thus concluding the proof of the claim.

Now, the proof of Lemma 3.1.1 is completed as follows.

$$2\Lambda(K,\mathbf{e},\Gamma,t) \geq (1/M(t)) \sum_{i=1}^{M(t)} \sum_{j=1}^{M(t)} P(A(i,j)\,|\,\mathbf{e}(t,i)) + P(A(j,i)\,|\,\mathbf{e}(t,j)) \tag{4}$$

$$\geq (1/M(t)) \sum_{i=1}^{M(t)} \sum_{\substack{j=1 \\ j\neq i}}^{M(t)} \sum_{\eta\epsilon Y^{kt}} \min\{P(\eta\,|\,\mathbf{e}(t,i)),P(\eta\,|\,\mathbf{e}(t,i))\} \tag{5}$$

$$\geq (1/2M(t)) \sum_{i=1}^{M(t)} \sum_{j=1}^{M(t)} P(B(i,j)\,|\,\mathbf{e}(t,i)) + P(B(j,i)\,|\,\mathbf{e}(t,j)) \tag{6}$$

$$= (1/M(t)) \sum_{i=1}^{M(t)} \sum_{j=1}^{M(t)} P(B(i,j) \mid \mathbf{e}(t,i))$$

$$= \lambda(K,\mathbf{e}(t)).$$

Here, (5) follows from (4) by the complementarity of A(i,j) and A(j,i) for $i \neq j$; in (6) we divide by 2 to account for the fact that, for $i \neq j$, B(i,j) and B(j,i) have in common those $\eta$ for which $P(\eta \mid \mathbf{e}(t,i)) = P(\eta \mid \mathbf{e}(t,j))$. □

The following sections of this chapter are devoted to finding outer bounds to the achievable rate region of sequential decoding (to be exact, of the stack algorithm with the particular tie-breaking rule described above) in various situations. These bounds are based on the fact that, if $\lambda(K,\mathbf{e}(t))$ grows without bound as t increases, then by Lemma 3.1.1, the average complexity of sequential decoding must, too, be unbounded.

## 3.2. The Cut-off Rate of Single-User Channels

The main result of this section is the proof that $R_0(K)$ is the cut-off rate of sequential decoding for any single-user discrete memoryless channel (DMC) K. This proof relies heavily on certain results about sphere-packing lower bounds to the probability of decoding error for block codes, which we review in the following subsection.

### 3.2.1. Sphere-Packing Lower Bounds

Probabilities of Error

Let $K=(P;X;Y)$ be a DMC and let $\mathbf{f}$ be a block code for this channel with rate R, block length N, and number of codewords M ($M=e^{NR}$). Denote the codewords of $\mathbf{f}$ by $\mathbf{f}(1),\ldots,\mathbf{f}(M)$. Let $\mathbf{d}=(Y_1,\ldots,Y_M)$ be a decoder for $\mathbf{f}$. Here, $Y_1,\ldots,Y_M$ are disjoint sets whose union is $Y^N$, and the decoder decides in favor of message i if the received word belongs to $Y_i$.

$P(Y_i^c \mid \mathbf{f}(i))$ is then the probability of decoding error for message i.

The average probability of decoding error is defined as

$$P_e(K,\mathbf{f},\mathbf{d}) = (1/M) \sum_{i=1}^{M} P(Y_i^c \mid \mathbf{f}(i)).$$

The maximum probability of decoding error is defined as

$$P_{e,max}(K,\mathbf{f},\mathbf{d}) = \max_{1 \leq i \leq M} P(Y_i^c \mid \mathbf{f}(i)).$$

$P_e(K,M,N)$ is defined as the minimum of $P_e(K,\mathbf{f},\mathbf{d})$ over all codes $\mathbf{f}$ with M codewords and block length N, and all decoders $\mathbf{d}$.

We shall give lower bounds to $P_e(K,\mathbf{f},\mathbf{d})$ and $P_{e,max}(K,\mathbf{f},\mathbf{d})$; but first more definitions are needed.

Compositions and the Sphere-Packing Exponent Function

A p.d. Q on X is said to be the composition of $\boldsymbol{\xi}\epsilon X^N$ iff, for each $\xi\epsilon X$, $NQ(\xi)$ equals the number of times $\xi$ appears in $\boldsymbol{\xi}$. A p.d. Q on X is said to be a composition class on $X^N$ iff $NQ(\xi)$ is integer-valued for each $\xi\epsilon X$. A code is called a fixed-composition code iff all of its codewords have the same composition.

For any channel K=(P;X;Y), any positive real number R, and any p.d. Q on X, the sphere-packing exponent, $E_{sp}(K,R,Q)$, is defined as

$$E_{sp}(K,R,Q) = \min_{V} D(V|P|Q)$$

$$\text{subject to } V(\eta|\xi) \geq 0 \text{ for each } \xi\epsilon X \text{ and } \eta\epsilon Y,$$

$$\sum_{\eta\epsilon Y} V(\eta|\xi) = 1 \text{ for each } \xi\epsilon X, \text{ and } R \geq I(Q;V).$$

Here, $D(V|P|Q) = \sum_{\xi\epsilon X}\sum_{\eta\epsilon Y} Q(\xi)V(\eta|\xi)\ln\{V(\eta|\xi) / P(\eta|\xi)\}$ and

$$I(Q;V) = \sum_{\xi\epsilon X}\sum_{\eta\epsilon Y} Q(\xi)V(\eta|\xi)\ln\{V(\eta|\xi) / \sum_{\zeta\epsilon X} Q(\zeta)V(\eta|\zeta)\}.$$

**Lemma 3.2.1.** Sphere-Packing Lower Bound for Fixed-Composition Codes Let K=(P;X;Y) be a channel, N be a positive integer, Q be a composition class on $X^N$, R and $\delta$ be positive real numbers. Let **f** be a fixed-composition code with composition Q, block length N, and number of codewords M. Suppose that $M \geq \exp N(R+\delta)$. Let **d** be a decoder for **f**. Then, for any such K, **f**, and **d**,

$$P_{e,max}(K,\mathbf{f},\mathbf{d}) > (1/2)\exp -N\{E_{sp}(K,R,Q)\ (1+\delta)\}$$

provided that $N > N_0(\delta,|X|,|Y|)$, for some function $N_0$. □

This is Theorem 5.3 in [16], and hence, its proof will be omitted here.

The explicit form of the function $N_0$ is not important for our purposes (it can be found in [16]); what is important is the fact that $N_0$ does not depend on Q.

**Corollary 3.2.1.**

For any K, N, Q, R, $\delta$, **f**, M, and **d** as in Lemma 3.2.1, satisfying the additional condition $(M-1)/2 \geq \exp N(R+\delta)$,

$$P_e(K,\mathbf{f},\mathbf{d}) > (1/4)\exp -N\{E_{sp}(K,R,Q)(1+\delta)\},$$

provided that $N > N_0(\delta,|X|,|Y|)$.

Proof. We make use of an idea of [17] (Eq. 4.41): If $(1/N)\ln[(M-1)/2] > R+\delta$ and $N > N_0(\delta,|X|,|Y|)$, then, by Lemma 3.2.1, at least half of the codewords of **f** have probability of error greater than

$$(1/2)\exp -N\{E_{sp}(K,R,Q)(1+\delta)\};$$

the corollary follows by noting that such codewords have probability of occurrence of at least one half. □

**Lemma 3.2.2.** <u>Some Properties of</u> $E_{sp}(K,R,Q)$

For fixed K=(P;X;Y) and Q, $E_{sp}(K,R,Q)$ is a convex, non-increasing function of $R \geq 0$. $E_{sp}(K,R,Q)$ is positive for $0 \leq R < I(Q;P)$ and zero for $R \geq I(Q;P)$. There is a rate $R_c(K,Q)$, called the <u>critical rate</u> for Q, which has the property that

$$R_c(K,Q) + E_{sp}(K,R_c(K,Q),Q) = E_0(K,Q), \text{ where, by definition,}$$

$$E_0(K,Q) = \min_{V} D(V|P|Q) + I(Q;V)$$

$$\text{s.t. } V(\eta|\xi) \geq 0 \text{ for all } \xi\epsilon X \text{ and } \eta\epsilon Y,$$

$$\sum_{\eta\epsilon Y} V(\eta|\xi) = 1 \text{ for all } \xi\epsilon X.$$

The assertions of this lemma are contained in Lemma 5.4 and Corollary 5.4 of [16]; hence, their proofs are omitted here.

**Lemma 3.2.3.** For any K and Q, $R_0(K) \geq E_0(K,Q) \geq R_0(K,Q)$.

Proof. We follow the hints given in problem 5.23 of [16]. The dependence of the functions on K will be suppressed in the following proof. First it will be shown that $R_0 \geq E_0(Q)$.

$$E_0(Q) = \min_{V} D(V|P|Q) + I(Q;V) \tag{1}$$

$$= \min_{V,U} \sum_{\xi \epsilon X} \sum_{\eta \epsilon Y} Q(\xi)V(\eta|\xi)\Big\{\ln\{V(\eta|\xi)/P(\eta|\xi)\} + \ln\{V(\eta|\xi)/U(\eta)\}\Big\}, \tag{2}$$

where U is a probability distribution on Y. (2) follows from (1) by noting that

$$I(Q;V) = \min_{U} \sum_{\xi \epsilon X} \sum_{\eta \epsilon Y} Q(\xi)V(\eta|\xi)\ln\{V(\eta|\xi)/U(\eta)\}, \tag{3}$$

which can be proved by considering the difference of the two sides in (3) for fixed U, and then using Jensen's inequality.

Now, note that

$$\sum_{\xi \epsilon X} \sum_{\eta \epsilon Y} Q(\xi)\, V(\eta|\xi)\Big\{\ln\{V(\eta|\xi)/P(\eta|\xi)\} + \ln\{V(\eta|\xi)/U(\eta)\}\Big\}$$

$$= -2 \sum_{\xi \epsilon X} Q(\xi) \sum_{\eta \epsilon Y} V(\eta|\xi) \ln\Big\{\sqrt{P(\eta|\xi)U(\eta)}/V(\eta|\xi)\Big\} \tag{4}$$

$$\geq -2 \sum_{\xi \epsilon X} Q(\xi) \ln\Big\{\sum_{\eta \epsilon Y} V(\eta|\xi)\Big[\sqrt{P(\eta|\xi)U(\eta)}/V(\eta|\xi)\Big]\Big\} \tag{5}$$

$$= -2 \sum_{\xi \epsilon X} Q(\xi) \ln\Big\{\sum_{\eta \epsilon Y} \sqrt{P(\eta|\xi)U(\eta)}\Big\}, \tag{6}$$

where (5) follows from (4) by Jensen's inequality; and equality holds in (5) if V is as follows.

$$V(\eta|\xi) = \frac{\sqrt{P(\eta|\xi)U(\eta)}}{\sum_{\tilde{\eta}\in Y}\sqrt{P(\tilde{\eta}|\xi)U(\tilde{\eta})}}$$

From (1)-(6), it follows that

$$E_0(Q) = \min_{U} -2\sum_{\xi\in X} Q(\xi)\ln\Big\{\sum_{\eta\in Y}\sqrt{P(\eta|\xi)U(\eta)}\Big\}. \qquad (7)$$

So, for any p.d. U on Y,

$$E_0(Q) \le -2\sum_{\xi\in X} Q(\xi)\ln\Big\{\sum_{\eta\in Y}\sqrt{P(\eta|\xi)U(\eta)}\Big\}. \qquad (8)$$

In particular, we may take U in (8) to be

$$U^*(\eta) = \frac{\Big\{\sum_{\xi\in X} Q^*(\xi)\sqrt{P(\eta|\xi)}\Big\}^2}{\sum_{\tilde{\eta}\in Y}\Big\{\sum_{\xi\in X} Q^*(\xi)\sqrt{P(\tilde{\eta}|\xi)}\Big\}^2} \quad \text{for each } \eta\in Y,$$

where Q* is a p.d. that maximizes $R_0(Q)$, i.e., $R_0(Q^*)=R_0$. By Theorem 5.6.5 of [12], Q* has the property that

$$\sum_{\eta\in Y}\sqrt{P(\eta|\xi)}\sum_{\zeta\in X} Q^*(\zeta)\sqrt{P(\eta|\zeta)} \ge \sum_{\eta\in Y}\Big\{\sum_{\zeta\in X} Q^*(\zeta)\sqrt{P(\eta|\zeta)}\Big\}^2 \qquad (9)$$

for each $\xi\in X$, with equality if $Q^*(\xi)>0$.

Substituting U* into (8), we get

$$E_0(Q) \leq -2 \sum_{\xi \epsilon X} Q(\xi) \ln\left\{ \sum_{\eta \epsilon Y} \sqrt{P(\eta \mid \xi) U^*(\eta)} \right\}$$

$$= -2 \sum_{\xi \epsilon X} Q(\xi) \ln \frac{\displaystyle\sum_{\eta \epsilon Y} \sqrt{P(\eta \mid \xi)} \sum_{\zeta \epsilon X} Q^*(\zeta) \sqrt{P(\eta \mid \zeta)}}{\sqrt{\displaystyle\sum_{\eta \epsilon Y} \left\{ \sum_{\zeta \epsilon X} Q^*(\zeta) \sqrt{P(\eta \mid \zeta)} \right\}^2}} \tag{10}$$

$$\leq R_0. \tag{11}$$

(11) follows by the property of Q* expressed in (9). This completes the proof of the first half of the lemma. We now prove that $E_0(Q) \geq R_0(Q)$ for all Q.

$$E_0(Q) = \min_U -2 \sum_{\xi \epsilon X} Q(\xi) \ln\left\{ \sum_{\eta \epsilon Y} \sqrt{P(\eta \mid \xi) U(\eta)} \right\} \tag{12}$$

$$\geq \min_U -2 \ln\left\{ \sum_{\xi \epsilon X} Q(\xi) \sum_{\eta \epsilon Y} \sqrt{P(\eta \mid \xi) U(\eta)} \right\} \tag{13}$$

$$= R_0(Q), \tag{14}$$

where (12) is just a restatement of (7); (13) follows by Jensen's inequality; and (14) follows by substituting the minimizing U, which is

$$U(\eta) = \frac{\left\{ \displaystyle\sum_{\xi \epsilon X} Q(\xi) \sqrt{P(\eta \mid \xi)} \right\}^2}{\displaystyle\sum_{\tilde{\eta} \epsilon Y} \left\{ \sum_{\xi \epsilon X} Q(\xi) \sqrt{P(\tilde{\eta} \mid \xi)} \right\}^2} \qquad \text{for each } \eta \epsilon Y. \; \square$$

**Corollary 3.2.2.** $\max_Q E_0(K,Q) = R_0(K)$ for all K.

Proof. By Lemma 3.2.3,

$$R_0(K,Q) \le E_0(K,Q) \le R_0(K);$$

hence,

$$\max_Q R_0(K,Q) \le \max_Q E_0(K,Q) \le R_0(K).$$

The proof follows by noting that $\max_Q R_0(K,Q) = R_0(K)$.

**Corollary 3.2.3.** $\max_Q R_c(K,Q) \le R_0(K)$ for all K.

Proof. $R_c(K,Q) \le E_0(K,Q)$ by Lemma 3.2.2, and $E_0(K,Q) \le R_0(K)$ by Lemma 3.2.3. Hence, $R_c(K,Q) \le R_0(K)$ for all K and Q.

### 3.2.2. A Lower Bound on $\lambda(K,f)$

**Lemma 3.2.4.** For any $K=(P;X;Y)$, any code $\mathbf{f}$ for K with M codewords and block length N, and any collection of integers $t, M_1, \ldots, M_t$ such that i) $t \ge 1$, ii) $M_i > 1$ for each $i \in \{1,\ldots,t\}$, and iii) $M-1 = \sum_{1 \le i \le t} (M_i - 1)$, one has

$$\lambda(K,\mathbf{f}) \ge P_e(K,M_1,N) + \cdots + P_e(K,M_t,N).$$

Proof. Fix K, $\mathbf{f}$, and $M_1,\ldots,M_t$. Let $\mathbf{f}(1),\ldots,\mathbf{f}(M)$ be the codewords of $\mathbf{f}$. Define

$$B(i,j) = \begin{cases} \{\boldsymbol{\eta} \in Y^N : P(\boldsymbol{\eta} \mid \mathbf{f}(i)) \le P(\boldsymbol{\eta} \mid \mathbf{f}(j))\} & \text{if } i \ne j; \\ \phi & \text{if } i = j. \end{cases}$$

For each $i \in \{1,\ldots,M\}$, define

$$\mathbf{P}_i = \left\{ (S_1,\ldots,S_t) : S_1 \cup \cdots \cup S_t = \{1,\ldots,M\},\ i \in S_j,\ |S_j| = M_j,\ j=1,\ldots,t \right\}.$$

It follows from the definition that, if $(S_1,\ldots,S_t)\epsilon \mathbf{P}_i$, then the sets $S_1,\ldots,S_t$ are mutually disjoint, except for i, which is common to all.

For each subset T of $\{1,\ldots,M\}$, define

$E_i(T)=\left\{\boldsymbol{\eta}\epsilon Y^N\text{: There exists } j\epsilon T \text{ such that } j\neq i \text{ and } P(\boldsymbol{\eta}\,|\,\mathbf{f}(i))\leq P(\boldsymbol{\eta}\,|\,\mathbf{f}(j))\right\}$.

Observe that, for any $\mathbf{S}=(S_1,\ldots,S_t)\epsilon \mathbf{P}_i$ and any $i\epsilon\{1,\ldots,M\}$,

$$\sum_{j=1}^{M} P(B(i,j)\,|\,\mathbf{f}(i)) \geq \sum_{k=1}^{t} P(E_i(S_k)\,|\,\mathbf{f}(i)).$$

So, for any p.d. $W_i$ on $\mathbf{P}_i$,

$$\sum_{j=1}^{M} P(B(i,j)\,|\,\mathbf{f}(i)) \geq \sum_{\mathbf{S}\epsilon \mathbf{P}_i} W_i(\mathbf{S}) \sum_{k=1}^{t} P(E_i(S_k)\,|\,\mathbf{f}(i)).$$

Take $W_i$ as the uniform distribution on $\mathbf{P}_i$ for each $i\epsilon\{1,\ldots,M\}$. Note that the cardinality of $\mathbf{P}_i$ equals $c := (M-1)!/(M_1-1)!\cdots(M_t-1)!$ .
Sum over all i to obtain

$$\lambda(K,\mathbf{C}) = (1/M)\sum_{i=1}^{M}\ \sum_{j=1}^{M} P(B(i,j)\,|\,\mathbf{f}(i)) \geq (1/cM)\sum_{i=1}^{M}\ \sum_{\mathbf{S}\epsilon \mathbf{P}_i}\ \sum_{k=1}^{t} P(E_i(S_k)\,|\,\mathbf{f}(i)).$$

Let $\alpha_k = (1/cM)\sum_{i=1}^{M}\ \sum_{\mathbf{S}\epsilon \mathbf{P}_i} P(E_i(S_k)\,|\,\mathbf{f}(i))$.

Now, $\lambda(K,\mathbf{C})\geq\alpha_1+\cdots+\alpha_t$. Clearly, the proof will be complete if we show that $\alpha_k\geq P_e(K,M_k,N)$.

Define $F(m)=\{D{:}D$ is a subset of $\{1,\ldots,M\}$ with m elements$\}$ and
$F_i(m)=\{D{:}D\epsilon F(m)$ and $i\epsilon D\}$.

$$\alpha_k = (1/cM) \sum_{i=1}^{M} \sum_{\mathbf{S}\in\mathbf{P}_i} P(E_i(S_k)\,|\,\mathbf{f}(i))$$

$$= (1/cM) \sum_{i=1}^{M} \sum_{D\in F_i(M_k)} \sum_{\mathbf{S}\in\mathbf{P}_i : S_k = D} P(E_i(S_k)\,|\,\mathbf{f}(i))$$

$$= (1/cM) \sum_{i=1}^{M} \sum_{D\in F_i(M_k)} P(E_i(D)\,|\,\mathbf{f}(i)) \sum_{\mathbf{S}\in\mathbf{P}_i : S_k = D} 1$$

$$= (1/cM) \sum_{i=1}^{M} \sum_{D\in F_i(M_k)} P(E_i(D)\,|\,\mathbf{f}(i)) \frac{(M-M_k)!\,(M_k-1)!}{(M_1-1)!\cdots(M_t-1)!}$$

$$= \frac{(M-M_k)!\,(M_k-1)!}{M!} \sum_{i=1}^{M} \sum_{D\in F_i(M_k)} P(E_i(D)\,|\,\mathbf{f}(i))$$

$$= \frac{(M-M_k)!\,(M_k-1)!}{M!} \sum_{D\in F(M_k)} \sum_{i\in D} P(E_i(D)\,|\,\mathbf{f}(i))$$

$$\geq \frac{(M-M_k)!\,(M_k-1)!}{M!} \sum_{D\in F(M_k)} M_k P_e(K,M_k,N)$$

$$= P_e(K,M_k,N). \ \square$$

**Corollary 3.2.4.** For any channel K=(P;X;Y), any code **f** for K with block length N and number of codewords M, and any integer H such that $M \geq 2H$, one has $\lambda(K,\mathbf{f}) > (M/2H)P_e(K,H,N)$.

Proof. Under the conditions of the corollary, integers $M_1,\ldots,M_t$ can be found such that $t>(M/2H)$ and $M_i \geq H$, for each i. The result follows from Lemma 3.2.4 by noting that $P_e(K,m_1,N)>P_e(K,m_2,N)$ for any pair of integers $m_1$ and $m_2$ such that $m_1>m_2$.

### 3.2.3. Proof that $R_0$ is the Cut-off Rate

**Lemma 3.2.5.** Let $\mathbf{f}_1,\mathbf{f}_2,..$ be an infinite sequence of block codes for a DMC $K=(P;X;Y)$. Let $N_i=ki$ be the block length of $\mathbf{f}_i$ for each i, where k is some fixed integer. Let $M_i$ be the number of codewords in $\mathbf{f}_i$. Suppose that $M_i>\exp N_i(R_0+\epsilon)$ for each i, where $\epsilon$ is a positive constant independent of i. Then, for all sufficiently large i, $(1/N_i)\ln\lambda(K,\mathbf{f}_i)>\epsilon/8$.

Proof. Let $\mathbf{g}_i$ be a subset of $\mathbf{f}_i$ with a fixed composition and with number of codewords at least as large as $M_i/(1+N_i)^{|X|}$. (There is no problem in assuming that $\mathbf{g}_i$ has this many codewords because $(1+N_i)^{|X|}$ is an upper bound on the number of composition classes on $X^{N_i}$.) Let $L_i$ be the number of codewords in $\mathbf{g}_i$, and let $Q_i$ be the composition of the codewords in $\mathbf{g}_i$.

Note that $\lambda(K,\mathbf{f}_i)\geq(L_i/M_i)\lambda(K,\mathbf{g}_i)$, a fact that will be used later in this proof.

Let $\delta=\epsilon/(8+4R_0(K))$.

It is tedious but conceptually straightforward to see that there is a function $\Omega(\epsilon,K,|X|,|Y|)$ such that for all $i>\Omega$ all of the following conditions hold simultaneously.

1. $(1/N_i)\ln L_i>R_0(K)+\epsilon/2$ (15)
2. $(1/N_i)\ln(L_i/8M_i)>-\epsilon/8$ (16)
3. $N_i>N_0(\delta,|X|,|Y|)$ (17)

4. There exist integers $H_i$ such that

   a) $L_i > 2H_i$ (18)

   b) $R_c(K,Q_i)+\delta < (1/N_i)\ln[(H_i-1)/2]$ (19)

   c) $R_c(K,Q_i)+2\delta > (1/N_i)\ln H_i$. (20)

The $N_0$ in (17) is the same as the $N_0$ in Lemma 3.3.1.

To see that (15) and (16) can be satisfied, recall the assumption on the size of $L_i$. To see how (18)-(20) can be satisfied, first note that, for large i, the right hand sides of (19) and (20) are almost identical; thus, (19) and (20) essentially require that $(1/N_i)\ln H_i$ be between $R_c(K,Q_i)+\delta$ and $R_c(K,Q_i)+2\delta$, a condition which can clearly be satisfied. Now, if $L_i$ are chosen to satisfy (16) and $H_i$ are chosen to satisfy (19) and (20), then, for all i sufficiently large, they also satisfy (18) in view of 1) $R_c(K,Q) \le R_0(K)$ (see Corollary 3.2.3) and 2) the relation $\delta = \epsilon/(8+4R_0(K))$.

Hereafter, suppose that i is larger than $\Omega$. Let $H_i$ be chosen so that (18)-(20) are satisfied. The rest of the proof is a simple consequence of the results established thus far.

$\lambda(K,\mathbf{f}_i) > (L_i/M_i)\lambda(K,\mathbf{g}_i)$ (true in general)

$> (L_i^2/(2M_iH_i))P_e(K,H_i,N_i)$ (by (18) and Corollary 3.2.4)

$> (L_i^2/(8M_iH_i))\exp{-N_i\{E_{sp}(K,R_c(K,Q_i),Q_i)(1+\delta)\}}$ (by (17), (19), and Corollary 3.2.1)

$= (L_i^2/(8M_iH_i))\exp{-N_i\{(1+\delta)[E_0(K,Q_i)-R_c(K,Q_i)]\}}$ (by Lemma 3.2.2)

$\ge (L_i^2/(8M_iH_i))\exp{-N_i\{(1+\delta)[R_0(K)-R_c(K,Q_i)]\}}$ (by Corollary 3.2.2)

(The concept of composition here has no relation to that in the previous section.)

**Lemma 3.3.2.** For any PRC K, and any block code $\mathbf{f}$ for K,

$$\lambda(K,\mathbf{f}) > (1/2)g(N)[-1+\exp -N\delta(K,M,\mathbf{Q})],$$

where N is the block length of $\mathbf{f}$; g(N) is as in Lemma 3.3.1; M is the parameter of $\mathbf{f}$ (if, say, $\mathbf{f}$ is an n-user code, then M is of the form $(M_1,...,M_n,N)$ where $M_i$ is the number of codewords in the $i^{th}$ component code and N is the common block length); $\mathbf{Q}$ is the composition of $\mathbf{f}$. The function $\delta(K,M,\mathbf{Q})$, as defined in §2.2, is the minimum of $R_0(K,\mathbf{Q},T)-R(T)$ over all T, where T is a non-empty subset of $\{1,...,n\}$ and R(T) is the sum, over $i \epsilon T$, of $(1/N)\ln M_i$.

The proof of Lemma 3.3.2 will be given following that of Theorem 3.3.1.

Proof of Theorem 3.3.1.
Let K be a PRC, and $\mathbf{f}$ be a tree code for K with parameter $M=(M_1,...,M_n,k)$, where n denotes the number of users. Let $R=(R_1,...,R_n)$ with $R_i=(1/k)\ln M_i$. Suppose that R does not belong to $\mathbf{R}_0(K)$. We will show that R does not belong to $\mathbf{R}(K)$, either.

Let $\mathbf{f}(i)$ be the block code obtained by truncating $\mathbf{f}$ at level i and let $\mathbf{Q}_i$ be the composition of $\mathbf{f}(i)$. The parameter of $\mathbf{f}(i)$, which is denoted by $M^i$, equals $(M_1{}^i,...,M_n{}^i,ki)$. The rate of $\mathbf{f}(i)$ thus equals $R=(R_1,...,R_n)$. Now, by Lemma 3.3.2,

$$\lambda(K,\mathbf{f}(i)) > (1/2)g(ki)[-1+\exp -ki\,\delta(K,M^i,\mathbf{Q}^i)]$$

$$> (1/2)g(ki)[-1+\exp -ki\,\Delta(K,M)],$$

where, by definition, $\Delta(K,M)=\sup\{\delta(K,M^i,\mathbf{Q}_i):i=1,2,3,...\}$.

$$\geq (L_i/8M_i)\exp N_i\{R_0(K)+\epsilon/2-R_c(K,Q_i)-2\delta-(1+\delta)R_0(K)+(1+\delta)R_c(K,Q_i)\}$$

$$\text{(by (15) and (20))}$$

$$= (L_i/8M_i)\exp N_i\{\epsilon/2-2\delta-\delta R_0(K)+\delta R_c(K,Q_i)\}$$

$$\geq (L_i/8M_i)\exp N_i\{\epsilon/2-2\delta-\delta R_0(K)\} \qquad \text{(since } R_c(K,Q_i)\geq 0\text{)}$$

$$= (L_i/8M_i)\exp N_i\epsilon/4 \qquad \text{(since } \delta = \epsilon/\{8+4R_0(K)\}\text{)}$$

$$> \exp N_i\epsilon/8 \qquad \text{(by (16))}. \ \square$$

**Theorem 3.2.1.** $R_0(K)$ is the cut-off rate of sequential decoding for any single-user DMC K.

Proof. For any single-user DMC K and any (M,k) tree code **e** for this channel, if $(1/k)\ln M>R_0(K)$, then Lemma 3.2.5 implies that $\lambda(K,\mathbf{e}(t))$ increases exponentially in increasing t. Hence, by Lemma 3.1.1, $\Lambda(K,\mathbf{e},\Gamma,t)$, too, increases exponentially in t regardless of what the metric $\Gamma$ is. It follows that rates above $R_0(K)$ are not achievable (in the sense of Def. 1.4.2). □

### 3.3 Proof that $R_0$=R for Pairwise Reversible Channels

A channel $K=(P;X_1,\ldots,X_n;Y)$ is said to be a pairwise reversible channel (PRC) iff for each $\xi_i,\zeta_i \epsilon X_i$, $i=1,\ldots,n$, and $\eta \epsilon Y$,

$$\sum_{\eta \epsilon Y} \sqrt{P(\eta|\xi_1,\ldots,\xi_n)P(\eta|\zeta_1,\ldots,\zeta_n)}\log(P(\eta|\xi_1,\ldots,\xi_n)/P(\eta|\zeta_1,\ldots,\zeta_n))=0.$$

(Here, $0\log 0=0$.)

PRC's were introduced by Shannon, Gallager, and Berlekamp in their study of zero-rate error exponents for block codes [17]. Some examples of PRC's are the two-user OR and erasure channels of §1.5. Our purpose in this section is to prove the following result.

**Theorem 3.3.1.** $\mathbf{R}_0(K)=\mathbf{R}(K)$ for any PRC K.

Recall that $\mathbf{R}_0(K)$ has already been shown to be an inner bound to $\mathbf{R}(K)$ for all K (Theorem 2.2.1). Thus, to prove that $\mathbf{R}_0(K)=\mathbf{R}(K)$ for a given K, it suffices to show that $\mathbf{R}_0(K)$ is an outer bound to $\mathbf{R}(K)$. The following result, taken from [17] without proof, is the key to proving this.

**Lemma 3.3.1.** For any PRC $K=(P;X_1,\ldots,X_n;Y)$, any positive integer N, and any pair of $\boldsymbol{\xi}\epsilon(X_1\times\cdots\times X_n)^N$ and $\boldsymbol{\zeta}\epsilon(X_1\times\cdots\times X_n)^N$,

$$\sum_{\boldsymbol{\eta}\epsilon Y^N} \min\{P(\boldsymbol{\eta}|\boldsymbol{\xi}),P(\boldsymbol{\eta}|\boldsymbol{\zeta})\} > g(N) \sum_{\boldsymbol{\eta}\epsilon Y^N} \sqrt{P(\boldsymbol{\eta}|\boldsymbol{\xi})P(\boldsymbol{\eta}|\boldsymbol{\zeta})}$$

where $g(N)=(1/4)\exp\{\sqrt{2N}\ln P_{min}\}$ and

$P_{min}=\min\{P(\eta|\xi):\eta\epsilon Y, \xi\epsilon(X_1\times\cdots\times X_n), \text{and } P(\eta|\xi)>0\}$.

($P_{min}$ is thus the smallest non-zero transition probability over K.) □

**Definition 3.4.1.** Let **f** be an (M,N) block code over a symbol alphabet X. A p.d. **Q** on $X^N$ is said to be the composition of **f** iff, for each $\boldsymbol{\xi}\epsilon X^N$, $M\mathbf{Q}(\boldsymbol{\xi})$ equals the number of times $\boldsymbol{\xi}$ appears as a codeword of **f**.

Since we assume that R does not belong to $\mathbf{R}_0(K)$, we have $\Delta(K,M)<0$. Therefore, $\lambda(K,\mathbf{f}(i))$ increases exponentially as i increases. This in turn implies, by Lemma 3.1.1, that $\Lambda(K,\mathbf{f},\Gamma,i)$ increases exponentially as i increases, regardless of what the metric is. This means that the expected number of nodes which reach the stack-top before the correct node at level i grows exponentially in increasing i. Hence, R does not belong to $\mathbf{R}(K)$. □

Proof of Lemma 3.3.2.
Let $K=(P;X_1,\ldots,X_n;Y)$ be a PRC and $\mathbf{f}$ be a $M=(M_1,\ldots,M_n,N)$ block code for K. Let $\mathbf{f}_i$ be the component code of $\mathbf{f}$ for user i, $i=1,\ldots,n$. Let the codewords of $\mathbf{f}_i$ be indexed by integers 1 through $M_i$, and the codewords of $\mathbf{f}$ by n-tuples of integers $(m_1,\ldots,m_n)$ where $m_i \epsilon \{1,\ldots,M_i\}$. The words index and message will be used interchangeably in what follows.

The codeword in $\mathbf{f}$ with index $(m_1,\ldots,m_n)$ corresponds to a collection of codewords, namely, codeword $m_i$ from code $\mathbf{f}_i$ for each $i \epsilon \{1,\ldots,n\}$. The codeword with index $m=(m_1,\ldots,m_n)$ will be denoted by $\mathbf{f}(m)$, as usual.

Recall that

$$\lambda(K,\mathbf{f}) = (1/H) \sum_m \sum_{\tilde{m}} P(B(m,\tilde{m} \,|\, \mathbf{f}(m)) \tag{1}$$

where $H=M_1 \cdots M_n$ is the total number of codewords, the summations run through all possible messages for $\mathbf{f}$, and $B(m,\tilde{m})$ is as defined in §3.1.

Now, by Lemma 3.3.1, for any distinct pair of m and $\tilde{m}$,

$$P(B(m,\tilde{m}) \,|\, \mathbf{f}(m)) + P(B(\tilde{m},m) \,|\, \mathbf{f}(\tilde{m})) \geq (g(N)/2) \sum_{\boldsymbol{\eta} \epsilon Y^N} \sqrt{P(\boldsymbol{\eta} \,|\, \mathbf{f}(m)) P(\boldsymbol{\eta} \,|\, \mathbf{f}(\tilde{m}))},$$

where the factor of 1/2 accounts for the fact that $B(m,\tilde{m})$ and $B(\tilde{m},m)$ have in common those $\boldsymbol{\eta} \epsilon Y^N$ for which $P(\boldsymbol{\eta} \,|\, \mathbf{f}(m)) = P(\boldsymbol{\eta} \,|\, \mathbf{f}(\tilde{m}))$.

Summing over all messages,

$$\sum_{m}\sum_{\tilde{m}} P(B_{m,\tilde{m}}\,|\,\mathbf{f}(m)) \geq (g(N)/2) \sum_{m}\ \sum_{\tilde{m}\neq m}\ \sum_{\boldsymbol{\eta}} \sqrt{P(\boldsymbol{\eta}\,|\,\mathbf{f}(m))P(\boldsymbol{\eta}\,|\,\mathbf{f}(\tilde{m}))}$$

$$= (g(N)/2)\,\{-H + \sum_{m}\ \sum_{\tilde{m}}\ \sum_{\boldsymbol{\eta}} \sqrt{P(\boldsymbol{\eta}\,|\,\mathbf{f}(m))P(\boldsymbol{\eta}\,|\,\mathbf{f}(\tilde{m}))}. \qquad (2)$$

This expression will now be simplified.

Let $\mathbf{Q}$ be the composition of $\mathbf{f}$, and $\mathbf{Q}_i$ be that of $\mathbf{f}_i$. The relationship between $\mathbf{Q}$ and $\mathbf{Q}_1,\ldots,\mathbf{Q}_n$ is a simple one: For any collection of $\boldsymbol{\xi}_i \epsilon X_i^N$, $i \epsilon \{1,\ldots,n\}$, $\mathbf{Q}(\boldsymbol{\xi}_1 \times\cdots\times \boldsymbol{\xi}_n) = \mathbf{Q}_1(\boldsymbol{\xi}_1)\cdots\mathbf{Q}_n(\boldsymbol{\xi}_n)$.

The following short-hand notation (which should be familiar by now) will be used in the rest of the proof. For any subset T of $\{1,\ldots,n\}$ and any $\boldsymbol{\xi}_1\times\cdots\times\boldsymbol{\xi}_n$, where $\boldsymbol{\xi}_i \epsilon X_i^N$, $\boldsymbol{\xi}_T$ will denote the collection of $\boldsymbol{\xi}_i$ for $i\epsilon T$; $\boldsymbol{\xi}$ will denote $\boldsymbol{\xi}_1\times\cdots\times\boldsymbol{\xi}_n$; $\mathbf{Q}(\boldsymbol{\xi}_T)$ will denote the product of $\mathbf{Q}_i(\boldsymbol{\xi}_i)$ over all $i\epsilon T$. $P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})$ and $P(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_T,\boldsymbol{\xi}_{T^c})$ will be used interchangeably.

For any message $m=(m_1,\ldots,m_n)$ and any subset T of $\{1,\ldots,n\}$, $\mathbf{T}_m(T)$ will denote the set of messages $\tilde{m}=(\tilde{m}_1,\ldots,\tilde{m}_n)$ for which $m_i=\tilde{m}_i$ for each $i\epsilon T^c$.

Now, we can proceed with the proof.

$$\sum_{m}\ \sum_{\tilde{m}}\ \sum_{\boldsymbol{\eta}} \sqrt{P(\boldsymbol{\eta}\,|\,\mathbf{f}(m))P(\boldsymbol{\eta}\,|\,\mathbf{f}(\tilde{m}))}$$

$$\geq \sum_{m}\ \sum_{\tilde{m}\epsilon\mathbf{T}_m(T)}\ \sum_{\boldsymbol{\eta}} \sqrt{P(\boldsymbol{\eta}\,|\,\mathbf{f}(m))P(\boldsymbol{\eta}\,|\,\mathbf{f}(\tilde{m}))}$$

Here, T is a fixed but arbitrary subset of $\{1,\ldots,n\}$.

$$= \sum_{m} \sum_{\boldsymbol{\zeta}_T} M(T)\, \mathbf{Q}(\boldsymbol{\zeta}_T) \sum_{\boldsymbol{\eta}} \sqrt{P(\boldsymbol{\eta} \,|\, \mathbf{f}(m))P(\boldsymbol{\eta} \,|\, \boldsymbol{\zeta}_T, \mathbf{f}(m)_{T^c})}$$

The summation over $\boldsymbol{\zeta}_T$ should be thought of as one summation for each element in T; the summation corresponding to an element i of T runs through all of $X_i^N$. Now, let M(T) denote the product of all $M_i$ for $i \in T$.

$$= H \sum_{\boldsymbol{\xi}} \mathbf{Q}(\boldsymbol{\xi}) M(T) \sum_{\boldsymbol{\zeta}_T} \mathbf{Q}(\boldsymbol{\zeta}_T) \sum_{\boldsymbol{\eta}} \sqrt{P(\boldsymbol{\eta} \,|\, \boldsymbol{\xi})P(\boldsymbol{\eta} \,|\, \boldsymbol{\zeta}_T, \boldsymbol{\xi}_{T^c})}$$

Recall that H is the total number of codewords in **f**. The summation over $\boldsymbol{\xi}$ runs through all of $(X_1 \times \cdots \times X_n)^N$.

$$= HM(T) \sum_{\boldsymbol{\xi}_{T^c}} \mathbf{Q}(\boldsymbol{\xi}_{T^c}) \sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T) \sum_{\boldsymbol{\zeta}_T} \mathbf{Q}(\boldsymbol{\zeta}_T) \sum_{\boldsymbol{\eta}} \sqrt{P(\boldsymbol{\eta} \,|\, \boldsymbol{\xi})P(\boldsymbol{\eta} \,|\, \boldsymbol{\zeta}_T, \boldsymbol{\xi}_{T^c})}$$

Note that $\mathbf{Q}(\boldsymbol{\xi}) = \mathbf{Q}(\boldsymbol{\xi}_{T^c})\mathbf{Q}(\boldsymbol{\xi}_T)$.

$$= HM(T) \sum_{\boldsymbol{\xi}_{T^c}} \mathbf{Q}(\boldsymbol{\xi}_{T^c}) \sum_{\boldsymbol{\eta}} \sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\sqrt{P(\boldsymbol{\eta} \,|\, \boldsymbol{\xi}_T, \boldsymbol{\xi}_{T^c})} \sum_{\boldsymbol{\zeta}_T} \mathbf{Q}(\boldsymbol{\zeta}_T)\sqrt{P(\boldsymbol{\eta} \,|\, \boldsymbol{\zeta}_T, \boldsymbol{\xi}_{T^c})}$$

$$= HM(T) \sum_{\boldsymbol{\xi}_{T^c}} \mathbf{Q}(\boldsymbol{\xi}_{T^c}) \sum_{\boldsymbol{\eta}} \left\{ \sum_{\boldsymbol{\xi}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\sqrt{P(\boldsymbol{\eta} \,|\, \boldsymbol{\xi})} \right\}^2$$

= $H\exp N(R(T) - R_0(K,\mathbf{Q},T))$, where R(T) is defined as (1/N)lnM(T).

We have thus proved that, for any non-empty subset T of {1,...,n},

$$\sum_{m} \sum_{\tilde{m}} \sum_{\boldsymbol{\eta}} \sqrt{P(\boldsymbol{\eta} \,|\, \mathbf{f}(m))P(\boldsymbol{\eta} \,|\, \mathbf{f}(\tilde{m}))} \geq H\exp N(R(T) - R_0(K,\mathbf{Q},T)). \qquad (3)$$

It follows that

$$\sum_{m} \sum_{\tilde{m}} \sum_{\boldsymbol{\eta}} \sqrt{P(\boldsymbol{\eta} \,|\, \mathbf{f}(m))P(\boldsymbol{\eta} \,|\, \mathbf{f}(\tilde{m}))} \geq H \exp N(\max_{T}\{R(T)-R_0(K,\mathbf{Q},T)\}). \qquad (4)$$

Noting that $\max_{T}\{R(T)-R_0(K,\mathbf{Q},T)]\} = -\delta(K,M,\mathbf{Q})$,

$$\sum_{m} \sum_{\tilde{m}} \sum_{\boldsymbol{\eta}} \sqrt{P(\boldsymbol{\eta} \,|\, \mathbf{f}(m))P(\boldsymbol{\eta} \,|\, \mathbf{f}(\tilde{m}))} \geq H \exp N(\max_{T}[R(T)-R_0(K,\mathbf{Q},T)]). \qquad (5)$$

Now, the lemma follows from (1), (2), and (5). □

## 3.4. A Lower Bound to the Ensemble Average of Computation in Sequential Decoding

**Theorem 3.4.1.** For any channel $K=(P;X_1,...,X_n;Y)$, any tree code ensemble $\mathbf{E}=\text{Ens}(M_1,...,M_n;k;X_1,...,X_n;\mathbf{Q}_1,...,\mathbf{Q}_n)$, any metric $\Gamma$ that can be used in sequential decoding of codes in **E**, and any positive integer t,

$$E\Lambda(K,\mathbf{e},\Gamma,t) \geq h(t)\exp -kt\delta(K,M,\mathbf{Q}),$$

where E denotes expectation (here, E is an averaging operation over all codes **e** in **E**); $h(t)=(g/\sqrt{t})+o(1/\sqrt{t})$ where g is a constant and $o(1/\sqrt{t})$ is a quantity which goes to zero faster than $1/\sqrt{t}$ as t goes to infinity; $M=(M_1,...,M_n,k)$; $\mathbf{Q}=(\mathbf{Q}_1,...,\mathbf{Q}_n)$; and, $\delta(K,M,\mathbf{Q})$ is as defined in §2.2.

Remarks

1) There are certain similarities between Theorem 3.4.1 and the results of §2.4, but neither is stronger than the other.

Theorem 2.4.1 and Lemma 2.4.1 together imply that, for branchwise additive metrics, the method of §2.1 cannot be used to prove the achievability of any point outside $\mathbf{R}_0$. The result here is much stronger: Theorem 3.4.1 states that the inability to prove achievability outside $\mathbf{R}_0$ is not due to a shortcoming of the particular method employed in §2.1, neither is it due to the restriction of the metrics to branchwise additive ones. It is because random-coding arguments over the class of ensembles we are considering in this thesis can not yield any achievable points outside $\mathbf{R}_0$; in this respect, the method of §2.1 can not be improved.

Theorem 2.4.1 gives an outer bound to what can be shown to be achievable by a given branchwise additive metric by using the method of §2.1. Theorem 3.4.1, on the other hand, implicitly deals only with the best possible metric.

2) In the one-user case, a result similar to Theorem 3.4.1 was proved by Gallager in a different context [18].

Proof of Theorem 3.4.1.

In view of Lemma 3.1.1, it is sufficient to prove that

$$E\lambda(K,\mathbf{e}(t)) \geq h(t)\exp -kt\delta(K,M,\mathbf{Q}).$$

Here, $\mathbf{e}(t)$ is the block code obtained by truncating the tree code $\mathbf{e}$ at level t, as defined in §3.1. We associate messages for $\mathbf{e}(t)$ with level-t nodes in $\mathbf{e}$. Now, by definition,

$$\lambda(K,\mathbf{e}(t)) = (1/M(t)) \sum_{u(..t)} \sum_{\tilde{u}(..t)} P(B(u(..t),\tilde{u}(..t))\,|\,\mathbf{e}u(..t)), \qquad (1)$$

where $M(t)$ is the total number of codewords in $\mathbf{e}(t)$, i.e., $M(t)=(M_1\cdots M_n)^t$; the sums are over all level-t nodes in $\mathbf{e}$; $\mathbf{e}u(..t)$ denotes the codeword in $\mathbf{e}(t)$ for message (node) $u(..t)$; and $B(u(..t),\tilde{u}(..t))$ is defined as follows.

$$B(u(..t),\tilde{u}(..t)) = \begin{cases} \{\boldsymbol{\eta}\epsilon Y^{kt}: P(\boldsymbol{\eta}\,|\,\mathbf{e}\tilde{u}(..t) \geq P(\boldsymbol{\eta}\,|\,\mathbf{e}u(..t))\} & u(..t)\neq u(..t); \\ \phi & \tilde{u}(..t)=u(..t). \end{cases}$$

Taking expectations of both sides of (1),

$$E\lambda(K,\mathbf{e}(t)) = (1/M(t)) \sum_{u(..t)} \sum_{\tilde{u}(..t)} EP(B(u(..t),\tilde{u}(..t))\,|\,\mathbf{e}u(..t)), \qquad (2)$$

$E\lambda(K,\mathbf{e}(t))$ can thus be lower-bounded by lower-bounding

$$EP(B(u(..t),\tilde{u}(..t))\,|\,\mathbf{e}u(..t)),$$

which is just the probability of the event that

$$\sum_{i=1}^{t} \ln[P(y(i)\,|\,\mathbf{e}\tilde{u}(i))/P(y(i)\,|\,\mathbf{e}u(i))] \geq 0. \qquad (3)$$

Here, $y(i)$ denotes the $i^{th}$ channel output block, and it is regarded as a random variable taking values in $Y^k$. Likewise, $\mathbf{e}$ is regarded as a random variable taking values in $\mathbf{E}$.

The distribution of $Z_i = \ln[P(y(i)|\mathbf{e}\tilde{u}(i)) / P(y(i)|\mathbf{e}u(i))]$ depends on the type of $\tilde{u}(..i)$ with respect to $u(..i)$. In order to simplify matters, let us suppose that the type of $\tilde{u}(..t)$ with respect to $u(..t)$ is $(T,...,T)$ for some non-empty subset $T$ of $\{1,...,n\}$. $Z_1,...,Z_t$ are then independent and identically-distributed. So, the probability of the event in (3), which is now just the probability that the sum of $t$ iid random variables, $Z_1,...,Z_t$, have a non-negative sum, can be lower-bounded by using the asymptotic form of the Chernoff bound, as given by equations 5.4.23 and 5.4.24 of [12]. To use the Chernoff bound, we note that the moment-generating function of $Z_1$, $E(\exp sZ_1)$, is as follows.

$$E(\exp sZ_1) = \sum_{\boldsymbol{\eta}} \sum_{\boldsymbol{\xi}_{T^c}} \sum_{\boldsymbol{\xi}_T} \sum_{\boldsymbol{\zeta}_T} \mathbf{Q}(\boldsymbol{\xi}_T)\mathbf{Q}(\boldsymbol{\zeta}_T)\mathbf{Q}(\boldsymbol{\xi}_{T^c})\mathbf{P}(\boldsymbol{\eta} \,|\, \boldsymbol{\xi}_T,\boldsymbol{\xi}_{T^c})^{1-s}\, \mathbf{P}(\boldsymbol{\eta} \,|\, \boldsymbol{\zeta}_T,\boldsymbol{\xi}_{T^c})^{s}$$

where we have used the notation of §3.3.

It can be verified easily that $E(\exp sZ_1)$ is a convex function of $s$ with a minimum at $s=1/2$; thus, the minimum value of $E(\exp sZ_1)$ equals

$$E(\exp(Z_1/2)) = \exp -kR_0(K,\mathbf{Q},T).$$

Now, the Chernoff bound states that

$$\Pr\{Z_1+\cdots+Z_t \geq 0\} \geq H(t)\exp -tkR_0(K,\mathbf{Q},T), \qquad (4)$$

where $H(t)$ is of the form $(\alpha/\sqrt{t}) + o(1/\sqrt{t})$ for some constant $\alpha$. (For the exact form of $H(t)$, see page 130 of [12].)

Note that $\exp\{k(t-1)R(T)\}$, where $R(T)=(1/k)\ln M(T)$, lower-bounds the number of level-$t$ nodes which are of type $(T,...,T)$ with respect to (any given) node $u(..t)$. Thus, it follows from (2), (3), and (4) that

$$E\lambda(K,\mathbf{e}(t)) \geq H(t)\exp\{-kR(T)\}\exp\{kt[R(T)-R_0(K,\mathbf{Q},T)]\}, \qquad (5)$$

which is true for any non-empty subset $T$ of $\{1,...,n\}$.

Now, lower-bounding $H(t)\exp\{-kR(T)$ by $h(t)=H(t)/(M_1\cdots M_t)$ and taking a T in (5) for which

$$R(T)-R_0(K,\mathbf{Q},T)=\max_{S:S \text{ is a non-empty subset of } \{1,\ldots,n\}}\{R(S)-R_0(K,\mathbf{Q},S)\}$$

$$= -\ \delta(K,M,\mathbf{Q}),$$

we obtain $$E\lambda(K,\mathbf{e}(t)) \geq h(t)\exp\{-kt\delta(K,M,\mathbf{Q}), \tag{6}$$

thus concluding the proof.

# Chapter 4

# NON-JOINT SEQUENTIAL DECODING

The sequential decoding procedure that we have been considering in the past chapters - joint sequential decoding (JSD), as it will be called in this chapter - requires a complete knowledge of all tree codes in the system on the part of a single processor. In this section, we shall consider what we call non-joint sequential decoding (NJSD) in which there is a separate sequential decoder for each user, the decoder for any given user working only on that user's tree code. (See Figure 4.1.) Our goal is to examine the achievable rate region of NJSD (to be defined presently) and compare it with that of JSD.

Consider a channel $K=(P;X_1,\ldots,X_n;Y)$ and suppose that user i employs a $(M_i,k)$ tree code $\mathbf{e}_i$, $i=1,\ldots,n$. Let $\mathbf{e}$ denote joint tree code for $\mathbf{e}_1,\ldots,\mathbf{e}_n$. NJSD in this situation uses n sequential decoders. The sequential decoder working on user i's tree code $\mathbf{e}_i$, which we denote by $SD_i$, uses a metric $\Gamma_i$ of the form

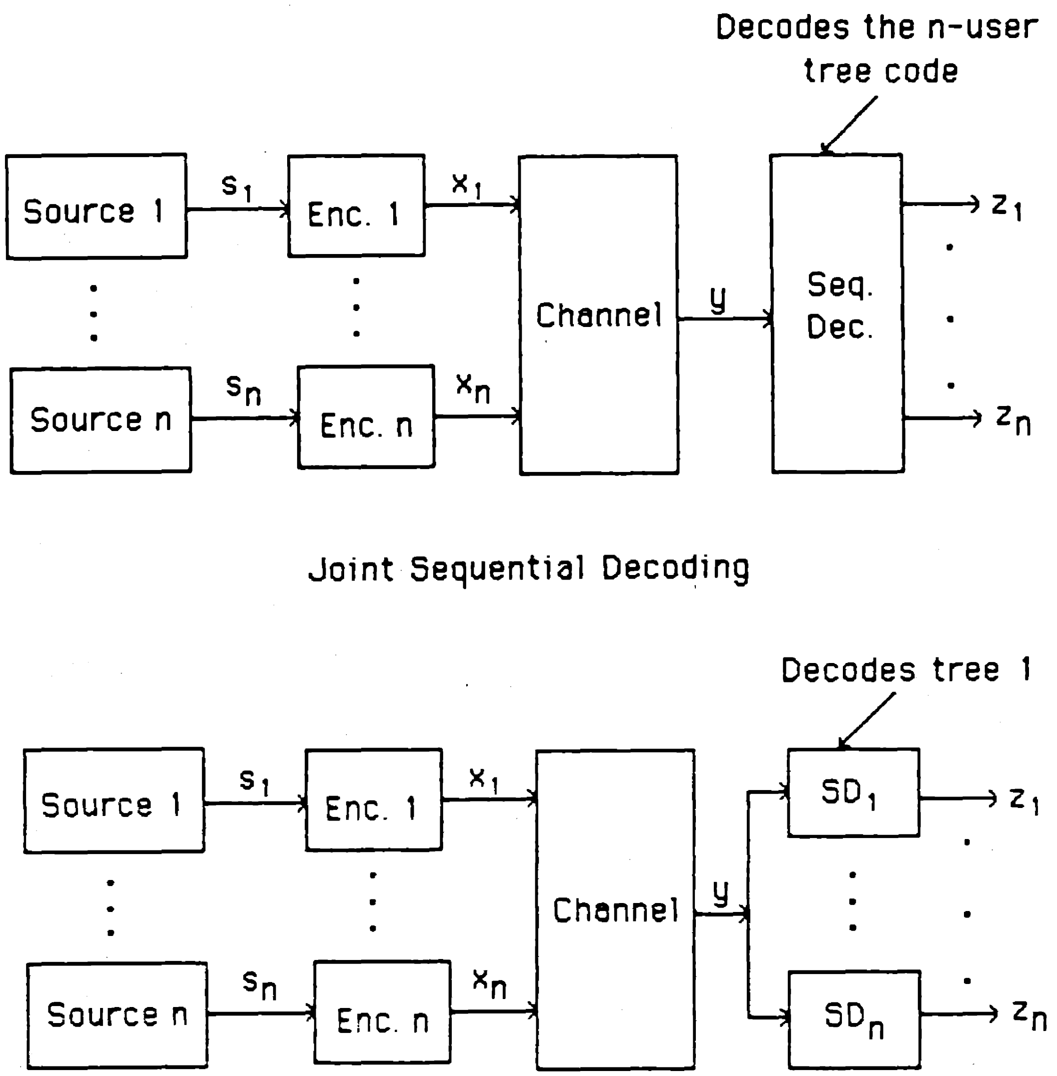
$$\Gamma_i : \bigcup_{h=1}^{\infty} (X_i^{hk} \times Y^{hk}) \longrightarrow [-\infty,\infty).$$

Note that the form of $\Gamma_i$ does not allow $SD_i$ to use any information about the codes of the other users.

Roughly speaking, achievability in NJSD requires that the average decoding complexity be finite for each $SD_i$, $i=1,\ldots,n$. What follows is a formalization of this idea.

## Achievability in Non-Joint Sequential Decoding

Let $C_{i,j}(K,\mathbf{e},\Gamma_i,s,y)$ be the number of nodes in $I_j(s_i)$, the $j^{th}$ incorrect subset of the correct path $s_i$ in $\mathbf{e}_i$, that reach the stack-top of $SD_i$.



Joint Sequential Decoding

Non-Joint Sequential Decoding

Figure 4.1. Joint and Non-Joint Sequential Decoding

(As usual, $s=s_1\times\cdots\times s_n$ denotes the correct path in $\mathbf{e}$, and $y$ denotes the channel output sequence.)

Let $C_{i,j}(K,\mathbf{e},\Gamma_i)$ be the value of $C_{i,j}(K,\mathbf{e},\Gamma_i,s,y)$ averaged over $s$ and $y$. Let $D_{i,L}(K,\mathbf{e},\Gamma_i)=\{C_{i,1}(K,\mathbf{e},\Gamma_i)+\cdots+C_{i,L}(K,\mathbf{e},\Gamma_i)\}/L$.

For large $L$, $D_{i,L}(K,\mathbf{e},\Gamma_i)$ can be interpreted as the average work $SD_i$ has to do to move one step along the correct path $s_i$.

A point $R=(R_1,\ldots,R_n)$ is said to be achievable by NJSD if there exists a finite constant $A$, $A=A(K,R)$, such that for any given $L$ there exist i) a code $\mathbf{e}$ with rate at least as large as $R$, and ii) metrics $\Gamma_1,\ldots,\Gamma_n$ such that

$$D_{1,L}(K,\mathbf{e},\Gamma_1)+\cdots+D_{n,L}(K,\mathbf{e},\Gamma_n)<A.$$

The achievable rate region of NJSD is defined as the closure of the set of all points achievable by NJSD, and is denoted by $\mathbf{R}_{nj}(K)$.

**Theorem 4.1.** $\mathbf{R}_{nj}(K)$ is inner-bounded by $\mathbf{R}_{nj,o}(K)$, which is defined as follows.

$$\mathbf{R}_{nj,o}(K) = \bigcup_{\mathbf{Q}} \mathbf{R}_{nj,o}(K,\mathbf{Q}),$$

where the union is over all $\mathbf{Q}=(\mathbf{Q}_1,\ldots,\mathbf{Q}_n)$ such that $\mathbf{Q}_i$ is a p.d. on $X_i^k$ for some k (k is the same for each i), and for any such $\mathbf{Q}$,

$$\mathbf{R}_{nj,o}(K,\mathbf{Q}) = \{(R_1,\ldots,R_n): 0\le R_i\le R_{nj,o}(K,\mathbf{Q},i) \text{ for each } i=1,\ldots,n\},$$

where $R_{nj,o}(K,\mathbf{Q},i) = -(1/k)\ln \sum_{\boldsymbol{\eta}\in Y^k} \left\{ \sum_{\boldsymbol{\xi}_i\in X_i^k} \mathbf{Q}_i(\boldsymbol{\xi}_i)\sqrt{P_i(\boldsymbol{\eta}\,|\,\boldsymbol{\xi}_i)}\right\}^2,$

and where

$$P_i(\boldsymbol{\eta} \mid \boldsymbol{\xi}_i) = \sum_{\boldsymbol{\zeta}_1} Q_1(\boldsymbol{\zeta}_1) \cdot\cdot \sum_{\boldsymbol{\zeta}_{i-1}} Q_{i-1}(\boldsymbol{\zeta}_{i-1}) \sum_{\boldsymbol{\zeta}_{i+1}} Q_{i+1}(\boldsymbol{\zeta}_{i+1}) \cdot\cdot \sum_{\boldsymbol{\zeta}_n} Q_n(\boldsymbol{\zeta}_n) P(\boldsymbol{\eta} \mid \boldsymbol{\zeta}_1,\ldots,\boldsymbol{\xi}_i,\ldots,\boldsymbol{\zeta}_n).$$

Proof. We use a random-coding argument that is essentially the same as the one in §2.1. Hence, details of the following proof are omitted.

Let $\mathbf{E}=\mathrm{Ens}(M_1,..,M_n;k;X_1,..,X_n;\mathbf{Q}_1,..,\mathbf{Q}_n)$ be an arbitrary ensemble such that $R_{nj,o}(K,\mathbf{Q},j)>(1/k)\ln M_j$ for each $j=1,\ldots,n$. To prove the theorem, it suffices to prove that there exist metrics $\Gamma_1,\ldots,\Gamma_n$ such that the expected value of $D_{1,L}(K,\mathbf{e},\Gamma_1)+\cdots+D_{n,L}(K,\mathbf{e},\Gamma_n)$ over $\mathbf{E}$ is uniformly bounded over all L. Simpler yet, it suffices to prove that, for any given i, there exists $\Gamma_i$ such that the expected value of $D_{i,L}(K,\mathbf{e},\Gamma_i)$ over $\mathbf{E}$ is uniformly bounded over all L. Without loss of generality, we may consider the expected value of $D_{1,L}(K,\mathbf{e},\Gamma_1)$ over $\mathbf{E}$, as we do next.

Let $\mathbf{E}_i=\mathrm{Ens}(M_i;k;X_i{}^k;\mathbf{Q}_i)$, $i=1,\ldots n$. Let E denote expectation over $\mathbf{E}$, and $E_i$ denote expectation over $\mathbf{E}_i$. Now,

$$\begin{aligned} ED_{1,L}(K,\mathbf{e},\Gamma_1) &= E_1\cdots E_n D_{1,L}(K,\mathbf{e},\Gamma_1) \\ &= E_1\{E_2\cdots E_n D_{1,L}(K,\mathbf{e},\Gamma_1)\} \\ &= E_1 D_L(K_1,\mathbf{e}_1,\Gamma_1), \end{aligned}$$

where $D_L$ is as in Def. 1.4.1, and $K_1=(P_1;X_1{}^k;Y^k)$ with $P_1$ as follows. For each $\boldsymbol{\xi}_1 \in X_1{}^k$ and $\boldsymbol{\eta} \in Y^k$,

$$P_1(\boldsymbol{\eta} \mid \boldsymbol{\xi}_1) = \sum_{\boldsymbol{\zeta}_2} \mathbf{Q}_2(\boldsymbol{\zeta}_2) \cdots \sum_{\boldsymbol{\zeta}_n} \mathbf{Q}_n(\boldsymbol{\zeta}_n) P(\boldsymbol{\eta} \mid \boldsymbol{\xi}_1,\boldsymbol{\zeta}_2,\ldots,\boldsymbol{\zeta}_n).$$

If $\Gamma_1$ is taken as $\mathrm{met}(K_1,1,\mathbf{Q}_1,B)$, with bias $B=\{R_0(K_1,\mathbf{Q}_1,\{1\})+\ln M_1\}/2$, then $E_1 D_L(K_1,\mathbf{e}_1,\Gamma_1)$ is uniformly bounded over all L by the results of §2.2.

<u>Remarks</u>

1) The branch metric for $\mathrm{met}(K_1,1,\mathbf{Q}_1,B)$ is as follows.

For each $\boldsymbol{\eta}\in Y^k$, $\boldsymbol{\xi}\in X_1{}^k$,

$$\gamma(\boldsymbol{\xi},\boldsymbol{\eta}) = \ln \frac{\sqrt{P_1(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}}{\sum\limits_{\boldsymbol{\zeta}\in X_1{}^k} \mathbf{Q}_1(\boldsymbol{\zeta})\sqrt{P_1(\boldsymbol{\eta}\,|\,\boldsymbol{\zeta})}} - B.$$

2) $\mathbf{R}_{nj,o}(K,\mathbf{Q})$ is also achievable if $SD_i$ uses the following Fano metric.

For each $\boldsymbol{\eta}\in Y^k$, $\boldsymbol{\xi}\in X_i{}^k$,

$$\gamma_F(\boldsymbol{\xi},\boldsymbol{\eta}) = \ln \frac{P_i(\boldsymbol{\eta}\,|\,\boldsymbol{\xi})}{\omega_i(\boldsymbol{\eta})} - \ln M_i,$$

where $\omega_i(\boldsymbol{\eta}) = \sum\limits_{\boldsymbol{\zeta}\in X_i{}^k} \mathbf{Q}_i(\boldsymbol{\zeta})P_i(\boldsymbol{\eta}\,|\,\boldsymbol{\zeta})$. □

One might think that $\mathbf{R}(K)$ must be at least as large as $\mathbf{R}_{nj}(K)$ for all $K$. This is not true. There is no general inclusion relationship between $\mathbf{R}$ and $\mathbf{R}_{nj}$, as illustrated by the following examples.

**Example 4.1.** A channel for which $\mathbf{R}$ is not contained in $\mathbf{R}_{nj}$.
Consider a channel $K=(P;X_1,X_2;Y_1\times Y_2)$ (Figure 4.2) where $X_1=X_2=Y_1=Y_2=\{0,1\}$ and the transition probabilities are as follows.

| | |
|---|---|
| $P((\xi,0)\,|\,(\xi,0)) = 1-\epsilon$ | $\xi=0,1$; |
| $P((\xi,1)\,|\,(\xi,0)) = \epsilon$ | $\xi=0,1$; |
| $P((\xi,1)\,|\,(\zeta,1)) = 1-\epsilon$ | $\xi=1$ and $\zeta=0$, or $\xi=0$ and $\zeta=1$; |
| $P((\xi,0)\,|\,(\zeta,1)) = \epsilon$ | $\xi=1$ and $\zeta=0$, or $\xi=0$ and $\zeta=1$; |

all other transitions have zero probability.

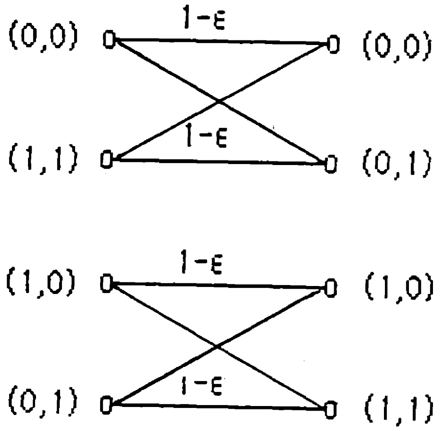


Figure 4.2. The two-user channel of Example 4.1.

Thus, in a sense, the input by the second user selects the channel for the first user. If $(\xi_1,\xi_2)$ is the channel input and $(\eta_1,\eta_2)$ is the channel output at a given time, the transition probabilities from $\xi_2$ to $\eta_2$ are the same as those of a binary symmetric channel with probability of error $\epsilon$. If $\xi_2=0$, one has $\eta_1=\xi_1$ with probability one; if $\xi_2=1$, then one has $\eta_1\neq\xi_1$ with probability one.

In order to decode the message of user 1, it is sufficient to decode that of user 2. So, any two-user rate (1 bit, $R_2$ bits) for which $R_2$ is smaller than the cut-off rate of a binary symmetric channel with probability of error $\epsilon$, namely $1-2\log_2\{\sqrt{\epsilon}+\sqrt{(1-\epsilon)}\}$ bits, is achievable by JSD.

If user 1 transmits at a rate of 1 bit, any decoder that decodes user 1's message correctly must produce (as a by-product) a correct decoding of user 2's message, whether or not we are interested in that message. Therefore, no two-user rate of the form (1 bit, $R_2$ bits) is achievable by NJSD if $R_2$ is positive. More precisely, if ($R_1$ bits, $R_2$ bits) is achievable by NJSD, then $R_1$ must be smaller than $1-\delta(R_2)$ bits, where $\delta$ is a function such that $\delta(R_2)>0$ for $R_2>0$.

**Example 4.2.** A channel for which $\mathbf{R}_{nj}$ is not contained in $\mathbf{R}$.
Consider a channel $K=(P;X_1,X_2;Y_1\times Y_2)$ (Figure 4.3) where $X_1=X_2=\{0,1\}$, $Y_1=Y_2=\{0,1,e\}$, and the transition probabilities are

$$P((\xi_1,\xi_2)\,|\,(\xi_1,\xi_2)) = 1-\epsilon$$

and $$P((e,e)\,|\,(\xi_1,\xi_2)) = \epsilon \quad \text{for each pair of } \xi_1\epsilon X_1,\ \xi_2\epsilon X_2.$$

The output symbol (e,e) is called an <u>erasure</u> and $\epsilon$ is called <u>the erasure probability</u>. We assume that $\epsilon$ satisfies $0<\epsilon<1$.

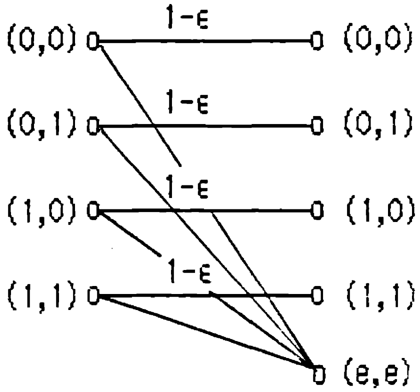


Figure 4.3. The two-user channel of Example 4.2.

An outer bound to $\mathbf{R}(K)$ is found by observing that, if $(R_1,R_2)$ belongs to $\mathbf{R}(K)$, then $R_1+R_2$ cannot be larger than $R_0(K_4)=-\ln\{(1+3\epsilon)/4\}$ nats, where $K_4$ is the <u>single-user</u> quaternary erasure channel, and $R_0(K_4)$ is the cut-off rate of $K_4$.

By Theorem 4.1, $\mathbf{R}_{nj}(K)$ is inner-bounded by $\mathbf{R}_{nj,o}(K,\mathbf{Q})$ for any $\mathbf{Q}$, in particular for $\mathbf{Q}^*=(\mathbf{Q}_1,\mathbf{Q}_2)$ where $\mathbf{Q}_1=\mathbf{Q}_2=$the uniform distribution on $\{0,1\}$. By simple calculation, $\mathbf{R}_{nj,o}(K,\mathbf{Q}^*)=\{(R_1,R_2):0\le R_1\le -\ln[(1+\epsilon)/2]$ nats, $0\le R_2\le -\ln[(1+\epsilon)/2]$ nats$\}$.*

* Actually, $\mathbf{R}_{nj,o}(K,\mathbf{Q}^*)=\mathbf{R}_{nj,o}(K)$; but we do not need this fact here.

Figure 4.4 shows the above bounds. We notice that there are points in the neighborhood of $(-\ln[(1+\epsilon)/2], -\ln[(1+\epsilon)/2])$ which belong to $\mathbf{R}_{nj}(K)$ but not to $\mathbf{R}(K)$, since $-2\ln[(1+\epsilon)/2] > -\ln[(1+3\epsilon)/4]$ for any $\epsilon$, $0<\epsilon<1$.

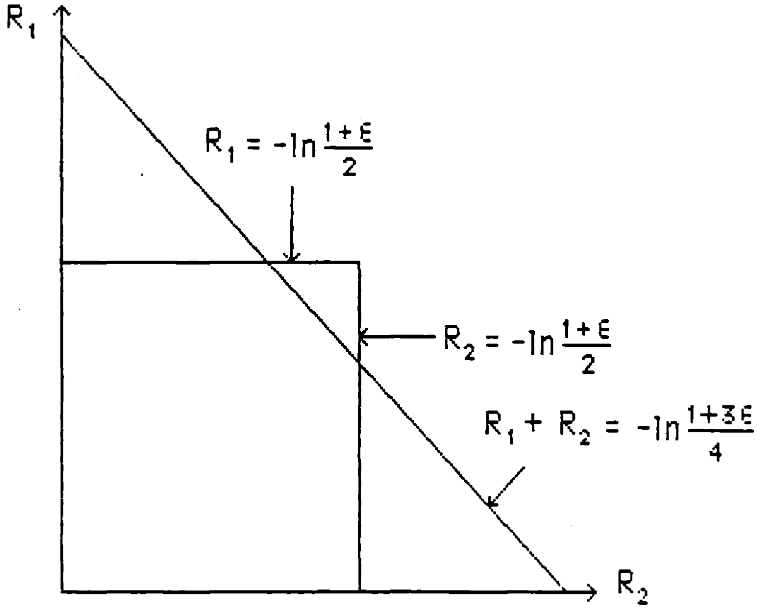

$R_1$
$R_1 = -\ln\frac{1+\epsilon}{2}$
$R_2 = -\ln\frac{1+\epsilon}{2}$
$R_1 + R_2 = -\ln\frac{1+3\epsilon}{4}$
$R_2$

Figure 4.4. Inner and outer bound regions of Example 4.2.

Complementary Remarks on Example 4.2

1) Example 4.2 may seem paradoxical: How can two sequential decoders, neither with a complete view of the system, achieve a point that is not achievable by JSD? This can be explained as follows.

Let $\mathbf{e}_1$ be the code for user 1, and $\mathbf{e}_2$ be the one for user 2. Let $\mathbf{e}$ be the joint tree code for $\mathbf{e}_1$ and $\mathbf{e}_2$. Let k be the number of channel symbols per branch.

The channel output here is a sequence of pairs of symbols: $(\eta_{11},\eta_{21})$, $(\eta_{12},\eta_{22})$, $(\eta_{13},\eta_{23})$,.... We shall denote the sequence $\eta_{11},\eta_{12},\eta_{13},\ldots$ by $y_1$. The first kt elements of $y_1$ will be denoted by $y_1(..t)$. $\eta_{21},\eta_{22},\eta_{23},\ldots$ will be denoted by $y_2$, and the first kt elements of $y_2$ by $y_2(..t)$.

A node $u_1(..t)$ in $\mathbf{e}_1$ is said to be <u>consistent</u> if $\mathbf{e}_1u_1(..t)$ agrees with $y_1(..t)$ in the unerased digits. A node $u_2(..t)$ in $\mathbf{e}_2$ is said to be consistent if $\mathbf{e}_2u_2(..t)$ agrees with $y_2(..t)$ in the unerased digits. A node $u_1(..t) \times u_2(..t)$ in $\mathbf{e}$ is said to be consistent if $u_1(..t)$ and $u_2(..t)$ are consistent. Let $W_1(y_1(..t))$, $W_2(y_2(..t))$, and $W(y_1 \times y_2(..t))$ denote the number of consistent level-t nodes in $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}$, respectively. Note the identity $W=W_1W_2$.

Conditional on $y_1(..t)$, all consistent level-t nodes in $\mathbf{e}_1$ are equally likely to be correct. Thus, $W_1(y_1(..t))/2$ is a lower bound to the number of level-t nodes in $\mathbf{e}_1$ that reach the stack-top of $SD_1$ in NJSD. (The reasoning here is the same as that leading to Lemma 3.1.1.) On the other hand, $W_1$ is an upper bound on the same number of nodes provided that $SD_1$ uses, as we assume that it does, a metric that assigns $-\infty$ to inconsistent nodes, thus preventing them from ever reaching the stack-top.

Similarly, the number of level-t nodes in $\mathbf{e}_2$ that reach the stack-top of $SD_2$ is lower-bounded by $W_2(y_2(..t))/2$ and upper-bounded by $W_2(y_2(..t))$. And the number of level-t nodes in $\mathbf{e}$ that reach the stack-top in JSD is lower-bounded by $W(y_1 \times y_2(..t))/2$ and upper-bounded by $W(y_1 \times y_2(..t))$.

What is of interest for our discussion is that 1) $W(y_1 \times y_2(..t))/2$ is a lower bound to the number of level-t nodes in $\mathbf{e}$ that are processed in JSD, and 2) $W_1(y(..t)) + W_2(y(..t))$ is an upper bound on the number of level-t nodes in $\mathbf{e}_1$ and $\mathbf{e}_2$ that are processed in NJSD. Since $W=W_1W_2$ and both $W_1$ and $W_2$ are at least 1, we have $W/2 \geq (W_1+W_2)/4$. It is thus clear that the complexity of JSD is greater than one fourth the combined complexity of $SD_1$ and $SD_2$. The conclusion that follows is that $\mathbf{R}(K)$ must be a subset of $\mathbf{R}_{nj}(K)$.

2) Example 4.2 was inspired by Massey's paper on sequential decoding for <u>single-user</u> M'ary erasure channels [15]. Massey observed that, if $M=2^L$, then an M'ary erasure channel decomposes into L completely correlated binary erasure channels (BEC), as illustrated in Figure 4.5 for L=2. The component BEC's are completely correlated in the sense that an erasure in one means an erasure in all.

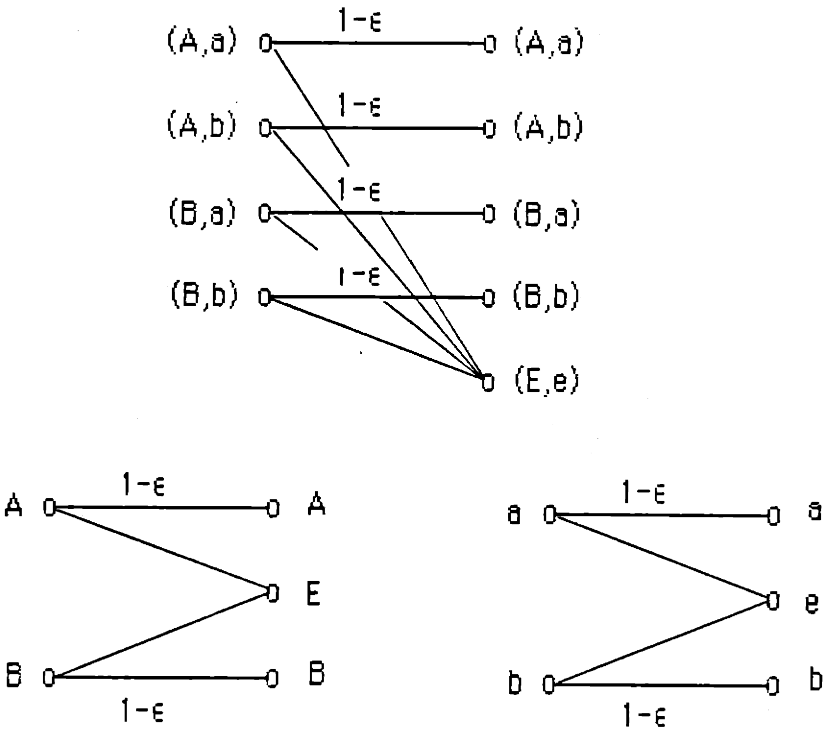


Figure 4.5. Decomposition of a quaternary erasure channel.

The cut-off rate of an M'ary erasure channel with erasure probability $\epsilon$ equals $R_0(M)=-\ln[\epsilon+(1-\epsilon)/M]$ nats. If one uses separate sequential decoders on each component BEC of a $2^L$'ary erasure channel, one can then achieve rates up to $LR_0(2)=-L\ln[(1+\epsilon)/2]$ nats. On the other hand, if sequential decoding is used directly on a $2^L$'ary erasure channel, then the achievable rates are upper-bounded by $R_0(2^L)$. But $LR_0(2)>R_0(2^L)$ for any $\epsilon$, $0<\epsilon<1$. In fact, $LR_0(2)/R_0(2^L)$ goes to infinity as L increases.

An explanation for this apparent peculiarity can be given in exactly the same way as has been done for Example 4.2. The conclusions that can be drawn from Massey's observation are that i) one cannot talk about a cut-off rate for single-user channels without being explicit about the sequential decoding scheme one has in mind, and ii) the cut-off rate of ordinary sequential decoding does not constitute a limit, even in an approximate sense, to rates at which reliable communication is possible in practice.

# Chapter 5

## SUGGESTIONS FOR FURTHER RESEARCH

1. Determine whether $\mathbf{R}(K)=\mathbf{R}_0(K)$ for all K.

2. Determine whether $\mathbf{R}$ is convex. Note that, if $\mathbf{R}$ is indeed convex, proving that it is convex does not necessarily require an explicit characterization of $\mathbf{R}$.

3. Determine whether $\mathbf{R}_0(K)=\text{convex-hull}\,\mathbf{R}_0(K,1)$ for all K.

4. Determine whether strong achievability (Def. 1.4.3) is equivalent to achievability (Def. 1.4.2).

5. The metric of §2.2 requires that, in order to maintain achievability as $\delta$, the distance between the desired rate and the "outer" boundary of $\mathbf{R}_0$, goes to zero, the number of channel symbols per branch increase without bound. Determine whether this requirement, which does not exist in the single-user case, is inherent in multi-user sequential decoding.

A result in this regard, which is not reported in this thesis, is that there is no metric that 1) satisfies the sufficient conditions of §2.1 over a region whose closure is $\mathbf{R}_0$, and 2) does not require the number of symbols per branch go to infinity as $\delta$ goes to zero.

6. A simulation study of multi-user sequential decoding may be done to obtain a better idea about its complexity. The analytical upper bounds of this thesis are useful for determining whether the average complexity is finite; but they are too weak to give an idea about the actual average complexity. Furthermore, a simulation study would provide information about the dynamic behavior of multi-user sequential decoders, a difficult subject to approach analytically.

7. The non-joint sequential decoding scheme of Chapter 4 is just one of several possible approaches to sequential decoding with multiple processors. It would be interesting to see what could be gained by letting the processors exchange information about their current estimates. Such schemes are not likely to be analytically tractable; but that should not deter one from exploring these potentially more powerful schemes.

## REFERENCES

1. C. E. Shannon, " Two-Way Communication Channels," Proc. 4th Berkeley Symp. Math. Stat. Prob., vol. 1, pp. 611-644, 1961.

2. R. Ahlswede, "Multi-Way Communication Channels," Proc. 2nd Int. Symp. on Inf. Th., Tsahkadsor, USSR, 1971.

3. H. Liao, "A Coding Theorem for Multiple Access Communications," Int. Symp. on Inf. Th., Asilomar, 1972.

4. T. Kasami and S. Lin, " Coding for a Multiple Access Channel," IEEE Trans. on Inf. Th., vol. IT-22, pp.129-137, March 1976.

5. D. Slepian and J. K. Wolf, " A Coding Theorem for Multiple Access Channels with Correlated Sources," Bell Syst. Tech. Jour., vol.52, pp.1037-1076, Sept. 1973.

6. R. G. Gallager, " A Perspective on Multiaccess Channels," IEEE Trans. on Inf. Th., vol IT-31, pp.124-142, March 1985.

7. J. M. Wozencraft, Sequential Decoding for Reliable Communications, Tech. Rept. 325, RLE, MIT, Cambridge, MA, 1957.

8. R. M. Fano, " A Heuristic Discussion of Sequential Decoding," IEEE Tran. on Inf. Th., vol. IT-9, pp.64-74, 1963.

9. K. Zigangirov, " Some Sequential Decoding Procedures," Problemy Peredachi Inf., vol. 2, pp.13-25, 1966.

10. F. Jelinek, " A Fast Sequential Decoding Algorithm Using a Stack," IBM Jour. Res. Dev., 13, pp. 675-685, 1969.

11. A. Viterbi and J. Omura, Principles of Digital Communication and Coding, McGraw Hill, 1979.

12. R. G. Gallager, Information Theory and Reliable Communication, John Wiley and Sons Inc., 1968.

13. J. M. Wozencraft and I. M. Jacobs, Principles of Communication Engineering, John Wiley and Sons Inc., 1965.

14. I. M. Jacobs and E. R. Berlekamp, " A Lowerbound to the Distribution of Computation for Sequential Decoding," IEEE Tran. on Inf. Th., vol. IT-13, pp.167-174, April 1967.

15. J. L. Massey, " Capacity, Cutoff Rate, and Coding for a Direct-Detection Optical Channel," IEEE Trans. Comm., vol. COM-29, pp.1615-1621, Nov.1981.

16. I. Csiszar and J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Channels, Academic Press, 1981.

17. C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, " Lowerbounds to Error Probability for Coding on Discrete Memoryless Channels," Parts I and II, Information and Control, 10, pp. 65-103, and pp. 522-552, 1967.

18. R. G. Gallager, " The Random Coding Bound Is Tight for the Average Code," IEEE Trans. on Inf. Th., pp 244-246, March 1973.

19. P. R. Halmos, Measure Theory, Second Printing, Springer-Verlag, 1974.