Diss. ETH No. 12656

Directed Information for Channels with Feedback

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Doctor of Technical Sciences

presented by
GERHARD KRAMER
M.Sc., University of Manitoba, Canada
born April 8, 1970
citizen of Canada

accepted on the recommendation of
Prof. Dr. J. L. Massey, examiner
Prof. Dr. A. J. Han Vinck, co-examiner

1998

Acknowledgments

I thank Professor James L. Massey for giving me the opportunity to do a doctoral dissertation with him, and for the many possibilities he created for me during my time at the ETH. I further thank Jim for the care and effort he took in forming my approach to research. I am especially grateful for his patience and perseverance in doing this, for it took me some time to understand and appreciate his method. But of the many facets of his teaching, it is his admirable character that made the deepest impression on me and gave the most important instruction. Thank you Jim!

I thank Professor George Moschytz for his exemplary guidance of the institute. I thank Professor Han Vinck for acting as the co-referee and for his suggestions.

The atmosphere at ISI was great. Several people contributed to this in a special way. My office-mate, Urs Loher, shared the moments of research excitement and frustration. Maybe CTDMA for the fourth generation? The "last" members of the Information Theory group became close colleagues: Richard De Moliner, Beat Keusch, Zsolt Kukorelly and Jossy Sayir. My only regret is that we didn't start our Friday afternoon "sessions" sooner! Sigi Wyrsch became a good friend (and ski companion) with whom I could discuss much more than research. I thank Carlo Harpes for the many cryptanalytic discussions, his comradeship, and for showing me Luxembourg. Volker Neugebauer helped ease my transition from Canada to Switzerland. The ISI runners, Markus Helfenstein, Fredy Neeser, Felix Tarköy and Christian Waldvogel kept me in shape (at least for a while). I greatly appreciated my contacts with Thomas Mittelholzer and Jürg Ruprecht.

There were several postdocs who added to the charm of my time at ISI. Alex Grant helped start those Friday afternoon sessions, and contributed to them and Appendix 5.B. Anne Canteaut corrected me on my improper use of the word “toque”. Xuduan Lin, Anne and Alex gave our group reason to celebrate by simply being part of it. Kenny Paterson motivated me to start juggling, but not with more than three balls. Lars Knudsen gave me some confidence in my cryptanalytic abilities.

The final word of gratitude must go to my closest loved ones, those who shared my ups and downs by long-distance. My parents Guenter and Martha Kramer listened, advised, and supported me in all my endeavours in Zürich. I am deeply grateful for that.

Abstract

The capacity regions of channels with feedback are investigated. The corresponding information rates are simplified by using the conditional independence of random variables. To establish conditional independence, use is made of d-separation in functional dependence graphs. A weaker condition called fd-separation is introduced and also shown to establish conditional independence in functional dependence graphs. Causally conditioned uncertainty and causally conditioned directed information are defined and used to express the capacity region of the two-way channel and the multiple-access channel with feedback. For both of these channels, rate regions whose points are approachable with error probability approaching zero are developed, including generalizations of Han's rate region for the two-way channel and generalizations of Cover and Leung's rate region for the multiple-access channel with feedback. Finally, feedback strategies are designed for the class of multiple-access channels for which one of the channel inputs is determined by the second channel input and the channel output. These strategies approach all rate points in the capacity region of these channels.

Keywords. feedback, functional dependence, statistical independence, causally conditioned uncertainty, directed information, two-way channel, multiple-access channel, strategy

Kurzfassung

Die Kapazitätsregionen von Kanälen mit Rückkopplung werden untersucht. Dabei werden die Informationsraten durch die Anwendung der bedingten Unabhängigkeit von Zufallsvariablen vereinfacht. Um die bedingte Unabhängigkeit festzustellen, wird d-Separation in Funktional-Abhängigkeits-Graphen benutzt. Eine schwächere Bedingung, die fd-Separation, wird eingeführt, welche ebenfalls die bedingte Unabhängigkeit in Funktional-Abhängigkeits-Graphen garantiert. Die kausal bedingte Unsicherheit und die kausal bedingte gerichtete Information werden definiert und zur Definition der Kapazitätsregionen des Zwei-Weg Kanals und des Mehrfachzugriffs-Kanals mit Rückkopplung benutzt. Für beide dieser Kanäle werden Raten-Regionen konstruiert, deren Raten-Punkte mit einer gegen Null strebenden Fehlerwahrscheinlichkeit angenähert werden können. Diese Raten-Regionen schliessen Verallgemeinerungen von Han's Raten-Region für den Zwei-Weg Kanal sowie Verallgemeinerungen von Cover und Leung's Raten-Region für den Mehrfachzugriffs-Kanal mit Rückkopplung ein. Abschliessend werden Strategien entworfen für die Klasse von Mehrfachzugriffs-Kanälen mit Rückkopplung für welche der eine Kanaleingang durch den zweiten Kanaleingang und den Kanalausgang bestimmt wird. Die vorgeschlagenen Strategien können alle Raten-Punkte in der Kapazitätsregion dieser Kanäle annähern.

Stichworte. Rückkopplung, funktionale Abhängigkeit, statistische Unabhängigkeit, kausalbedingte Unsicherheit, gerichtete Information, Zwei-Weg Kanal, Mehrfachzugriffs-Kanal, Strategie

Contents

Notation and Terminology

Random Variables.

- Random variables are written with uppercase letters, and values they take on with the corresponding lowercase letters.
- The probability distribution of random variables is sometimes denoted by only p when the arguments of p specify the distribution. For example, the value $p_{XY}(x, y)$ of the joint distribution p_{XY} of the random variables X and Y is written simply as $p(x, y)$.

Sequences of Symbols.

- The semi-infinite sequence of symbols $X_1, X_2, X_3, \ldots$ is sometimes denoted by X. The context will make clear whether X is a symbol or a sequence of symbols.
- Subscripts on a symbol are used to denote the symbol's source and/or to denote the symbol's position in a sequence. For example, X_{22} could mean "the output *sequence* of the twenty-second encoder" (although we never consider a problem with 22 encoders) or "the *22nd random variable* in the sequence X", or even "the *second* random variable in the sequence X_2 generated by the *second* encoder." Again, the context will make clear which of these interpretations is in use.
- Superscripts denote finite length sequences of symbols, e.g.,

$$\begin{aligned} x^N &= x_1, x_2, \ldots, x_N \\ y^{2..N} &= y_2, y_3, \ldots, y_N \\ Z_1^{4..N} &= Z_{14}, Z_{15}, \ldots, Z_{1N}. \end{aligned}$$

In the last example the context $Z_1^{4..N}$ makes clear that Z_{14} is interpreted as "the fourth random variable in the sequence Z_1^N".

- We consider only positive integer subscripts for symbols. However, for simplicity of notation, we sometimes allow non-positive subscripts. For example, we write "X^nY^{n-1} for $n \geq 1$" as a shorthand for "X_1 for $n = 1$ and X^nY^{n-1} for $n > 1$."

- An underlined letter denotes a vector of symbols, e.g.,

$$\underline{A} = [A_1\ A_2\ \cdots\ A_N].$$

Uncertainty and Mutual Information. The usual notation for uncertainty, or entropy, and mutual information is used.

- $H(X)$ denotes the uncertainty of a (discrete or continuous) random variable X. $H(X|Y)$ denotes the uncertainty of X conditioned on the random variable Y.

- $I(X;Y)$ denotes the mutual information between the random variables X and Y. $I(X;Y|Z)$ denotes the mutual information between X and Y conditioned on the random variable Z.

Achievability and Approachability. We often use the terms "achievable" and "approachable" for transmission rates, and the term *rate point* for a rate or a rate pair.

- A rate point is said to be *achievable* if one can send at this rate point with arbitrarily small positive error probability.

- A rate point is said to be *approachable* if one can send at rate points arbitrarily close to this rate point with arbitrarily small positive error probability. [The term "close" refers to some natural distance, e.g., Euclidean distance.]

For example, if the capacity of a binary symmetric channel with crossover probability p, $0 < p < 0.5$, is C, then any rate *less* than C is achievable, but one cannot achieve C. On the other hand, C is approachable.

Chapter 1

Introduction

The information theory of channels with feedback poses challenging problems which often have unexpected solutions. Already the first result in this area, that feedback cannot increase the capacity of memoryless channels [1], was surprising. Several other surprising *theoretical* results followed before the lack of *practical* results left researchers with a relatively pessimistic attitude towards the subject. This pessimism is amply summarized in the following excerpt from a review article [2, page 736] in a series commemorating the 25th anniversary of Shannon's landmark paper [3].

> Two of the areas we discussed in this survey may be dying in their present form: adaptive equalization and feedback communication. ... In the case of feedback communications the basic challenges of reality have not been met. Except for certain space communication uses, where noiseless feedback is well approximated, the theoretical results do not show how to improve real systems. Whether this is the fault of the real systems, or their designers, or the theory, or the theorists remains to be seen.

For another statement in the same spirit, see the indented quotation in Section 1.1.

Perhaps because the available theory seemed to lack application to the real problems, Shannon chose feedback as the subject of the first "Shannon Lecture" which he presented at the IEEE International Symposium in Information Theory in 1973. The relatively important role that feedback plays in *real* communications systems would lead one to expect that *theory* should play an important role in designing a system. Even today, "it is rather surprising that, after almost fifty years since its birth as a science, information theory has only rarely been applied to communication systems that incorporate feedback" [4].

1.1 Channels with Feedback

We briefly review in two parts the history of information theory for channels with feedback. This first part discusses single-user communications, while the second part considers multi-user communications.

Single-User Channels

The first result in the information theory of feedback channels is due to Shannon [1] who proved that even complete noiseless feedback of the output symbols of a memoryless channel to the transmitter does *not* increase capacity. This result is surprising, since one might expect that feedback should help.

In fact, feedback does help. Feedback reduces the *complexity* of the encoding and decoding required to achieve a specified error probability. Horstein demonstrated this by designing a simple sequential feedback strategy for approaching the capacity of discrete memoryless channels [5]. Subsequently, Schalkwijk and Kailath [6, 7] published the remarkable result that, for the Gaussian channel, one can achieve a *double*-exponential decrease in the error probability with the number of channel uses, provided only that the transmission rate is below capacity. For channels *without* feedback, only a single-exponential error decrease is possible [8, Chapter 5]. (At about the same time as the work reported in [6, 7], Horstein's strategy was applied to the Gaussian channel by Zigangirov [9]. However, Zigangirov does not mention the double-exponential decrease in error probability.)

A number of subsequent papers refined the results of these early papers. A partial list of the publications during the late 1960's can be found in the review article by Lucky [2]. One interesting result is that the double-exponential decrease of [6, 7] becomes only single-exponential when the maximum power of each transmitted symbol is limited to some finite value [10].

Since the mid-seventies, there have been few papers on information theory for single-user memoryless feedback channels, although there have been several papers on single-user channels *with* memory *and* feedback [11, 12, 13, 14, 15]. We cite a survey article from 1973 which, still today, summarizes this situation well [16, page 259].

> Considerable research was also devoted in the sixties to coding for two-way channels, or, more precisely, to forward channels aided by possible retransmission over a relatively noiseless feedback channel. While, as was shown by Shannon [1], capacity cannot be increased in this way, some remarkable increases in the negative exponent of the error probability appeared possible [5, 6]. Unfortunately, most of the advantage was lost as soon as a reasonable level of noise was introduced in the feedback channel, or a severe peak power constraint was imposed on the forward channel. Thus the area had become less fashionable by the end of the decade, largely because of the lack of a real application.

Multi-User Channels

The information theory of *multi-user* channels with feedback, and of multi-user channels in general, began with another paper of Shannon's. Shannon in [17] introduced the two-way channel and mentions the multiple-access channel (MAC). The capacity *region* of the two-way channel and the two-user MAC is a set of approachable rate pairs (R_1, R_2), where R_1 and R_2 are the rates at which the two users are transmitting their respective data. Calculating the capacity region of the two-way channel is today, almost forty years later, still an unsolved problem. Some partial results can be found in [18, 19, 20, 21, 22, 23, 24]. The capacity region of the MAC *without* feedback was determined in [25, 26, 27].

The first result for multiple-access channels *with* feedback was another surprise, again because of Shannon's original result in [1]. Gaarder and Wolf [28] demonstrated that feedback *can* increase the capacity of a memoryless MAC with feedback. There followed a flurry of activity in the area, culminating in the achievable rate region of Cover and Leung [18, 29]. Subsequently, Willems [30] determined that this rate region *is* the capacity region of a certain class of MACs with feedback. Ozarow [31] found the capacity region of the two-user additive white Gaussian noise MAC with feedback, which region is larger than Cover and Leung's region. Other interesting results can be found in [32, 33, 34, 35].

1.2 Directed Information

The lack of results for calculating the capacity of channels with feedback suggests that one should bring feedback explicitly into the theory. As pointed out by Massey [36], Marko was one of the first to do this. Marko in [37] considered the problem of two coupled systems and introduced the idea of giving *direction* to information. This led to several definitions such as *free information*, *directed transinformation*, *coincidence* and *stochastical degree of synchronization*. Although we will not consider these definitions further, Marko's work motivated the results of this dissertation.

Marko's definition of directed transinformation was later refined by Massey [36], who also introduced the notation $I(X^N \to Y^N)$ for the *directed information* flowing from the length N sequence of random variables X^N to the length N sequence of random variables Y^N. This reformulation of Marko's ideas was the starting point of our work.

1.3 Outline

This dissertation further develops the ideas of Marko and Massey and applies them to channels with feedback. We concentrate on the two multi-user channels originally introduced by Shannon, namely the two-way channel and the multiple-access channel. Other channels such as

the broadcast channel, the interference channel and the relay channel are not dealt with here (see, e.g. [18]). Our aim is to see to what extent directed information can serve as a basis for the information theory of channels with feedback.

This dissertation is organized as follows. We begin in Chapter 2 with a *graphical* technique for establishing the conditional independence of random variables. We make heavy use of this technique in the remaining chapters. Our interest in graphical techniques was motivated by the usefulness of graphs for visualizing how the coding for multi-user channels is done, especially when the feedback causes additional dependencies between random variables.

Chapter 3 states the definitions that we need for the later chapters and reviews the results of Marko [37] and Massey [36]. We introduce *causally conditioned uncertainty*, which is a refinement of Marko's *free information*. A natural consequence of causal conditioning and directed information is what we call *causally conditioned directed information*. It is this quantity that we make most use of in subsequent chapters.

Chapter 4 applies the definitions of Chapter 3 to the two-way channel. We reformulate Shannon's general solution for the capacity region of the two-way channel in terms of causally conditioned directed information. We also show that *concatenated codes* provide a framework for classifying and generalizing the coding techniques of Shannon [17] and Han [22].

Chapter 5 applies the definitions and results of Chapters 3 and 4 to the two-user multiple-access channel with feedback (MAC-FB). We give a general solution for the capacity region of the MAC-FB in terms of causally conditioned directed information. This solution is a limiting expression in the number of uses of the channel, as is Shannon's solution for the two-way channel. Thus, our solution *is* useful for calculating approachable rate points, but it does *not* seem useful for calculating points on the boundary of the capacity region. Other results in Chapter 5 are a generalization of the rate region of Cover and Leung [29], and examples of discrete MAC-FBs for which the capacity region is strictly larger than that rate region.

While Chapters 4 and 5 deal with *random coding* techniques, Chapter 6 considers simple feedback strategies. We design feedback strategies

for that class of MAC-FBs for which Willems determined the capacity region [38]. We also show that these strategies can approach any rate point in the capacity region of these channels.

Chapter 2

Using Functional Dependence Graphs to Establish Conditional Statistical Independence

The independence of random variables plays a central role in many theorems of information theory. The task of determining whether two random variables are independent is, however, fraught with pitfalls. In this chapter, we consider a graphical technique for establishing the conditional independence of random variables that are determined by a set of independent random variables.

The use of graphical techniques for establishing conditional independence is not a new idea and there are several approaches for doing this. Pearl presented a graphical criterion called *d-separation* which establishes conditional independence in *Bayesian Networks* [39]. The "*d*" denotes "directional" [40, reprint] or "direction-dependent" [39, p. 116]. A Bayesian network is an acyclic directed graph (usually called a "directed acyclic graph" or "DAG") in which the vertices represent random variables and the directed branches "signify the existence of direct causal influences between the linked variables" [39, p. 117]. More precisely, the graph represents a factorization of the joint probability distribution, or

an expansion of the joint uncertainty (entropy), of the random variables under study [39, 41, 42].

A second approach uses an acyclic directed graph to represent *functional dependencies* [43, 44]. This approach allows fewer probability distributions than the first one because functional dependence is more restrictive than Markovian dependence. However, the additional restriction facilitates proofs of results that are not true in general when only Markovian dependencies are guaranteed. The graphical criterion specified in [43, 44] for establishing conditional independence is identical to that of [39].

The application of d-separation to a directed graph *with* cycles has also received attention. Spirtes mentions this problem in [45, Section 12.1] and treats it in some detail when the graph represents linear dependencies in [46]. Koster [47] has shown that some of the results from [41] carry over to what he calls "reciprocal graphs", which include directed graphs with cycles. Finally, Pearl and Dechter in [44] proved that d-separation is a valid criterion for establishing conditional independence in directed graphs with cycles when the graph represents functional dependencies and the random variables involved are discrete and finite.

The graphical method considered here is the same as that of [39, 44]. We also consider *functional dependence graphs* with cycles and prove that d-separation is a valid criterion for establishing conditional independence. In addition, we remove the limitations on the types of random variables used; we do not require that they be discrete or finite. Our approach relies on the equivalence of the d-separation criterion with being able to bring the functional dependence graph into a special form for which conditional independence is easily proved. We also introduce an extension of the d-separation criterion which we call the *fd-separation criterion*, the "fd" denoting "functional dependence".

This chapter is organized as follows. In Section 2.1 we describe our model and prove conditional independence results for certain systems of random variables whose functional dependence graphs have cycles. In Section 2.2 we prove that d-separation and fd-separation establish conditional independence in functional dependence graphs with cycles. Finally, in Section 2.3 we give several examples from information theory to demonstrate the utility of the technique.

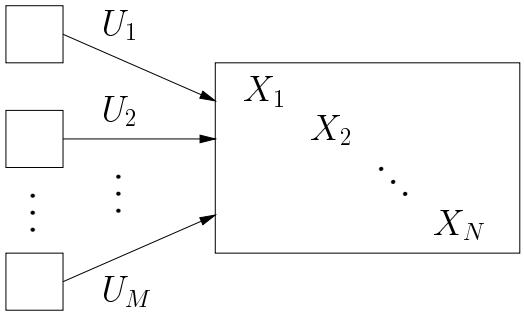


Figure 2.1: *The considered problem. The source random variables* $U_1, U_2, \ldots U_M$ *are independent and determine the random variables* $X_1, X_2, \ldots X_N$.

2.1 Functional Dependence Graphs with Independent Sources

2.1.1 Preliminaries

The problem considered is depicted in Figure 2.1 (see also [44]), where M *independent* random variables $U_1, U_2, \ldots, U_M$ determine the N random variables $X_1, X_2, \ldots, X_N$. One can interpret U_1, $U_2, \ldots, U_M$ as *source* random variables, and we shall refer to them as such. We will refer to the random variables $X_1, X_2, \ldots, X_N$ as *secondary* random variables.

The functional dependence of the secondary random variables on the source random variables can be depicted via a directed graph called the *functional dependence graph*. A functional dependence graph for a set of N functions in $M + N$ variables is a directed graph whose vertices represent the random variables, and for which a branch is drawn from one vertex to another if the former vertex's random variable is required as an argument of the function defining the latter vertex's random variable. There are M vertices that have no branches entering into them, and these are called *source* vertices. The source vertices represent independent random variables and we require that at least one such source vertex exists, i.e., $M > 0$. Furthermore, we require that the N functions suffice to determine *all* the secondary random variables in terms of the source random variables, i.e., given values of the M source random variables the functional equations have a unique solution for the values of the secondary random variables. This last

requirement plays a central role in our proofs.

To clarify the above definition, we give some examples. Suppose that the source random variables U_1 and U_2 and the secondary random variables X_1 and X_2 satisfy the equations

$$X_1 = f_1(U_1, U_2) \tag{2.1}$$
$$X_2 = f_2(U_1, U_2). \tag{2.2}$$

Then U_1 and U_2 determine X_1 and X_2. The functional dependence graph corresponding to this set of equations is depicted in Figure 2.2. The vertices represent the random variables and each branch entering a vertex represents an argument of the functions f_1 and f_2. We have drawn the source vertices as hollow circles to distinguish them from the solid circles used for the vertices representing the secondary random variables. The requirement that there be at least one source vertex is met, and certainly both X_1 and X_2 are determined by U_1 and U_2.

Suppose now that U_1 and U_2 determine X_1 and X_2 via the equations

$$X_1 = g_1(U_1, X_2) \tag{2.3}$$
$$X_2 = g_2(U_2, X_1). \tag{2.4}$$

The graph corresponding to this set of equations is depicted in Figure 2.3. Note that one must be able to rewrite the equations (2.3) and (2.4) in the form (2.1) and (2.2), respectively, because X_1 and X_2 by hypothesis are determined by U_1 and U_2. Thus, the graph of Figure 2.2 could also be used for this case, but the graph of Figure 2.3 incorporates extra knowledge about the system under study. Note that there is a directed cycle in the graph of Figure 2.3. This cycle indicates in general that the source random variables U_1 and U_2 may not determine X_1 and X_2. Whether U_1 and U_2 determine X_1 and X_2 when the functional dependence graph is that of Figure 2.3 depends on the specific functions g_1 and g_2. Later examples will make this clear.

2.1.2 Conditional Independence of Source Random Variables

The graph of Figure 2.3 will now be considered in detail. In particular, we are interested in answering the question: Are the source random

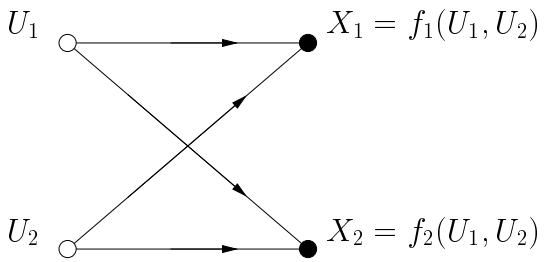


Figure 2.2: *A functional dependence graph for $M = N = 2$.*

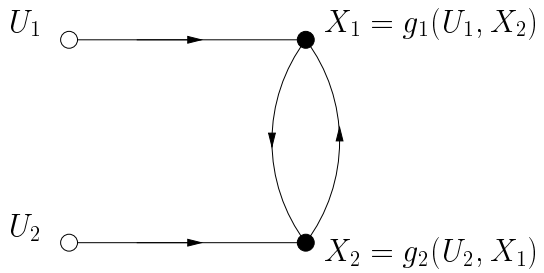


Figure 2.3: *A functional dependence graph with a cycle.*

variables U_1 and U_2 independent when *conditioned* on X_1 and X_2? This conditional independence may be expressed as

$$I(U_1; U_2 \mid X_1 X_2) = 0 \tag{2.5}$$

or, equivalently, as

$$H(U_1 \mid U_2 X_1 X_2) = H(U_1 \mid X_1 X_2), \tag{2.6}$$

where $I(A; B \mid C)$ is the mutual information between the random variables A and B conditioned on the random variable C, and $H(A \mid B)$ is the uncertainty (or entropy) of A conditioned on B. Before proving that (2.6) holds for the situation in Figure 2.3, we first prove the following.

Proposition 2.1 *Let U_1 and U_2 be two independent random variables and g_1 and g_2 functions such that the random variables $X_1 = g_1(U_1, X_2)$ and $X_2 = g_2(U_2, X_1)$ are determined by U_1 and U_2. Then the event $\{X_1 = x_1, X_2 = x_2\}$ occurs if and only if the event $\{g_1(U_1, x_2) = x_1, g_2(U_2, x_1) = x_2\}$ occurs.*

Proof: Occurrence of the event $\{X_1 = x_1, X_2 = x_2\}$ implies occurrence of the event $\{g_1(U_1, x_2) = x_1, g_2(U_2, x_1) = x_2\}$. We note that

occurrence of the event $\{g_1(U_1, x_2) = x_1, g_2(U_2, x_1) = x_2\}$ restricts the values that the pair (U_1, U_2) can take on to some set $\mathcal{U}$. But if $(U_1, U_2) = (u_1, u_2) \in \mathcal{U}$ we must have $\{X_1 = x_1, X_2 = x_2\}$ as this is *a* solution of $\{g_1(u_1, X_2) = X_1, g_2(u_2, X_1) = X_2\}$ and this solution is unique by hypothesis. □

This proposition allows us to show very simply that (2.6) indeed holds for the functional dependence graph of Figure 2.3. For any possible event $\{X_1 = x_1, X_2 = x_2\}$ we may write

$$\begin{aligned} H(U_1 \mid U_2, X_1 = x_1, X_2 = x_2) & \\ &= H(U_1 \mid U_2, g_1(U_1, x_2) = x_1, g_2(U_2, x_1) = x_2) \\ &= H(U_1 \mid g_1(U_1, x_2) = x_1, g_2(U_2, x_1) = x_2) \\ &= H(U_1 \mid X_1 = x_1, X_2 = x_2), \end{aligned} \tag{2.7}$$

where the first equality holds by the proposition, the second by the independence of U_1 and U_2, and the third by the proposition. By averaging both sides of equation (2.7) using $p(x_1, x_2)$, we obtain (2.6). Thus, the random variables U_1 and U_2 are independent when conditioned on X_1 and X_2.

2.1.3 Conditional Independence of Secondary Random Variables

We now consider the rather complex functional dependence graph $\mathcal{G}^*$ shown in Figure 2.4 which, as we shall see in Section 2.2, has the most general properties of interest to us. We are now interested in whether X_3 and X_4 are conditionally independent given U_3, X_1 and X_2, i.e., in whether $I(X_3; X_4 \mid U_3 X_1 X_2) = 0$. We first write

$$X_1 = g_1(U_1, U_3, X_2, X_3) = g_1(U_1, U_3, X_2, g_3(U_1, U_3, X_1, X_2)) \tag{2.8}$$

$$X_2 = g_2(U_2, U_3, X_1, X_4) = g_2(U_2, U_3, X_1, g_4(U_2, U_3, X_1, X_2)), \tag{2.9}$$

and then proceed by using the following proposition, whose proof is omitted as it is very similar to that of Proposition 2.1.

Proposition 2.2 *Let U_1, U_2 and U_3 be three independent random variables and g_1, g_2, g_3 and g_4 be functions such that the random variables*

$$X_1 = g_1(U_1, U_3, X_2, g_3(U_1, U_3, X_1, X_2))$$

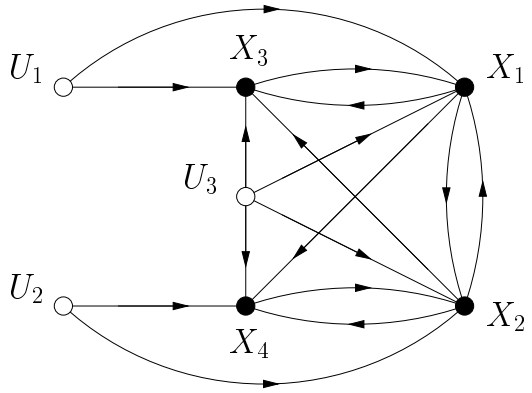


Figure 2.4: *The functional dependence graph $\mathcal{G}^*$.*

and

$$X_2 = g_2(U_2, U_3, X_1, g_4(U_2, U_3, X_1, X_2))$$

are determined by U_1, U_2 and U_3. Then the event $\{X_1 = x_1, X_2 = x_2\}$ occurs if and only if the event

$$\begin{aligned}\{&g_1(U_1, U_3, x_2, g_3(U_1, U_3, x_1, x_2)) = x_1,\\ &\quad g_2(U_2, U_3, x_1, g_4(U_2, U_3, x_1, x_2)) = x_2\}\end{aligned}$$

occurs.

This proposition may be used to show that $I(U_1; U_2 \mid U_3 X_1 X_2) = 0$ in a manner similar to how Proposition 2.1 was used in the previous section. Furthermore,

$$\begin{aligned} I(U_1; U_2 \mid U_3 X_1 X_2) &= I(U_1 X_3; U_2 X_4 \mid U_3 X_1 X_2) && (2.10)\\ &\geq I(X_3; X_4 \mid U_3 X_1 X_2), && (2.11)\end{aligned}$$

where the equality follows from (2.8) and (2.9), and the inequality follows from the fact that mutual information (conditioned or not) between sets of random variables cannot be increased by removing random variables from either set. Thus, because $I(U_1; U_2 \mid U_3 X_1 X_2) = 0$ and information cannot be negative, we conclude that $I(X_3; X_4 \mid U_3 X_1 X_2) = 0$.

The graphs of Figures 2.3 and 2.4 are actually more general than they seem at first glance. Nothing prevents U_1, U_2, U_3, X_1, X_2, X_3 and X_4 from being *vector-valued* random variables. Thus, if any functional

dependence graph can be put in the form of Figure 2.4 by collecting random variables into random vectors, we have, as in (2.10),

$$I(\,\underline{U}_1\,\underline{X}_3\,;\,\underline{U}_2\,\underline{X}_4\mid\underline{U}_3\,\underline{X}_1\,\underline{X}_2\,)=0. \tag{2.12}$$

2.2 Conditional Independence, d-Separation and fd-Separation

The development of the previous section will be used to prove that the d-separation criterion of [39] establishes the conditional independence of random variables in functional dependence graphs. We first define what is meant by d-separation.

Definition: (cf. Pearl [39, p. 117]) Consider a functional dependence graph $\mathcal{G}$ and random vectors $\underline{A}$, $\underline{B}$ and $\underline{C}$ whose respective components form disjoint subsets of the vertices of $\mathcal{G}$. The vector $\underline{C}$ is said to *d-separate* $\underline{A}$ from $\underline{B}$ if there is no path between a vertex in $\underline{A}$ and a vertex in $\underline{B}$ after the following manipulations of the graph have been performed.

1. Consider the subgraph $\mathcal{G}_{\underline{D}}$ of $\mathcal{G}$ containing those branches and vertices encountered when moving *backwards* one or more branches starting from vertices in $\underline{A}$ or $\underline{B}$ or $\underline{C}$. We call $\mathcal{G}_{\underline{D}}$ the graph *relevant* to $\underline{D}=[\underline{A}\,\underline{B}\,\underline{C}]$ as it contains all source random variables defining $\underline{D}$.

2. In $\mathcal{G}_{\underline{D}}$ delete all branches coming *out* of vertices in $\underline{C}$, obtaining the graph $\mathcal{G}_{\underline{D}|\underline{C}}$.

3. Remove the arrows on the branches of $\mathcal{G}_{\underline{D}|\underline{C}}$ to obtain the undirected graph $\mathcal{G}^u_{\underline{D}|\underline{C}}$, possibly with parallel branches between vertices.

For example, in the graph $\mathcal{G}^*$ of Figure 2.4, the vector $[U_3\ X_1\ X_2]$ d-separates U_1 and U_2. The graphs $\mathcal{G}_{\underline{D}}$, $\mathcal{G}_{\underline{D}|\underline{C}}$ and $\mathcal{G}^u_{\underline{D}|\underline{C}}$ for this example are the same as the respective $\tilde{\mathcal{G}}$, $\tilde{\mathcal{G}}_{\underline{D}|\underline{C}}$ and $\tilde{\mathcal{G}}^u_{\underline{D}|\underline{C}}$ in Figure 2.6, up to a relabelling of the vertices.

A few comments are in order here. We first caution that, by definition in a functional dependence graph the source random variables must

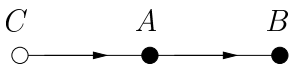


Figure 2.5: *A functional dependence graph for which C fd-separates A and B, but does not d-separate A and B.*

determine all other random variables. If they do not, then the technique of this section may not work as will be shown in Section 2.3.2. The purpose of Step 1 is to ensure that only that part of the graph relevant to the random variables under consideration is retained. Steps 1 to 3 may alternatively be performed by using the so-called *moral graph* [41].

One can extend the d-separation rule for functional dependence graphs. One may additionally delete all branches coming out of any secondary vertex that has no incoming branches. This is possible because these vertices must be determined by the conditioning random variables and may thus also be included in the conditioning. We call this extension the fd-separation criterion, for "functional dependence" separation, and define it explicitly.

Definition: Consider a functional dependence graph $\mathcal{G}$ and random vectors $\underline{A}$, $\underline{B}$ and $\underline{C}$ whose respective components form disjoint subsets of the vertices of $\mathcal{G}$. The vector $\underline{C}$ is said to *fd-separate* $\underline{A}$ from $\underline{B}$ if there is no path between a vertex in $\underline{A}$ and a vertex in $\underline{B}$ after the following manipulations of the graph have been performed.

1. Consider the subgraph $\mathcal{G}_{\underline{D}}$ of $\mathcal{G}$ containing those branches and vertices encountered when moving backwards one or more branches starting from vertices in $\underline{A}$ or $\underline{B}$ or $\underline{C}$.

2. In $\mathcal{G}_{\underline{D}}$ delete all branches coming out of vertices in $\underline{C}$ and *successively delete all branches coming out of secondary vertices having no incoming branches.* Denote the resulting graph by $\mathcal{G}_{\underline{D}|\underline{C}}$.

3. Remove the arrows on the branches of $\mathcal{G}_{\underline{D}|\underline{C}}$ to obtain the undirected graph $\mathcal{G}^u_{\underline{D}|\underline{C}}$, possibly with parallel branches between vertices.

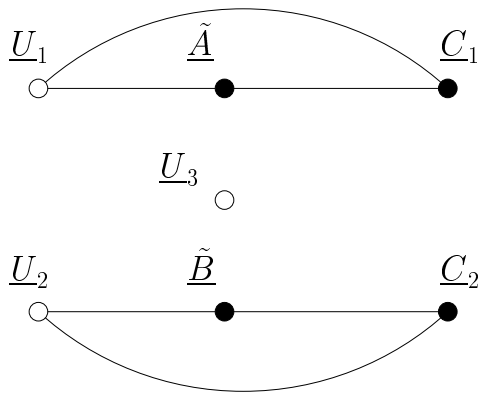
Note that d-separation implies fd-separation but not conversely as shown by the simple graph in Figure 2.5. We now link d-separation with the graph $\mathcal{G}^*$ in Figure 2.4. In the following lemma, by *grouping* the vertices $A_1, A_2, \ldots, A_L$ we mean deleting the branches between these

vertices and then considering them as a single vertex, i.e., the branches coming into and going out of $A_1, A_2, \ldots, A_L$ become the respective branches coming into and going out of the the single vertex.

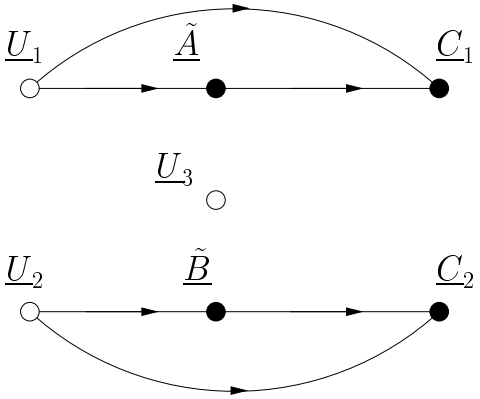
Lemma 2.1 *Let $\mathcal{G}$ be a functional dependence graph, and $\underline{A}$, $\underline{B}$ and $\underline{C}$ be random vectors whose respective components form disjoint subsets of the vertices of $\mathcal{G}$. Let $\underline{C}_1$, $\underline{C}_2$ and $\underline{U}_3$ be random vectors whose components form disjoint subsets of all the components of $\underline{C}$, where $\underline{U}_3$ contains the source vertices (if any) in $\underline{C}$. Then the vector $\underline{C}$ d-separates $\underline{A}$ and $\underline{B}$ in $\mathcal{G}$ if and only if $\mathcal{G}$ can, by grouping vertices after Step 1 of the definition of d-separation, be made into a subgraph of the graph $\tilde{\mathcal{G}}$ in Figure 2.6c. To construct $\tilde{\mathcal{G}}$, the components of $\underline{A}$ were made components of $\underline{U}_1$ or $\tilde{\underline{A}}$ and the components of $\underline{B}$ were made components of $\underline{U}_2$ or $\tilde{\underline{B}}$.*

Proof: Consider the graph $\tilde{\mathcal{G}}$ shown in Figure 2.6c. One may check that $\underline{C} = [\underline{C}_1\ \underline{C}_2\ \underline{U}_3]$ d-separates $[\underline{U}_1\ \tilde{\underline{A}}]$ and $[\underline{U}_2\ \tilde{\underline{B}}]$. Thus, if $\mathcal{G}$ can be made a subgraph of $\tilde{\mathcal{G}}$, then $\underline{C}$ d-separates the random vectors $[\underline{U}_1\ \tilde{\underline{A}}]$ and $[\underline{U}_2\ \tilde{\underline{B}}]$. Hence $\underline{C}$ d-separates $\underline{A}$ and $\underline{B}$.

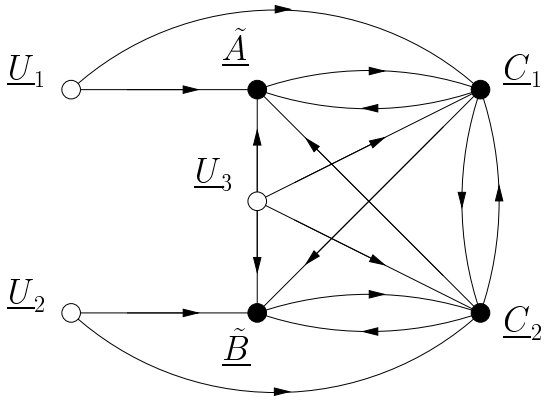
Conversely, assume that $\underline{C}$ d-separates $\underline{A}$ and $\underline{B}$. This means that $\underline{A}$ and $\underline{B}$ are in unconnected parts of the undirected graph $\mathcal{G}^u_{\underline{D}|\underline{C}}$ after step 3 in the procedural definition of d-separation. We can thus group all the vertices in $\mathcal{G}^u_{\underline{D}|\underline{C}}$ into seven vectors (see Figure 2.6a). All source vertices in $\underline{C}$ are grouped in $\underline{U}_3$. All secondary vertices in $\underline{C}$ connected to vertices in $\underline{A}$ are grouped in $\underline{C}_1$. All remaining source vertices connected to vertices in $\underline{A}$ are grouped in $\underline{U}_1$, and all remaining (secondary) vertices connected to $\underline{A}$, including those in $\underline{A}$ itself, are grouped in $\tilde{\underline{A}}$. Any remaining (secondary) vertices in $\underline{C}$ are grouped in $\underline{C}_2$. All remaining source vertices are grouped in $\underline{U}_2$ and all remaining (secondary) vertices are grouped in $\tilde{\underline{B}}$. The resulting graph must then be a subgraph of $\tilde{\mathcal{G}}^u_{\underline{D}|\underline{C}}$ in Figure 2.6a. Thus after Step 2, the graph must have been a subgraph of $\tilde{\mathcal{G}}_{\underline{D}|\underline{C}}$ in Figure 2.6b. At Step 2, any branches emanating from $\underline{C}_1$, $\underline{C}_2$ and $\underline{U}_3$ were removed so that the most general form of the graph $\tilde{\mathcal{G}}$ after Step 1 is given in Figure 2.6c. We see that this graph is a relabelled version of $\mathcal{G}^*$ in Figure 2.4. Of course, not all of the branches and vertices need be included in Figures 2.6a and 2.6c, so that $\tilde{\mathcal{G}}$ could be a proper subgraph of $\mathcal{G}^*$. □



a) The graph $\tilde{\mathcal{G}}^u_{\underline{D}|\underline{C}}$.



b) The directed graph $\tilde{\mathcal{G}}_{\underline{D}|\underline{C}}$.



c) The directed graph $\tilde{\mathcal{G}}$: a relabelled version of $\mathcal{G}^*$.

Figure 2.6: *The steps of the graphical technique are retraced to determine the most general form of a graph for which d-separation implies conditional independence.*

This lemma, along with equation (2.12), shows that d-separation implies conditional independence in functional dependence graphs. We state this as a theorem.

Theorem 2.1 *Let $\mathcal{G}$ be a functional dependence graph, and let $\underline{A}$, $\underline{B}$ and $\underline{C}$ be vector random variables having vertices in $\mathcal{G}$ whose respective components form disjoint subsets of the graph's vertices. If $\underline{C}$ d-separates $\underline{A}$ and $\underline{B}$ in $\mathcal{G}$, then*

$$I(\underline{A}\,;\underline{B} \mid \underline{C}) = 0.$$

Proof: By Lemma 2.1, we can make the graph $\mathcal{G}_{\underline{D}}$ relevant to $\underline{D} = [\underline{A}\ \underline{B}\ \underline{C}]$ a subgraph of $\tilde{\mathcal{G}}$ in Figure 2.6. But then equation (2.12) guarantees that

$$I(\,\underline{U}_1\,\tilde{\underline{A}}\,;\,\underline{U}_2\,\tilde{\underline{B}} \mid \underline{C}\,) = 0. \tag{2.13}$$

Since the components of $\underline{A}$ and $\underline{B}$ are in the respective vectors $[\underline{U}_1\ \tilde{\underline{A}}]$ and $[\underline{U}_2\ \tilde{\underline{B}}]$, we obtain $I(\underline{A}\,;\underline{B} \mid \underline{C}) = 0$ by expanding (2.13). □

It is now easy to prove that fd-separation also implies conditional independence in functional dependence graphs. The reason is that if $\underline{C}$ fd-separates $\underline{A}$ and $\underline{B}$ in $\mathcal{G}$, then $[\underline{C}\ \underline{D}]$ d-separates $\underline{A}$ and $\underline{B}$ in $\mathcal{G}$, where the components of $\underline{D}$ are all random variables not in $\underline{C}$ which are determined by $\underline{C}$. We state this as a Corollary to Theorem 2.1.

Corollary 2.1 *Let $\mathcal{G}$ be a functional dependence graph and let $\underline{A}$, $\underline{B}$ and $\underline{C}$ be vector random variables having as components vertices in $\mathcal{G}$ that contain disjoint subsets of the graph's vertices. If $\underline{C}$ fd-separates $\underline{A}$ and $\underline{B}$ in $\mathcal{G}$, then*

$$I(\underline{A}\,;\underline{B} \mid \underline{C}) = 0.$$

Proof: Since $[\underline{C}\ \underline{D}]$ d-separates $\underline{A}$ and $\underline{B}$ in $\mathcal{G}$, where $\underline{D}$ contains all random variables not in $\underline{C}$ which are determined by $\underline{C}$, $I(\underline{A};\underline{B} \mid \underline{C}\,\underline{D}) = 0$ by Theorem 2.1. But $I(\underline{A}\,;\underline{B} \mid \underline{C}\,f(\underline{C})) = I(\underline{A};\underline{B} \mid \underline{C})$ for any function $f(\cdot)$. This proves the corollary. □

Theorem 2.1 and Corollary 2.1 state that d-separation and fd-separation are *sufficient* for establishing conditional independence in functional dependence graphs. A natural question to ask is if these conditions are *necessary*. The example of Figure 2.5 shows that d-separation is not necessary. A simple example showing that fd-separation is also not necessary is the set of equations over the reals

$$X_1 = U_1 + U_2 \tag{2.14}$$
$$X_2 = U_1 - U_2 \tag{2.15}$$

whose functional dependence graph is shown in Figure 2.2. In this graph $[X_1\ X_2]$ does not fd-separate U_1 and U_2. But because U_1 and U_2 are are determined by X_1 and X_2 we have $I(U_1; U_2 \mid X_1\, X_2) = 0$.

2.3 Applications of the Graphical Technique

We give four examples to demonstrate the usefulness of d-separation. The first example shows that allowing cycles in functional dependence graphs can be helpful. The second demonstrates that if full determination is not guaranteed, d-separation may not imply conditional independence. The third and fourth examples show how functional dependence graphs can be applied to cryptography and multi-user information theory.

2.3.1 Showing Conditional Independence by Rewriting Equations

Consider the following two equations in $\mathbb{Z}_8$, the ring of integers modulo 8:

$$X_1 = 6U_1 + 6U_2 \tag{2.16}$$
$$X_2 = 4U_1 + 6U_2. \tag{2.17}$$

Obviously, X_1 and X_2 are determined by U_1 and U_2. The functional dependence graph for this set of equations is shown in Figure 2.2. Suppose we now want to know whether $I(U_1; U_2 \mid X_1 X_2) = 0$. This is

not obvious from (2.16) and (2.17). But by rewriting the equations we obtain

$$X_1 = 2U_1 + X_2 \tag{2.18}$$
$$X_2 = 2U_2 + 2X_1, \tag{2.19}$$

whose functional dependence graph is shown in Figure 2.3. Thus, as shown in Section 2.1.2, U_1 and U_2 are indeed independent when conditioned on X_1 and X_2. Note that this is true even though U_1 and U_2 are not determined by X_1 and X_2.

2.3.2 Non-determination for Graphs with Cycles

Assume now that we have the following two equations in $\mathbb{Z}_3$:

$$X_1 = U_1^2 \cdot X_2$$
$$X_2 = U_2^2 \cdot X_1.$$

One may check that U_1 and U_2 do not determine X_1 and X_2, e.g., by setting $U_1 = U_2 = 1$. Hence a functional dependence graph cannot be constructed. However, if one ignores the non-uniqueness of X_1 and X_2 and attempts to make a "functional dependence graph", one arrives at the graph of Figure 2.3. Now suppose that the source random variables are independent and uniformly distributed in $\mathbb{Z}_3$, and that the probability distribution $p_{U_1U_2X_1X_2}$ specifies that $U_1 \cdot U_2 = 2$ if and only if $X_1 = 1$. Then one finds that U_1 and U_2 determine X_1 and X_2 yet $I(U_1; U_2 \mid X_1 = 1, X_2 = 1) = 1$ bit. Thus, if the functions available do not ensure that the source random variables determine all other random variables, d-separation may not imply conditional independence. It was already noted by Spirtes [46] that d-separation in directed graphs with cycles may not imply conditional independence unless extra restrictions are introduced. In our case, as in [44], these needed restrictions are the independence of the source random variables and the full determination of the secondary random variables by the source random variables using the available functions.

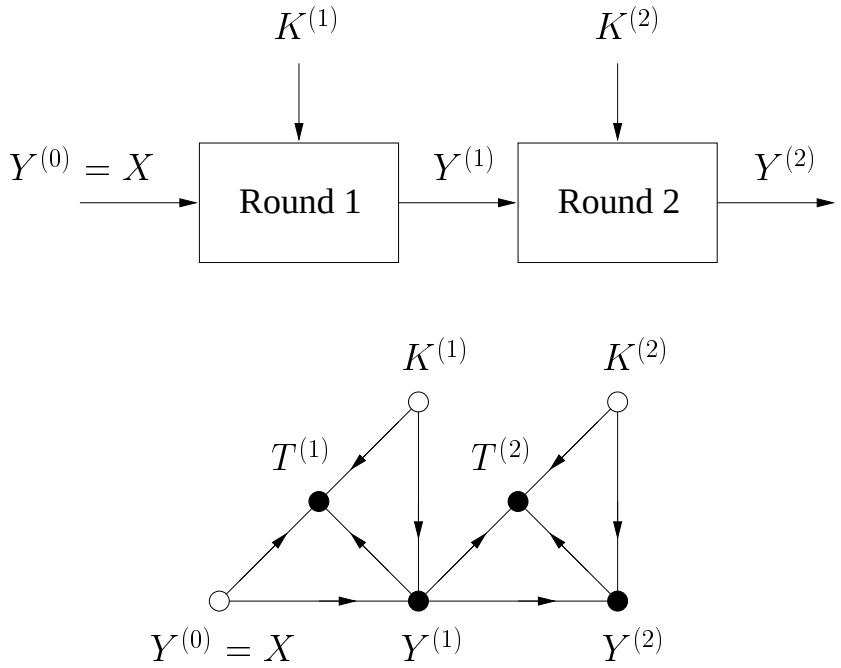


Figure 2.7: *The structure of an iterated block cipher and the functional dependence graph for linear cryptanalysis.*

2.3.3 Linear Cryptanalysis

The structure of a two-round iterated block cipher is shown in Figure 2.7. This particular cipher encrypts the plaintext X by two successive applications of a keyed round function with a different key in each round. The i-th round output $Y^{(i)}$ is a function of the i-th round input $Y^{(i-1)}$ and the i-th round key $K^{(i)}$. The plaintext and the round keys are assumed to be independent random variables. [The round keys are usually obtained in practice via a *key-scheduling algorithm*, which causes the round keys to be dependent. Furthermore, the key is usually kept fixed for many encryptions. However, it is still useful to consider randomly chosen keys, as explained in [49].]

A general technique for cryptanalyzing block ciphers, called *linear cryptanalysis*, was developed by Matsui [50], and generalized and formalized by Harpes et al. [49]. For this technique, *threefold sums* $T^{(i)} = f_i(Y^{(i-1)}, Y^{(i)}, K^{(i)})$ are generated for each round (see Figure 2.7), and then added together. For the ensuing cryptanalysis it is important that these threefold sums be independent. This can in fact be guaranteed if each round threefold sum $T^{(i)}$ is independent of the corresponding round input $Y^{(i-1)}$. To show this, we consider the functional depen-

dence graph defining the threefold sums in Figure 2.7. We see that $Y^{(1)}$ d-separates $T^{(1)}$ and $T^{(2)}$ so that

$$I(T^{(1)}; T^{(2)} \mid Y^{(1)}) = 0. \tag{2.20}$$

Thus

$$\begin{aligned} H(T^{(2)}) &= H(T^{(2)} \mid Y^{(1)}) \\ &= H(T^{(2)} \mid T^{(1)}Y^{(1)}), \end{aligned}$$

where the first equality follows by the specified independence of $T^{(i)}$ and $Y^{(i-1)}$, and the second equality follows from (2.20). Thus, $T^{(2)}$ is independent of $[T^{(1)}, Y^{(1)}]$, and thus independent of $T^{(1)}$. It is easy to check that, for more than two rounds, the $T^{(i)}$ are all independent when they have the property that they are independent of their round inputs. [Several important ciphers such as IDEA and SAFER (as well as DES if one considers two "DES rounds" as constituting one round) have this property when their round keys are chosen independently and uniformly at random and the threefold sums used are homomorphic with respect to the group operation by which the key is inserted [49, Section 2.6].]

2.3.4 The Multiple-Access Channel with Feedback

In Chapter 5, we consider the two-user memoryless multiple-access channel with feedback. For this channel, User k wishes to transmit B_k data bits $U_k^{B_k}$ to the receiver by using the channel N times. Each user may use his entire data sequence to choose his inputs X_{kn}, $n = 1 \ldots N$, to the channel. Additionally, via a feedback link, the users see the previous outputs of the channel and may also use these to choose subsequent channel inputs. The channel maps the inputs X_{1n} and X_{2n}, and possibly the noise Z_n, to the output Y_n. The random variables $U_1^{B_1}$, $U_2^{B_2}$ and Z_n, $n = 1 \ldots N$, are required to be independent. The functional dependence graph up to time $n = 3$ for this problem is shown in Figure 2.8.

Willems [30] determined the capacity region of a certain class of memoryless multiple-access channels with feedback. An important step in his proof required showing that the two channel inputs X_{1n} and X_{2n} are independent when conditioned on the so-called *auxiliary random variable* $V_n = [X_1^{n-1}, Y^{n-1}]$. This independence is easily shown by our

technique. Using the d-separation criterion of Section 2.2 for $n = 3$ in Figure 2.8, we obtain the undirected graph of Figure 2.9. Because the random variables X_{13} and X_{23} are in unconnected components of this graph, we know that

$$I(X_{1n}; X_{2n} \mid X_1^{n-1} Y^{n-1}) = 0 \tag{2.21}$$

for $n = 3$. It is easy to see that (2.21) holds for arbitrary n. Thus, $I(X_{1n}; X_{2n} \mid V_n) = 0$, as was to be shown.

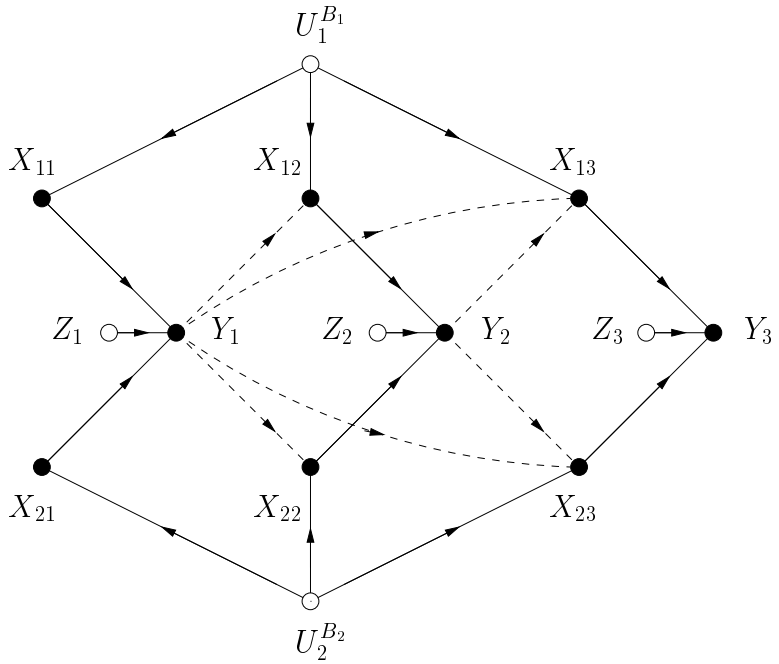


Figure 2.8: *The functional dependence graph for three uses of the two-user memoryless multiple-access channel with feedback. The dependence due to the feedback is shown by dashed lines.*

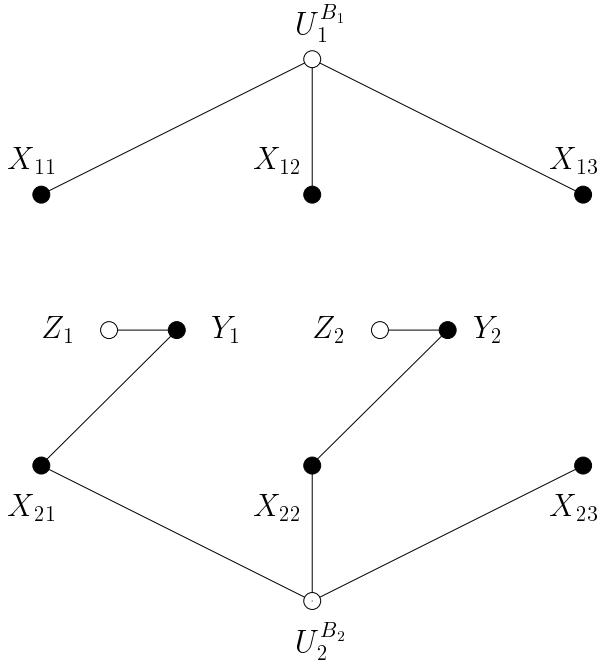


Figure 2.9: *The undirected graph obtained from the directed graph in Figure 2.8, and used to show that* $I(X_{13}; X_{23} \mid X_1^2 Y^2) = 0$.

Chapter 3

Causal Conditioning and Directed Information

The concept of associating *direction* with information flow was introduced by Marko [37] to deal with channels where feedback couples the data transmission of two users. Marko defined the "directed transinformation" flowing from one user to another by assuming stationarity and dealing with information rates. We follow here the more general approach of Massey [36] who defined *directed information* based on finite-length sequences.

This chapter is organized as follows. In Section 3.1 we introduce the concept of *causally conditioned uncertainty* and use it to express directed information and to define *causally conditioned directed information*. Section 3.2 derives some basic properties of causally conditioned uncertainty and directed information, including chain rules and stationarity properties. In Section 3.3 we review the problem considered by Marko in [37], and in Section 3.4 we review some of the results of Massey [36] for discrete channels with feedback.

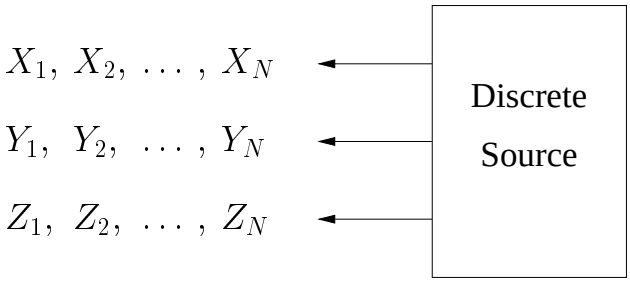


Figure 3.1: *A discrete source producing the sequences X^N, Y^N and Z^N.*

3.1 Definitions

3.1.1 Causal Conditioning

Consider a discrete source which emits three length N sequences of random variables X^N, Y^N and Z^N, as depicted in Figure 3.1. We define the uncertainty, or entropy, of the sequence X^N *causally conditioned* on the sequence Y^N as

$$H(X^N \| Y^N) := \sum_{n=1}^{N} H(X_n | X^{n-1} Y^n). \tag{3.1}$$

This definition is closely related to Marko's *free information*, cf. equation (8) in [37]. $H(X^N \| Y^N)$ differs from the conditional uncertainty $H(X^N | Y^N)$ only in that Y^n replaces Y^N in each term on the right of (3.1). Our term "causal" reflects the conditioning on *past and present* values of the sequence Y^N only. Note that this interpretation assumes that the sequences are "synchronized", i.e., that the nth terms in the sequences X^N and Y^N occur "at the same time", and that the nth terms occur "before" the $(n+1)$st terms.

In the above definition, nothing prevents X and/or Y from being a vector-valued random variable. For example,

$$\begin{aligned} H(X^N \| Y^N Z^N) &= \sum_{n=1}^{N} H(X_n | X^{n-1} Y^n Z^n) \\ H(X^N Y^N \| Z^N) &= \sum_{n=1}^{N} H(X_n Y_n | X^{n-1} Y^{n-1} Z^n) \end{aligned}$$

$$H(W^N X^N \| Y^N Z^N) = \sum_{n=1}^{N} H(W_n X_n | W^{n-1} X^{n-1} Y^n Z^n)$$

and similar expressions are already defined by (3.1). Less obvious is how to "mix" causal conditioning and the usual conditioning. We will use the notational convention that conditioning is done from left to right. Thus, for instance,

$$H(X^N | Y^N \| Z^N) := H(X^N Y^N \| Z^N) - H(Y^N \| Z^N) \tag{3.2}$$

where the causal conditioning on Z^N is applied after the usual conditioning on Y^N. Similarly,

$$H(X^N \| Y^N | Z^N) := \sum_{n=1}^{N} H(X_n | X^{n-1} Y^n Z^N) \tag{3.3}$$

where now the causal conditioning on Y^N is applied before the usual conditioning on Z^N. Note that, unlike causal conditioning, usual conditioning does not require synchronization of sequences; for instance

$$H(X^N \| Y^N | Z) = \sum_{n=1}^{N} H(X_n | X^{n-1} Y^n Z). \tag{3.4}$$

is well defined.

3.1.2 Directed Information

The directed information $I(X^N \to Y^N)$ flowing from a sequence X^N to a sequence Y^N was introduced by Massey [36] and in our notation can be written as

$$I(X^N \to Y^N) = H(Y^N) - H(Y^N \| X^N). \tag{3.5}$$

Equivalently,

$$I(X^N \to Y^N) = \sum_{n=1}^{N} I(X^n; Y_n | Y^{n-1}), \tag{3.6}$$

which is the definition given in [36]. This definition is closely related to Marko's *directed transinformation*, cf. equations (15) and (16) in [37].

As in the definition of causally-conditioned uncertainty, this definition differs from the mutual information $I(X^N;Y^N)$ only in that X^n replaces X^N in each term on the right of (3.6). [The name *mutual information* and the notation $I(X;Y)$ was introduced by Fano [51, Chapter 2].] Note that whereas $I(X^N;Y^N) = I(Y^N;X^N)$, in general $I(X^N \to Y^N) \neq I(Y^N \to X^N)$.

We next define the directed information $I(X^N \to Y^N \| Z^N)$ flowing from X^N to Y^N when *causally conditioned* on the sequence Z^N as

$$I(X^N \to Y^N \| Z^N) := H(Y^N \| Z^N) - H(Y^N \| X^N Z^N). \tag{3.7}$$

Equivalently,

$$I(X^N \to Y^N \| Z^N) = \sum_{n=1}^{N} I(X^n; Y_n | Y^{n-1} Z^n). \tag{3.8}$$

This definition differs from that of the conditional mutual information $I(X^N;Y^N|Z^N)$ only in that X^n and Z^n replace X^N and Z^N in each term on the right of (3.8). Combinations of causal conditioning and usual conditioning are defined with the notational convention adopted above for (3.2) and (3.3), i.e., the conditioning is performed from left to right.

3.1.3 Directed Information Rates

In many applications, directed information (3.5) increases linearly with N. An important parameter is then the rate of this increase, i.e., the "directed information rate" or the "directed information per letter". For this purpose, we define the following per-letter uncertainties and informations:

$$\begin{array}{lcl} H_N(X) & = & \frac{1}{N} H(X^N) \\ H_N(X\|Y) & = & \frac{1}{N} H(X^N \| Y^N) \\ I_N(X \to Y) & = & \frac{1}{N} I(X^N \to Y^N) \\ I_N(X \to Y \| Z) & = & \frac{1}{N} I(X^N \to Y^N \| Z^N). \end{array} \tag{3.9}$$

In most cases of interest, the terms in (3.9) have limits as N tends to infinity. We denote these limits by

$$\begin{array}{lcl} H_\infty(X) & = & \lim_{N\to\infty} \frac{1}{N} H(X^N) \\ H_\infty(X\|Y) & = & \lim_{N\to\infty} \frac{1}{N} H(X^N\|Y^N) \\ I_\infty(X\to Y) & = & \lim_{N\to\infty} \frac{1}{N} I(X^N\to Y^N) \\ I_\infty(X\to Y\|Z) & = & \lim_{N\to\infty} \frac{1}{N} I(X^N\to Y^N\|Z^N). \end{array} \tag{3.10}$$

The first of these limits, $H_\infty(X)$, is the usual "information rate" or "entropy rate" of the source producing the sequence X_1,X_2,X_3,... [48, page 63]. If the source is *stationary*, then [48, page 64]

$$H_\infty(X) = \lim_{n\to\infty} H(X_n|X^{n-1}). \tag{3.11}$$

Similar results hold for the other limits in (3.10), and will be developed in Section 3.2.2.

3.2 Properties of Directed Information

3.2.1 Basic Properties

We begin with two simple bounds, one on causally-conditioned uncertainty and the other on directed information.

Property 3.1 (A Bound on Causally-Conditioned Uncertainty)

$$H(X^N|Y^N) \le H(X^N\|Y^N) \le H(X^N) \tag{3.12}$$

with equality on the left if and only if $H(X_n|X^{n-1}Y^n) = H(X_n|X^{n-1}Y^N)$ *for all* $n = 1, 2, \ldots, N$, *and with equality on the right if and only if* $I(X_n; Y^n|X^{n-1}) = 0$ *for all* $n = 1, 2, \ldots, N$.

Proof: Since conditioning cannot increase uncertainty [48, page 27], the nth term in the sum $H(X^N|Y^N) = \sum_{n=1}^N H(X_n|X^{n-1}Y^N)$ is not more than the nth term in the sum $H(X^N\|Y^N) = \sum_{n=1}^N H(X_n|X^{n-1}Y^n)$. This proves the inequality on the left of (3.12) and that equality holds if and only if the individual terms are equal. Similarly, the nth term

in the sum $H(X^N \| Y^N) = \sum_{n=1}^{N} H(X_n | X^{n-1} Y^n)$ is not more than the nth term in the sum $H(X^N) = \sum_{n=1}^{N} H(X_n | X^{n-1})$. This proves the inequality on the right of (3.12). Equality holds if and only if $H(X_n | X^{n-1}) = H(X_n | X^{n-1} Y^n)$ or equivalently $I(X_n; Y^n | X^{n-1}) = 0$ for all $n = 1, 2, \ldots, N$. □

One might be tempted to guess that equality holds on the right in (3.12) only if X^N and Y^N are independent. However, this is not true, as the following example demonstrates. Let X_1 and Y_1 be independent binary random variables with $\Pr(X_1 = 0) = \Pr(X_1 = 1) = 1/2$ and $\Pr(Y_1 = 0) = \Pr(Y_1 = 1) = 1/2$ and let $X_2 = X_1$ and $Y_2 = X_1 \oplus Y_1$ where $\oplus$ denotes addition modulo 2. Then it is easy to check that $H(X^2 \| Y^2) = H(X^2) = 1$ bit, but X^2 and Y^2 are not independent since $I(X^2; Y^2) = 1$ bit.

Property 3.2 (A Bound on Directed Information) *(Massey [36])*

$$0 \leq I(X^N \to Y^N) \leq I(X^N; Y^N), \tag{3.13}$$

with equality on the left if and only if $I(X^n; Y_n | Y^{n-1}) = 0$ *for all* $n = 1, 2, \ldots, N$, *and with equality on the right if and only if* $H(Y_n | Y^{n-1} X^n) = H(Y_n | Y^{n-1} X^N)$ *for all* $n = 1, 2, \ldots, N$.

Proof: Both the bounds and the conditions for equality follow immediately from Property 3.1. □

It is not true that equality holds on the left in (3.13) only if X^N and Y^N are independent. The right inequality is interesting because, as we will see in Section 3.4, directed information exactly specifies the information transfer between the sender's information sequence and the output sequence of discrete channels with feedback. The right inequality suggests that directed information may be useful in obtaining new upper bounds on the capacity of such channels [36].

We now consider two sequence chain rules for directed information.

Property 3.3 (Chain Rules for Directed Information)

$$\begin{aligned} I(X^N Y^N \to Z^N) &= I(X^N \to Z^N) + I(Y^N \to Z^N \| X^N) && (3.14)\\ I(X^N \to Y^N Z^N) &= I(X^N \to Y^N \| DZ^N) + && (3.15)\\ &\quad I(X^N \to Z^N \| Y^N), && (3.16) \end{aligned}$$

where $DZ^N := 0, Z_1, Z_2, \ldots, Z_{N-1}$ *is the delay of* Z^N *by one time instant (with discard of the last component).*

Proof: The first equation follows from

$$\begin{aligned} I(X^N Y^N \to Z^N) &= H(Z^N) - H(Z^N \| X^N Y^N) \\ &= \left[H(Z^N) - H(Z^N \| X^N)\right] + \\ &\quad \left[H(Z^N \| X^N) - H(Z^N \| X^N Y^N)\right]. \end{aligned}$$

The second follows from

$$\begin{aligned} &I(X^n; Y_n Z_n | Y^{n-1} Z^{n-1}) = \\ &\qquad I(X^n; Y_n | Y^{n-1} Z^{n-1}) + I(X^n; Z_n | Y^n Z^{n-1}). \end{aligned}$$

□

Finally, although $I(X^N \to Y^N) \neq I(Y^N \to X^N)$ in general, it would be intuitively pleasing if $I(X^N \to Y^N) + I(Y^N \to X^N) = I(X^N; Y^N)$ as suggested by a result of Marko [37, Equation (14)]. However, this relation does not hold in general. In fact, we have

Property 3.4 (Sum of Oppositely Directed Informations)

$$\begin{aligned} &I(X^N \to Y^N) + I(Y^N \to X^N) = \\ &\qquad I(X^N; Y^N) + I(X^N \to Y^N \| DX^N). \end{aligned} \tag{3.17}$$

By symmetry, this implies $I(X^N \to Y^N \| DX^N) = I(Y^N \to X^N \| DY^N)$.

Proof: We can write

$$\begin{aligned} I(X^N \to Y^N) + I(Y^N \to X^N) &= H(X^N) + H(Y^N) - \\ &\quad \left[H(X^N \| Y^N) + H(Y^N \| X^N)\right]. \end{aligned}$$

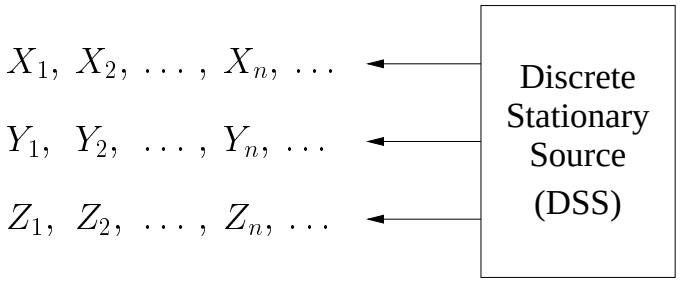


Figure 3.2: *A discrete stationary source producing the sequences X, Y and Z.*

The term in square brackets is

$$\sum_{n=1}^{N} \left[H(X_n|X^{n-1}Y^n) + H(Y_n|Y^{n-1}X^n) \right],$$

and $H(X_n|X^{n-1}Y^n) = H(X_nY_n|X^{n-1}Y^{n-1}) - H(Y_n|X^{n-1}Y^{n-1})$. The sum of the terms $H(X_nY_n|X^{n-1}Y^{n-1})$ gives $H(X^NY^N)$, and combining the remaining terms yields (3.17). □

3.2.2 Stationarity Properties

We now wish to consider the case when the sequences X^N, Y^N, and Z^N are the output of a discrete stationary source (Figure 3.2). A *discrete stationary source* (DSS) is a device that emits a sequence $U_1, U_2, U_3, \ldots$ of discrete random variables such that, for every $n \geq 1$ and $L \geq 1$, the random vectors $[U_1 \ \ldots \ U_L]$ and $[U_{n+1} \ \ldots \ U_{n+L}]$ have the same probability distribution. This means that for every window length L along the DSS output sequences, one sees the same statistical behavior regardless of where the window is placed along the DSS output sequences [52, page 2.31]. In Figure 3.2, we have chosen $U_n = (X_n, Y_n, Z_n)$.

We are interested in the various information rates of a DSS. We first consider the quantities $H(X_L|X^{L-1}Y^L)$ for $L \geq 1$, and prove the following result.

Property 3.5 (Causally-Conditioned Uncertainty of DSS Sequences) *If $X_1, X_2, X_3, \ldots$ and $Y_1, Y_2, Y_3, \ldots$ are the output sequences of a DSS, then:*

1. $H(X_L|X^{L-1}Y^L) \leq H_L(X\|Y)$ *for all* $L \geq 1$.

2. $H(X_L|X^{L-1}Y^L)$ *is non-increasing as* L *increases.*

3. $H_L(X\|Y)$ *is non-increasing as* L *increases.*

4. $\lim_{L\to\infty} H(X_L|X^{L-1}Y^L) = \lim_{L\to\infty} H_L(X\|Y)$, *i.e., both of these limits exist and have the same value* $H_\infty(X\|Y)$.

These four properties of a DSS are well known to hold if the DSS has a single output or, equivalently, if the Y sequence is independent of the X sequence (see, e.g., [52, page 2.32]). The proof of the more general case is given in Appendix 3.A.

Property 3.5 can now be applied to calculate the directed information rates between the output sequences of a DSS.

Property 3.6 (Directed Information Rates of a DSS) *If* X *and* Y *are output sequences of a DSS, then* $\lim_{L\to\infty} I_L(X \to Y)$ *exists and is given by*

$$I_\infty(X \to Y) = H_\infty(Y) - H_\infty(Y\|X). \tag{3.18}$$

If Z *is also an output sequence of the DSS, then* $\lim_{L\to\infty} I_L(X \to Y\|Z)$ *exists and is given by*

$$I_\infty(X \to Y\|Z) = H_\infty(Y\|Z) - H_\infty(Y\|XZ). \tag{3.19}$$

Proof: Consider (3.18). By Property 3.5, $H_\infty(Y)$ and $H_\infty(Y\|X)$ exist. But the limit of a real sequence whose elements are the term-by-term differences of the elements of two real convergent sequences exists. Furthermore, this limit is the difference of the two limits of the original sequences [53, page 223]. Thus, $\lim_{L\to\infty} I_L(X \to Y) = \lim_{L\to\infty}[H_L(Y) - H_L(Y\|X)] = \lim_{L\to\infty} H_L(Y) - \lim_{L\to\infty} H_L(Y\|X)$. This proves (3.18). Equation (3.19) is proved in the same manner. □

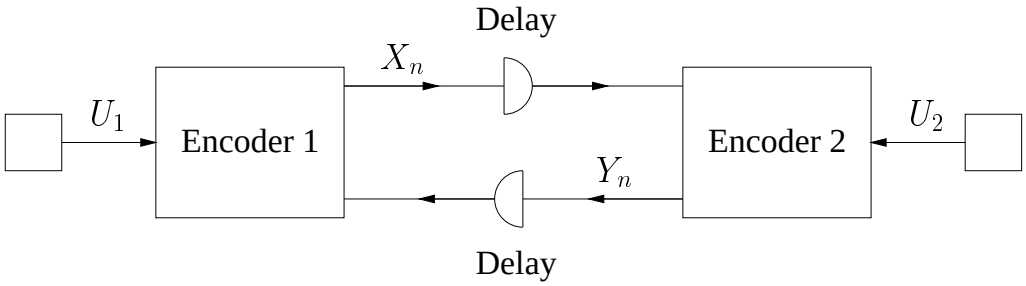


Figure 3.3: *Marko's two-user problem. The encoders may use all past observed values to encode their next transmitted symbol.*

3.3 Marko's Two-User Problem

We apply the definitions and results of the previous sections to the problem for which Marko originally introduced directed information rates [37]. The model and functional dependence graph for Marko's two-user problem are depicted in Figures 3.3 and 3.4. There are two users transmitting U_1 and U_2 to one another by encoding these independent random variables into the sequences $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$. After a small delay, each user receives the symbol of the other user and may use all the past received symbols to choose the next transmitted symbol, i.e.,

$$X_n = f_n(U_1, Y^{n-1}) \tag{3.20}$$
$$Y_n = g_n(U_2, X^{n-1}), \tag{3.21}$$

where $f_1, f_2, \ldots$ and $g_1, g_2, \ldots$ are functions chosen by the users. The functional dependence graph for this problem is depicted in Figure 3.4.

If we are interested in calculating $I(U_1; Y^N)$, we first expand

$$\begin{aligned} H(Y^N|U_1) &= \sum_{n=1}^{N} H(Y_n|Y^{n-1}U_1) \\ &= \sum_{n=1}^{N} H(Y_n|Y^{n-1}X^nU_1) \\ &= \sum_{n=1}^{N} H(Y_n|Y^{n-1}X^n) \\ &= H(Y^N \| X^N), \end{aligned} \tag{3.22}$$

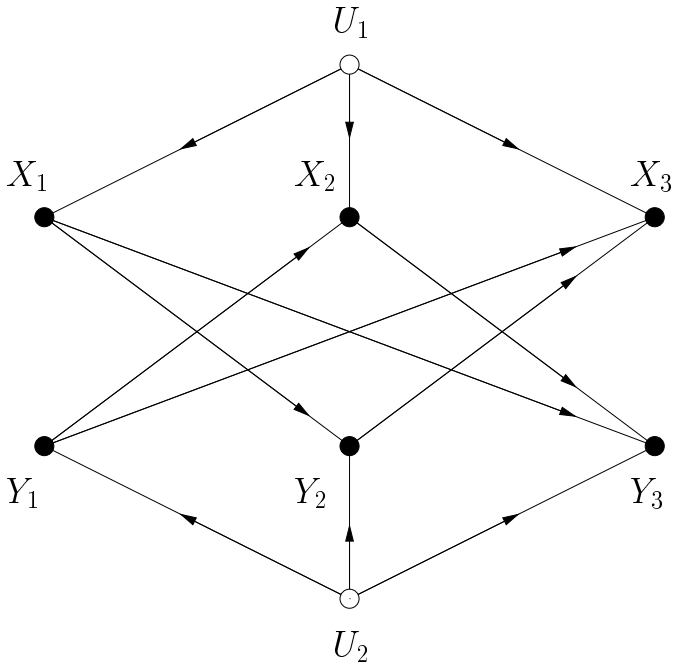


Figure 3.4: *The functional dependence graph for the first three uses of the encoders in Marko's two-user problem.*

where the second equality follows from (3.20), the third equality from the functional dependence graph, and the fourth equality by definition. Inserting (3.22) into $I(U_1;Y^N) = H(Y^N) - H(Y^N|U_1)$ we find that

$$I(U_1;Y^N) = H(Y^N) - H(Y^N\|X^N) = I(X^N \to Y^N),$$

and similarly $I(U_2;X^N) = I(Y^N \to X^N)$. Thus, we can remove U_1 and U_2 from the analysis, which simplifies matters.

Marko in [37] assumed stationarity. He then defined the first three information rates in (3.10) and used these to show the intuitively pleasing result [37, Equation (14)]

$$I_\infty(X \to Y) + I_\infty(Y \to X) = \lim_{N\to\infty} \frac{1}{N} I(X^N;Y^N). \tag{3.23}$$

This result follows from Property 3.4 and the functional dependence graph, which ensures that $I(X_n;Y_n|X^{n-1}Y^{n-1}) = 0$ because $X^{n-1}Y^{n-1}$ d-separates X_n from Y_n. Marko further defined and gave bounds on quantities such as "residual entropy" and "stochastical degree of synchronization". He also showed how these quantities, and directed transinformation in particular, may be applied to study the social behavior of monkeys. For example, Marko was able to quantify the information

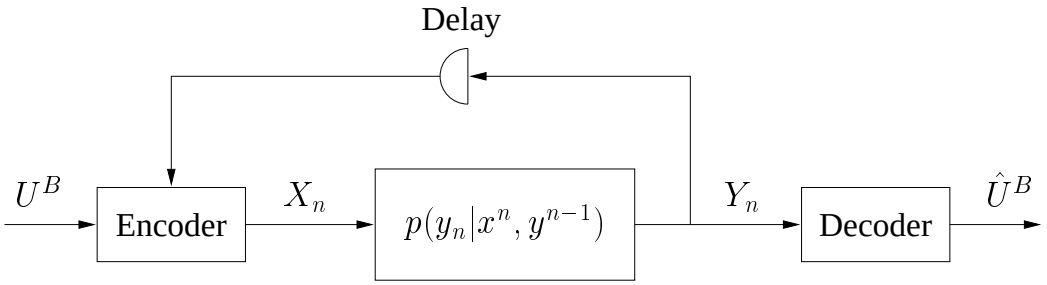


Figure 3.5: *The discrete channel with feedback.*

flow from one monkey to another when typical behavioral activities were used to form sequences of "random" variables.

We will not consider Marko's problem any further. As shown in the next section, directed information has applications beyond Marko's two-user problem, e.g. to data transmission over the discrete channel with feedback.

3.4 The Discrete Channel with Feedback

A discrete channel is defined by a discrete input alphabet $\mathcal{X}$, a discrete output alphabet $\mathcal{Y}$ and the probability distributions

$$p(y_n|x^n y^{n-1})$$

for all x_n in $\mathcal{X}$, y_n in $\mathcal{Y}$ and $n \geq 1$. The discrete channel with feedback is depicted in Figure 3.5, and its functional dependence graph for $n = 3$ is shown in Figure 3.6. In these figures, U^B is the sender's B-bit information sequence and $\hat{U}^B$ is the receiver's estimate of U^B. X_n is the nth input into the channel and Y_n is the nth output of the channel. The sender generates the input symbol X_n using U^B and Y^{n-1}, i.e.,

$$X_n = f_n(U^B, Y^{n-1}) \tag{3.24}$$

for some functions $f_1, f_2, \ldots$.

As a first step in applying directed information to the discrete channel with feedback, we combine Theorems 1 and 3 of [36] as a lemma. In this lemma, the channel was said by Massey to be *used without feedback* if

$$p(x_n|x^{n-1}y^{n-1}) = p(x_n|x^{n-1}). \tag{3.25}$$

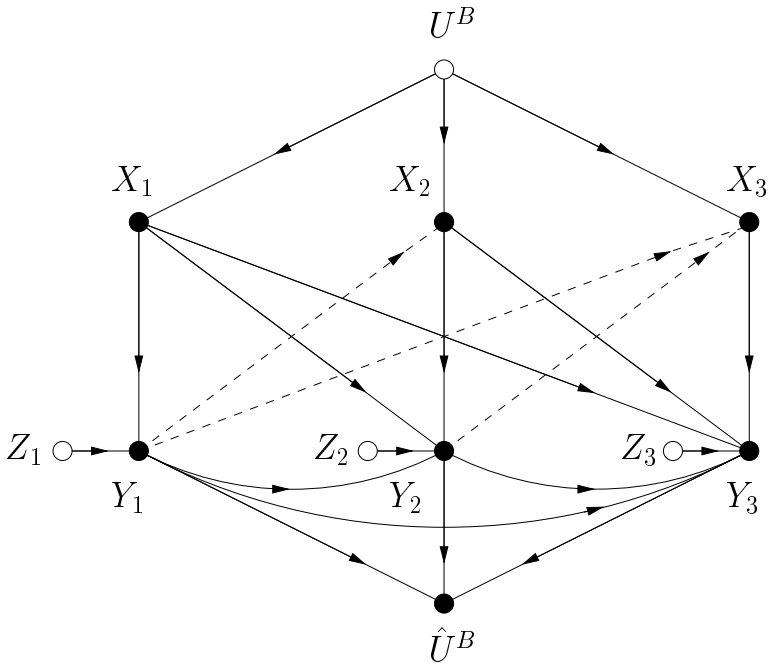


Figure 3.6: *The functional dependence graph for the first three uses of the the discrete channel with feedback. The feedback links are drawn using dashed lines.*

Lemma 3.1 *(Massey [36]) If U^B is the information bit sequence into the encoder, and if X^N and Y^N are the input and output sequence, respectively, of a discrete channel, then*

$$I(U^B;Y^N) = I(X^N \to Y^N) \leq I(X^N;Y^N)$$

with equality on the right if the channel is used without feedback.

Proof: We expand

$$\begin{aligned} I(U^B;Y^N) &= \sum_{n=1}^{N} \left[H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}U^B)\right] \\ &= \sum_{n=1}^{N} \left[H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}U^B X^n)\right] \\ &= \sum_{n=1}^{N} \left[H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1}X^n)\right] \\ &= I(X^N \to Y^N), \end{aligned}$$

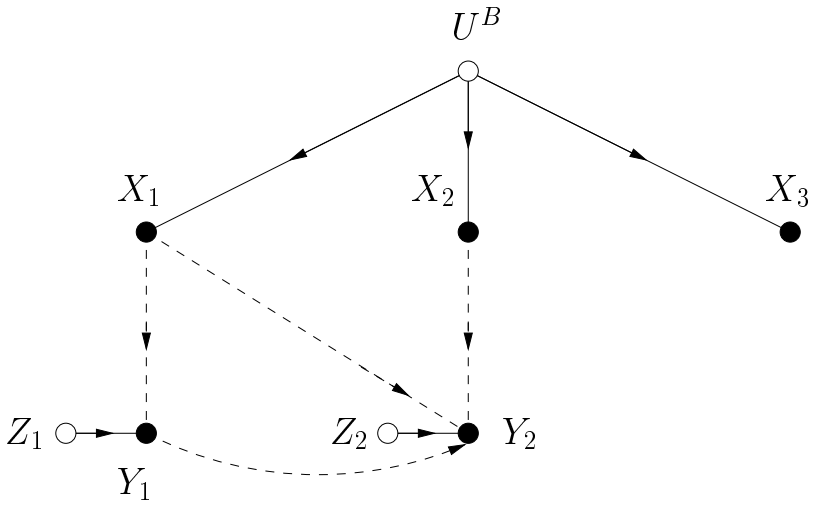


Figure 3.7: *The functional dependence graph relevant to Y^2 and X^3 for the discrete channel without feedback. The branches out of the conditioning random variables in $I(X_3; Y_2|Y_1X^2)$ are shown using dashed lines.*

where the second and third equalities follow from the functional dependence graph relevant to U^B and Y^n.
The inequality in the lemma is just Property 3.2. For equality, Property 3.2 requires $H(Y_n|Y^{n-1}X^n) = H(Y_n|Y^{n-1}X^N)$, all n, $1 \leq n \leq N$, which is equivalent to (3.25). We now show that this condition is indeed met when the channel is used without feedback in an operational sense. The functional dependence graph relevant to Y^n and X^N, for any $N \geq 1$ and any n such that $1 \leq n \leq N$, can be used for this purpose. For example, Figure 3.7 depicts the functional dependence graph relevant to Y^2 and X^3. By cutting the branches out of the conditioning random variables Y_1 and X^2 we see that $I(X_3; Y_2|Y_1X^2) = 0$, or $H(Y_2|Y_1X^2) = H(Y_2|Y_1X^3)$. By constructing the appropriate functional dependence graph for any other $N \geq 1$ and n such that $1 \leq n \leq N$, one sees that $H(Y_n|Y^{n-1}X^n) = H(Y_n|Y^{n-1}X^N)$. □

The inequality in Lemma 3.1 may actually be *strict* when feedback is used. Thus, as mentioned in [36], directed information can give stronger bounds than mutual information on the information transfer between U^B and Y^N.

We will not pursue the analysis of the discrete channel with feedback further. Instead, we will concentrate on multi-user channels in Chap-

ters 4 and 5, where we show that the capacity region of some important channels are characterized completely by causally conditioned directed information rates.

3.A Appendix: Proof of a Stationarity Property

We now prove Property 3.5 in Section 3.2.2. The steps of this proof are virtually identical to those given in [52, Section 2.4].

Proof: We first prove part 2.

$$H(X_{L+1}|X^L Y^{L+1}) \leq H(X_{L+1}|X^{2\,..\,L} Y^{2\,..\,L+1}) = H(X_L|X^{L-1}Y^L)$$

where the inequality follows because removing conditioning cannot decrease uncertainty, and where the equality follows from the stationarity of the source. This proves part 2.
Next, consider part 1. From the definition of $H_L(X\|Y)$ we have

$$L \cdot H_L(X\|Y) = \sum_{k=1}^{L} H(X_k|X^{k-1}Y^k).$$

From part 2, we know that $H(X_k|X^{k-1}Y^k) \geq H(X_L|X^{L-1}Y^L)$ for $1 \leq k \leq L$. Thus,

$$L \cdot H_L(X\|Y) \geq L \cdot H(X_L|X^{L-1}Y^L),$$

which proves part 1.
We now prove part 3 using parts 1 and 2. From the definition of $H_L(X\|Y)$ and $H_{L+1}(X\|Y)$, we obtain

$$(L+1) \cdot H_{L+1}(X\|Y) = L \cdot H_L(X\|Y) + H(X_{L+1}|X^L Y^{L+1}).$$

But

$$H(X_{L+1}|X^L Y^{L+1}) \leq H(X_L|X^{L-1}Y^L) \leq H_L(X\|Y)$$

where the first inequality follows from part 2 and the second from part 1. Thus,

$$(L+1) \cdot H_{L+1}(X\|Y) \leq (L+1) \cdot H_L(X\|Y),$$

which proves part 3.
The limits in part 4 must exist because both $H(X_L|X^{L-1}Y^L)$ and

$H_L(X\|Y)$ are non-increasing with L and lower bounded by zero. Furthermore, part 1 implies that

$$\lim_{L\to\infty} H(X_L|X^{L-1}Y^L) \leq \lim_{L\to\infty} H_L(X\|Y) = H_\infty(X\|Y). \tag{3.26}$$

It remains to show that equality holds in (3.26). We first consider

$$(L+n)\cdot H_{L+n}(X\|Y) = L\cdot H_L(X\|Y) + \sum_{k=L+1}^{L+n} H(X_k|X^{k-1}Y^k).$$

By part 2, the n terms in the sum cannot be larger than the first term in the sum. Thus,

$$(L+n)\cdot H_{L+n}(X\|Y) \leq L\cdot H_L(X\|Y) + n\cdot H(X_{L+1}|X^LY^{L+1}).$$

Dividing both sides by $L+n$ we obtain

$$H_{L+n}(X\|Y) \leq \frac{L}{L+n}\cdot H_L(X\|Y) + \frac{n}{L+n}\cdot H(X_{L+1}|X^LY^{L+1}).$$

Now letting n approach infinity gives

$$H_\infty(X\|Y) \leq H(X_{L+1}|X^LY^{L+1}),$$

so that

$$H_\infty(X\|Y) \leq \lim_{L\to\infty} H(X_{L+1}|X^LY^{L+1}).$$

This, together with (3.26), proves part 4. □

Chapter 4

Directed Information for the Two-Way Channel

This chapter applies the definitions and results of Chapter 3 to the two-way channel. We show that the capacity region of the two-way channel can be expressed in terms of causally conditioned directed information rates. Most of this chapter is concerned with coding techniques, and we concentrate on finding simple *inner bounds* to the capacity region of the two-way channel. We use *concatenated codes* to describe and generalize the two-way channel coding techniques of Shannon [17] and Han [22], and we review the bootstrapping technique of Schalkwijk [20, 21]. For *outer bounds* to the capacity region, we refer to Shannon [17, Section 9], Zhang, Berger and Schalkwijk [23], and Hekstra and Willems [24].

This chapter is organized as follows. Section 4.1 introduces the model of the two-way channel and describes the structure of the codes for this channel. In Section 4.2 we show that Shannon's general solution for the capacity of the two-way channel can be written as a limiting expression involving causally-conditioned directed information rates. Section 4.3 describes steady-state and concatenated coding techniques. Finally, Section 4.4 applies the inner bounds of Section 4.3 to certain two-way channels.

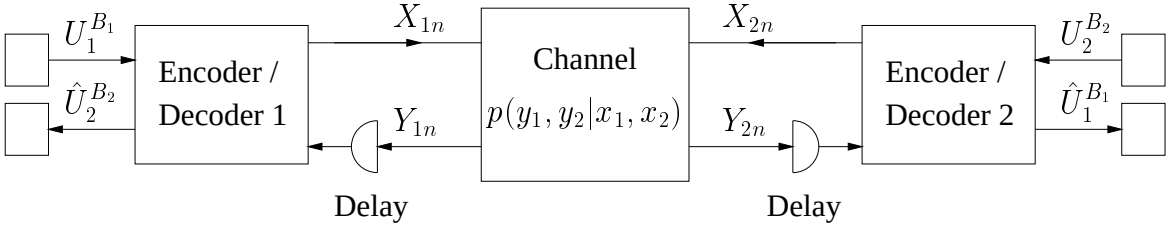


Figure 4.1: *The two-way channel.*

4.1 Model and Adaptive Codewords

Shannon introduced the two-way channel in [17]. The discrete memoryless two-way channel is defined by the discrete input alphabets $\mathcal{X}_1$ and $\mathcal{X}_2$, the discrete output alphabets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ and the probability distribution $p(y_1, y_2|x_1, x_2)$ where

$$p(y_{1n}, y_{2n}|x_1^n, y_1^{n-1}, x_2^n, y_2^{n-1}) = p_{Y_1Y_2|X_1X_2}(y_{1n}, y_{2n}|x_{1n}, x_{2n}). \quad (4.1)$$

The two-way channel is depicted in Figure 4.1. User 1 transmits the information sequence $U_1^{B_1}$ of B_1 bits, and User 2 transmits the information sequence $U_2^{B_2}$ of B_2 bits. The sequences $U_1^{B_1}$ and $U_2^{B_2}$ are independent and have entropy B_1 and B_2 bits, respectively. The channel is assumed to be used N times so that the transmission rate-pair of the users is $(R_1, R_2) = (B_1/N, B_2/N)$ bits per use. The symbols input to the channel at time n are

$$X_{1n} = f_{1n}(U_1^{B_1}, Y_1^{n-1}) \quad (4.2)$$
$$X_{2n} = f_{2n}(U_2^{B_2}, Y_2^{n-1}), \quad (4.3)$$

i.e., they are functions of the information bit sequences of the two users and the previous outputs Y_1^{n-1} and Y_2^{n-1} of the channel. After completing their transmissions over the channel, Users 1 and 2 output their respective decisions $\hat{U}_2^{B_2}$ and $\hat{U}_1^{B_1}$ for $U_2^{B_2}$ and $U_1^{B_1}$.

Equations (4.2) and (4.3) specify the form of the code one may use for the two-way channel. The code consists of *adaptive codewords* whose symbols at time n depend on the feedback Y_1^{n-1} or Y_2^{n-1}. Two examples of such codes for a two-way channel are shown in Figure 4.2 (see also [54, Sections 9.1 and 9.4]).

We will denote the length N adaptive codeword for Users 1 and 2 by the rooted trees $\mathbf{a}_1^N(u_1^{B_1})$ and $\mathbf{a}_2^N(u_2^{B_2})$ where $u_1^{B_1}$ and $u_2^{B_2}$ are the

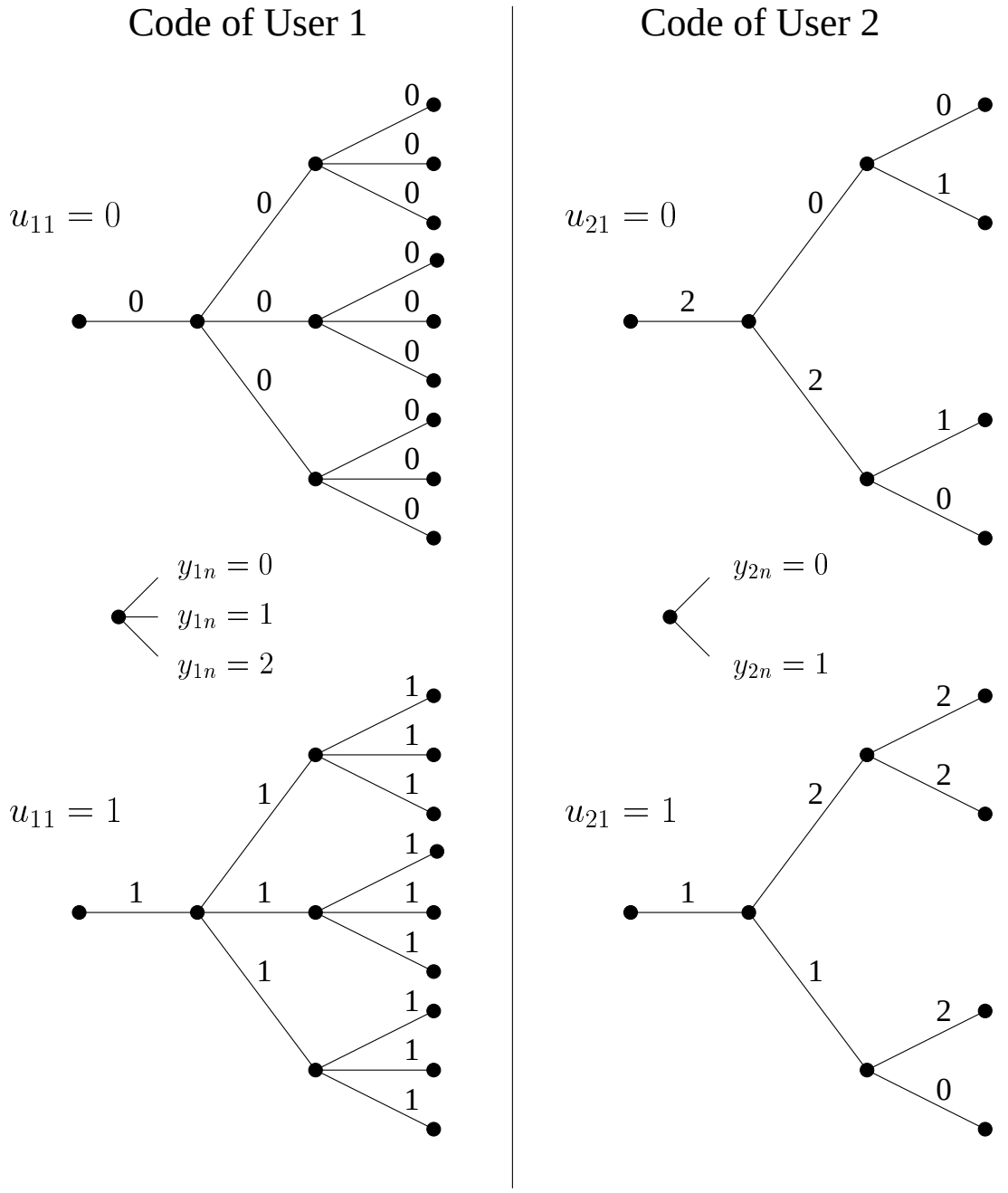


Figure 4.2: *Codes with adaptive codewords for a two-way channel. The symbol alphabets are* $\mathcal{X}_1 = \{0, 1\}$, $\mathcal{X}_2 = \{0, 1, 2\}$ *and* $\mathcal{Y}_1 = \mathcal{X}_2$, $\mathcal{Y}_2 = \mathcal{X}_1$. *An example of such a channel is* $Y_1 = X_1 \cdot X_2$ *and* $Y_2 = Y_1 \mod 2$ *where the multiplication is performed over the integers. Both users are trying to send* $B_1 = B_2 = 1$ *bit of information using the channel* $N = 3$ *times. The code for User 1 is a repeat code, while User 2's adaptive codewords were randomly chosen from the* $3 \cdot 3^2 \cdot 3^4 = 2187$ *possible adaptive codewords. As an example of how the codes function, assume that User 2's information bit is 0. Then he sends* $x_{21} = 2$ *for the first use of the channel. If* $y_{21} = 0$ *he would next send* $x_{22} = 0$, *and so on.*

corresponding information sequences. If J_1 is the cardinality of $\mathcal{Y}_1$, then there are J_1 branches leaving each vertex at depth i from the root in $\mathbf{a}_1^N(u_1^{B_1})$ for $i = 1, 2, \ldots, N-1$ while a single branch leaves the root. Each branch in the tree $\mathbf{a}_1^N(u_1^{B_1})$ is labelled with a symbol in $\mathcal{X}_1$. The meaning is that the actual transmitted sequence X_1^N when $U_1^{B_1} = u_1^{B_1}$ will be the path in $\mathbf{a}_1^N(u_1^{B_1})$ from the root to the leaf corresponding to the actual value y_1^{N-1} of the feedback to User 1. For example, consider the code of User 2 in Figure 4.2. The adaptive codewords $\mathbf{a}_2^3(0)$ and $\mathbf{a}_2^3(1)$ are the two trees on the right in Figure 4.2. We will sometimes denote the rooted trees specifying adaptive codewords in the manner

$$\begin{aligned} \mathbf{a}_2^3(0) &= [\,2\,,\,02\,,\,0110\,] \\ \mathbf{a}_2^3(1) &= [\,1\,,\,21\,,\,2220\,]. \end{aligned}$$

The meaning is for instance that $\mathbf{a}_{21}(0) = 2$ is the symbol on the branch leaving the root of the tree $\mathbf{a}_2^3(0)$ (cf. Fig. 4.2), that $\mathbf{a}_{22}(0) = [0, 2]$ is the ordered list of symbols leaving the vertex at depth 1, and that $\mathbf{a}_{23}(0) = [0, 1, 1, 0]$ is the ordered list of symbols leaving the vertices at depth 2 in the tree. The adaptive codewords $\mathbf{a}_1^N(u_1^{B_1})$ and $\mathbf{a}_2^N(u_2^{B_2})$ are all selected *before* the channel is used and are known by both users. The functional dependence graph of the two-way channel which includes the adaptive codewords $\mathbf{A}_1^N(U_1^{B_1})$ and $\mathbf{A}_2^N(U_2^{B_2})$ is depicted in Figure 4.3. In this figure, we have denoted the adaptive codewords simply as A_1^N and A_2^N where the dependence on the information sequences $U_1^{B_1}$ and $U_2^{B_2}$ is understood. The random variables $Z_1, Z_2, \ldots$ represent noise, i.e.,

$$\begin{aligned} Y_{1n} &= f(X_{1n}, X_{2n}, Z_n) \\ Y_{2n} &= g(X_{1n}, X_{2n}, Z_n) \end{aligned}$$

for some functions f and g.

4.2 Shannon's General Solution for the Capacity Region

Shannon in [17, Section 15] described the capacity region of the two-way channel. His description was as a limiting expression of inner bound regions which we develop in Section 4.2.2. First, however, we review Shannon's outer bound.

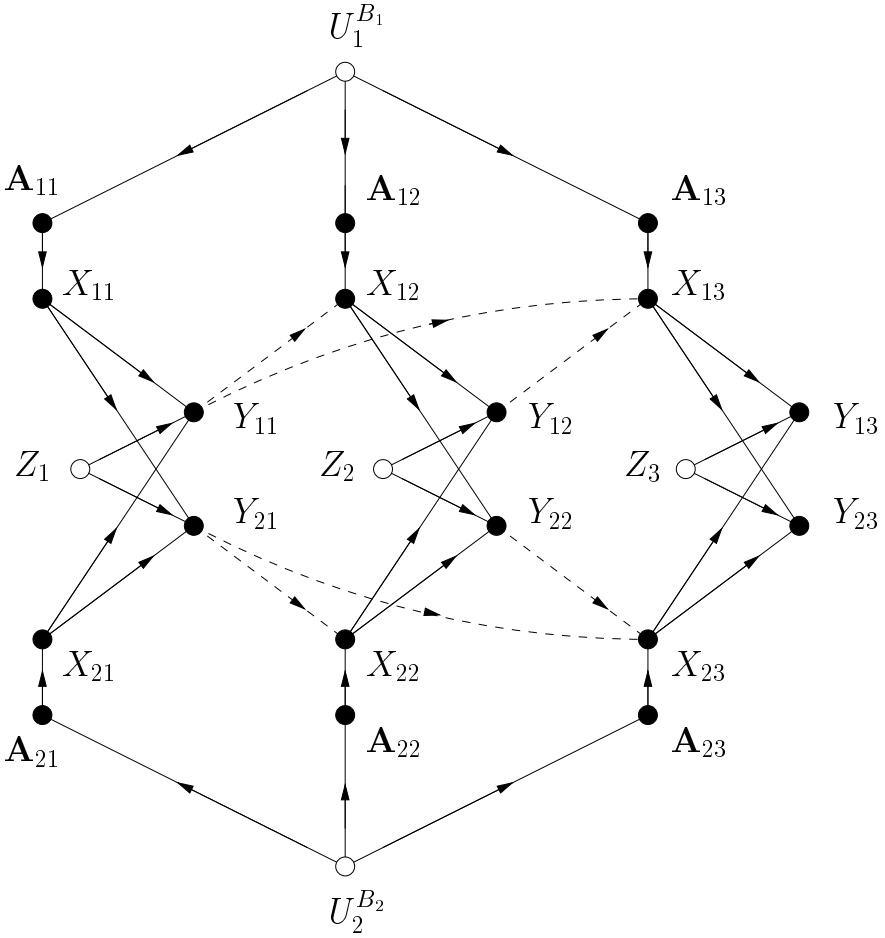


Figure 4.3: *The functional dependence graph for three uses of the two-way channel including the adaptive codewords. The feedback links are drawn using dashed lines.*

4.2.1 An Outer Bound

Consider the data transmission from User 1 to User 2. Let the probability that User 2 incorrectly decodes the bth bit of $U_1^{B_1}$ be denoted by $P_{1b} = \Pr(\hat{U}_{1b} \neq U_{1b})$. The *average* bit error probability of User 2 is then

$$\overline{P}_1 = \frac{1}{B_1} \sum_{b=1}^{B_1} P_{1b}, \tag{4.4}$$

and the error probabilities P_{2b} and $\overline{P}_2$ of User 1 are similarly defined. The *capacity region* $\mathcal{C}_{\mathrm{TWC}}$ of the two-way channel is the set of *approachable* rate-pairs (R_1, R_2), i.e., those rate-pairs which one can approach with arbitrarily small positive $\overline{P}_1$ and $\overline{P}_2$. [See the "Notation and Ter-

minology" section for comments concerning "achievable" and "approachable".]

In Appendix 4.A we show that $\overline{P}_1$ satisfies (see Equations (4.49) and (4.50))

$$h(\overline{P}_1) \geq 1 - \frac{1}{B_1} I(\mathbf{A}_1^N; X_2^N Y_2^N). \tag{4.5}$$

where $h(\cdot)$ is the binary entropy function [48, p. 14]. Furthermore, we show that (see Equation (4.51))

$$I(\mathbf{A}_1^N; X_2^N Y_2^N) = I(\mathbf{A}_1^N \to Y_2^N \| X_2^N). \tag{4.6}$$

In Appendix 4.A these two results are used to prove the following lemma.

Lemma 4.1 (An Outer Bound to the Capacity Region) *The capacity region $\mathcal{C}_{\mathrm{TWC}}$ of the discrete memoryless two-way channel is contained within the closure $\mathcal{C}_{\mathrm{TWC}}^{\mathrm{OUT}}$ of the set of rate-pairs (R_1, R_2) such that*

$$\begin{aligned} R_1 &= I_L(\mathbf{A}_1 \to Y_2 \| X_2) && (4.7) \\ R_2 &= I_L(\mathbf{A}_2 \to Y_1 \| X_1), && (4.8) \end{aligned}$$

where L is a positive integer, and where $\mathbf{A}_1^L$ and $\mathbf{A}_2^L$ are independent adaptive codewords.

We remark that the region $\mathcal{C}_{\mathrm{TWC}}^{\mathrm{OUT}}$ as specified in the lemma is automatically convex because allowing L to be any positive integer implicitly allows time-sharing (cf. [17, Section 8]).

4.2.2 Inner Bounds

Shannon in [17, Sections 6-8] showed that the convex hull of rate-pairs (R_1, R_2) such that

$$\begin{aligned} R_1 &= I(X_1; Y_2 \mid X_2) && (4.9) \\ R_2 &= I(X_2; Y_1 \mid X_1) && (4.10) \end{aligned}$$

is contained within the capacity region, where X_1 and X_2 are independent. We call this rate region Shannon's *First* Inner Bound Region to

distinguish it from the other inner bound regions he derived. To derive his other inner bounds, Shannon considered adaptive codewords such as those of Figure 4.2 to be the input *symbols* of a *derived* channel consisting of L consecutive uses of the two-way channel [17, Sections 15]. Shannon then suggested forming long codewords by concatenating many, say M, randomly chosen "symbols" $\mathbf{a}_1^L$. The random codes formed in this manner are used to prove the following lemma in Appendix 4.B.

Lemma 4.2 (Shannon's Lth Inner Bound Region) *The convex hull of the set of rate-pairs (R_1, R_2) such that (in our notation)*

$$R_1 = I_L(\mathbf{A}_1 \to Y_2 \| X_2) \tag{4.11}$$

$$R_2 = I_L(\mathbf{A}_2 \to Y_1 \| X_1) \tag{4.12}$$

is contained within the capacity region of the discrete memoryless two-way channel, where $\mathbf{A}_1^L$ and $\mathbf{A}_2^L$ are independent adaptive codewords and L is some positive integer.

We call the region $\mathcal{R}_L$ corresponding to the integer L in Lemma 4.2 *Shannon's Lth Inner Bound Region* to the capacity region of the two-way channel. Setting $L = 1$ in (4.11) and (4.12) yields (4.9) and (4.10) because $X_{11} = f(\mathbf{A}_{11})$ and $X_{21} = g(\mathbf{A}_{21})$ for some functions f and g, and $H(Y_1|X_{11}X_{21}\mathbf{A}_{11}\mathbf{A}_{21}) = H(Y_1|X_{11}X_{21})$.

Shannon next proved the following theorem in [17, Section 15]. For this theorem, $\lim_{L\to\infty} \mathcal{R}_L$ denotes the set of limit points of all convergent sequences (R_{11}, R_{21}), (R_{12}, R_{22}), $(R_{13}, R_{23}), \ldots$, where (R_{1L}, R_{2L}) is a point in $\mathcal{R}_L$.

Theorem 4.1 (The Capacity Region) *The capacity region $\mathcal{C}_{\text{TWC}}$ of the discrete memoryless two-way channel is the region $\mathcal{C}_{\text{TWC}}^{\text{OUT}}$ of Lemma 4.1. Furthermore, if $\mathcal{R}_L$ is Shannon's Lth Inner Bound Region, then*

$$\mathcal{C}_{\text{TWC}} = \lim_{L\to\infty} \mathcal{R}_L. \tag{4.13}$$

Proof: The region $\mathcal{C}_{\text{TWC}}^{\text{OUT}}$, being the *closure* of the set $\{\mathcal{R}_L : L \text{ finite}\}$, is (see [53, page 743])

$$\mathcal{C}_{\text{TWC}}^{\text{OUT}} = \{\mathcal{R}_L : L \text{ finite}\} \bigcup \lim_{L\to\infty} \mathcal{R}_L. \tag{4.14}$$

From Lemma 4.2 all points in $\{\mathcal{R}_L : L \text{ finite}\}$ are approachable, and so are all points in $\lim_{L\to\infty} \mathcal{R}_L$ because these points are limits of convergent sequences in $\{\mathcal{R}_L : L \text{ finite}\}$. This establishes that $\mathcal{C}_{\text{TWC}} = \mathcal{C}_{\text{TWC}}^{\text{OUT}}$. To prove (4.13) it remains to show that $\{\mathcal{R}_L : L \text{ finite}\} \subseteq \lim_{L\to\infty} \mathcal{R}_L$, i.e., that for any point (R_1, R_2) in $\mathcal{R}_L$ one can generate a sequence in $\{\mathcal{R}_L : L \text{ finite}\}$ whose limit is (R_1, R_2). The trick Shannon used in [17, Section 15] was to write $\ell = mL + j$ for some integers $m \geq 0$ and $0 \leq j < L$, and then to randomize the code by repeating m times the distributions for (R_1, R_2), and by sending no information in the last j uses of the channel. This means that $(R_{1\ell}, R_{2\ell}) = mL/(mL + j) \cdot (R_1, R_2)$ lies in $\mathcal{R}_\ell$. Moreover, $\lim_{m\to\infty}(R_{1(mL+j)}, R_{2(mL+j)}) = (R_1, R_2)$. This proves the theorem. □

4.2.3 The Common-Output Two-Way Channel

A two-way channel is called a *common-output* two-way channel if $Y_1 = Y_2$, in which case we denote the common output by Y. For this channel X_{1n} is determined by $\mathbf{A}_1^n$ and Y^{n-1}, and X_{2n} is determined by $\mathbf{A}_2^n$ and Y^{n-1}. This means that

$$\begin{aligned} H(Y^N \| X_2^N \mathbf{A}_1^N) &= H(Y^N \| X_2^N \mathbf{A}_1^N X_1^N) \\ &= H(Y^N \| X_2^N X_1^N), \end{aligned} \tag{4.15}$$

where both steps follow from the functional dependence graph of the two-way channel. Similarly, $H(Y^N \| X_1^N \mathbf{A}_2^N) = H(Y^N \| X_1^N X_2^N)$, so that

$$\begin{aligned} I(\mathbf{A}_1^N \to Y^N \| X_2^N) &= I(X_1^N \to Y^N \| X_2^N), \\ I(\mathbf{A}_2^N \to Y^N \| X_1^N) &= I(X_2^N \to Y^N \| X_1^N). \end{aligned} \tag{4.16}$$

We thus have the following corollary to Theorem 4.1.

Corollary 4.1 *The capacity region $\mathcal{C}_{\text{TWC}}$ of the discrete memoryless common-output two-way channel is the closure of the set of rate-pairs (R_1, R_2) such that*

$$R_1 = I_L(X_1 \to Y \| X_2) \tag{4.17}$$

$$R_2 = I_L(X_2 \to Y \| X_1) \tag{4.18}$$

where L is a positive integer and where $p(x_{1\ell}, x_{2\ell}|x_1^{\ell-1}, x_2^{\ell-1}, y^{\ell-1})$ factors as

$$q(x_{1\ell}|x_1^{\ell-1}, y^{\ell-1}) \cdot q(x_{2\ell}|x_2^{\ell-1}, y^{\ell-1}) \tag{4.19}$$

for all $\ell = 1, 2, \ldots, L$.

Proof: Consider the functional dependence graph in Figure 4.3, and combine Y_1 and Y_2 into a common vertex Y. By cutting the appropriate branches we see that $X_1^{\ell-1}Y^{\ell-1}$ d-separates $X_{1\ell}$ from X_2^{ℓ} for all ℓ. Similarly, $X_2^{\ell-1}Y^{\ell-1}$ d-separates $X_{2\ell}$ from X_1^{ℓ} for all ℓ. From these results we obtain the factorization (4.19). Equations (4.17) and (4.18) follow from Lemmas 4.1 and 4.2, and from $H(Y_n|X_{1n}X_{2n}\mathbf{A}_1^n\mathbf{A}_2^n) = H(Y_n|X_{1n}X_{2n})$. □

Equation (4.19) actually simplifies the random coding. Rather than having to generate the symbols of the ℓth branch of the tree jointly via $p(\mathbf{a}_{1n}|\mathbf{a}_1^{n-1})$, one can generate all the symbols on the branches coming out of the same vertex of the adaptive codeword *independently* using the same distribution $p_{X_{1n}|X_1^{n-1}Y_1^{n-1}}(\cdot|x_1^{n-1}y_1^{n-1})$.

The rates given by Equations (4.17) and (4.18) are *upper* bounds on the rates given in Lemma 4.2, as demonstrated in Section 4.4.2. In fact, Equations (4.17) and (4.18) (in a different form) were used in [23] to derive outer bounds to the capacity region of the general two-way channel. This suggests that, by using the rates of Lemma 4.2, it should be possible to obtain outer bounds on $\mathcal{C}_{\text{TWC}}$ that are tighter than those of [23].

4.3 Coding Techniques

The previous section used codes with M repetitions of adaptive codewords of fixed length L, and let M become large. However, Lemma 4.4 in Appendix 4.B, as well as the entire development in this appendix, requires only that $N = L \cdot M$ become large and says nothing about the relation between L and M. In this section we first consider the case $M = 1$ and $L = N$ for which we develop a steady-state approachable rate region. We then consider *concatenated codes* and show that the codes of Shannon [17] and Han [22] are special cases of such codes.

4.3.1 Steady-State Coding

Consider the case $M = 1$ and $L = N$. The coding distribution $q_{\mathbf{A}_1^N}$ can be specified by making $\mathbf{a}_{1n}$ dependent on the state σ_{1n} of User 1 at time n, where σ_{1n} is some function of $\mathbf{a}_1^{n-1}$, i.e., $q(\mathbf{a}_{1n}|\mathbf{a}_1^{n-1}) = q(\mathbf{a}_{1n}|\sigma_{1n})$ where $\sigma_{1n} = f_n(\mathbf{a}_1^{n-1})$ for some f_n. User 2 uses the same coding technique. We will consider only distributions using a finite number of states and we call $\sigma_n = (\sigma_{1n}, \sigma_{2n})$ the *system state*.

Each of the random state sequences Σ_1, Σ_2 and Σ is a Markov chain. As an additional simplification, we will consider only *ergodic* Markov chains, i.e., Markov chains for which all states are *recurrent* and *aperiodic*.[1] Ergodic Markov chains are guaranteed to have a unique stationary distribution, which is also their *steady-state* distribution [55, Section 4.2]. It follows from Lemma 4.4 in Appendix 4.B that, as N increases, the distribution of the random variables at successive times will approach the steady-state distribution of $(\Sigma_n, \mathbf{A}_{1n}, \mathbf{A}_{2n}, Y_{1n}, Y_{2n})$. Thus, the steady-state information rates

$$R_1 = I_\infty(\mathbf{A}_1 \to Y_2 \| X_2) \tag{4.20}$$

$$R_2 = I_\infty(\mathbf{A}_2 \to Y_1 \| X_1), \tag{4.21}$$

which exist by Property 3.6 of a DSS, are approachable. Note that, because N is finite, we need not worry about the problem of decoding at regular intervals, which we would have to deal with if a semi-infinite code were used.

We will consider two approaches to steady-state coding with a finite number of states in Section 4.4. The first is *memory* μ coding, in which each user chooses the next *channel input* symbol dependent on only the past μ observed *channel input and output* symbols. The second approach uses distributions where the system state returns to the starting state after a finite number of uses of the channel.

[1] One could use the more general notion of "ergodic" to allow *periodic* Markov chains and this would include time-sharing [55, p. 118]. However, for the channels and rates we will consider, time-sharing is not crucial. Moreover, one can include time-sharing by simply adding a convex hull operation to the rate region achieved with ergodic chains.

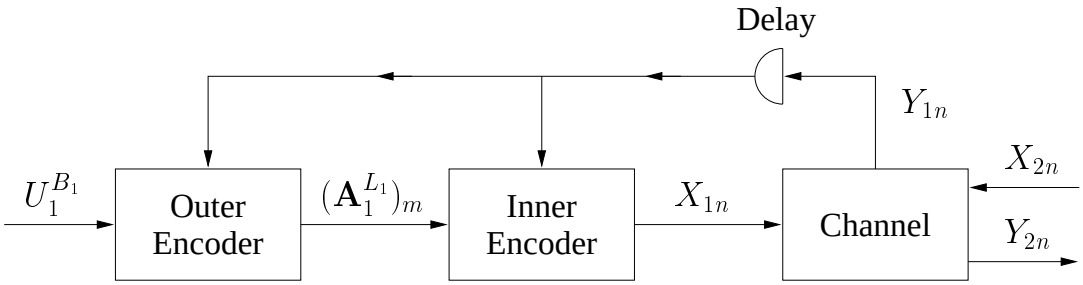


Figure 4.4: *A concatenated code for User 1. The inner code has length L_1 and the outer code has length L_2. Each symbol $\mathbf{B}_{1m}$ of the outer codeword $\mathbf{B}_1^{L_2}$ chooses one of the inner codewords $(\mathbf{A}_1^{L_1})_m$ based on the past output $(Y_1^{L_1})^{m-1}$.*

4.3.2 Concatenated Codes

We use *concatenated codes* [56, p. 278] as a framework for classifying and generalizing the coding techniques of Shannon and Han [22]. A concatenated code consists of an *inner* code having codewords of length L_1, and an *outer* code having codewords of length L_2 (see Figure 4.4). The labels of the branches of the inner adaptive codewords are channel input symbols, while the labels of the branches of the outer adaptive codewords are codewords from the inner code. In other words, the outer code is a code for the "overall" channel formed by the inner encoder, the two-way channel, and the inner decoder. We write $\mathbf{a}_1^{L_1}$ and $\mathbf{b}_1^{L_2}$ to denote an adaptive codeword of User 1's inner code and outer code, respectively. Furthermore, we write $(X_{1\ell})_m$ to denote the ℓth inner-code symbol of the mth outer-code symbol, and similarly for $(X_{2\ell})_m$, $(Y_{1\ell})_m$ and $(Y_{2\ell})_m$. By repeating the concatenated codes many times, Shannon's Lth Inner Bound Region becomes the convex hull of rate pairs (R_1, R_2) such that

$$R_1 = I_{L_2}(\mathbf{B}_1 \to (Y_2^{L_1}) \parallel (X_2^{L_1})) \tag{4.22}$$

$$R_2 = I_{L_2}(\mathbf{B}_2 \to (Y_1^{L_1}) \parallel (X_1^{L_1})), \tag{4.23}$$

where $\mathbf{B}_1$ and $\mathbf{B}_2$ are independent.

The most general concatenated code has adaptive codewords for both the inner code and the outer code. However, we will consider only the cases where only the inner code *or* only the outer code is adaptive, but not both.

Non-Adaptive Outer Code

A non-adaptive outer code has no feedback to the outer-codeword symbols. This is precisely Shannon's coding technique for the two-way channel with $L_1 = L$ and $L_2 = M$. The approachable rates for finite L_1 and large L_2 are thus given by Shannon's L_1th Inner Bound Region.

Non-Adaptive Inner Code

A non-adaptive inner code has no feedback to the inner-codeword symbols, so we simply set $(\mathbf{A}_1^{L_1})_m = (X_1^{L_1})_m$ and $(\mathbf{A}_2^{L_1})_m = (X_2^{L_1})_m$ for all $m = 1, 2, \ldots, L_2$. We further assume that, for $\ell = 1, \ldots, L_1$, all $(X_{1\ell})_m$ have the same distribution conditional on $\mathbf{B}_1^m$ and $(Y_1^{L_1})^{m-1}$, and all $(X_{2\ell})_m$ have the same distribution conditional on $\mathbf{B}_2^m$ and $(Y_2^{L_1})^{m-1}$. With this, the terms in the sum (3.8) defining (4.22) become

$$L_1 \cdot I(\mathbf{B}_1^m; (Y_{2\ell})_m \mid (X_{2\ell})_m (X_2^{L_1})^{m-1} (Y_2^{L_1})^{m-1}), \tag{4.24}$$

where ℓ is any of the integers $1, 2, \ldots, L_1$, and similarly for (4.23).

Next, consider the simpler random coding technique where $\mathbf{B}_{1m}$ and $\mathbf{B}_{2m}$ are length L_1 vectors for all m. Furthermore, assume that the ℓth symbol of each inner code's codeword was generated using only the ℓth symbol of all previous inner and outer codewords , i.e.,

$$\begin{aligned} &p((x_{1\ell})_m \mid \mathbf{b}_1^m, (x_1^{L_1})^{m-1}, (y_1^{L_1})^{m-1}) \\ &\quad = p((x_{1\ell})_m \mid (b_{1\ell})^m, (x_{1\ell})^{m-1}, (y_{1\ell})^{m-1}), \end{aligned} \tag{4.25}$$

and similarly for User 2. In this case (4.24) simplifies to

$$L_1 \cdot I((B_{1\ell})^m; (Y_{2\ell})_m \mid (X_{2\ell})^m (Y_{2\ell})^{m-1}), \tag{4.26}$$

where ℓ is any of the integers $1, 2, \ldots, L_1$.

Han's Coding Technique

The steady-state code introduced by Han in [22] is a special case of a non-adaptive inner code satisfying (4.25) and the corresponding simplification for User 2. Furthermore, the dependence of the inner code

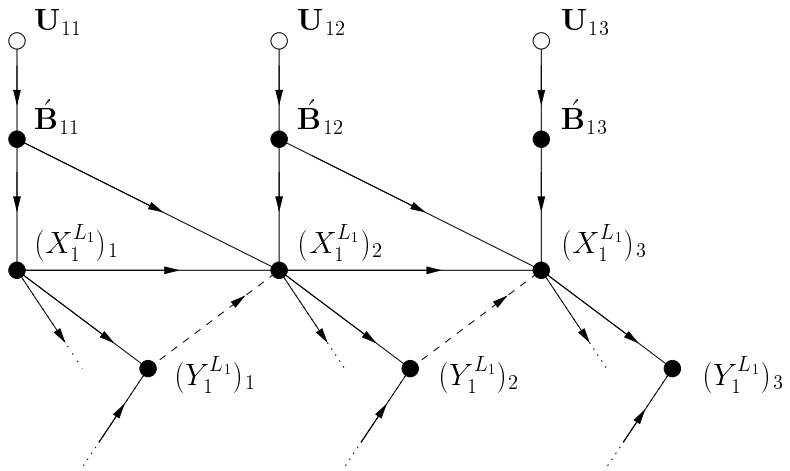


Figure 4.5: *The dependence of the inner code on the information sequence for Han's coding technique. The symbols* $\mathbf{U}_{1m}$*,* $m = 1, \ldots, L_2$*, are independent, and the inner code is nonadaptive. Only part of the functional dependence graph is shown.*

on the information symbols is produced in the manner depicted in the functional dependence graph of Figure 4.5. The information sequence $U_1^{B_1}$ is divided into L_2 independent blocks of bits and we denote the mth block by $\mathbf{U}_{1m}$, $m = 1, \ldots, L_2$. Each block $\mathbf{U}_{1m}$ determines a symbol (or codeword) $\acute{\mathbf{B}}_{1m}$. The inner codeword $(X_1^{L_1})_m$ depends on the present symbol $\acute{\mathbf{B}}_{1m}$ and the previous symbol $\acute{\mathbf{B}}_{1(m-1)}$ of the codeword $\acute{\mathbf{B}}_1^{L_2}$, as well as on the previous channel inputs $(X_1^{L_1})_{m-1}$ and outputs $(Y_1^{L_1})_{m-1}$. One may thus interpret $(X_1^{L_1})_m$ as being generated using a memory 1 distribution.

Due to the explicit dependence of $(X_1^{L_1})_m$ on $(X_1^{L_1})_{m-1}$, we cannot interpret the codeword $\acute{\mathbf{B}}_1^{L_2}$ as the outer codeword of a concatenated code. However, we can say that Han's coding technique is a special case of the concatenated coding technique where $\acute{\mathbf{B}}_1^m$ is a function of the outer codeword symbols $\mathbf{B}_1^m$, i.e., $\acute{\mathbf{B}}_1^m = f_m(\mathbf{B}_1^m)$ for some $f_m(\cdot)$, and similarly for User 2. This means that

$$I_\infty(B_1 \to Y_2 \,\|\, X_2) \;\geq\; I_\infty(\acute{B}_1 \to Y_2 \,\|\, X_2) \tag{4.27}$$

$$I_\infty(B_2 \to Y_1 \,\|\, X_1) \;\geq\; I_\infty(\acute{B}_2 \to Y_1 \,\|\, X_1), \tag{4.28}$$

where for simplicity we have dropped the subscript ℓ and the brackets in our notation and where the inequalities follow from the data processing inequality (see Appendix 4.A). The approachability of the rate point defined by (4.22) and (4.23) thus implies that the information rate point

defined by the right sides of (4.27) and (4.28) is also approachable. In Appendix 4.C, we show that these two information rates overbound the rates of Han [22], i.e., that

$$I_\infty(\acute{B}_1 \to Y_2 \parallel X_2) \geq I(\tilde{\acute{B}}_1; X_2 Y_2 \tilde{\acute{B}}_2 \tilde{X}_2 \tilde{Y}_2) \tag{4.29}$$

$$I_\infty(\acute{B}_2 \to Y_1 \parallel X_1) \geq I(\tilde{\acute{B}}_1; X_1 Y_1 \tilde{\acute{B}}_1 \tilde{X}_1 \tilde{Y}_1). \tag{4.30}$$

where $(X_1, X_2, Y_1, Y_2, \tilde{\acute{B}}_1, \tilde{\acute{B}}_2, \tilde{X}_1, \tilde{X}_2, \tilde{Y}_1, \tilde{Y}_2)$ has the same steady-state distribution as

$$\Big((X_{1\ell})_m, (X_{2\ell})_m, (Y_{1\ell})_m, (Y_{2\ell})_m, \acute{B}_{1(m-1)}, \acute{B}_{2(m-1)}, (X_{1\ell})_{m-1}, (X_{2\ell})_{m-1}, (Y_{1\ell})_{m-1}, (Y_{2\ell})_{m-1} \Big)$$

for any $\ell = 1, 2, \ldots, L_1$. Thus, the rate point defined by the right sides of (4.27) and (4.28) must also be approachable. The reason that the directed information rates *overbound* the rates of [22] is that Han uses a decoder which makes a decision on $\acute{B}_{1(m-1)}$ as soon as the channel outputs corresponding to the mth outer-code symbol are received rather than waiting until more (ideally *all*) channel output symbols are received. The information rate point specified in [22] is nonetheless useful because it is usually much simpler to calculate than the directed information rate point.

4.3.3 Bootstrapping with Slepian-Wolf Source Coding

For the steady-state coding techniques described above, there may be states for which the uncertainty about the other user's next symbol is smaller than the information rate of the coding technique. When this happens, the following method introduced by Schalkwijk [21] will improve the transmission rate.

The approach is to use the steady-state diagram simultaneously many times, just like a non-adaptive inner code. However, whenever a state is reached in which the uncertainty about the other user's next symbol is smaller than the coding technique's information rate, both users discontinue transmission and store their next channel input symbol in large buffers. The channel symbols in the buffers of the two users

are *correlated* and can be compressed to their entropy using Slepian-Wolf source coding [57]. Furthermore, the compressed symbol streams can be transmitted across the channel at the information rate of the original algorithm, resulting in a reduction in the number of channel uses and hence an increase in the transmission rate. By passing to the steady-state, the rate for User 1 becomes

$$R_1 = \frac{I_1}{(1 - p_{\mathcal{S}_1}) + \sum_{\sigma_1 \in \mathcal{S}_1} p(\sigma_1) H(X_2 | \Sigma_1 = \sigma_1)/R_1}, \tag{4.31}$$

where I_1 is the information rate of the basic algorithm for User 1, $\mathcal{S}_1$ is the set of states σ_1 with $H(X_2|\Sigma_1 = \sigma_1) < R_1$, and $p_{\mathcal{S}_1} = \sum_{\sigma_1 \in \mathcal{S}_1} p(\sigma_1)$. Rewriting (4.31), we obtain

$$R_1 = \frac{I_1 - \sum_{\sigma_1 \in \mathcal{S}_1} p(\sigma_1) H(X_2 | \Sigma_1 = \sigma_1)}{1 - p_{\mathcal{S}_1}}, \tag{4.32}$$

and a similar expression holds for R_2.

4.4 Case Studies

We use the coding techniques described above to develop lower bounds on the equal-rate capacity point of two two-way channels. The first channel is the noiseless binary multiplying channel (BMC), and the second channel is a noisy BMC.

4.4.1 The Binary Multiplying Channel

The binary multiplying channel (BMC) has the common output $Y = X_1 \cdot X_2$, where $\mathcal{X}_1 = \mathcal{X}_2 = \{0, 1\}$ (see Figure 4.6). This channel was introduced in [17, Section 13], and has been studied extensively by Schalkwijk [20, 21] and others (see the references in Meeuwissen [58]). We have not found improvements over the best approachable rates previously obtained, e.g., [21], but this section serves nonetheless to demonstrate how the bounds of this chapter may be applied to a common-output two-way channel.

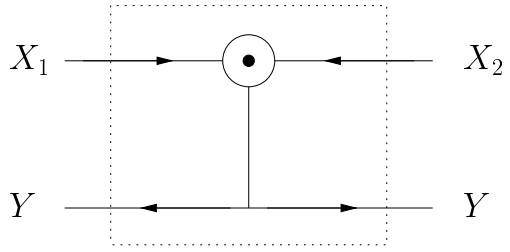


Figure 4.6: *The binary multiplying channel.*

Shannon's Inner Bounds

We first consider the case $L = 1$ in Corollary 4.1, which was originally considered in [17, Section 13]. Setting $q_{X_1}(0) = q_1$ and $q_{X_2}(0) = q_2$ the rates (4.9) and (4.10) become

$$R_1 = (1 - q_2)\, h(q_1) \tag{4.33}$$

$$R_2 = (1 - q_1)\, h(q_2). \tag{4.34}$$

The best equal-rate point is obtained by setting $q_1 = q_2 = 0.2965$, which yields $R_1 = R_2 = 0.61695$. We will consider only coding distributions which are the same for both users, and thus we consider only equal-rate points.

Setting $L = 2$ gives no improvement for the equal-rate point. However, the error exponent $E_{G1}^L(R) = \max_{q_{\mathbf{A}^L}} E_{G1}^L(R, (q_{\mathbf{A}^L})^2)$ obtained from Equation (4.61) increases, as shown in Figure 4.7. The error exponents in this figure are for the case where both users use the same coding technique so that no time-sharing is allowed. This is the reason that the error exponent is finite for rates less than 0.5, even though zero error probability is possible for such rates by using time-sharing.

Setting $L = 3$, one finds rate improvements over $L = 1$ coding. The best distribution we found was

$$\begin{array}{llllll} q & = 0.3277, & & & & \\ q_{00} & = 0.0517, & q_{10} & = 0.6480, & q_{11} & = 0.2965, \\ q_{0000} & = 0, & q_{0010} & = 0.3098, & q_{0011} & = 0, \\ q_{1000} & = 0.2975, & q_{1010} & = 1, & q_{1011} & = 1, \\ q_{1100} & = 0.2965, & q_{1101} & = 0.2965, & q_{1111} & = 0.2965, \end{array} \tag{4.35}$$

where we have set $q = q_{X_{11}}(0)$, $q_{ij} = q_{X_{12}|X_{11}Y_1}(0|i,j)$ and $q_{ijkl} = q_{X_{13}|X_{11}Y_1X_{12}Y_2}(0|i,j,k,l)$. If User 2 codes using the same distribution, the equal-rate point $R_1 = R_2 = 0.61964$ is approachable.

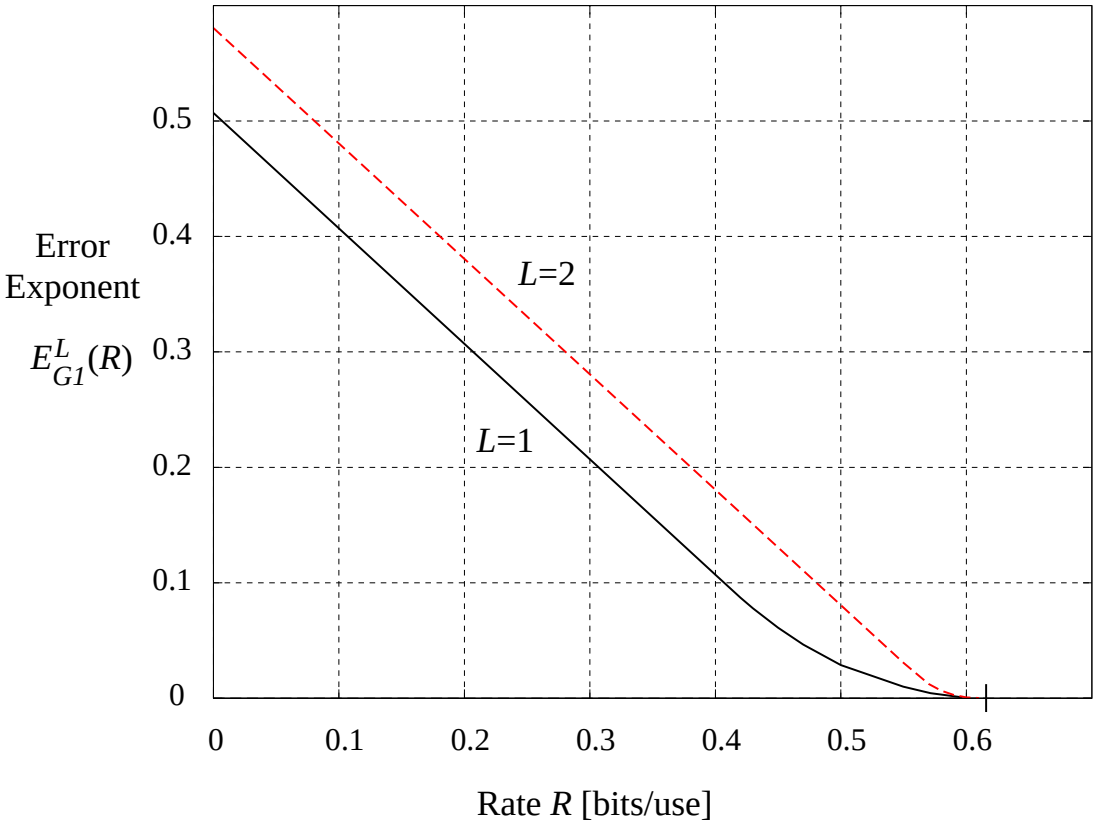


Figure 4.7: *The symmetric coding error exponent* $E_{G1}^{L}(R)$ *for the binary multiplying channel and* $L = 1, 2$.

Steady-State Inner Bounds

It is difficult to calculate steady-state information rates for the reasons discussed in Appendix 4.D. Because of these problems, we could calculate only lower and upper *bounds* on the information rates for the steady-state memory μ coding techniques.

Steady-state coding with memory $\mu = 1$ gives no improvement over Shannon's $L = 1$ equal-rate point of 0.61695 (even when a simple *upper* bound on the information rates is used). Nonexhaustive searches over the distributions for memory $\mu = 2$ also did not yield improvements.

Schalkwijk proposed a steady-state coding technique where the system state returns to the starting state after at most three uses of the channel [20]. The best equal-rate point he found for this strategy was $R_1 = R_2 = 0.61914$, which is not as good as our $R_1 = R_2 = 0.61964$ for $L = 3$ coding.

Non-adaptive Inner Codes

The rates of the coding technique of Han [22] are overbounded by the directed information rates in (4.22) and (4.23). Furthermore, because the BMC is a common-output channel, these respective directed information rates are the same as those in (4.17) and (4.18). Thus, we will not obtain better rates than for the memory $\mu = 1$ or $\mu = 2$ coding techniques.

Bootstrapping

Schalkwijk improved the rates of his steady-state coding technique by using the bootstrapping trick described in Section 4.3.3. The best equal-rate point he found in [21] was $R_1 = R_2 = 0.63056$. The best approachable equal-rate point we are aware of is $R_1 = R_2 = 0.63072$ [58]. The best equal-rate *outer* bound to date is $R_1 = R_2 = 0.64628$ [24].

One can also use the bootstrapping technique to improve the rates of Shannon's inner bound codes. For example, for $L = 3$ we calculated that the rate point $R_1 = R_2 = 0.62449$ is approachable by modifying the distribution of (4.35).

4.4.2 A Noisy Binary Multiplying Channel

We next consider the noisy BMC shown in Figure 4.8. The noise random variables Z_1 and Z_2 are independent and for our example we will use $\delta = 0.1$. We again use the coding techniques developed in this chapter.

Shannon's Inner Bounds

We first consider the case $L = 1$ in Lemma 4.2. Setting $q_{X_1}(0) = q_1$ and $q_{X_2}(0) = q_2$ the rates (4.9) and (4.10) become

$$R_1 = (1 - q_2)\left[h\left(q_1(1-\delta)\right) - q_1 h(\delta)\right] \tag{4.36}$$

$$R_2 = (1 - q_1)\left[h\left(q_2(1-\delta)\right) - q_2 h(\delta)\right]. \tag{4.37}$$

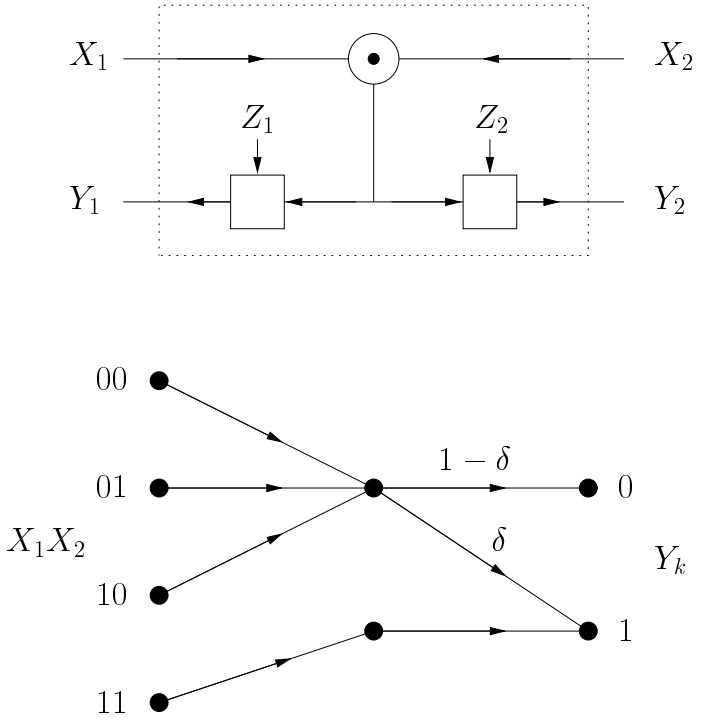


Figure 4.8: *A noisy binary multiplying channel. The noise random variables Z_1 and Z_2 are independent, and $k = 1, 2$.*

The best equal-rate point is obtained by setting $q_1 = q_2 = 0.2794$, which yields $R_1 = R_2 = 0.49184$.

For $L = 2$, there are $2^{1+2} = 8$ adaptive codewords. As for the noiseless BMC, we did not find a rate improvement over the $L = 1$ case. We can, however, demonstrate that the rates in (4.17) and (4.18) may be strictly larger than the rates in (4.11) and (4.12), and are thus only *outer bounds* on these approachable rates. Consider the adaptive codeword distribution

$$p(\mathbf{a}_k^2) = \begin{cases} 1/2 & \text{if } \mathbf{a}_k^2 = [\,0\,,\,10\,] \\ 1/2 & \text{if } \mathbf{a}_k^2 = [\,1\,,\,01\,] \\ 0 & \text{otherwise,} \end{cases}$$

for $k = 1, 2$. Straightforward calculations yield

$$\begin{aligned} & I_2(\mathbf{A}_1 \to Y_2 \| X_2) &=& \ 0.28889 \\ \text{and} \quad & I_2(X_1 \to Y_2 \| X_2) &=& \ 0.36901, \end{aligned}$$

which is substantially greater. Of course, this does not mean that the rate pair $(R_1, R_2) = (0.36901, 0.36901)$ is not approachable for this channel, as we have seen.

For $L = 3$, there are $2^{1+2+4} = 128$ adaptive codewords for each user. Due to the complexity of the distribution search for this case, we were not able to do an exhaustive search or to obtain a rate improvement over the $L = 1$ case.

Non-adaptive Inner Codes

We consider the coding technique of Han [22] with binary $\acute{B}_1$ and $\acute{B}_2$. User 1 has $2^3 = 8$ states $\sigma_1 = (\tilde{\acute{b}}_1, \tilde{x}_1, \tilde{y}_1)$ and User 2 has 8 states $\sigma_2 = (\tilde{\acute{b}}_2, \tilde{x}_2, \tilde{y}_2)$. The overall system has 64 states $\sigma = (\sigma_1, \sigma_2)$.

As one might expect from our results for the memory 1 coding for the noiseless BMC, we found no rate improvements over the rate point $R_1 = R_2 = 0.49184$ by using Han's coding technique. However, to demonstrate how directed information rates can be used to improve Han's rates, we consider the distribution where $p_{\acute{\mathbf{B}}_1}(0) = p_{\acute{\mathbf{B}}_2}(0) = 0.2794$ and $q_{ijk} = \Pr(X_1 \neq \acute{\mathbf{B}}_1 | \Sigma_1 = (i,j,k)) = \Pr(X_2 \neq \acute{\mathbf{B}}_2 | \Sigma_2 = (i,j,k))$. Setting $q_{0jk} = 0.1$ and $q_{1jk} = 0$ for all j, k, we obtain (see (4.29) and (4.30))

$$I(\tilde{\acute{B}}_1; X_2 Y_2 \tilde{\acute{B}}_2 \tilde{X}_2 \tilde{Y}_2) = 0.39453. \tag{4.38}$$

Because both users are coding with the same distributions the rate point $(0.39453, 0.39453)$ is approachable with Han's decoding technique. But in (4.67) we could have used the tighter bound

$$\begin{aligned} H(\acute{B}_{1(m-1)} | X_2^m Y_2^m \acute{B}_1^{m-2}) \quad &\leq \quad H(\acute{B}_{1(m-1)} | X_{2(m-2)} X_{2(m-1)} X_{2m} \\ &\qquad Y_{2(m-2)} Y_{2(m-1)} Y_{2m} \acute{B}_{1(m-2)}), \end{aligned} \tag{4.39}$$

which leads to the steady-state approachable rate

$$\begin{aligned} I(\acute{B}_{1(m-1)}; X_{2m} Y_{2m} X_{2(m-1)} Y_{2(m-1)} \\ \acute{B}_{1(m-2)} X_{2(m-2)} Y_{2(m-2)}) \quad &= \quad 0.41991. \end{aligned} \tag{4.40}$$

Thus, the rate point $(0.41991, 0.41991)$ is approachable with a decoding technique that waits until all blocks have been received before making a decision, rather than decoding each information block after the next block is received as Han chose to analyze.

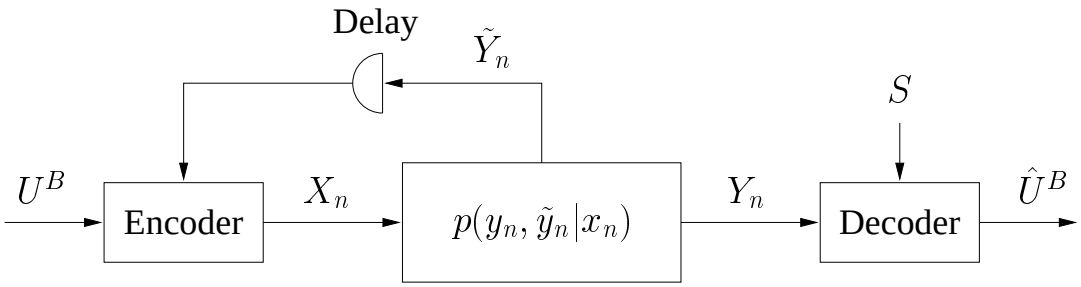


Figure 4.9: *A single-user channel model with noisy feedback, or side information at the transmitter, and side information S at the receiver.*

4.A Appendix: Outer Bounds on the Capacity Region

Preliminaries

We use the single-user model of Figure 4.9 to derive a lower bound on the error probabilities for the two-way channel and (in Chapter 5) for the multiple-access channel with feedback. In this model, the sender is transmitting the information bit sequence U^B with entropy B bits across a channel with probability distribution $p(y, \tilde{y}|x)$. The transmitter uses the channel N times and at time n may use all available feedback symbols to encode, i.e.,

$$X_n = f_n(U^B, \tilde{Y}^{n-1}). \tag{4.41}$$

The receiver's estimate $\hat{U}^B$ of U^B is a function of the output of the channel Y^N and the appropriate side information S.

We will show that the receiver's average error probability

$$\overline{P} = \frac{1}{B}\sum_{b=1}^{B} P_b, \tag{4.42}$$

where $P_b = \Pr(\hat{U}_b \neq U_b)$, can be bounded using the mutual information $I(U^B; \hat{U}^B)$. To do this, we will need several basic results.

First, we use the terminology of [8, Section 4.4] and say that a real-valued function $f(\cdot)$ of a vector is *convex-*$\cap$ over a convex region $\mathcal{R}$

of vector space if, for all $\underline{x}$ in $\mathcal{R}$, $\underline{y}$ in $\mathcal{R}$, and α, $0 < \alpha < 1$, the function satisfies

$$f(\alpha\underline{x} + (1-\alpha)\underline{y}) \geq \alpha\, f(\underline{x}) + (1-\alpha)\, f(\underline{y}). \tag{4.43}$$

If the inequality is reversed for all such $\underline{x}, \underline{y}$, and α, $f(\cdot)$ is convex-$\cup$.

Second, *Jensen's inequality* [52, p. 4.42] states that for a convex-$\cap$ function $f(\cdot)$ defined over a convex region $\mathcal{R}$ and any random vector $\underline{X}$ taking values in $\mathcal{R}$

$$f\left(\mathrm{E}[\underline{X}]\right) \geq \mathrm{E}\left[f(\underline{X})\right]. \tag{4.44}$$

For example, the binary entropy function $h(\cdot)$, where $h(p) = -p\log(p) - (1-p)\log(1-p)$, is convex-$\cap$ over the convex region $\mathcal{R} = [0,1]$, so that $h(E[X]) \geq E[h(X)]$ for any X taking on values in the interval $[0,1]$.

Next, *Fano's inequality* [8, p. 78] relates error probability to uncertainty. It states that if $\hat{U}$ is an estimate of U, where both $\hat{U}$ and U take on values in the same discrete alphabet $\mathcal{U}$, then the error probability $P_e = \Pr(\hat{U} \neq U)$ satisfies $h(P_e) + P_e \log(|\mathcal{U}| - 1) \geq H(U|\hat{U})$, where $|\mathcal{U}|$ is the number of values in $\mathcal{U}$. For example, if U is binary then Fano's inequality becomes

$$h(P_e) \geq H(U|\hat{U}). \tag{4.45}$$

Finally, by the *data processing inequality* [48, p. 32], $I(X;Z) \leq I(X;Y)$ for any Markov chain $X - Y - Z$, where the notation $X - Y - Z$ means $I(X;Z|Y) = 0$. This implies that in Figure 4.9 we may write

$$I(U^B;\hat{U}^B) \leq I(U^B;Y^N S), \tag{4.46}$$

because the estimate $\hat{U}^B$ was specified to be a function of Y^N and S.

Bounding the Bit Error Probability

As a first step to lower bounding the error probability $\overline{P}$, we derive a *generalized Fano's inequality* [59, p.4].

Lemma 4.3 (Generalized Fano's Inequality) *If the binary sequence $\hat{U}^B$ is an estimate of the binary sequence U^B, then the average bit error probability $\overline{P} = \frac{1}{B}\sum_{b=1}^{B} P_b$, where $P_b = \Pr(\hat{U}_b \neq U_b)$, satisfies*

$$h(\overline{P}) \geq \frac{1}{B} H(U^B|\hat{U}^B). \tag{4.47}$$

Proof: We write the chain of inequalities

$$\begin{aligned} h(\overline{P}) = h\left(\frac{1}{B}\sum_{b=1}^{B} P_b\right) &\geq \frac{1}{B}\sum_{b=1}^{B} h(P_b) \\ &\geq \frac{1}{B}\sum_{b=1}^{B} H(U_b|\hat{U}_b) \\ &\geq \frac{1}{B}\sum_{b=1}^{B} H(U_b|\hat{U}^B U^{b-1}) \\ &= \frac{1}{B} H(U^B|\hat{U}^B), \end{aligned}$$

where we have used (4.44), (4.45), that conditioning cannot increase uncertainty, and the chain rule for uncertainty. □

Next, we expand $H(U^B|\hat{U}^B) = H(U^B) - I(U^B;\hat{U}^B)$ and note that $H(U^B)$=B bits. Thus, using (4.47) and (4.46), $\overline{P}$ satisfies

$$\begin{aligned} h(\overline{P}) &\geq 1 - \frac{1}{B} I(U^B;\hat{U}^B) \\ &\geq 1 - \frac{1}{B} I(U^B;Y^N S). \end{aligned} \tag{4.48}$$

Application to the Two-Way Channel

Equation (4.48) can be applied to the two-way channel by considering the information transmission from User 1 to User 2 as corresponding to Figure 4.9 with $B = B_1$, $\overline{P} = \overline{P}_1$, $U^B = U_1^{B_1}$, $X^N = X_1^N$, $Y^N = Y_2^N$, and $\tilde{Y}^N = Y_1^N$. The side information at the receiver is $S = [U_2^{B_2}\ X_2^N]$. Thus,

$$h(\overline{P}_1) \geq 1 - \frac{1}{B_1} I(U_1^{B_1}; U_2^{B_2} X_2^N Y_2^N). \tag{4.49}$$

Manipulating the information in (4.49),

$$\begin{aligned} I(U_1^{B_1}; U_2^{B_2} X_2^N Y_2^N) &= I(U_1^{B_1} \mathbf{A}_1^N; U_2^{B_2} X_2^N Y_2^N) \\ &= I(\mathbf{A}_1^N; U_2^{B_2} X_2^N Y_2^N) \\ &= I(\mathbf{A}_1^N; X_2^N Y_2^N), \end{aligned} \tag{4.50}$$

where the last two equalities follow from d-separation in the functional dependence graph of Figure 4.3. We expand (4.50) as

$$\begin{aligned} I(\mathbf{A}_1^N; X_2^N Y_2^N) &= \sum_{n=1}^{N} I(\mathbf{A}_1^N; X_{2n} Y_{2n} | X_2^{n-1} Y_2^{n-1}) \\ &= \sum_{n=1}^{N} I(\mathbf{A}_1^N; Y_{2n} | X_2^n Y_2^{n-1}) \\ &= \sum_{n=1}^{N} I(\mathbf{A}_1^n; Y_{2n} | X_2^n Y_2^{n-1}) \\ &= I(\mathbf{A}_1^N \to Y_2^N \| X_2^N), \end{aligned} \tag{4.51}$$

where the second and third equalities follow from d-separation in the functional dependence graph of Figure 4.3. Inserting (4.51) into (4.49), we obtain

$$h(\overline{P}_1) \geq 1 - \frac{1}{B_1} I(\mathbf{A}_1^N \to Y_2^N \| X_2^N), \tag{4.52}$$

which we use to prove the following proposition.

Proposition 4.1 *Consider a discrete memoryless two-way channel that is used N times and for which a fixed choice of codes for User 1 and User 2 is made, where the information rate of User 1 is $R_1 = B_1/N$ bits per use. Then if there is some $\epsilon > 0$ such that*

$$R_1 \geq I_L(\mathbf{A}_1 \to Y_2 \| X_2) + \epsilon \tag{4.53}$$

for all positive integers L and all $f_{1\ell}(\cdot)$, $f_{2\ell}(\cdot)$ in (4.2), (4.3), $1 \leq \ell \leq L$, the average bit error probability $\overline{P}_1$ of User 2 satisfies

$$\overline{P}_1 \geq h^{-1}\left(\frac{\epsilon}{\epsilon + \log(|\mathcal{Y}_2|)}\right), \tag{4.54}$$

where $|\mathcal{Y}_2|$ is the size of Users 2's channel output alphabet and $h^{-1}(\cdot)$ is the "inverse" binary entropy function taking on values between 0 and 1/2.

Proof: The condition (4.53) for *all* L implies that $R_1 = I_N(\mathbf{A}_1 \to Y_2\|X_2) + \delta$ for some $\delta \geq \epsilon$. We now manipulate (4.52) as

$$\begin{aligned} h(\overline{P}_1) &\geq 1 - \frac{1}{B_1} I(\mathbf{A}_1^N \to Y_2^N \| X_2^N) \\ &= 1 - \frac{1}{R_1} I_N(\mathbf{A}_1 \to Y_2 \| X_2) \\ &= 1 - \frac{I_N(\mathbf{A}_1 \to Y_2\|X_2)}{I_N(\mathbf{A}_1 \to Y_2\|X_2) + \delta} \\ &= \frac{\delta}{\delta + I_N(\mathbf{A}_1 \to Y_2\|X_2)}. \end{aligned}$$

But since $f(x) = x/(x+c)$ is nondecreasing with x for $x > 0$ and $c \geq 0$, we have $\delta/(\delta + c) \geq \epsilon/(\epsilon + c)$ for $c = I_N(\mathbf{A}_1 \to Y_2\|X_2)$. Furthermore, from (3.7) and (3.12) we obtain $I_N(\mathbf{A}_1 \to Y_2\|X_2) \leq H_N(Y_2)$ and, from a basic bound on uncertainty, $H_N(Y_2) \leq \log(|\mathcal{Y}_2|)$ (see, e.g., [48, p. 27]). Combining these results yields (4.54). □

Proposition 4.1 may also be applied to R_2 and $\overline{P}_2$ by interchanging the subscripts "1" and "2" at the appropriate places. Further, Proposition 4.1 allows the number of channel uses N to be any positive integer, which lets us prove the following lemma which was stated in Section 4.2.1. For this lemma we need to add a *closure* operation because the definition of $\mathcal{C}_{\text{TWC}}$ includes those rate-pairs which one gets arbitrarily close to. For example, the information rate $I_N(\mathbf{A}_1 \to Y_2\|X_2)$ may approach, but never reach, a limit as the block length N increases.

Lemma 4.1 (An Outer Bound to the Capacity Region) *The capacity region $\mathcal{C}_{\text{TWC}}$ of the discrete memoryless two-way channel is contained within the closure $\mathcal{C}_{\text{TWC}}^{\text{OUT}}$ of the set of rate-pairs (R_1, R_2) such that*

$$R_1 = I_L(\mathbf{A}_1 \to Y_2\|X_2) \tag{4.55}$$

$$R_2 = I_L(\mathbf{A}_2 \to Y_1\|X_1), \tag{4.56}$$

where L is a positive integer, and where $\mathbf{A}_1^L$ and $\mathbf{A}_2^L$ are independent adaptive codewords.

Proof: First, the functional dependence graph specifies that $\mathbf{A}_1^L$ and $\mathbf{A}_2^L$ are independent. Next, any rate-pair lying outside the closure $\mathcal{C}_{\text{TWC}}^{\text{OUT}}$

must satisfy $R_1 \geq I_L(\mathbf{A}_1 \to Y_2 \| X_2) + \epsilon$ or $R_2 \geq I_L(\mathbf{A}_2 \to Y_1 \| X_1) + \epsilon$ or both for some $\epsilon > 0$ and all $L \geq 1$. For such a rate-pair, Proposition 4.1 guarantees that no codes can achieve an error probability below that specified by equation (4.54), which is positive for $\epsilon > 0$. Thus, this rate-pair is not approachable and not in $\mathcal{C}_{\text{TWC}}$. $\square$

4.B Appendix: Random Coding and Maximum-Likelihood Decoding

The codes for Users 1 and 2 are generated as specified in Section 4.2.2. We first consider the data transmission from User 1 to User 2, and assume that User 1 wishes to transmit one of $\lceil 2^{NR_1} \rceil$ messages. For each of these messages we generate a length $N = L \cdot M$ codeword by concatenating M randomly generated adaptive codewords of length L. All adaptive codewords are chosen with the same probability distribution $q_{\mathbf{A}_1^L}$. We will use the notation $\underline{\mathbf{a}}_1^L$ to denote a length N codeword for User 1; the corresponding random variable is $\underline{\mathbf{A}}_1^L$. The coding for User 2 is done in an analogous fashion with $\lceil 2^{NR_2} \rceil$ messages and the codewords $\underline{\mathbf{a}}_2^L$.

Bounding the Block Error Probability

The receivers use maximum-likelihood decoders where the maximum-likelihood (ML) rule for User 2's decoder is: choose the message i if i is (any one of) the index(es) that maximizes

$$p_{\underline{X}_2^L \underline{Y}_2^L | \underline{\mathbf{A}}_1^L} \left(\underline{x}_2^L, \underline{y}_2^L \mid \underline{\mathbf{a}}_1^L(i) \right), \tag{4.57}$$

where $\underline{\mathbf{a}}_1^L(i)$ is the codeword for the ith message of User 1. We would now like to bound the average *bit* error probability $\overline{P}_1$ using the average *block* error probability $\overline{P}_{B1}$ of User 2. We do this by noting that a bit error means that a block error occurred and a block error means that there must be at least one bit error, but no more that B_1 bit errors among the decoded information bits. Thus, $\overline{P}_1 \leq \overline{P}_{B1} \leq B_1 \overline{P}_1$.

We use the same approach as that taken by Gallager [8, Chapter 5] to bound the average error probability of the ML decoder over the ensemble of codes generated by M uses of the distribution $q(\mathbf{a}_1^L)$. The resulting bound is an obvious modification of Theorem 5.6.1 in [8], and it states that, when the message u_1 is sent, the average *block* error

probability $\overline{P}_{B1}(u_1)$ over the ensemble of codewords satisfies

$$\overline{P}_{B1}(u_1) \leq \left(\lceil 2^{NR_1} \rceil - 1\right)^{\rho} \sum_{\underline{x}_2^L, \underline{y}_2^L} \left[\sum_{\underline{\mathbf{a}}_1^L} q(\underline{\mathbf{a}}_1^L)\, p(\underline{x}_2^L, \underline{y}_2^L | \underline{\mathbf{a}}_1^L)^{1/(1+\rho)} \right]^{1+\rho}, \tag{4.58}$$

for any choice of ρ, $0 \leq \rho \leq 1$. Using the memoryless nature of the channel and $\lceil 2^{NR_1} \rceil - 1 < 2^{NR_1}$, we arrive at a bound corresponding to Gallager's Theorem 5.6.2 [8]:

$$\overline{P}_{B1}(u_1) \leq 2^{-N\left[E_{o1}^L(\rho, q_{\mathbf{A}_1^L \mathbf{A}_2^L}) - \rho R_1\right]}, \tag{4.59}$$

where

$$E_{o1}^L(\rho, q_{\mathbf{A}_1^L \mathbf{A}_2^L}) = -\frac{1}{L} \log_2 \sum_{x_2^L, y_2^L} \left[\sum_{\mathbf{a}_1^L} q(\mathbf{a}_1^L)\, p(x_2^L, y_2^L \mid \mathbf{a}_1^L)^{1/(1+\rho)} \right]^{1+\rho}, \tag{4.60}$$

and $q_{\mathbf{A}_1^L \mathbf{A}_2^L} = q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L}$. The random coding exponent is thus

$$E_{G1}^L(R_1, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L}) = \max_{0 \leq \rho \leq 1} \left[E_{o1}^L(\rho, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L}) - \rho R_1 \right]. \tag{4.61}$$

If $E_{G1}^L(R_1, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L}) > 0$, the bound (4.59) guarantees that $\overline{P}_{B1}(u_1)$ can be made to approach zero by increasing the block length N. But since there must exist a code with $P_{B1}(u_1) \leq \overline{P}_{B1}(u_1)$ there is a code for which $P_{B1}(u_1)$ is small if $\overline{P}_{B1}(u_1)$ is small. One can now show that there is also a code having the same error exponent $E_{G1}^L(R_1, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L})$ for all u_1 (see Corollary 2 in [8, Section 5.6]). Furthermore, one can prove the following lemma using the same steps as in the proof of Theorem 5.6.3 in [8, Section 5.6] together with Equation (4.51).

Lemma 4.4 *For the discrete memoryless two-way channel and adaptive codeword probability distributions $q_{\mathbf{A}_1^L}$ and $q_{\mathbf{A}_2^L}$, the random coding exponents $E_{G1}^L(R_1, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L})$ and $E_{G2}^L(R_2, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L})$ are convex-$\cup$, decreasing, positive functions of the respective rates R_1 and R_2 for*

$$\begin{aligned} 0 \leq \quad & R_1 \quad < I_L(\mathbf{A}_1 \to Y_2 \| X_2) \\ 0 \leq \quad & R_2 \quad < I_L(\mathbf{A}_2 \to Y_1 \| X_1). \end{aligned} \tag{4.62}$$

By fixing L and letting M become large, this lemma immediately implies Lemma 4.2 in Section 4.2.2. The convex hull operation in Lemma 4.2 can be included in the approachable rate region because one may use time-sharing. By setting $M = 1$ and letting L become large, Lemma 4.4 implies the approachability of the steady-state rates in equations (4.20) and (4.21).

4.C Appendix: Information Rates for Non-Adaptive Inner Codes

We bound the information on the left side of in (4.29) by bounding the terms in the sum (3.8) defining this directed information

$$\begin{aligned} I(\acute{B}_1^m; Y_{2m} \mid X_2^m Y_2^{m-1}) &= \sum_{\ell=1}^{m} I(\acute{B}_{1\ell}; Y_{2m} | X_2^m Y_2^{m-1} \acute{B}_1^{\ell-1}) \\ &\geq I(\acute{B}_{1(m-1)}; Y_{2m} \mid X_2^m Y_2^{m-1} \acute{B}_1^{m-2}) \\ &\quad + I(\acute{B}_{1m}; Y_{2m} \mid X_2^m Y_2^{m-1} \acute{B}_1^{m-1}). \quad (4.63) \end{aligned}$$

Next, we expand the first term of (4.63) as

$$H(\acute{B}_{1(m-1)} | X_2^{m-1} Y_2^{m-1} \acute{B}_1^{m-2}) - H(\acute{B}_{1(m-1)} | X_2^m Y_2^m \acute{B}_1^{m-2}) \quad (4.64)$$

and the second term as

$$H(\acute{B}_{1m}) - H(\acute{B}_{1m} | X_2^m Y_2^m \acute{B}_1^{m-1}), \quad (4.65)$$

where we have used the functional dependence graph (see Figure 4.5) to simplify both (4.64) and (4.65). But in the steady state,

$$H(\acute{B}_{1(m-1)} | X_2^{m-1} Y_2^{m-1} \acute{B}_1^{m-2}) = H(\acute{B}_{1m} | X_2^m Y_2^m \acute{B}_1^{m-1})$$

and $H(\acute{B}_{1m}) = H(\acute{B}_{1(m-1)})$ so that the sum in (4.63) becomes

$$H(\acute{B}_{1(m-1)}) - H(\acute{B}_{1(m-1)} | X_2^m Y_2^m \acute{B}_1^{m-2}). \quad (4.66)$$

To obtain Han's information rates [22], we bound the second term of (4.66) as

$$\begin{aligned} &H(\acute{B}_{1(m-1)} | X_2^m Y_2^m \acute{B}_1^{m-2}) \\ &\quad = H(\acute{B}_{1(m-1)} | X_2^m Y_2^m \acute{B}_1^{m-2} \acute{B}_2^m) \\ &\quad \leq H(\acute{B}_{1(m-1)} | X_{2(m-1)} X_{2m} Y_{2(m-1)} Y_{2m} \acute{B}_{2(m-1)}), \quad (4.67) \end{aligned}$$

where we have used the functional dependence graph (Figure 4.5) for the first step. Thus, (4.66) can be overbounded by

$$I(\acute{B}_{1(m-1)}; X_{2m} Y_{2m} \acute{B}_{2(m-1)} X_{2(m-1)} Y_{2(m-1)}). \quad (4.68)$$

This information rate corresponds to the rate derived by Han in [22]. The left side of (4.30) can be bounded in the same manner.

4.D Appendix: Bounding Information Rates

The calculation of a steady-state directed information rate becomes difficult when the state diagrams of the random variables have many states. For example, consider memory 1 random coding for the Binary Multiplying Channel (Section 4.4.1). User 1's conditional distributions are a function of his past channel input $\tilde{X}_1$ and the past channel output $\tilde{Y}$. Thus, User 1 has three states which we denote by $(\tilde{X}_1, \tilde{Y}) \in \{(0,0),(1,0),(1,1)\}$. User 2 codes in the same manner. Because the common output is determined by the inputs, the system state diagram has only four states, which we denote by $(\tilde{X}_1, \tilde{X}_2) \in \{(0,0),(0,1),(1,0),(1,1)\}$.

Now consider $I_\infty(X_1 \to Y \| X_2)$. Because $H(Y|X_1X_2) = 0$, we need only to calculate

$$H_\infty(Y \parallel X_2) = \lim_{\ell \to \infty} H(Y_\ell \mid X_2^\ell Y^{\ell-1}).$$

A difficulty arises because knowing only $X_2^\ell Y^{\ell-1}$ does not tell us which state User 1 is in at time ℓ. As a result, the state diagram for User 1 given $X_2^\ell Y^{\ell-1}$ grows exponentially with ℓ. Instead of trying to find an exact solution for $H_\infty(Y \parallel X_2)$, we used the upper and lower bounds

$$H(Y_\ell|X_2^\ell Y^{\ell-1}\Sigma_{\ell-D}) \le H_\infty(Y\|X_2) \le H(Y_\ell|X_2^{\ell-D..\ell}Y^{\ell-D..\ell-1}), \tag{4.69}$$

both of which can be shown to converge to $H_\infty(Y\|X_2)$ as D increases (see [48, Section 4.4]). We used $D = 2$ for our calculations.

Chapter 5

Directed Information for the Multiple-Access Channel with Feedback

This chapter applies the definitions and results of Chapters 3 and 4 to the multiple-access channel with feedback (MAC-FB). Just as for the two-way channel, we show that the capacity region of the MAC-FB can be expressed in terms of causally-conditioned directed information rates. We are primarily interested in finding *inner* bounds to the capacity region. Our main result is a generalization of Cover and Leung's rate region [29].

This chapter is organized as follows. Section 5.1 introduces the model of the MAC-FB. Section 5.2 develops a general solution for the capacity region of the MAC-FB in terms of causally-conditioned directed information rates. The resulting expression is similar to the capacity region of the common-output two-way channel. Section 5.3 reviews the Cover-Leung region and presents directed information generalizations of this rate region. Finally, Section 5.4 applies the bounds of this chapter to several interesting MAC-FBs.

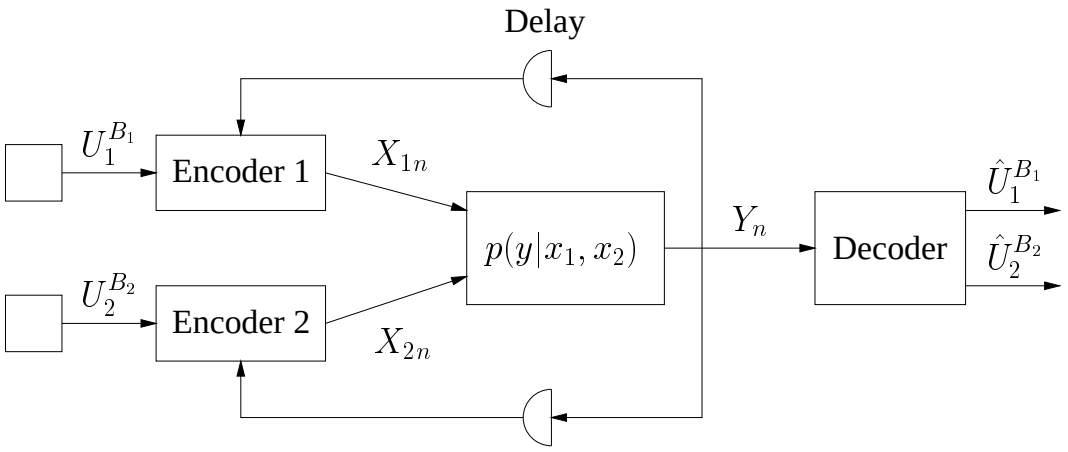


Figure 5.1: *The multiple-access channel with feedback.*

5.1 Model and Adaptive Codes

The discrete memoryless multiple-access channel with feedback is defined by the discrete input alphabets $\mathcal{X}_1$ and $\mathcal{X}_2$, the discrete output alphabet $\mathcal{Y}$ and the conditional probability distribution $p(y|x_1, x_2)$ where

$$p(y_n|x_1^n, x_2^n, y^{n-1}) = p_{Y|X_1X_2}(y_n|x_{1n}, x_{2n}). \tag{5.1}$$

The MAC-FB is depicted in Figure 5.1 and its functional dependence graph in Figure 5.2. This functional dependence graph is in fact identical to the functional dependence graph of the common-output two-way channel. Users 1 and 2 transmit the respective information bit sequences $U_1^{B_1}$ and $U_2^{B_2}$. The sequences $U_1^{B_1}$ and $U_2^{B_2}$ are independent and have entropy B_1 bits and B_2 bits, respectively. The channel is used N times so that the transmission rate is $(R_1, R_2) = (B_1/N, B_2/N)$ bits per use. The symbols input to the channel at time n are

$$X_{1n} = f_{1n}(U_1^{B_1}, Y^{n-1}) \tag{5.2}$$
$$X_{2n} = f_{2n}(U_2^{B_2}, Y^{n-1}). \tag{5.3}$$

After the transmissions over the channel are completed, the receiver outputs the decisions $\hat{U}_1^{B_1}$ and $\hat{U}_2^{B_2}$ for $U_1^{B_1}$ and $U_2^{B_2}$.

Just as for the two-way channel, equations (5.2) and (5.3) specify the form of the code one may use. The code consists of *adaptive codewords* whose symbols at time n depend on the feedback Y^{n-1}. However, one difference from Figure 4.2 is that now both Users' adaptive codewords have J branches coming out of each vertex, where J is the cardinality of

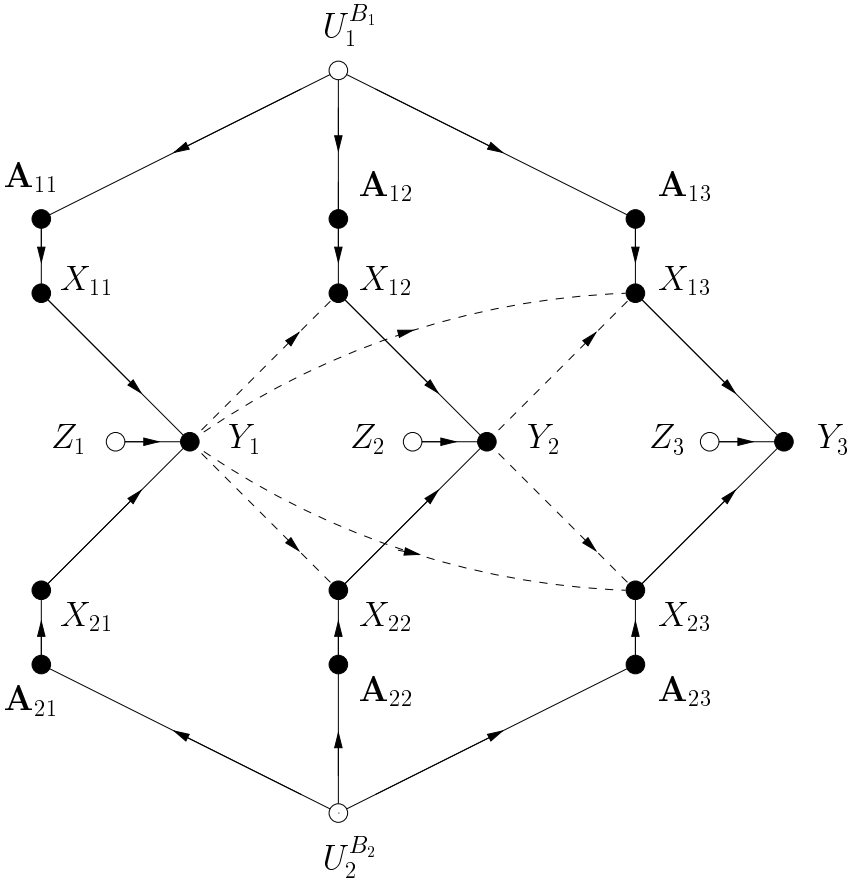


Figure 5.2: *The functional dependence graph of the multiple-access channel with feedback including the adaptive codewords. The feedback links are drawn using dashed lines.*

$\mathcal{Y}$. Furthermore, as for the common-output two-way channel, one can generate the symbols of the branches coming out of the same vertex *independently* using the same distribution $p_{X_{1n}|X_1^{n-1}Y_1^{n-1}}(\cdot|x_1^{n-1}y_1^{n-1})$, cf. the comments below Corollary 4.1.

5.2 A General Solution for the Capacity Region

This section gives a general solution for the capacity region of the MAC-FB. The development parallels that of Section 4.2, and we therefore only outline the proofs for the outer and inner bounds. The details are given in Appendix 5.A.

5.2.1 An Outer Bound

The two average error probabilities that the receiver wishes to make small are

$$\overline{P}_k = \frac{1}{B_k} \sum_{b=1}^{B_k} P_{kb}, \tag{5.4}$$

where $P_{kb} = \Pr(\hat{U}_{kb} \neq U_{kb})$ and $k = 1, 2$. We will also consider the overall average error probability

$$\begin{aligned} \overline{P}_{12} &= \frac{B_1}{B_1 + B_2}\overline{P}_1 + \frac{B_2}{B_1 + B_2}\overline{P}_2. \\ &= \frac{1}{B_1 + B_2}\left(\sum_{b=1}^{B_1} P_{1b} + \sum_{b=1}^{B_2} P_{2b}\right). \end{aligned} \tag{5.5}$$

The *capacity region* $\mathcal{C}_{\rm FB}$ of the MAC-FB is the set of *approachable* rate-pairs (R_1, R_2), i.e., those rate-pairs which one can approach with arbitrarily small positive $\overline{P}_1$ and $\overline{P}_2$. In Appendix 5.A, we show that $\overline{P}_1$ satisfies (see Equation (5.22))

$$h(\overline{P}_1) \geq 1 - \frac{1}{B_1} I(X_1^N \to Y^N \| X_2^N). \tag{5.6}$$

By symmetry, we may exchange the subscripts "1" and "2" in (5.6) to bound $\overline{P}_2$. Furthermore, $\overline{P}_{12}$ satisfies (see Equation (5.23))

$$h(\overline{P}_{12}) \geq 1 - \frac{1}{B_1 + B_2} I(X_1^N X_2^N \to Y^N). \tag{5.7}$$

As one might expect, these bounds lead to the following lemma which corresponds to Lemma 4.1 of Chapter 4 and is proved in Appendix 5.A.

Lemma 5.1 (An Outer Bound to the Capacity Region) *The capacity region $\mathcal{C}_{\rm FB}$ of the discrete memoryless multiple-access channel with feedback is contained within the closure $\mathcal{C}_{\rm FB}^{\rm OUT}$ of the set of rate-pairs (R_1, R_2) such that*

$$\begin{array}{ccccc} 0 & \leq & R_1 & \leq & I_L(X_1 \to Y \| X_2) \\ 0 & \leq & R_2 & \leq & I_L(X_2 \to Y \| X_1) \\ 0 & \leq & R_1 + R_2 & \leq & I_L(X_1 X_2 \to Y), \end{array} \tag{5.8}$$

where L is a positive integer and where $p(x_{1\ell}, x_{2\ell}|x_1^{\ell-1}, x_2^{\ell-1}, y^{\ell-1})$ factors as

$$q(x_{1\ell}|x_1^{\ell-1}, y^{\ell-1}) \cdot q(x_{2\ell}|x_2^{\ell-1}, y^{\ell-1}) \tag{5.9}$$

for $\ell = 1, 2, \ldots, L$.

5.2.2 Inner Bounds

The inner bounds are obtained in the same manner as Shannon's Lth Inner Bound Region of Chapter 4. As Shannon did for the two-way channel, we consider the adaptive codewords as *symbols* of a *derived* channel consisting of L consecutive uses of the MAC-FB. We then form long codewords by concatenating M randomly chosen "symbols" $\mathbf{a}_1^L$. The random codes formed in this manner are used to prove the following lemma in Appendix 5.A.

Lemma 5.2 (Directed Information Lth Inner Bound Region) *The convex hull of the set of rate-pairs (R_1, R_2) such that*

$$\begin{array}{ccccc} 0 & \leq & R_1 & \leq & I_L(X_1 \to Y \| X_2) \\ 0 & \leq & R_2 & \leq & I_L(X_2 \to Y \| X_1) \\ 0 & \leq & R_1 + R_2 & \leq & I_L(X_1 X_2 \to Y) \end{array} \tag{5.10}$$

is contained within the capacity region of the discrete memoryless multiple access channel with feedback, where L is a positive integer and $p(x_{1\ell}, x_{2\ell}|x_1^{\ell-1}, x_2^{\ell-1}, y^{\ell-1})$ factors as in (5.9) for $\ell = 1, 2, \ldots, L$.

We call the region $\mathcal{R}_L$ corresponding to the integer L in Lemma 5.2 the *Directed Information Lth Inner Bound Region* to the capacity region of the MAC-FB. We can now state the following theorem whose proof is virtually identical to the proof of Theorem 4.1 and is omitted. The set $\lim_{L\to\infty} \mathcal{R}_L$ has the same meaning here as for Theorem 4.1.

Theorem 5.1 (The Capacity Region) *The capacity region $\mathcal{C}_{\text{TWC}}$ of the discrete memoryless multiple-access channel with feedback is $\mathcal{C}_{\text{FB}}^{\text{OUT}}$ of Lemma 5.1. Furthermore, if $\mathcal{R}_L$ is the Directed Information Lth Inner Bound Region, then*

$$\mathcal{C}_{\text{FB}} = \lim_{L\to\infty} \mathcal{R}_L. \tag{5.11}$$

5.3 Generalizing the Cover-Leung Region

One could now consider similar coding techniques as for the two-way channel in Section 4.3. The finite-state coding will function in an analogous manner as for the two-way channel and the same concatenated coding techniques can be used. However, to obtain better results, we will concentrate on Cover and Leung's superposition coding technique. We first review and then generalize the Cover-Leung technique using directed information. The proofs of the lemmas are given in Appendix 5.B

5.3.1 Cover and Leung's Rate Region

The coding technique specified by Cover and Leung in [29] involves *superposing* a random variable U onto a random variable V which simply means that the result of the superposition is a function of U and V. The coding is done in $B+1$ blocks of length N and the functional dependence graph of the coding technique is shown in Figure 5.3. In this graph the information sequences are divided into B independent blocks of equal length, which we denote by $\mathbf{U}_{kb}$ for $k = 1, 2$ and $b = 1, 2, ...B$.

For the first block, both users agree on a common v_1^N and superpose their respective information sequences $\mathbf{U}_{11}$ and $\mathbf{U}_{21}$ onto v_1^N to obtain the codewords X_{11}^N and X_{21}^N. The transmission rate of both users is assumed to be slow enough so that both users can understand what the other user had sent. However, the receiver may not be able to understand what the users had sent and thus both users cooperate in the second block to resolve the receiver's ambiguity. This cooperation is taken into account by the random variables V_{12}^N and V_{22}^N, which coincide if neither user erroneously decodes the other user's information. However, the transmitters do not solely resolve the receiver's ambiguity in the second block, but also superpose new information $\mathbf{U}_{12}$ and $\mathbf{U}_{22}$ onto the respective V_{12}^N and V_{22}^N. This process is repeated up to the Bth block. In the $(B+1)$st block the remaining ambiguity of the receiver is resolved without superposing new information. By making N and B large enough, Cover and Leung [29] showed that all points inside the following rate region are achievable.

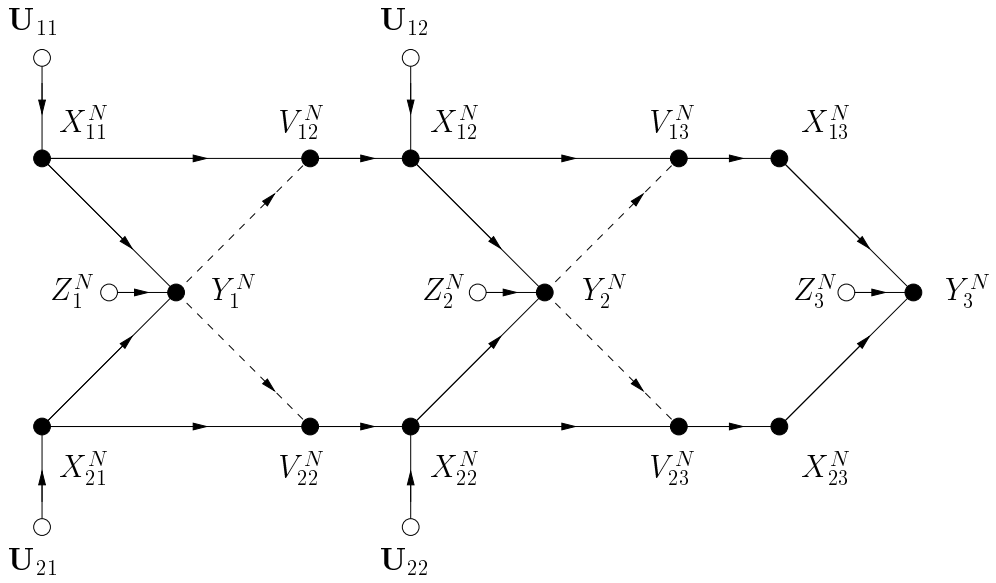


Figure 5.3: *The functional dependence graph for the coding technique of Cover and Leung. Here $B = 2$ so that 3 blocks of channel inputs are transmitted. For proper operation of the technique V_{1b}^N must equal V_{2b}^N for all $b = 2, 3, \ldots, B+1$.*

Lemma 5.3 (The Cover-Leung Region) *The set $\mathcal{R}^{CL}$ of rate-pairs (R_1, R_2) such that*

$$\begin{array}{ccccc} 0 & \leq & R_1 & \leq & I(X_1; Y|X_2V) \\ 0 & \leq & R_2 & \leq & I(X_2; Y|X_1V) \\ 0 & \leq & R_1 + R_2 & \leq & I(X_1X_2; Y) \end{array} \tag{5.12}$$

is contained within the capacity region of the discrete memoryless multiple access channel with feedback, where V is a discrete random variable such that

$$p(x_1, x_2, y|v) = q(x_1|v) \cdot q(x_2|v) \cdot p(y|x_1, x_2). \tag{5.13}$$

Note that, as in Lemma 5.2, no convex hull operation is necessary here because V may serve as the time sharing random variable. The cardinality of V may be limited without loss of generality to $\max(|\mathcal{X}_1| \cdot |\mathcal{X}_2| + 1, |\mathcal{Y}| + 2)$, where $|\mathcal{X}_1|$, $|\mathcal{X}_2|$ and $|\mathcal{Y}|$ are the cardinalities of the random variables X_1, X_2 and Y, respectively [30].

The Cover-Leung Region is actually the capacity region for some interesting channels. For example, Willems [30] showed that $\mathcal{C}_{\text{FB}} = \mathcal{R}^{CL}$

for the class of MAC-FBs where one of the channel inputs is determined by the other channel input and the channel output. We will consider these channels in Chapter 6 and develop feedback strategies for them. Other channels for which $\mathcal{C}_{\mathrm{FB}} = \mathcal{R}^{CL}$ were found by Carleial [32], Csiszár and Körner [60, page 299] and Willems and van der Meulen [61] (all three of these references consider the same class of channels as in [30] but with feedback to one user only) and by Hekstra and Willems [24, Section VII].

5.3.2 Generalizations

We are interested in finding approachable rates which lie outside the Cover-Leung region. One approach is to increase L in (5.10), but the increase in the size of $\mathcal{R}_L$ is usually slow as L increases. Instead, we consider the natural generalization of the Cover-Leung Region using directed information, and prove the following lemma in Appendix 5.B. In this lemma $I_L(X_1 \to Y \| X_2 | V)$ denotes $\frac{1}{L} \cdot I(X_1^L \to Y^L \| X_2^L | V)$.

Lemma 5.4 (Generalization of the Cover-Leung Region) *The set $\mathcal{R}_L^{CL}$ of rate-pairs (R_1, R_2) such that*

$$\begin{array}{ccccc} 0 & \leq & R_1 & \leq & I_L(X_1 \to Y \| X_2 | V) \\ 0 & \leq & R_2 & \leq & I_L(X_2 \to Y \| X_1 | V) \\ 0 & \leq & R_1 + R_2 & \leq & I_L(X_1 X_2 \to Y) \end{array} \tag{5.14}$$

is contained within the capacity region of the discrete memoryless multiple access channel with feedback, where L is a positive integer and where V is a discrete random variable such that $p(x_{1\ell}, x_{2\ell}, y_\ell | v, x_1^{\ell-1}, x_2^{\ell-1}, y^{\ell-1})$ factors as

$$q(x_{1\ell} | v, x_1^{\ell-1}, y^{\ell-1}) \cdot q(x_{2\ell} | v, x_2^{\ell-1}, y^{\ell-1}) \cdot p(y_\ell | x_{1\ell} x_{2\ell}) \tag{5.15}$$

for $\ell = 1, 2, \ldots, L$.

A special case of the above lemma is obtained by replacing V by V^L and choosing the coding distributions in the manner specified in the following corollary to Lemma 5.4.

Corollary 5.1 *The set of rate-pairs* (R_1, R_2) *such that*

$$\begin{array}{ccccc} 0 & \leq & R_1 & \leq & I_L(X_1 \to Y \| X_2 V) \\ 0 & \leq & R_2 & \leq & I_L(X_2 \to Y \| X_1 V) \\ 0 & \leq & R_1 + R_2 & \leq & I_L(X_1 X_2 \to Y) \end{array} \tag{5.16}$$

is contained within $\mathcal{R}_L^{CL}$*, where* L *is a positive integer and where* V^L *is a sequence of discrete random variables such that* $p(x_{1\ell}, x_{2\ell}, y_\ell \,|\, v^L, x_1^{\ell-1}, x_2^{\ell-1}, y^{\ell-1})$ *factors as*

$$q(x_{1\ell}|v^\ell, x_1^{\ell-1}, y^{\ell-1}) \cdot q(x_{2\ell}|v^\ell, x_2^{\ell-1}, y^{\ell-1}) \cdot p(y_\ell|x_{1\ell}x_{2\ell}) \tag{5.17}$$

for $\ell = 1, 2, \ldots, L$.

Note that $\mathcal{R}_1^{CL} = \mathcal{R}^{CL}$. Also, it is clear that $\mathcal{R}_L \subseteq \mathcal{R}_L^{CL}$, and thus from Theorem 5.1 that $\lim_{L\to\infty} \mathcal{R}_L^{CL} = \mathcal{C}_{\rm FB}$. In the next section, we give several examples of MAC-FBs for which $\mathcal{R}_1^{CL} \subset \mathcal{R}_L^{CL}$ when $L > 1$ by using Corollary 5.1.

5.4 Case Studies

This section applies the results above to three MAC-FBs. We first consider the extension of Lemma 5.4 to the white Gaussian MAC-FB and show that one can approach rate points outside $\mathcal{R}^{CL}$. This is not surprising, as Ozarow [31] has shown that the capacity region of the white Gaussian MAC-FB is larger than $\mathcal{R}^{CL}$. We then consider two examples of *discrete* MAC-FBs for which $\mathcal{R}^{CL}$ is not the capacity region, which are the first such examples of which we are aware. Both of these examples are noisy Binary Adder Channels (BACs)

5.4.1 The Additive White Gaussian Noise Channel

We use the white Gaussian MAC-FB model of [31] with signal powers $P_1 = P_2 = 10$ and noise variance $\sigma^2 = 1$. The equal-rate points of $\mathcal{R}^{CL}$ use the probability distributions defined by

$$\begin{aligned} X_{11} &= a_1 \cdot U_{11} + b_1 \cdot \tilde{V}_1 \\ X_{21} &= a_1 \cdot U_{21} + b_1 \cdot \tilde{V}_1, \end{aligned} \tag{5.18}$$

where U_{11}, U_{21} and $\tilde{V}_1$ are independent Gaussian random variables with unit variance. Optimizing over the multiplying factors we find that $a_1 = 2.1525$ and $b_1 = 2.3166$, and that the rate point $R_1 = R_2 = 0.8643$ nats per use is approachable, which is the equal-rate point on the boundary of the Cover-Leung region.

For $L = 2$, we use the probability distribution where X_{11} and X_{21} have the same form as in (5.18), and

$$\begin{aligned} X_{12} &= a_2 \cdot U_{12} + b_2 \cdot \tilde{V}_2 + c_2 \cdot X_{11} + d_2 \cdot Y_1 \\ X_{22} &= a_2 \cdot U_{22} + b_2 \cdot \tilde{V}_2 - c_2 \cdot X_{21} - d_2 \cdot Y_1, \end{aligned} \tag{5.19}$$

where U_{11}, U_{12}, U_{21}, U_{22}, $\tilde{V}_1$ and $\tilde{V}_2$ are independent Gaussian random variables with unit variance. Optimizing over the multiplying factors, we obtain $a_1 = 2.1510$, $b_1 = 2.3180$, $a_2 = 1.5232$, $b_2 = 0$, $c_2 = 1.7335$, $d_2 = -0.83909$, and we find that $R_1 = R_2 = 0.8710$ nats per use is approachable, which is outside the Cover-Leung region. Ozarow found that the equal-rate point on the boundary of the capacity region is $R_1 = R_2 = 0.8905$ nats per use.

5.4.2 A BAC with Additive $(0, 1)$ Noise

We next consider the noisy BAC whose output is $Y = X_1 + X_2 + Z$ where X_1, X_2 and Z are binary $(0, 1)$ random variables and $\Pr(Z = 0) = \Pr(Z = 1) = 1/2$ (see Figure 5.4). In Appendix 5.C, we show that the equal-rate point on the boundary of $\mathcal{R}^{CL}$ has the probability distribution $p_{VX_1X_2Y}$ with $\Pr(V = 0) = \Pr(V = 1) = 1/2$ and $\Pr(X_1 \neq V) = \Pr(X_2 \neq V) = 1 - 1/\sqrt{2}$. The resulting rate point is $R_1 = R_2 = h(1/\sqrt{2})/2 = 0.43621$ bits per use, where $h(\cdot)$ is the binary entropy function.

For $L \to \infty$, we use a steady-state coding technique with memory 1 (described in Appendix 5.D) and find that the equal-rate point $R_1 = R_2 = 0.43879$ bits per use is approachable, which is outside the Cover-Leung region. The distribution used to obtain this rate point is given in Appendix 5.D. The equal-rate capacity point of this noisy BAC is not known.

This noisy two-user channel can be viewed as the three-user noiseless BAC where the third user is sending data at rate 1. Thus, we see that

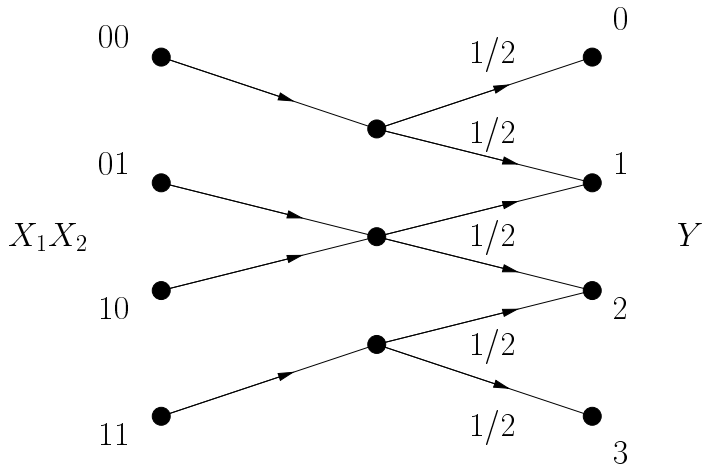


Figure 5.4: *A binary adder channel with additive* $(0, 1)$ *noise.*

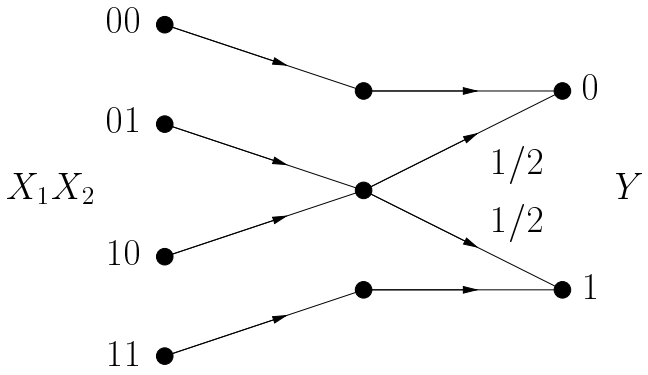


Figure 5.5: *A binary adder channel with noise if the users' inputs to the channel are different.*

the capacity region of the three-user noiseless BAC is strictly larger than the straightforward generalization of the Cover-Leung rate region to three-user channels.

5.4.3 A BAC with Different-Input Noise

Finally, we consider the noisy BAC shown in Figure 5.5 in which the "noise" enters only when the two users send different binary digits. In Appendix 5.C, we show that the equal-rate point on the boundary of $\mathcal{R}^{CL}$ has the probability distribution $p_{VX_1X_2Y}$ with $\Pr(V = 0) = \Pr(V = 1) = 1/2$ and $\Pr(X_1 \neq V) = \Pr(X_2 \neq V) = 0.3044$. The resulting rate point is $R_1 = R_2 = 0.28826$ bits per use.

For $L \to \infty$, we again use a steady-state coding technique with memory 1 and find that the equal-rate point $R_1 = R_2 = 0.28911$ bits per use is approachable, which lies outside the Cover-Leung region. The distribution used to obtain this rate point is given in Appendix 5.D. The equal-rate capacity point of this noisy BAC is not known.

5.A Appendix: Bounds on the Capacity Region

An Outer Bound

We derive bounds on the receiver's average bit error probabilities $\overline{P}_1$, $\overline{P}_2$ and $\overline{P}_{12}$ defined by (5.4) and (5.5). We first consider $\overline{P}_1$. The receiver's estimate $\hat{U}_1^{B_1}$ of $U_1^{B_1}$ is a function of the output of the channel Y^N. We thus have the situation of Figure 4.9 with $B = B_1$, $\overline{P} = \overline{P}_1$, $U^B = U_1^{B_1}$, $X^N = X_1^N$, and $\tilde{Y}^N = Y^N$. We thus use (4.48) to bound

$$h(\overline{P}_1) \geq 1 - \frac{1}{B_1} I(U_1^{B_1}; Y^N). \tag{5.20}$$

But $I(U_1^{B_1}; Y^N) \leq I(U_1^{B_1}; X_2^N Y^N)$, and

$$\begin{aligned} I(U_1^{B_1}; X_2^N Y^N) &= H(X_2^N Y^N) - H(X_2^N Y^N | U_1^{B_1}) \\ &= \sum_{n=1}^{N} H(Y_n X_{2n} | Y^{n-1} X_2^{n-1}) - H(Y_n X_{2n} | Y^{n-1} X_2^{n-1} U_1^{B_1}) \\ &= \sum_{n=1}^{N} H(Y_n | Y^{n-1} X_2^n) - H(Y_n | Y^{n-1} X_2^n U_1^{B_1}) \\ &= \sum_{n=1}^{N} H(Y_n | Y^{n-1} X_2^n) - H(Y_n | X_{1n} X_{2n}) \\ &= I(X_1^N \to Y^N \| X_2^N), \end{aligned} \tag{5.21}$$

where the third and fourth equalities follow from the functional dependence graph of the MAC-FB. Inserting (5.21) into (5.20), we find that $\overline{P}_1$ satisfies

$$h(\overline{P}_1) \geq 1 - \frac{1}{B_1} I(X_1^N \to Y^N \| X_2^N). \tag{5.22}$$

By symmetry, we may exchange the subscripts "1" and "2" in (5.22) to bound $\overline{P}_2$.

Next, consider $\overline{P}_{12}$. The receiver's estimate $(\hat{U}_1^{B_1}, \hat{U}_2^{B_2})$ of $(U_1^{B_1}, U_2^{B_2})$ is also function of the output of the channel Y^N. We let U^B be the concatenation of $U_1^{B_1}$ and $U_2^{B_2}$, so that we have the situation of Figure 4.9

with $B = B_1 + B_2$, $\overline{P} = \overline{P}_{12}$, $X^N = (X_1, X_2)^N$, and $\tilde{Y}^N = Y^N$. We thus use (4.48) to bound

$$h(\overline{P}_{12}) \geq 1 - \frac{1}{B_1 + B_2} I(U_1^{B_1} U_2^{B_2}; Y^N). \tag{5.23}$$

Expanding the information in (5.23),

$$\begin{aligned} I(U_1^{B_1} U_2^{B_2}; Y^N) &= \sum_{n=1}^{N} H(Y_n|Y^{n-1}) - H(Y_n|Y^{n-1} U_1^{B_1} U_2^{B_2}) \\ &= \sum_{n=1}^{N} H(Y_n|Y^{n-1}) - H(Y_n|X_{1n} X_{2n}) \\ &= I(X_1^N X_2^N \to Y^N) \end{aligned} \tag{5.24}$$

and inserting this into (5.23), we find that $\overline{P}_{12}$ satisfies

$$h(\overline{P}_{12}) \geq 1 - \frac{1}{B_1 + B_2} I(X_1^N X_2^N \to Y^N). \tag{5.25}$$

The bounds (5.22) and (5.25) may be combined to prove the following proposition, cf. Proposition 4.1.

Proposition 5.1 *Consider a discrete memoryless multiple-access channel with feedback that is used N times and for which a fixed choice of codes for User 1 and User 2 is made, where the information rates of Users 1 and 2 are $R_1 = B_1/N$ and $R_2 = B_2/N$ bits per use, respectively. Then if there is some $\epsilon > 0$ such that any one of the following inequalities is satisfied*

$$\begin{aligned} R_1 &\geq I_L(X_1 \to Y \| X_2) + \epsilon \\ R_2 &\geq I_L(X_2 \to Y \| X_1) + \epsilon \\ R_1 + R_2 &\geq I_L(X_1 X_2 \to Y) + \epsilon \end{aligned} \tag{5.26}$$

for all positive integers L and all $f_{1\ell}(\cdot)$, $f_{2\ell}(\cdot)$ in (5.2), (5.3), $1 \leq \ell \leq L$, the sum of the bit error probabilities $\overline{P}_1 + \overline{P}_2$ of the receiver satisfies

$$\overline{P}_1 + \overline{P}_2 \geq h^{-1}\left(\frac{\epsilon}{\epsilon + \log(|\mathcal{Y}|)}\right), \tag{5.27}$$

where $|\mathcal{Y}|$ is the size of the receiver's channel output alphabet and $h^{-1}(\cdot)$ is the "inverse" binary entropy function taking on values between 0 and 1/2.

Proof: Since $\overline{P}_{12}$ is the average of $\overline{P}_1$ and $\overline{P}_2$, either $\overline{P}_1$ or $\overline{P}_2$ cannot be made to approach zero if $\overline{P}_{12}$ cannot be made to approach zero. This leads to the sum rate bound in (5.26). The bound (5.27) follows by combining the three bounds on $\overline{P}_1$, $\overline{P}_2$ and $\overline{P}_1 + \overline{P}_2$ obtained like (4.54) from the three bounds in (5.26). The remaining steps of the proof are the same as the proof of Proposition 4.1 and are omitted. □

Like Proposition 4.1, Proposition 5.1 allows the number of channel uses N to be any positive integer. This lets us prove the following lemma that corresponds to Lemma 4.1 in Appendix 4.A.

Lemma 5.5 (An Outer Bound to the Capacity Region) *The capacity region $\mathcal{C}_{\rm FB}$ of the discrete memoryless multiple-access channel with feedback is contained within the closure $\mathcal{C}_{\rm FB}^{\rm OUT}$ of the set of rate-pairs (R_1, R_2) such that*

$$\begin{array}{ccccc} 0 & \leq & R_1 & \leq & I_L(X_1 \to Y \| X_2) \\ 0 & \leq & R_2 & \leq & I_L(X_2 \to Y \| X_1) \\ 0 & \leq & R_1 + R_2 & \leq & I_L(X_1 X_2 \to Y). \end{array} \tag{5.28}$$

where L is a positive integer and where $p(x_{1\ell}, x_{2\ell} | x_1^{\ell-1}, x_2^{\ell-1}, y^{\ell-1})$ factors as

$$q(x_{1\ell} | x_1^{\ell-1}, y^{\ell-1}) \cdot q(x_{2\ell} | x_2^{\ell-1}, y^{\ell-1}) \tag{5.29}$$

for $\ell = 1, 2, \ldots, L$.

Proof: The factorization (5.29) follows from the functional dependence graph of the MAC-FB. The rest of the proof uses the same steps as the proof for Lemma 4.1 and is omitted. □

Like $\mathcal{C}_{\rm TWC}^{\rm OUT}$, the region $\mathcal{C}_{\rm FB}^{\rm OUT}$ is automatically convex because allowing L to be any positive integer implicitly allows time-sharing.

Inner Bounds

The coding is done in the same manner as in Appendix 4.B and we will use the same notation. The ML rule is: choose the message pair (i, j)

if (i, j) is (any one of) the index pair(s) that maximizes

$$p_{\underline{Y}_2^L|\underline{\mathbf{A}}_1^L\underline{\mathbf{A}}_2^L}\left(\underline{y}^L \mid \underline{\mathbf{a}}_1^L(i)\underline{\mathbf{a}}_2^L(j)\right), \tag{5.30}$$

where $\underline{\mathbf{a}}_1^L(i)$ and $\underline{\mathbf{a}}_2^L(j)$ are the adaptive codewords for the respective messages i and j of Users 1 and 2 (cf. Equation (4.57)). Note that the channel output $\underline{y}^L$ specifies which path $\underline{x}_k^L(\underline{y}^{L-1})$ through the adaptive codeword $\underline{\mathbf{a}}_k^L$ was taken for $k = 1, 2$. Thus, we could also have stated the ML rule (5.30) using $\underline{x}_1^L$ and $\underline{x}_2^L$ rather than $\underline{\mathbf{a}}_1^L$ and $\underline{\mathbf{a}}_2^L$.

The *block* error probability over the ensemble of codes when the message pair (u_1, u_2) is sent is denoted by $\overline{P}_B(u_1, u_2)$. We follow the same approach as Gallager in [62, Section IIA], who in turn used results of Slepian and Wolf [57]. We denote by (u_1, u_2) the message pair that was sent. Let E_1 be the event that the decoded message pair $(\hat{u}_1, \hat{u}_2) = (i, u_2)$ where $i \neq u_1$, E_2 be the event that $(\hat{u}_1, \hat{u}_2) = (u_1, j)$ where $j \neq u_2$, and E_{12} be the event that $(\hat{u}_1, \hat{u}_2) = (i, j)$ where both $i \neq u_1$ and $j \neq u_2$. As these events are mutually exclusive, we have

$$\overline{P}_B(u_1, u_2) = \overline{P}(\mathrm{E}_1) + \overline{P}(\mathrm{E}_2) + \overline{P}(\mathrm{E}_{12}), \tag{5.31}$$

where $\overline{P}(\mathrm{E}_1)$, $\overline{P}(\mathrm{E}_2)$, and $\overline{P}(\mathrm{E}_{12})$ are the probabilities of the events E_1, E_2 and E_{12} averaged over the ensemble of codes.

We first bound $\overline{P}(\mathrm{E}_{12})$ by considering the message pair (u_1, u_2) to be the message of a single user to a single input channel with input alphabet $\mathcal{X}_1 \times \mathcal{X}_2$. There are $(\lceil 2^{NR_1} \rceil - 1)(\lceil 2^{NR_2} \rceil - 1)$ message pairs that cause the event E_{12} so that, using Theorem 5.6.1 in [8],

$$\begin{aligned} \overline{P}(\mathrm{E}_{12}) \quad \leq \quad & \left((\lceil 2^{NR_1} \rceil - 1)(\lceil 2^{NR_2} \rceil - 1)\right)^{\rho} \cdot \\ & \sum_{\underline{y}^L} \left[\sum_{\underline{\mathbf{a}}_1^L, \underline{\mathbf{a}}_2^L} q(\underline{\mathbf{a}}_1^L) q(\underline{\mathbf{a}}_2^L)\, p(\underline{y}^L | \underline{\mathbf{a}}_1^L, \underline{\mathbf{a}}_2^L)^{1/(1+\rho)} \right]^{1+\rho} \end{aligned} \tag{5.32}$$

for any choice of ρ, $0 \leq \rho \leq 1$.

Next, as in [62], we may condition E_1 on the event that the codeword of User 2 was $\underline{\mathbf{a}}_2^L(u_2)$, in which case we have a single-user problem with the channel $p(\underline{y}^L | \underline{\mathbf{a}}_1^L, \underline{\mathbf{a}}_2^L(u_2))$. But the event E_1 implies that

$$p_{\underline{Y}_2^L|\underline{\mathbf{A}}_1^L\underline{\mathbf{A}}_2^L}\left(\underline{y}^L \mid \underline{\mathbf{a}}_1^L(i)\,\underline{\mathbf{a}}_2^L(u_2)\right) \geq p_{\underline{Y}_2^L|\underline{\mathbf{A}}_1^L\underline{\mathbf{A}}_2^L}\left(\underline{y}^L \mid \underline{\mathbf{a}}_1^L(u_1)\,\underline{\mathbf{a}}_2^L(u_2)\right)$$

for some $i \neq u_1$, so that we may overbound $\overline{P}(\mathrm{E}_1|\text{User 2 sent } \underline{\mathbf{a}}_2^L(u_2))$ as

$$\overline{P}(\mathrm{E}_1|\text{User 2 sent } \underline{\mathbf{a}}_2^L(u_2)) \leq \left(\lceil 2^{NR_1} \rceil - 1\right)^{\rho} \cdot \sum_{\underline{y}^L} \left[\sum_{\underline{\mathbf{a}}_1^L} q(\underline{\mathbf{a}}_1^L)\, p(\underline{y}^L|\underline{\mathbf{a}}_1^L, \underline{\mathbf{a}}_2^L(u_2))^{1/(1+\rho)} \right]^{1+\rho}. \tag{5.33}$$

Averaging this over all $\underline{\mathbf{a}}_2^L(u_2)$, we obtain

$$\overline{P}(\mathrm{E}_1) \leq \left(\lceil 2^{NR_1} \rceil - 1\right)^{\rho} \cdot \sum_{\underline{\mathbf{a}}_2^L, \underline{y}^L} q(\underline{\mathbf{a}}_2^L) \left[\sum_{\underline{\mathbf{a}}_1^L} q(\underline{\mathbf{a}}_1^L)\, p(\underline{y}^L|\underline{\mathbf{a}}_1^L, \underline{\mathbf{a}}_2^L)^{1/(1+\rho)} \right]^{1+\rho}. \tag{5.34}$$

The bound for $\overline{P}(\mathrm{E}_2)$ is obtained by interchanging the subscripts "1" and "2".

The bounds (5.32) and (5.34) may be simplified by using the memoryless nature of the channel and $\lceil 2^{NR} \rceil - 1 < 2^{NR}$. Following the same approach as in Appendix 4.B, we find that

$$\begin{aligned} \overline{P}(\mathrm{E}_1) &\leq 2^{-N\left[E_{o1}^L(\rho, q_{\mathbf{A}_1^L \mathbf{A}_2^L}) - \rho R_1\right]} \\ \overline{P}(\mathrm{E}_2) &\leq 2^{-N\left[E_{o2}^L(\rho, q_{\mathbf{A}_1^L \mathbf{A}_2^L}) - \rho R_2\right]} \\ \overline{P}(\mathrm{E}_{12}) &\leq 2^{-N\left[E_{o12}^L(\rho, q_{\mathbf{A}_1^L \mathbf{A}_2^L}) - \rho (R_1+R_2)\right]} \end{aligned} \tag{5.35}$$

where $E_{o1}^L(\rho, q_{\mathbf{A}_1^L \mathbf{A}_2^L})$, $E_{o2}^L(\rho, q_{\mathbf{A}_1^L \mathbf{A}_2^L})$ and $E_{o12}^L(\rho, q_{\mathbf{A}_1^L \mathbf{A}_2^L})$ are given by

$$\begin{aligned} &-\frac{1}{L}\log_2 \sum_{\mathbf{a}_2^L, y^L} q(\mathbf{a}_2^L) \left[\sum_{\mathbf{a}_1^L} q(\mathbf{a}_1^L)\, p(y^L \mid \mathbf{a}_1^L, \mathbf{a}_2^L)^{1/(1+\rho)} \right]^{1+\rho}, \\ &-\frac{1}{L}\log_2 \sum_{\mathbf{a}_1^L, y^L} q(\mathbf{a}_1^L) \left[\sum_{\mathbf{a}_2^L} q(\mathbf{a}_2^L)\, p(y^L \mid \mathbf{a}_1^L, \mathbf{a}_2^L)^{1/(1+\rho)} \right]^{1+\rho}, \\ &-\frac{1}{L}\log_2 \sum_{y^L} \left[\sum_{\mathbf{a}_1^L, \mathbf{a}_2^L} q(\mathbf{a}_1^L)\, q(\mathbf{a}_2^L)\, p(y^L \mid \mathbf{a}_1^L, \mathbf{a}_2^L)^{1/(1+\rho)} \right]^{1+\rho}, \end{aligned} \tag{5.36}$$

respectively, and $q_{\mathbf{A}_1^L\mathbf{A}_2^L} = q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L}$. We further define the error exponents $E_{G1}^L(R_1, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L})$, $E_{G2}^L(R_2, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L})$ and $E_{G12}^L(R_1+R_2, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L})$ by the respective

$$\begin{aligned}
&\max_{0\le\rho\le 1}\left[E_{o1}^L(\rho, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L}) - \rho R_1\right], \\
&\max_{0\le\rho\le 1}\left[E_{o2}^L(\rho, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L}) - \rho R_2\right], \\
&\max_{0\le\rho\le 1}\left[E_{o12}^L(\rho, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L}) - \rho(R_1+R_2)\right].
\end{aligned} \tag{5.37}$$

If all three error exponents are positive, then $\overline{P}_B(u_1, u_2)$ can be made to approach zero by increasing the block length N. One can further show there is a code having the error exponents (5.37) for all (u_1, u_2) pairs. On can also derive the following identities from the functional dependence graph of the MAC-FB

$$\begin{aligned}
I(\mathbf{A}_1^L; \mathbf{A}_2^L Y^L) &= I(X_1^L \to Y^L \| X_2^L) \\
I(\mathbf{A}_2^L; \mathbf{A}_1^L Y^L) &= I(X_2^L \to Y^L \| X_1^L) \\
I(\mathbf{A}_1^L \mathbf{A}_2^L; Y^L) &= I(X_1^L X_2^L \to Y^L).
\end{aligned} \tag{5.38}$$

Combining these results, we can prove the following lemma using the same steps as in the proof of Theorem 5.6.3 in [8].

Lemma 5.6 *For the discrete memoryless multiple-access channel with feedback and adaptive codeword probability distributions $q_{\mathbf{A}_1^L}$ and $q_{\mathbf{A}_2^L}$, the random coding exponents $E_{G1}^L(R_1, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L})$, $E_{G2}^L(R_2, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L})$ and $E_{G12}^L(R_1+R_2, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L})$ are convex-$\cup$, decreasing, positive functions of the respective rates R_1, R_2 and R_1+R_2 for*

$$\begin{aligned}
0 &\le R_1 < I_L(X_1 \to Y \| X_2) \\
0 &\le R_2 < I_L(X_2 \to Y \| X_1) \\
0 &\le R_1 + R_2 < I_L(X_1 X_2 \to Y).
\end{aligned} \tag{5.39}$$

By fixing L and letting M become large, this lemma immediately implies Lemma 5.2 in Section 5.2.2.

5.B Appendix: Proof for the Generalization of the Cover-Leung Region

To prove the approachability of all points in $\mathcal{R}_L^{CL}$ for $L > 1$ one could use the same steps as in the proof for $L = 1$ given in [29]. The only difference is that now the *symbols* are *adaptive codewords*, just like Shannon's coding for the two-way channel. However, rather than using typical sequence decoders as in [29], we will use maximum-likelihood decoders.

Code Generation

The transmitters generate their adaptive codewords by agreeing on a common V code having $\lceil 2^{NR_V} \rceil$ codewords of length M. Each of these codewords is generated by choosing M symbols independently using the distribution p_V. Next, for *each* of the $\lceil 2^{NR_V} \rceil$ codewords of the V code Users 1 and 2 choose a code having $\lceil 2^{NR_1} \rceil$ and $\lceil 2^{NR_2} \rceil$ codewords of length M, respectively. The M symbols of these two codes consist of length L adaptive codewords that are generated using the distributions $p(x_{1\ell}|v, x_1^{\ell-1}, y^{\ell-1})$ and $p(x_{2\ell}|v, x_2^{\ell-1}, y^{\ell-1})$ independently for each of the M symbols, $\ell = 1, \ldots, L$. Both users and the receiver know the codes.

The First Block

We consider data transmission in $B+1$ blocks of length $N = L \cdot M$, each block consisting of M adaptive codewords of length L. The transmitters divide their information sequences into B independent blocks of equal length, which we denote by $\mathbf{U}_{kb}$ for $k = 1, 2$ and $b = 1, 2, ..., B$.

Consider data transmission in the first block. Both senders and the receiver agree on an initial choice v_1^M of the codeword from the V code, where the subscript "1" denotes the first block of transmission. The senders next superpose their respective information sequences $\mathbf{U}_{11}$ and $\mathbf{U}_{21}$ onto $\underline{v}_1^M$ to obtain the M length L adaptive codewords $\underline{\mathbf{A}}_{11}^L$ and $\underline{\mathbf{A}}_{21}^L$. By modifying Lemma 4.4 in Chapter 4 to include the knowledge

of V_1^M and using the probability distribution in (5.15), we find that if

$$\begin{array}{rcccl} 0 & \leq & R_1 & < & I_L(X_1 \to Y \| X_2 | V) \\ 0 & \leq & R_2 & < & I_L(X_2 \to Y \| X_1 | V) \end{array} \tag{5.40}$$

then the error exponents $E_{G1}^L(R_1)$ and $E_{G2}^L(R_2)$, modified to include the knowledge of V_1^M, are positive. (For convenience, we have dropped the probability distributions in our notation $E_{G1}^L(R_1, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L})$ and $E_{G2}^L(R_2, q_{\mathbf{A}_1^L} \cdot q_{\mathbf{A}_2^L})$.) Thus, both senders can decode $\mathbf{U}_{11}$ and $\mathbf{U}_{21}$ with vanishing error probability as M or N increases. The receiver, however, may not be able to decode correctly these information sequences. Instead, he forms a list $\mathcal{L}_1$ of the L_s most likely information sequence pairs using v_1^M and $\underline{Y}_1^L$. We must now bound the probability $P_{\mathcal{L}}$ that the senders' information sequence pair $(\mathbf{u}_{11}, \mathbf{u}_{21})$ is not in $\mathcal{L}_1$, an event called a *list error*, cf. [8, Problem 5.20].

The List-Decoding Error Probability

We bound $P_{\mathcal{L}}$ by dividing the $(\mathbf{i}, \mathbf{j})$ pairs competing with $(\mathbf{u}_{11}, \mathbf{u}_{21})$ into three classes. The first class contains those pairs for which $\mathbf{i} \neq \mathbf{u}_{11}$ but $\mathbf{j} = \mathbf{u}_{21}$, the second class those pairs for which $\mathbf{i} = \mathbf{u}_{11}$ but $\mathbf{j} \neq \mathbf{u}_{21}$, and the third class those pairs for which $\mathbf{i} \neq \mathbf{u}_{11}$ and $\mathbf{j} \neq \mathbf{u}_{21}$. Let $\mathrm{E}_1(1)$ be the event that at least one element in $\mathcal{L}$ is from the first class of pairs, $\mathrm{E}_2(1)$ be the event that at least one element in $\mathcal{L}$ is from the second class of pairs, and $\mathrm{E}_3(L_s)$ be the event that *all* elements in $\mathcal{L}$ are from the third class of pairs. Then, using the union bound [63, page 264], we have

$$\begin{aligned} \overline{P}_{\mathcal{L}} &= \overline{P}\left(\mathrm{E}_1(1) \cup \mathrm{E}_2(1) \cup \mathrm{E}_3(L_s)\right) \\ &\leq \overline{P}\left(\mathrm{E}_1(1)\right) + \overline{P}\left(\mathrm{E}_2(1)\right) + \overline{P}\left(\mathrm{E}_3(L_s)\right), \end{aligned} \tag{5.41}$$

where $\overline{P}_{\mathcal{L}}$ and $\overline{P}(\mathrm{E})$ denote the respective averages of $P_{\mathcal{L}}$ and $\Pr(\mathrm{E})$ over the ensemble of codes, where E is an event.

The event $\mathrm{E}_1(1)$ implies that there was a pair $(\mathbf{i}, \mathbf{u}_{21})$ with $\mathbf{i} \neq \mathbf{u}_{11}$ which has a larger probability of having been sent than $(\mathbf{u}_{11}, \mathbf{u}_{21})$. But the average probability of this latter event has the same error exponent $E_{G1}^L(R_1)$ as User 2's average probability of incorrectly decoding User 1's information sequence given v_1^M. Thus, $\overline{P}(\mathrm{E}_1(1))$ has an error exponent at least as large as $E_{G1}^L(R_1)$ which is positive for the rates in (5.40).

Similarly, the error exponent of $\overline{P}(\mathrm{E}_2(1))$ is at least as large as $E_{G2}^L(R_2)$ which is also positive for those R_2 in (5.40).

We bound $\overline{P}(\mathrm{E}_3(L_s))$ by noting that there are $\binom{(M_1-1)(M_2-1)}{L_s}$ ways to build a list of length L_s using the $(M_1-1)(M_2-1)$ pairs in the third class. This number can be upper bounded by using $\binom{n}{k} \leq 2^{nh(k/n)}$, where $h(\cdot)$ is the binary entropy function [48, page 284]. Equivalently, $\log_2\left(\binom{n}{k}\right) \leq -k\log_2(k/n) - (n-k)\log_2(1-k/n)$. We use the same steps as in [8, Problem 5.20] to allow $0 \leq \rho \leq L_s$, where $\rho = L_s\rho_0$, and to bound

$$\begin{aligned}\log_2\left(\overline{P}(\mathrm{E}_3(L_s))\right) &\leq \log_2\left(\binom{(M_1-1)(M_2-1)}{L_s}^{\rho_0} 2^{-NE_{o12}(\rho)}\right) \\ &\leq -\rho\Big[\log_2(L_s/2^{N(R_1+R_2)}) + (2^{N(R_1+R_2)}/L_s - 1) \\ &\quad \cdot \log_2(1 - L_s/2^{N(R_1+R_2)})\Big] - NE_{o3}^L(\rho) \qquad (5.42)\end{aligned}$$

where we have used $M_1 - 1 < 2^{NR_1}$ and $M_2 - 1 < 2^{NR_2}$ and have set

$$E_{o3}^L(\rho, p_V \cdot q_{\mathbf{A}_1^L|V} \cdot q_{\mathbf{A}_1^L|V}) = -\frac{1}{L}\log_2 \sum_v p(v) \sum_{y^L} \left[\sum_{\mathbf{a}_1^L, \mathbf{a}_2^L} q(\mathbf{a}_1^L|v)\, q(\mathbf{a}_1^L|v)\, p(y^L \mid \mathbf{a}_1^L, \mathbf{a}_2^L, v)^{1/(1+\rho)}\right]^{1+\rho}. \qquad (5.43)$$

We now set $L_s = 2^{N\lambda}$ and note that the second term in square brackets in (5.42) approaches $-1/\ln(2)$ as N increases, and is lowerbounded by this number if $\lambda \leq R_1 + R_2$. Thus,

$$\overline{P}(\mathrm{E}_3(L_s)) \leq 2^{-NE_{G3}^L(R_1+R_2, \lambda, p_V \cdot q_{\mathbf{A}_1^L|V} \cdot q_{\mathbf{A}_1^L|V})} \qquad (5.44)$$

for all ρ, $0 \leq \rho \leq L_s$, where $E_{G3}^L(R_1 + R_2, \lambda, p_V \cdot q_{\mathbf{A}_1^L|V} \cdot q_{\mathbf{A}_1^L|V})$ is

$$\max_{0 \leq \rho \leq L_s} \left[E_{o3}^L(\rho, p_V \cdot q_{\mathbf{A}_1^L|V} \cdot q_{\mathbf{A}_1^L|V}) - \rho\left[R_1 + R_2 - \lambda + 1/(N\ln(2))\right]\right]. \qquad (5.45)$$

Furthermore, by using the same steps as in the proof of Theorem 5.6.3 in [8], we find that (5.45) is positive for

$$0 \leq R_1 + R_2 - \lambda + 1/(N\ln(2)) < I_L(X_1X_2 \to Y|V). \qquad (5.46)$$

Thus, since $1/\ln(2) < 2$, if we choose

$$\lambda = R_1 + R_2 - I_L(X_1X_2 \to Y|V) + 2/N \tag{5.47}$$

and $\lambda \leq R_1 + R_2$, then the error exponent (5.45) is positive. This, along with the conditions in (5.40), guarantees that the list-decoding error probability $\overline{P}_{\mathcal{L}}$ vanishes as N increases.

The Second Block

For the second block of transmission, both senders can form the same list as the receiver as they both know v_1^M and $\underline{Y}_1^L$. Furthermore, if they have correctly estimated $\mathbf{U}_{11}$ and $\mathbf{U}_{21}$ and $(\mathbf{u}_{11}, \mathbf{u}_{21})$ is in $\mathcal{L}_1$, then they both know which pair in the list $\mathcal{L}_1$ is the correct one. To transmit this knowledge to the receiver, they choose the *same* length M codeword V_2^M from the V code. The V code needs to have rate $\lambda \cdot L$ bits per V symbol due to the size of $\mathcal{L}_1$. The transmitters then superpose their new information sequences $\mathbf{U}_{12}$ and $\mathbf{U}_{22}$ onto V_2^M to obtain the next channel inputs $\underline{X}_{12}^L$ and $\underline{X}_{22}^L$.

After receiving the second block of output $\underline{Y}_2^L$, both senders again decode $\mathbf{U}_{12}$ and $\mathbf{U}_{22}$ with the same error exponents $E_{G1}^L(R_1)$ and $E_{G2}^L(R_2)$ as before. The receiver now tries to determine which codeword V_2^M was sent by using $\underline{Y}_2^L$. By considering each of the M blocks of $\underline{Y}_2^L$ corresponding to an adaptive codeword of User 1 as an output symbol of the derived channel from V to the receiver, we may directly apply Theorem 5.6.4 in [8] to conclude that if

$$\lambda \cdot L = R_V < I(V; \underline{Y}), \tag{5.48}$$

then the error exponent $E_G(R_V)$ is positive. Combining (5.47) with (5.48), we have the two conditions

$$\begin{aligned} \lambda &= R_1 + R_2 - I_L(X_1X_2 \to Y|V) + 2/N \\ &< I(V; \underline{Y})/L & (5.49) \\ \lambda &\leq R_1 + R_2 & (5.50) \end{aligned}$$

for decoding V_2^M reliably *and* for ensuring that the pair $(\mathbf{u}_{11}, \mathbf{u}_{21})$ is in the list $\mathcal{L}_1$. These two conditions may be rewritten as

$$\begin{aligned} R_1 + R_2 &< I_L(X_1X_2 \to Y) - 2/N & (5.51) \\ 2/N &\leq I_L(X_1X_2 \to Y|V), & (5.52) \end{aligned}$$

where we have used $I(V;\underline{Y})/L + I_L(X_1X_2 \to Y|V) = I_L(X_1X_2 \to Y)$.

The condition (5.52) is met for large enough N as long as $I_L(X_1X_2 \to Y|V) > 0$. But $I_L(X_1X_2 \to Y|V) = 0$ implies that $H(Y^L|V)/L = H_L(Y\|X_1X_2)$, which in turn implies that $I_L(X_1 \to Y\|X_2|V) = 0$ and $I_L(X_2 \to Y\|X_1|V) = 0$, or $R_1 = R_2 = 0$. Thus, for all interesting rate points, the condition (5.52) is satisfied.

Once the receiver has determined V_2^M, he can again create a list $\mathcal{L}_2$ of the same size $2^{N\lambda}$ as before. The rest of the operation of the algorithm simply repeats the procedure described above. To bound the probability of an error occurring in the $B+1$ blocks, we may use the union bound, which limits the error probability to at most $4B$ times the maximum error probability corresponding to the four decoding error exponents per block (the first block has only the three list-decoding error exponents while last block has only the V code error exponent). By fixing B and letting N approach infinity, we may approach any rate point in the region bounded by

$$\begin{array}{ccccc} 0 & \leq & R_1 & \leq & \frac{B}{B+1} I_L(X_1 \to Y\|X_2|V) \\ 0 & \leq & R_2 & \leq & \frac{B}{B+1} I_L(X_2 \to Y\|X_1|V) \\ 0 & \leq & R_1 + R_2 & \leq & \frac{B}{B+1} I_L(X_1X_2 \to Y). \end{array} \tag{5.53}$$

We then let B become large to approach all rate points given in Lemma 5.4.

5.C Appendix: Equal-Rate Points on the Cover-Leung Region Boundary

For both channels considered in Sections 5.4.2 and 5.4.3, we define $q_{1v} = Pr(X_1 = 0|V = v)$ and $q_{2v} = Pr(X_2 = 0|V = v)$ for all v.

A BAC with Additive $(0, 1)$ Noise

Consider the channel model of Figure 5.4. We use a binary V with $\Pr(V = 0) = \Pr(V = 1) = 1/2$, and $\Pr(X_1 \neq V) = \Pr(X_2 \neq V) = q$. The Cover-Leung rate $R = R_1 = R_2$ is then bounded by (see Lemma 5.3)

$$\begin{aligned} R &\leq h(q)/2 && (5.54)\\ 2R &\leq h\left([q^2 + (1-q)^2]/2\right). && (5.55) \end{aligned}$$

The best q is obtained when $q = [q^2 + (1-q)^2]/2$, or $q = 1 - 1/\sqrt{2}$, and yields $R = h(1/\sqrt{2})/2 \approx 0.43621$ bits per use.

For general $p_{VX_1X_2Y}$, straightforward manipulations of the informations result in

$$\begin{aligned} I(X_1; Y|X_2V) &= \sum_v p(v)\, h(q_{1v})/2 && (5.56)\\ I(X_2; Y|X_1V) &= \sum_v p(v)\, h(q_{2v})/2 && (5.57)\\ I(X_1X_2; Y) &= h^{(4)}\left(p_Y(0), p_Y(1), p_Y(2), p_Y(3)\right) - 1, && (5.58) \end{aligned}$$

where $h^{(4)}(p_1, p_2, p_3, p_4) = \sum_{i=1}^{4} -p_i \log(p_i)$, and

$$\begin{aligned} p_Y(0) &= \sum_v p(v)\, q_{1v}q_{2v}/2\\ p_Y(1) &= \sum_v p(v)\, \left[q_{1v}q_{2v} + q_{1v}(1-q_{2v}) + (1-q_{1v})q_{2v}\right]/2\\ p_Y(2) &= \sum_v p(v)\, \left[(1-q_{1v})(1-q_{2v})\right.\\ &\qquad \left. + q_{1v}(1-q_{2v}) + (1-q_{1v})q_{2v}\right]/2\\ p_Y(3) &= \sum_v p(v)\, (1-q_{1v})(1-q_{2v})/2. \end{aligned}$$

We would like to show that a binary V is best for equal-rate points. As a first bound, we use the convexity of the entropies and Jensen's inequality (see Appendix 4.A) to write

$$\begin{aligned} I(X_1X_2;Y) &\leq h^{(4)}\left(\frac{p_Y(0)+p_Y(3)}{2}, \frac{p_Y(1)+p_Y(2)}{2}, \right. \\ &\qquad \left. \frac{p_Y(1)+p_Y(2)}{2}, \frac{p_Y(0)+p_Y(3)}{2}\right) - 1 \\ &= h\left(p_Y(0)+p_Y(3)\right) \\ &= h\left(\sum_v p(v)\,(1-t_v)/2\right), \end{aligned} \tag{5.59}$$

where $t_v := q_{1v}(1-q_{2v}) + (1-q_{1v})q_{2v}$.

The bounds (5.56), (5.57) and (5.59) are virtually identical with the bounds (3) and (5) in [38]. We may thus use the same function $\phi(\cdot)$ defined there, namely

$$\phi(t) = \begin{cases} \left(1-\sqrt{1-2t}\right)/2 & \text{for } 0 \leq t \leq 1/2, \\ \left(1-\sqrt{2t-1}\right)/2 & \text{for } 1/2 < t \leq 1. \end{cases} \tag{5.60}$$

In [38] it is shown that the composite function $h(\phi(\cdot))$ is symmetrical around $t = 1/2$ and convex-$\cap$ in t for $0 \leq t \leq 1$. Following the same steps as in equation (8) of [38],

$$R \leq h\left(\phi(\bar{t})\right)/2, \tag{5.61}$$

where $\bar{t} = \sum_v p(v)\,t_v$. Combining (5.61) and (5.59), we find that R satisfies

$$R \leq \min_{0\leq t\leq 1/2}\{h(\phi(t))/2\,,\; h\left((1-t)/2\right)/2\}, \tag{5.62}$$

or, by setting $q = \phi(t)$ so that $t = 2\,q(1-q)$,

$$R \leq \min_{0\leq q\leq 1/2}\{h(q)/2\,,\; h\left([q^2+(1-q)^2]/2\right)/2\}, \tag{5.63}$$

which is the same as the bounds (5.54) and (5.55). Thus, the rate point $\left(h(1/\sqrt{2})/2, h(1/\sqrt{2})/2\right) \approx (0.43621, 0.43621)$ bits per use lies on the boundary of the Cover-Leung region.

A BAC with Different-Input Noise

Consider the channel model of Figure 5.5. We again use a binary V with $\Pr(V=0) = \Pr(V=1) = 1/2$ and $\Pr(X_1 \neq V) = \Pr(X_2 \neq V) = q$. The Cover-Leung rate $R = R_1 = R_2$ is then bounded by

$$\begin{aligned} R &\leq q\, h\left(\frac{1-q}{2}\right) + (1-q)\, h\left(\frac{q}{2}\right) - 2q(1-q) && (5.64)\\ 2R &\leq 1 - 2q(1-q). && (5.65) \end{aligned}$$

The best $q \approx 0.3044$ yields $R \approx 0.28826$ bits per use.

For general $p_{VX_1X_2Y}$,

$$\begin{aligned} H(Y|X_1V) &= \sum_v p(v) \left[q_{1v} h\left(\frac{1-q_{2v}}{2}\right) + (1-q_{1v}) h\left(\frac{q_{2v}}{2}\right)\right] && (5.66)\\ H(Y|X_2V) &= \sum_v p(v) \left[q_{2v} h\left(\frac{1-q_{1v}}{2}\right) + (1-q_{2v}) h\left(\frac{q_{1v}}{2}\right)\right] && (5.67)\\ H(Y) &= h\left(\sum_v p(v)\,(q_{1v}+q_{2v})/2\right) && (5.68)\\ H(Y|X_1X_2) &= \sum_v p(v)\,\left[q_{1v}(1-q_{2v}) + (1-q_{1v})q_{2v}\right]. && (5.69) \end{aligned}$$

We would again like to show that a binary V is best for equal-rate points. We begin with $I(X_1X_2;Y)$. Using $(q_{1v}-q_{2v})^2 \geq 0$, we bound $2\,q_{1v}q_{2v} \leq q_{1v}^2 + q_{2v}^2$, so that

$$\begin{aligned} H(Y|X_1X_2) &\geq \sum_v p(v) \left[q_{1v} + q_{2v} - q_{1v}q_{2v} - (q_{1v}^2 + q_{2v}^2)/2\right] \\ &= \sum_v p(v) \left[q_{1v} + q_{2v} - 2\left(\frac{q_{1v}+q_{2v}}{2}\right)^2\right] \\ &= \sum_v p(v)\; 2\,\overline{q}_v(1-\overline{q}_v) && (5.70) \end{aligned}$$

where $\overline{q}_v := (q_{1v}+q_{2v})/2$. Using $H(Y) \leq 1$ and setting $\tau_v := 2\overline{q}_v(1-\overline{q}_v)$, we find that $I(X_1X_2;Y)$ satisfies

$$I(X_1X_2;Y) \leq 1 - \sum_v p(v)\,\tau_v. \qquad (5.71)$$

To bound $(I(X_1;Y|X_2V)+I(X_2;Y|X_1V))/2$, we prove the following lemmas.

Lemma 5.7 *The function*

$$f(x,y) = \frac{1-x}{2}\left[h\left(\frac{y}{2}\right) - y\right] + \frac{1-y}{2}\left[h\left(\frac{x}{2}\right) - x\right] \tag{5.72}$$

is overbounded by the function

$$g(x,y) = \left(1 - \frac{x+y}{2}\right)\left[h\left(\frac{x+y}{4}\right) - \frac{(x+y)}{2}\right] \tag{5.73}$$

for $0 \le x \le 1$ *and* $0 \le y \le 1$.

Proof: We may restrict our attention to $x \le y$ because $f(\cdot,\cdot)$ and $g(\cdot,\cdot)$ are symmetric about the line $x = y$, i.e., $f(x,y) = f(y,x)$ and $g(x,y) = g(y,x)$. The function defined by $(1-x)[h(y/2) - y]$ is *not* convex-$\cap$ in the pair (x,y). However, consider the line $x + y = c$ where c is a constant. The function $g(\cdot,\cdot)$ is constant on this line and we claim that the function defined by $f_c(x) = (1-x)[h((c-x)/2) - (c-x)]$ *is* convex-$\cap$ in x for $0 \le x \le c/2$ and $0 \le c \le 2$. The second derivative of $f_c(\cdot)$ with respect to x is given by

$$f_c''(x) = \frac{-1}{\ln(2)}\frac{1-x}{(c-x)[2-(c-x)]} + \log_2\left(\frac{2-(c-x)}{c-x}\right) - 2. \tag{5.74}$$

This value is clearly nonpositive if $c - x \ge 2/5$ as then the second term is not greater than 2. This is in turn guaranteed if $c \ge 4/5$ because $c - x \ge c/2$. For $c < 4/5$, we use $2 - (c-x) \le 2$ and $\log_2(x) \le \log_2(4/c) + (cx/4 - 1)/\ln(2)$ to bound

$$f_c''(x) \le \frac{x[3 + 2\ln(c)] - [1 + c + 2c\ln(c)]}{2\ln(2)(c-x)}. \tag{5.75}$$

The numerator of (5.75) is less than zero for $0 \le x \le c/2$ and $0 \le c < 4/5$, as may easily be checked. Thus $f_c(\cdot)$ is convex-$\cap$ as claimed and we may use Jensen's inequality to bound

$$f(x, c-x) \le (1-c)/2\,[h(c/4) - c/2] = g(c, c-x), \tag{5.76}$$

for all x, $y = c - x$, and c of interest. This proves the lemma. □

Lemma 5.8 *The function*

$$\psi(t) = \phi(t)\, h\left(\frac{1-\phi(t)}{2}\right) + (1-\phi(t))\, h\left(\phi(t)/2\right), \tag{5.77}$$

defined by (5.60) is convex-$\cap$ over the region $0 \le t \le 1$.

Proof: Denoting the first derivative of $\psi(\cdot)$ by $\psi'(\cdot)$ and similarly for $\phi'(\cdot)$, we find that

$$\begin{aligned}\psi'(t) = &-\phi(t)\frac{\phi'(t)}{2\ln(2)}\ln\left(\frac{1+\phi(t)}{1-\phi(t)}\right) + \phi'(t)h\left(\frac{1-\phi(t)}{2}\right)\\ &+(1-\phi(t))\frac{\phi'(t)}{2\ln(2)}\ln\left(\frac{2-\phi(t)}{\phi(t)}\right) - \phi'(t)h\left(\phi(t)/2\right).\end{aligned} \tag{5.78}$$

Denoting the second derivatives by $\psi''(\cdot)$ and $\phi''(\cdot)$, we have

$$\begin{aligned}\psi''(t) = &-\frac{\{\phi'(t)\}^2}{\ln(2)}\left[\frac{\phi(t)}{1-\{\phi(t)\}^2} + \frac{(1-\phi(t))}{\phi(t)[2-\phi(t)]}\right.\\ &\left.+\ln\left(\frac{1+\phi(t)}{1-\phi(t)}\cdot\frac{2-\phi(t)}{\phi(t)}\right)\right]\\ &+\frac{\phi''(t)}{\ln(2)}\left[[1-\phi(t)]\ln\left(\frac{2-\phi(t)}{\phi(t)}\right) - \phi(t)\ln\left(\frac{1+\phi(t)}{1-\phi(t)}\right)\right]\\ &+\phi''(t)\left[h\left(\frac{1-\phi(t)}{2}\right) - h\left(\phi(t)/2\right)\right]\end{aligned} \tag{5.79}$$

All terms in square brackets are positive because $0 \le \phi(t) \le 1/2$ for $0 \le t \le 1$. Furthermore, in [38] it was shown that $\phi''(t) \le 0$ for $0 \le t < 1/2$ and $1/2 < t \le 1$. Thus, $\psi''(t) \le 0$ for $0 \le t < 1/2$ and $1/2 < t \le 1$. It remains to consider the discontinuity at $t = 1/2$. As in [38], we must show that $\lim_{t\uparrow 1/2}\psi'(t) \ge 0$. Setting $p = \sqrt{1-2t}$, or $\phi(t) = (1-p)/2$, and using $\phi'(t) = 1/2p$ for $0 \le t < 1/2$, we can rewrite this limit as

$$\begin{aligned}\lim_{p\downarrow 0}\quad &\frac{1}{2p}\left[\frac{1+p}{4\ln(2)}\ln\left(\frac{3+p}{1-p}\right) - \frac{1-p}{4\ln(2)}\ln\left(\frac{3-p}{1+p}\right)\right.\\ &\left.+h\left(\frac{1+p}{4}\right) - h\left(\frac{1-p}{4}\right)\right]\\ = \quad &\frac{4+3\ln(3)}{12\ \ln(2)} + \frac{h'(1/4)}{4} \approx 1.2734.\end{aligned} \tag{5.80}$$

Combining (5.79) and (5.80), and noting that $\psi(\cdot)$ is symmetrical round $1/2$ because $\phi(\cdot)$ is symmetrical round $1/2$, we see that $\psi(\cdot)$ is convex-$\cap$ over $0 \leq t \leq 1$. $\square$

We now return to bounding the information rates. Expanding, we can write

$$\begin{aligned}(I(X_1;Y|X_2V) + I(X_2;Y|X_1V))/2 = \sum_v p(v)\{ \\ \frac{q_{1v}}{2}\left[h\left(\frac{1-q_{2v}}{2}\right) - (1-q_{2v})\right] + \frac{q_{2v}}{2}\left[h\left(\frac{1-q_{1v}}{2}\right) - (1-q_{1v})\right] \\ + \frac{1-q_{1v}}{2}\left[h(q_{2v}/2) - q_{2v}\right] + \frac{1-q_{2v}}{2}\left[h(q_{1v}/2) - q_{1v}\right]\Bigg\}. \end{aligned} \tag{5.81}$$

Consider the first two terms and the last two terms in the curly brackets in (5.81). We can apply Lemma 5.7 with $x = 1 - q_{1v}$ and $y = 1 - q_{2v}$ to the first two terms, and with $x = q_{1v}$ and $y = q_{2v}$ to the last two terms, to obtain

$$R \leq \sum_v p(v)\left\{\overline{q}_v\left[h((1-\overline{q}_v)/2) - (1-\overline{q}_v)\right] + (1-\overline{q}_v)\left[h(\overline{q}_v/2) - \overline{q}_v\right]\right\}, \tag{5.82}$$

where we have used $R \leq I(X_1;Y|X_2V)$, $R \leq I(X_2;Y|X_1V)$ and $\overline{q}_v = (q_{1v} + q_{2v})/2$. Using $\overline{q}_v = \phi(2\,\overline{q}_v(1-\overline{q}_v))$ and $2\,\overline{q}_v(1-\overline{q}_v) = \tau_v$, we may rewrite (5.82) as

$$\begin{aligned} R &\leq \sum_v p(v)\left\{\phi(\tau_v)h((1-\phi(\tau_v))/2) + (1-\phi(\tau_v))h(\phi(\tau_v)/2) - \tau_v\right\} \\ &\leq \phi(\overline{\tau})h((1-\phi(\overline{\tau}))/2) + (1-\phi(\overline{\tau}))h(\phi(\overline{\tau})/2) - \overline{\tau}, \end{aligned} \tag{5.83}$$

where $\overline{\tau} := \sum_v p(v)\,\tau_v$ and where we have used the convexity of $\psi(\cdot)$ (cf. Lemma 5.8) and Jensen's inequality for the second step. Combining (5.83) and (5.71), we find that R satisfies

$$R \leq \min_{0 \leq t \leq 1/2}\{\phi(t)h((1-\phi(t))/2) + (1-\phi(t))h(\phi(t)/2) - t\,,\ 1-t\}, \tag{5.84}$$

or, upon setting $\phi(t) = q$ so that $t = 2\,q(1-q)$, that

$$\begin{aligned} R \leq \min\nolimits_{0 \leq q \leq 1/2} \quad & \{qh((1-q)/2) + (1-q)h(q/2) - 2\,q(1-q), \\ & 1 - 2\,q(1-q)\}, \end{aligned} \tag{5.85}$$

which is the same as the bounds (5.64) and (5.65). Thus, the rate point $(R_1, R_2) \approx (0.28826, 0.28826)$ bits per use lies on the boundary of the Cover-Leung region.

5.D Appendix: Coding Distributions for the Case Studies

A BAC with Additive $(0,1)$ Noise

We consider the channel model of Figure 5.4 and use a memory 1 coding technique where $\sigma_1 = (\tilde{x}_1, \tilde{y}_1)$ and $\sigma_2 = (\tilde{x}_2, \tilde{y}_2)$, where the tilde denotes the symbol of the previous channel use. The users' state diagrams have six states while the system state diagram has eight states. We use the probability distribution $p_V(0) = p_V(1) = 1/2$, $\Pr(X_k \neq V \mid \tilde{X}_k = i, \tilde{Y} = j) = q_{ij}^{(k)}$. The best $q_{ij}^{(k)}$ we found were

$$\begin{array}{llllll} q_{00}^{(1)} & = 0.2584, & q_{01}^{(1)} & = 0.7148, & q_{02}^{(1)} & = 1, \\ q_{11}^{(1)} & = 0, & q_{12}^{(1)} & = 1 - 0.7148, & q_{13}^{(1)} & = 0.2584 \\ & & & & & \\ q_{00}^{(2)} & = 0.2584, & q_{01}^{(2)} & = 1 - 0.7148, & q_{02}^{(2)} & = 0, \\ q_{11}^{(2)} & = 1, & q_{12}^{(2)} & = 0.7148, & q_{13}^{(2)} & = 0.2584. \end{array} \tag{5.86}$$

The resulting system state probability distribution is

$$\begin{array}{llllll} p_{000} & = 0.1486, & p_{001} & = 0.1486, & p_{011} & = 0.1014, \\ p_{012} & = 0.1014, & p_{101} & = 0.1014, & p_{102} & = 0.1014, \\ p_{112} & = 0.1486, & p_{113} & = 0.1486, & & \end{array} \tag{5.87}$$

where $p_{ijk} = \Pr(\Sigma = (\tilde{X}_1, \tilde{X}_2, \tilde{Y}) = (i,j,k))$. Bounds corresponding to (4.69) for the steady-state entropies are (all quantities in bits per use)

$$\begin{aligned} H(Y_\ell | \Sigma_\ell V^\ell X_2^\ell) &= 1.33811 \leq \\ H(Y_\ell | \Sigma_{\ell-1} V^\ell X_2^\ell Y^{\ell-1}) &= 1.42003 \leq \\ H(Y_\ell | \Sigma_{\ell-2} V^\ell X_2^\ell Y^{\ell-1}) &= 1.43879 \leq \\ H_\infty(Y \| X_2 V) &\leq \\ H(Y_\ell | \Sigma_{2(\ell-2)} V^{\ell-2..\ell} X_2^{\ell-2..\ell} Y^{\ell-2..\ell-1}) &= 1.44559 \leq \\ H(Y_\ell | \Sigma_{2(\ell-1)} V^{\ell-2..\ell} X_2^{\ell-1..\ell} Y_{\ell-1}) &= 1.44772 \leq \\ H(Y_\ell | \Sigma_{2\ell} V_\ell X_{2\ell}) &= 1.45586 \end{aligned} \tag{5.88}$$

and

$$\begin{aligned} H(Y_\ell|\Sigma_\ell) &= 1.86319 \leq \\ H(Y_\ell|Y^{\ell-1}\Sigma_{\ell-1}) &= 1.87673 \leq \\ H(Y_\ell|Y^{\ell-1}\Sigma_{\ell-2}) &= 1.87758 \leq \\ H_\infty(Y) &\leq \\ H(Y_\ell|Y_{\ell-2}Y_{\ell-1}) &= 1.87764 \leq \\ H(Y_\ell|Y_{\ell-1}) &= 1.87765 \leq \\ H(Y_\ell) &= 1.87783. \end{aligned} \tag{5.89}$$

Because $H(Y|X_1X_2) = 1$ and because both users have the same directed information rates, we have

$$\begin{aligned} I_\infty(X_1 \to Y\|X_2V) &\geq 1.43879 - 1 = 0.43879 \\ I_\infty(X_2 \to Y\|X_1V) &\geq 1.43879 - 1 = 0.43879 \\ I_\infty(X_1X_2 \to Y) &\geq 1.87758 - 1 = 2(0.43879). \end{aligned}$$

Thus, $R_1 = R_2 = 0.43879$ bits per use is approachable. This is beyond the rate point $R_1 = R_2 = 0.43621$ bits per use that lies on the boundary of the Cover-Leung region.

A BAC with Different-Input Noise

We consider the channel model of Figure 5.5 and again use a memory 1 coding technique where $\sigma_1 = (\tilde{x}_1, \tilde{y}_1)$ and $\sigma_2 = (\tilde{x}_2, \tilde{y}_2)$ and where the tilde denotes the symbol of the previous channel use. The users' state diagrams have four states while the system state diagram has six states. We again use the probability distribution $p_V(0) = p_V(1) = 1/2$, $\Pr(X_k \neq V \mid \tilde{X}_k = i, \tilde{Y} = j) = q_{ij}^{(k)}$. The best $q_{ij}^{(k)}$ we found were

$$\begin{aligned} q_{00}^{(1)} &= 0.2197, & q_{01}^{(1)} &= 0, \\ q_{10}^{(1)} &= 0.7236, & q_{11}^{(1)} &= 0.4169 \\ \\ q_{00}^{(2)} &= 0.4169, & q_{01}^{(2)} &= 0.7236, \\ q_{10}^{(2)} &= 0, & q_{11}^{(2)} &= 0.2197. \end{aligned} \tag{5.90}$$

The resulting system state probability distribution is

$$\begin{aligned} p_{000} &= 0.2891, & p_{010} &= 0.1054, & p_{011} &= 0.1054, \\ p_{100} &= 0.1054, & p_{101} &= 0.1054, & p_{111} &= 0.2891, \end{aligned} \tag{5.91}$$

where $p_{ijk} = \Pr(\Sigma = (\tilde{X}_1, \tilde{X}_2, \tilde{Y}) = (i,j,k))$. Bounds corresponding to (4.69) for the steady-state entropies are (all quantities in bits per use)

$$\begin{aligned}
H(Y_\ell|\Sigma_\ell V^\ell X_2^\ell) &= 0.66971 \leq \\
H(Y_\ell|\Sigma_{\ell-1} V^\ell X_2^\ell Y^{\ell-1}) &= 0.70677 \leq \\
H(Y_\ell|\Sigma_{\ell-2} V^\ell X_2^\ell Y^{\ell-1}) &= 0.71089 \leq \\
H_\infty(Y\|X_2V) &\leq \\
H(Y_\ell|\Sigma_{2(\ell-2)} V^{\ell-2..\ell} X_2^{\ell-2..\ell} Y^{\ell-2..\ell-1}) &= 0.71151 \leq \\
H(Y_\ell|\Sigma_{2(\ell-1)} V^{\ell-2..\ell} X_2^{\ell-1..\ell} Y_{\ell-1}) &= 0.71195 \leq \\
H(Y_\ell|\Sigma_{2\ell} V_\ell X_{2\ell}) &= 0.71552. \qquad (5.92)
\end{aligned}$$

For the specified distribution, $H_\infty(Y) = 1$ and $H(Y|X_1X_2) = 0.42178$, so that

$$\begin{aligned}
I_\infty(X_1 \to Y\|X_2V) &\geq 0.71089 - 0.42178 = 0.28911 \\
I_\infty(X_2 \to Y\|X_1V) &\geq 0.71089 - 0.42178 = 0.28911 \\
I_\infty(X_1X_2 \to Y) &\geq 1 - 0.42178 = 2(0.28911).
\end{aligned}$$

Thus, $R_1 = R_2 = 0.28911$ bits per use is approachable. This is beyond the rate point $R_1 = R_2 = 0.28826$ bits per use that lies on the boundary of the Cover-Leung region.

Chapter 6

Feedback Strategies for a Class of Two-User Multiple-Access Channels

6.1 The Capacity Region

There is as at present no simple way to calculate the boundary points of the capacity region of the general two-user discrete memoryless multiple-access channel (MAC) with feedback. However, for the class of channels where one of the channel inputs, say X_1, is determined by the other input X_2 and the channel output Y, i.e., $H(X_1|X_2Y) = 0$, the feedback capacity region was shown by Willems [30] to be the rate region $\mathcal{R}^{CL}$ obtained by Cover and Leung [29]. This rate region is the set of nonnegative rate pairs (R_1, R_2) for which (see Lemma 5.3)

$$\begin{aligned} R_1 &\leq I(X_1; Y \mid X_2 V) && (6.1)\\ R_2 &\leq I(X_2; Y \mid X_1 V) && (6.2)\\ R_1 + R_2 &\leq I(X_1 X_2; Y), && (6.3) \end{aligned}$$

where V is any discrete random variable with cardinality $|\mathcal{V}|$ at most $\min(|\mathcal{X}_1|\cdot|\mathcal{X}_2|+1, |\mathcal{Y}|+2)$ and where the probability distribution $p_{VX_1X_2Y}$

of the random variables V, X_1, X_2 and Y factors as

$$p_{VX_1X_2Y} = p_V \cdot p_{X_1|V} \cdot p_{X_2|V} \cdot p_{Y|X_1X_2}. \tag{6.4}$$

For example, the binary adder channel with feedback (BAC-FB) falls into this class of channels. For the BAC-FB, one need consider only $|\mathcal{V}| \leq 5$ and, if one chooses $\Pr(V = 0) = \Pr(V = 1) = 1/2$ and $\Pr(X_1 \neq V) = \Pr(X_2 \neq V) = 0.23766$, one finds that the rate pair $(0.7911, 0.7911)$ can be approached [18] and is the equal-rate point on the boundary of the capacity region [38].

The rate region $\mathcal{R}^{CL}$ was derived in [29] using *block Markov superposition encoding*, a technique which relies on the use of randomly chosen block codes. Alternatively, one can consider constructive strategies which exploit feedback strategies for single-user channels. This approach was taken by Vinck [64, 65, 66], who found that for the BAC-FB the rate pair $(0.7909, 0.7909)$ can be approached by using a repetition strategy of Schalkwijk [67]. In fact, it can be shown that by choosing a better strategy, one can approach the rate point $(0.7911, 0.7911)$ [68].

We generalize the strategies of [64, 65, 68]. The structure of our strategies is based on the superposition coding structure of Cover and Leung [29], except that now the (single-user) *random codes* have been replaced by (single-user) feedback *strategies*.

This chapter is organized as follows. In Section 6.2 we consider the BAC-FB in detail to motivate the final structure of our strategies. We follow an approach similar to Vinck's [64, 65] to show that the equal-rate point on the boundary of the capacity region of the BAC-FB can be approached with a simple feedback strategy. We then extend this strategy to approach any rate point in the capacity region of channels for which $H(X_1|X_2Y) = 0$ *and* $H(X_2|X_1Y) = 0$, which includes the BAC-FB. Finally, in Section 6.4, the strategies are extended to channels for which $H(X_1|X_2Y) = 0$ *or* $H(X_2|X_1Y) = 0$.

6.2 Strategies for the Binary Adder Channel

The first strategy we shall consider for the BAC-FB is depicted in Figure 6.1. The two users independently generate the information bit se-

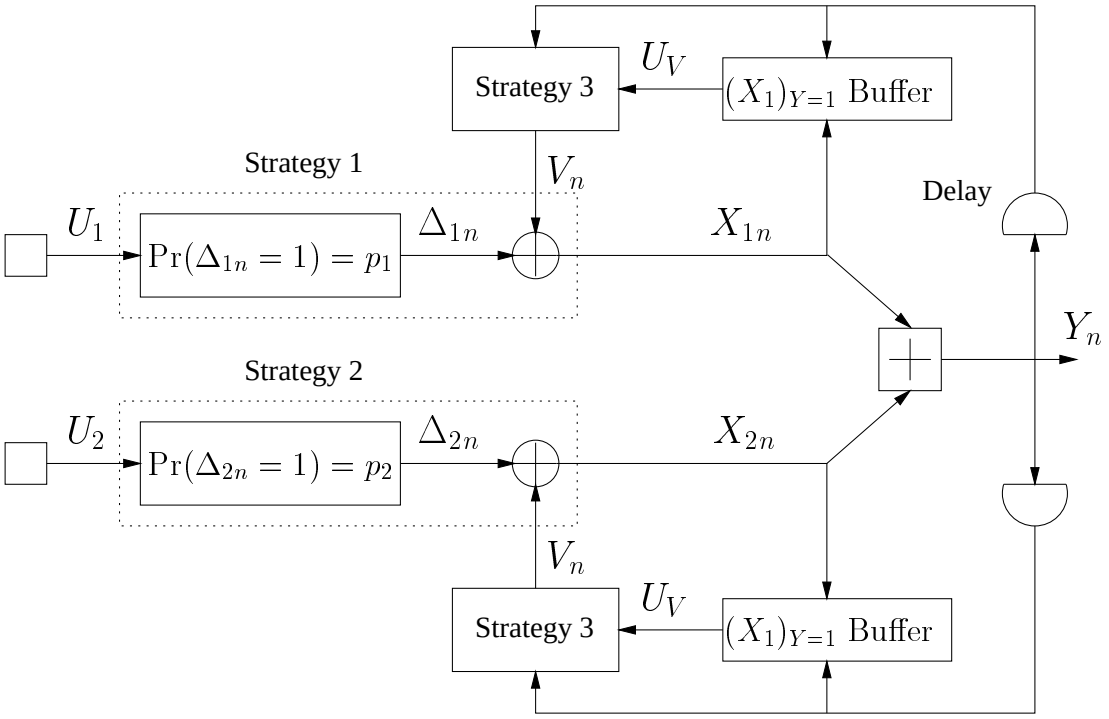


Figure 6.1: *A simple strategy for the noiseless binary adder channel with feedback. The ambiguous X_{1n} form the source sequence U_V.*

quences U_1 and U_2 and encode these into sequences Δ_1 and Δ_2 of independent and identically distributed bits so that $\Pr(\Delta_{1n} = 1) = p_1$ and $\Pr(\Delta_{2n} = 1) = p_2$, all n. For example, User 1 (and User 2) can map his information bits onto a point in the unit interval as in [5], and then divide this interval into one part with *a priori* probability p_1 (and p_2) and another with *a priori* probability $1 - p_1$ (and $1 - p_2$). The bits Δ_{1n} and Δ_{2n} are then added modulo 2 to V_n to obtain the inputs X_{1n} and X_{2n} to the channel. The output Y_n of the channel is fed back to both transmitters (with a small delay) so that both know X_{1n} and X_{2n} after receiving the feedback Y_n. The receiver, however, does not know X_{1n} or X_{2n} when $Y_n = 1$. To resolve these ambiguities, whenever $Y_n = 1$ the transmitters append X_{1n}, which both know, to a bit stream U_V and encode this stream into the sequence of bits V by using a single-user feedback strategy, e.g. [5, 69, 70, 71, 72].

6.2.1 The Equal-Rate Point

We first consider the equal-rate points of the BAC-FB capacity region. For this case, we set $p_1 = p_2 = p$ and the analysis of the strategy

becomes particularly simple. As long as the buffer for the ambiguous X_1 is stable, the transmitters can resolve all the ambiguities. But the rate at which bits enter this buffer is $\Pr(Y_n = 1)$ and we can empty the buffer at the capacity $C(p)$ of the channel from V to Y. Thus, the condition for buffer stability is simply $\Pr(Y_n = 1) < C(p)$. It is easily checked that $\Pr(Y_n = 1) = 2p(1-p)$ and that the channel from V to Y is a binary symmetric erasure channel with capacity (see also Vinck [64])

$$C(p) = \left[p^2 + (1-p)^2\right] \cdot \left[1 - h\left(\frac{p^2}{p^2 + (1-p)^2}\right)\right], \tag{6.5}$$

where $h(\cdot)$ is the binary entropy function. We find that $2p(1-p) < C(p)$ when $0 \le p < p^* = 0.23766$, which limits the transmission rates of the users to values less than $h(p^*) = 0.7911$. It follows that as long as the strategy used to transmit V approaches the capacity of the V to Y channel, one can approach equal-rate pairs up to $(0.7911, 0.7911)$.

To show that the condition $2p(1-p) = C(p)$ exactly determines the equal-rate point on the boundary of the capacity region, we must consider a result of Willems [38], who showed that the equal-rate boundary point is $(h(p), h(p))$ where p satisfies $0 \le p \le 1/2$ and

$$h(p) = \frac{1}{2}\left[\, h(2p(1-p)) + p^2 + (1-p)^2 \,\right]. \tag{6.6}$$

But if we expand the right side of (6.5) and set $\pi = p^2 + (1-p)^2$ we obtain

$$\begin{aligned} C(p) &= \pi + p^2 \log(p^2/\pi) + (1-p)^2 \log((1-p)^2/\pi) \\ &= \pi + h(\pi) + (1-\pi)\log(1-\pi) + p^2 \log(p^2) \\ &\quad + (1-p)^2 \log((1-p)^2) \\ &= \left[p^2 + (1-p)^2\right] + h(2p(1-p)) + 2p(1-p) - 2h(p), \end{aligned} \tag{6.7}$$

where the third step follows from $1 - \pi = 2p(1-p)$. Setting $C(p) = 2p(1-p)$ we obtain (6.6), as desired.

6.2.2 Other Rate Points

We next consider the case $p_1 \neq p_2$ for which $\Pr(Y_n = 1) = p_1(1-p_2) + p_2(1-p_1)$ and the V-to-Y channel is a binary symmetric erasure channel

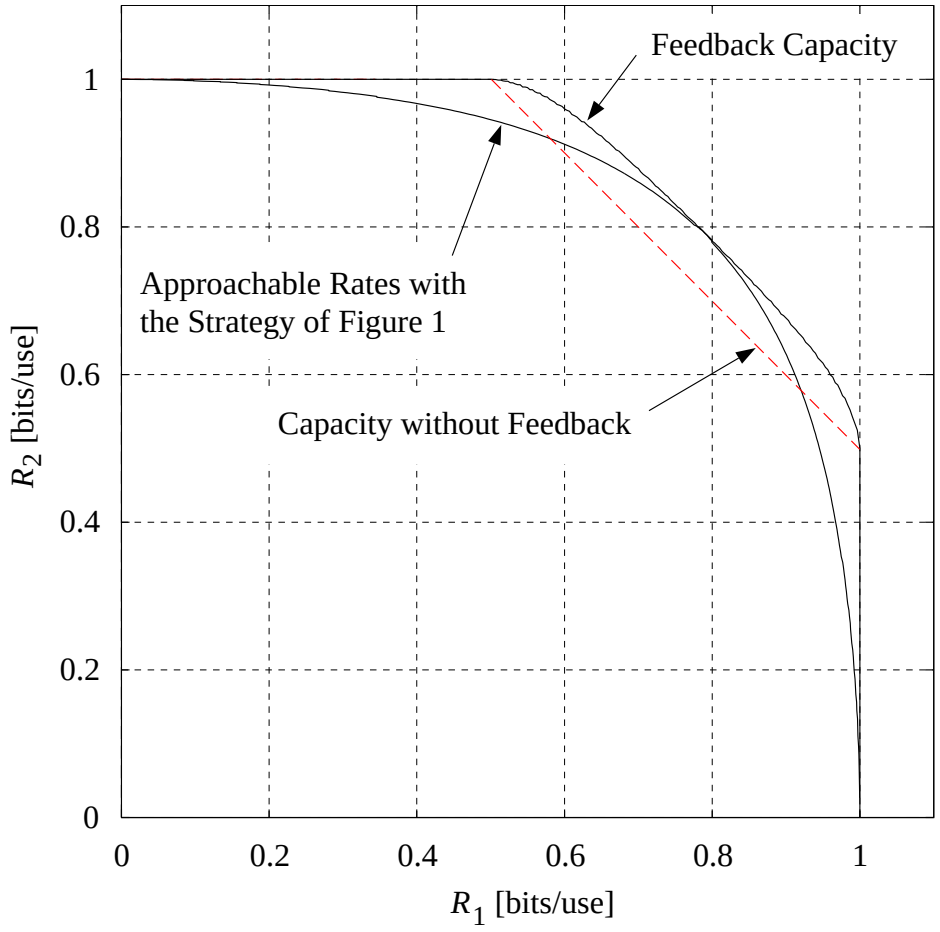


Figure 6.2: *Rates approachable with error probability approaching zero for the BAC-FB using the strategy of Figure 6.1.*

with capacity $C(p_1, p_2)$ given by

$$[p_1p_2 + (1-p_1)(1-p_2)] \cdot \left[1 - h\left(\frac{p_1p_2}{p_1p_2 + (1-p_1)(1-p_2)}\right)\right]. \quad (6.8)$$

The approachable rate points are those $(h(p_1), h(p_2))$ where (p_1, p_2) satisfies $\Pr(Y_n = 1) \leq C(p_1, p_2)$ and they are plotted in Figure 6.2. We see that the strategy of Figure 6.1 approaches the equal-rate point on the boundary of the capacity region, but it does *not* do so for any other (nontrivial) point on the boundary. For example, consider the boundary point $(1, 0.5)$. In this case we need $h(p_1) = 1$ and $h(p_2) = 0.5$, or $p_1 = 0.5$ and $p_2 \approx 0.11$. For these values, $\Pr(Y_n = 1) = 0.5$ but $C(0.5, 0.11) = 0.25$.

Of course, one may now combine the strategy of Figure 6.1 with other known strategies. For example, the rate point $(1, 0.5)$ can be

approached if User 2 uses a selective-repeat automatic repeat request (ARQ) strategy. Thus, one can approach any rate point inside the convex hull of the points $(1, 0.5)$, $(0.5, 1)$ as well as those points on the approachable rate curve in Figure 6.2. To approach even better rate points, we will modify the encoding structure.

Adding Source Coding

In the strategy of Figure 6.1, the uncertainty that the receiver has about X_{1n} when $Y_n = 1$ depends on p_1 and p_2. This uncertainty, after decoding V, is

$$H(X_1|V, Y = 1) = h\left(\frac{p_1(1-p_2)}{p_1(1-p_2) + (1-p_1)p_2}\right). \tag{6.9}$$

For example, if $p_1 = p_2$ we obtain $H(X_1|V, Y = 1) = 1$, so that U_V is a sequence of uniformly random bits as far as the receiver is concerned. On the other hand, if $p_1 = 0.5$ and $p_2 \approx 0.11$ as in the previous section, then $H(X_1|V, Y = 1) = 0.5$. Thus, it is not surprising that the strategy of Figure 6.1 does not approach the rate point $(1, 0.5)$, as one is trying to send a binary sequence U_V which has entropy 0.5 bits per symbol without compressing it first.

We thus modify the encoding structure to that shown in Figure 6.3. This structure includes a data compression algorithm for a discrete memoryless source (DMS), e.g., an arithmetic source encoder [73, page 61]. In fact, this box was implicit in the strategy described in [65] for ternary channels. Note also that if we use versions of [71, 72] for the channel coding, both the "Compress" and "Strategy 3" boxes in Figure 6.3 contain source encoders.

We can now show that the encoding structure of Figure 6.3 approaches every rate point $(h(p_1), h(p_2))$ for which the pair (p_1, p_2) satisfies $\Pr(Y_n = 1) \cdot H(X_1|V, Y = 1) \leq C(p_1, p_2)$. The resulting set of approachable rates matches those obtained in (6.1)-(6.4) for a binary V and $\Pr(X_1 \neq V) = p_1$, $\Pr(X_2 \neq V) = p_2$. However, we do not know whether this $p_{VX_1X_2Y}$ distribution approaches all the rate points in the BAC-FB capacity region. Thus, we prefer to introduce a more general strategy that will also allow us to generalize our approach to other channels.

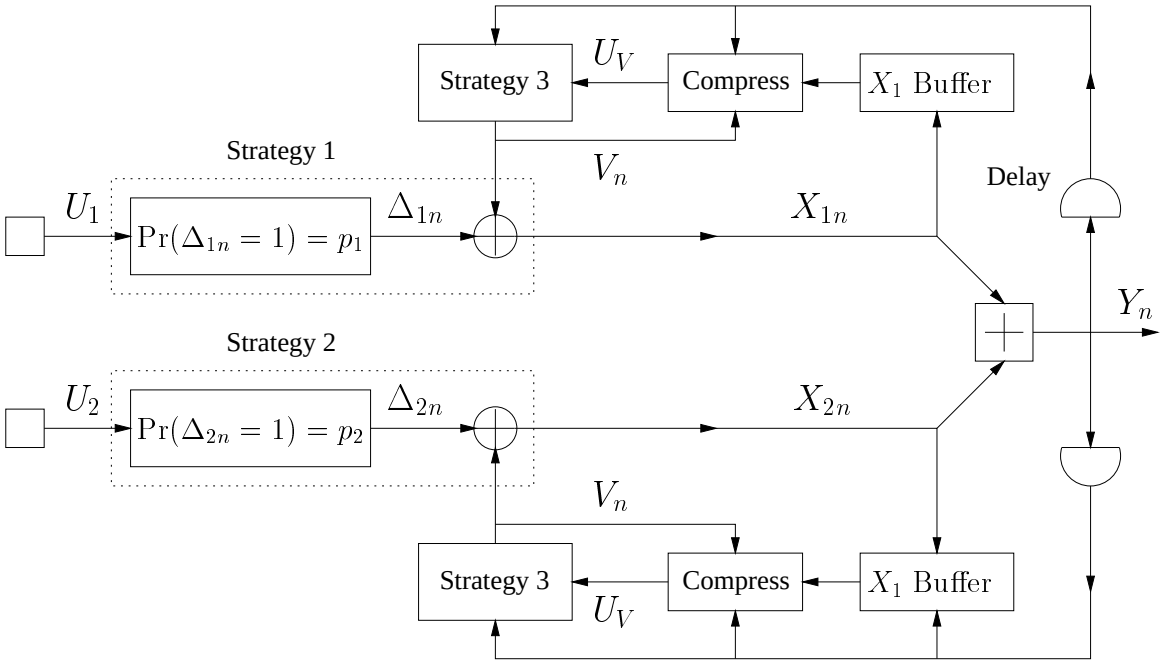


Figure 6.3: *A second strategy for the BAC-FB. This strategy includes the first one as a special case. The sequence of X_{1n} is compressed to form the source sequence U_V by using the knowledge of the V and Y sequences and the distributions $p_{X_1|VY}(\cdot|v_n, y_n)$.*

6.3 Strategies for Channels with $H(X_1|X_2Y) = H(X_2|X_1Y) = 0$

The third structure we shall consider in detail is depicted in Figure 6.4. This strategy differs from that of Figure 6.3 in that the X_{1n} and X_{2n} are directly generated from U_1, U_2 and V_n, without creating Δ_{1n} and Δ_{2n} as in Figure 6.1.

We explain how the various boxes in Figure 6.4 operate. User 1 and User 2's "X_1 Buffer" box collects the sequence of X_{1n}. The contents of this buffer are passed to the "Compress" box that contains a source-coding scheme which can compress a DMS to its entropy given the side information $V_n = v_n$ and $Y_n = y_n$, e.g., an arithmetic source encoder. The resulting output is the sequence of bits U_V.

The "Strategy 3" box contains a single-user feedback strategy that can approach the rate $I(V;Y)$ on the V-to-Y channel when the probability distribution $p_V(\cdot)$ is fixed. One may, e.g., use the Horstein-like strate-

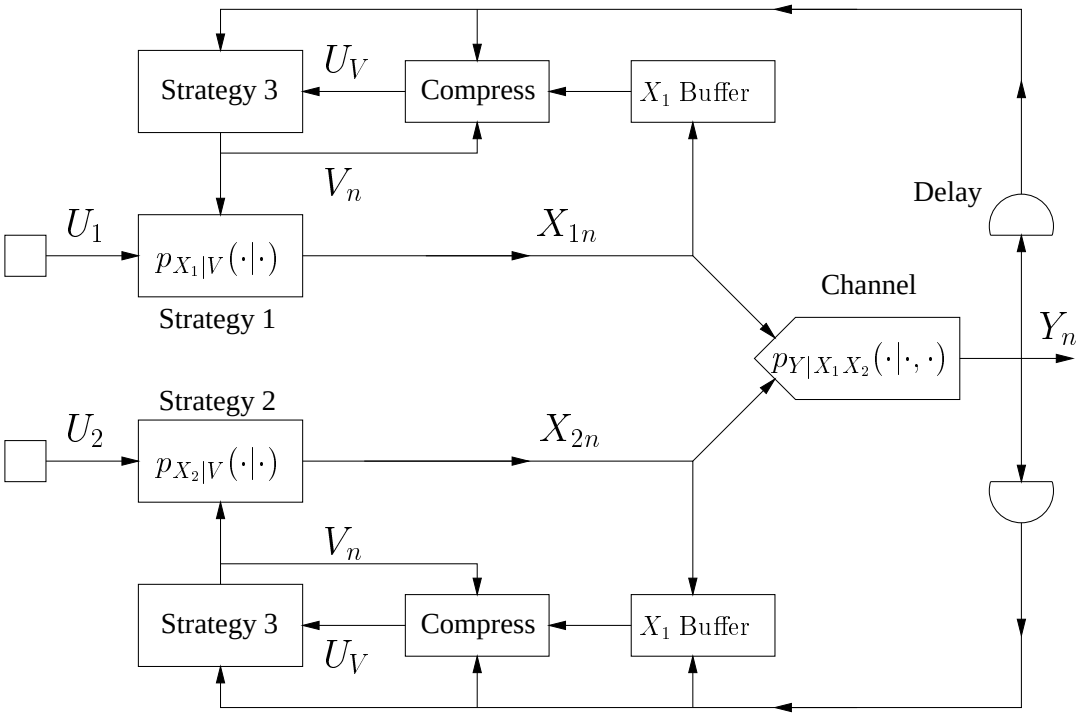


Figure 6.4: *An encoding structure for channels for which* $H(X_1|X_2Y) = H(X_2|X_1Y) = 0$. *The* X_{1n} *and* X_{2n} *are now directly formed from* U_1, U_2 *and* V_n, *without creating* Δ_{1n} *and* Δ_{2n} *as in Figure 6.3.*

gies of [5, 69, 70], but as far as we know no one has proved that these approach $I(V;Y)$ for arbitrary discrete memoryless channels (DMCs) – even though it seems clear that they do. Of course, one may always resort to the source coding strategies of [71, 72].

The "$p_{X_1|V}(\cdot|\cdot)$" and "$p_{X_2|V}(\cdot|\cdot)$" boxes operate like Horstein's strategy but without feedback. For example, if X_1 takes on K_1 values, then the "$p_{X_1|V}(\cdot|\cdot)$" box maps the information bits U_1 onto a point in the unit interval and divides this interval into K_1 parts with *a priori* probabilities specified by $p_{X_1|V}(\cdot|v_n)$. The channel input X_{1n} is then chosen according to the interval in which U_1 lies. User 2 generates X_{2n} in an analogous fashion. The probability distribution $p_{VX_1X_2Y}$ will then have the desired form given in (6.4).

The receiver operates by first decoding the U_V sequence, from which the V sequence can be generated. Using the V and Y sequences, the receiver can decompress U_V to form the X_1 sequence. Because $H(X_2|X_1Y) = 0$, the receiver can also generate the X_2 sequence. Finally, the bit sequences U_1 and U_2 are obtained.

6.3.1 A Condition for Achievable Rates

Using the strategy of Figure 6.4, we find that the U_V buffer will tend to empty as long as the senders can, at rate $I(V;Y)$, resolve the receiver's ambiguities, which have rate $H(X_1|YV)$. The condition for the U_V buffer to be stable is thus

$$H(X_1|YV) < I(V;Y). \tag{6.10}$$

Rewriting this condition by adding $I(X_1X_2;Y|V)$ to both sides of (6.10) and simplifying using $H(X_1|X_2V) = H(X_1|V)$ and $H(X_1|X_2Y) = 0$, we obtain the condition

$$I(X_1;Y|X_2V) + I(X_2;Y|X_1V) < I(X_1X_2;Y). \tag{6.11}$$

Thus, for all $p_{VX_1X_2Y}$ for which the sum-rate bound (6.3) is unnecessary in the bounds (6.1)-(6.3), the strategy of Figure 6.4 approaches all rate points in the rectangle defined by (6.1) and (6.2). For example, all the rate points achieved by the strategy of Figure 6.3 obey (6.11). Those $p_{VX_1X_2Y}$ which do not satisfy (6.11) are treated next.

6.3.2 Mixed Strategies

For those $p_{VX_1X_2Y}$ for which the condition (6.11) is not satisfied, the bounds (6.1)-(6.3) determine a pentagon as in Figure 6.5. As in this figure, we will use the shorthand I_1, I_2 and I_{12} for $I(X_1;Y|X_2V)$, $I(X_2;Y|X_1V)$ and $I(X_1X_2;Y)$, respectively.

We consider the queueing rate R_Q for the U_V buffer, i.e., the arrival rate minus the departure rate. From (6.10) and (6.11), we have

$$R_Q = I_1 + I_2 - I_{12}. \tag{6.12}$$

Suppose that in a first mode of operation both senders send their data in a block of length N ("Mode 1" in Figure 6.6). Then after this block of data has been sent, the U_V buffer will have about R_QN bits in it. The users can now, instead of sending new data in the next block, empty the buffer by sending at the "superuser" rate I_{12}, i.e., they switch to a transmission mode dedicated to emptying the U_V buffer ("Mode 2" in Figure 6.6). In fact, if the capacity of the superuser channel is larger

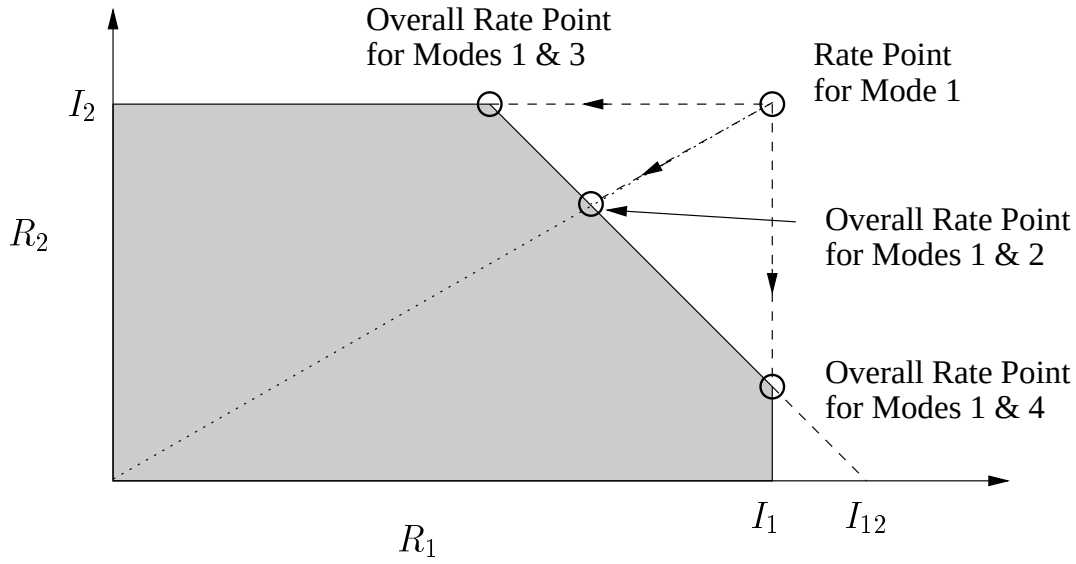


Figure 6.5: *The rate region defined by the bounds (6.1)-(6.3) for a particular $p_{VX_1X_2Y}$. In this figure, $I_1 = I(X_1;Y|X_2V)$, $I_2 = I(X_2;Y|X_1V)$ and $I_{12} = I(X_1X_2;Y)$. The operating points for the modes of operation of Figure 6.6 are circled.*

than $I_{12} = I(X_1X_2;Y)$, the users will be able to transmit at a higher rate than that specified by the bounds (6.1)-(6.3). For example, this occurs in the Gaarder-Wolf scheme [28] when the users are in the mode of cooperatively reconciling ambiguities at the receiver. The overall information rates for this dual-mode strategy are

$$R_1 = \frac{\#\text{Bits}}{\#\text{Uses}} = \frac{N \cdot I_1}{N + NR_Q/I_{12}} = \frac{I_{12}}{I_1 + I_2} \cdot I_1, \tag{6.13}$$

$$R_2 = \frac{I_{12}}{I_1 + I_2} \cdot I_2. \tag{6.14}$$

This rate point is labelled in Figure 6.5 as the "Overall Rate Point for Modes 1 & 2". We remark that Vinck [65] used a dual-mode strategy to empty the U_V buffer for approaching the total cooperation line of the noiseless ternary adder channel with a *binary* V. But one can also use a *ternary* V to approach the total cooperation line without resorting to two modes of operation, as was shown by Lalive d'Epinay and Giger [74]. In fact, any rate point approached by time-sharing the modes of operation in Figure 6.6 can be approached by using only Mode 1 with a new V which has time sharing built into it. This is shown in the appendix to this chapter.

The above demonstrates that one can approach one of the points

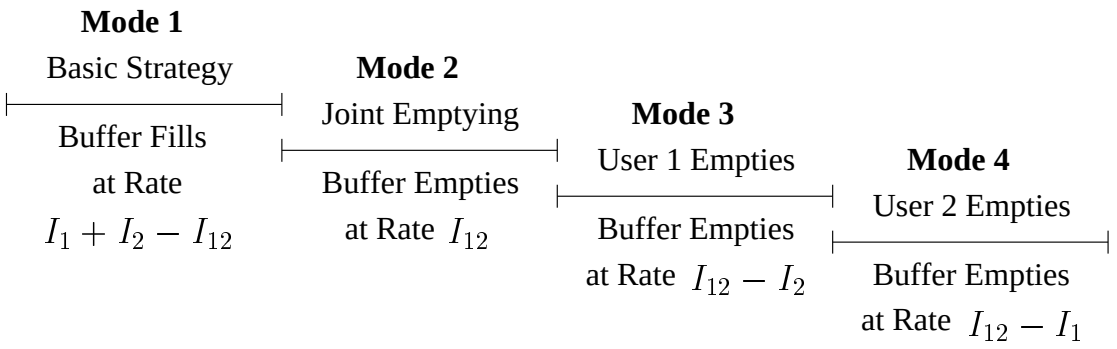


Figure 6.6: *Four modes of operation to approach any point in the rate region shown in Figure 6.5.*

on the boundary of the pentagon with a dual-mode strategy. We now show that the outermost corner points of the pentagon can also be so approached by a dual-mode strategy. The idea is simply to let *one* of the users empty the U_V buffer in the second block ("Mode 3" or "Mode 4" in Figure 6.6) while the other user continues to send data as in the first block and ignores the buffer bits that the other user is sending (Mode 1). For example, if during the second block User 1 empties the U_V buffer by himself (Mode 3) while User 2 is sending data (Mode 1), then the U_V buffer empties at rate I_1 but is simultaneously filling up at rate R_Q. Thus, User 1's overall information rate is

$$R_1 = \frac{\#\text{Bits}}{\#\text{Uses}} = \frac{N \cdot I_1}{N + NR_Q/(I_1 - R_Q)} = I_{12} - I_2, \tag{6.15}$$

while $R_2 = I_2$. The resulting rate point is labelled in Figure 6.5 as the "Overall Rate Point for Modes 1 & 3". The "Overall Rate Point for Modes 1 & 4" is obtained by interchanging the roles of Users 1 and 2.

The above shows that any point on the diagonal line of the pentagon in Figure 6.5 can be approached: one simply time-shares the four modes of operation in Figure 6.6. This, along with the result of Section 6.3.1, shows that the strategy of Figure 6.4 used in several modes can approach any point in the rate region $\mathcal{R}^{CL}$. Thus, this strategy can approach any point in the capacity region of channels with $H(X_1|X_2Y) = H(X_2|X_1Y) = 0$.

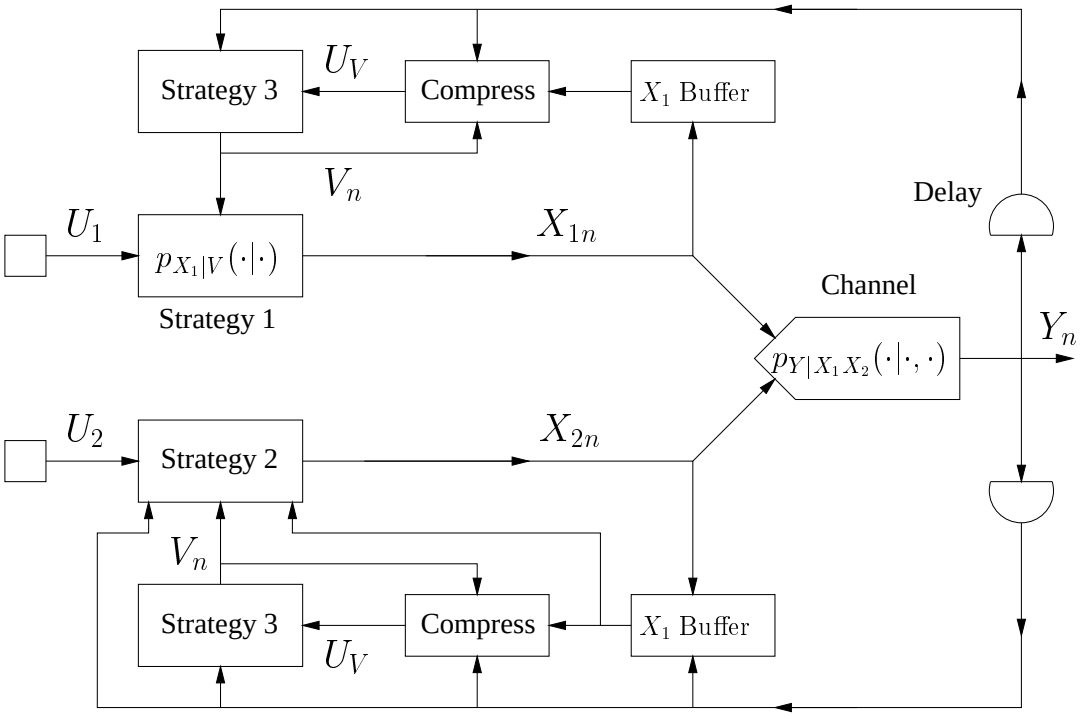


Figure 6.7: *An encoding structure for channels for which $H(X_1|X_2Y) = 0$ but $H(X_2|X_1Y) > 0$. User 2 must now use a true strategy to transmit his data, rather than just a "$p_{X_2|V}(\cdot|\cdot)$" box as in Figure 6.4.*

6.4 Strategies for $H(X_1|X_2Y) = 0$ Channels

The results of the previous section are easily extended to the case where $H(X_1|X_2Y) = 0$ but $H(X_2|X_1Y) > 0$. The only new difficulty is that the receiver cannot determine X_{2n} from X_{1n} and Y_n alone. User 2 will now have to send his "new" data using a true feedback strategy rather than just a "$p_{X_2|V}(\cdot|\cdot)$" box. This is depicted in Figure 6.7. The box labelled "Strategy 2" sends the U_2 sequence by assuming that the receiver has decoded the U_V sequence and formed the V and X_1 sequences from U_V. Since the receiver knows V and X_1, the information rate of the resulting U_2-to-Y channel is simply $I(X_2; Y|X_1V)$, as required.

The Binary Switching Multiple-Access Channel

The Binary Switching Multiple-Access Channel (BS-MAC) [65] is an example of a channel with $H(X_2|X_1Y) = 0$ but $H(X_1|X_2Y) > 0$

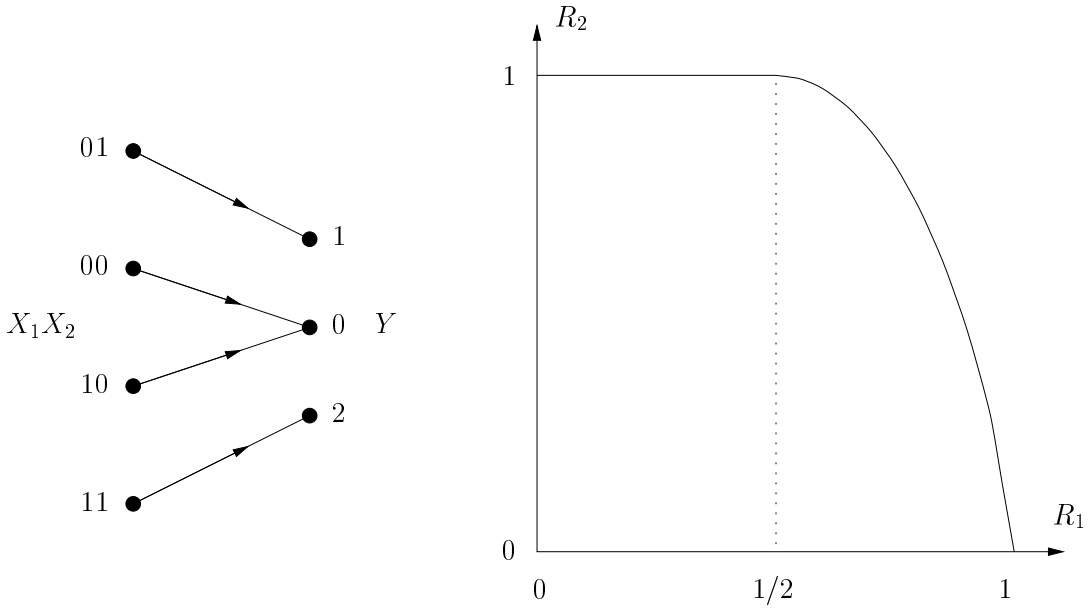


Figure 6.8: *The two-user binary switching multiple access channel and its capacity region with and without feedback.*

(we have exchanged the roles of X_1 and X_2 here to conform with [65] and [75]). Both users' input alphabets are $\{0, 1\}$ and the output is $Y = (X_1 + X_2) \cdot X_2$ (see Figure 6.8). The capacity region $\mathcal{C}_{\rm FB}$ of this channel is the same with or without feedback and is the set of rate pairs (R_1, R_2) for which [65]

$$0 \le R_2 \le = \begin{cases} 1 & \text{for } 0 \le R_1 \le 1/2, \\ h(R_1) & \text{for } 1/2 < R_1 \le 1. \end{cases} \tag{6.16}$$

A plot of the capacity region is given in Figure 6.8. In fact, the zero-error capacity region of the BS-MAC (with and without feedback) also coincides with $\mathcal{C}_{\rm FB}$ [75].

The strategy of Figure 6.7, with the roles of User 1 and 2 reversed, is particularly simple for this channel. User 2 codes as in Figure 6.1 by converting his information bit stream U_2 into a sequence $\Delta_2 = X_2$ of independent identically distributed bits where $\Pr(X_{2n} = 0) = p$, all n. User 1 then sees a Binary Erasure Channel (BEC) with erasure probability p, and makes Strategy 2 a selective-repeat ARQ strategy to approach the capacity $1 - p$ of the BEC. Strategy 3 is not used at all because $H(X_2|Y) = 0$. The approachable rate region is thus the convex hull of the points $(0, 0)$, $(0, 1)$ and those (R_1, R_2) for which $0 \le p \le 1/2$. This coincides with the capacity region $\mathcal{C}_{\rm FB}$. This example demonstrates

the importance of making Strategy 2 a *feedback strategy*, and not simply a "$p_{X_2|V}(\cdot|\cdot)$" box as in Figure 6.4.

Remarks

The MACs for which $H(X_1|X_2Y) = 0$ or $H(X_2|X_1Y) = 0$ are not the only discrete channels for which the capacity region is known to be the rate region of Cover and Leung. For example, it is known that $\mathcal{R}^{CL}$ is the capacity region of the channels considered in this chapter but with feedback to one user only [32, 60, 61]. Another class of channels for which $\mathcal{R}^{CL}$ is the capacity region was found by Hekstra and Willems [24, Section VII].

It does not seem that the strategies presented here can be extended to achieve all points in the capacity region of these other channels. The difficulty is that the structure of Figure 6.7 relies heavily on User 2 being able to form the U_V sequence in a simple manner, i.e., without *decoding* a code. On the other hand, the only nontrivial continuous MAC for which the feedback capacity region is known is the additive white Gaussian noise MAC [31] and there is a simple feedback strategy which approaches all points in the capacity region of this channel. Whether there are simple capacity-approaching feedback strategies for all MACs with feedback remains an open question.

6.A Appendix: Approaching Capacity with a Single Mode of Operation

The four modes of Figure 6.6 can be combined into a single mode by increasing the cardinality of the random variable V as follows. Given $p_{VX_1X_2Y}$, define the four random variables

$$\begin{aligned} V^{(1)} &= V \\ V^{(2)} &= [V\ X_1\ X_2] \\ V^{(3)} &= [V\ X_1] \\ V^{(4)} &= [V\ X_2]. \end{aligned}$$

We replace V with the random variable $\tilde{V}$ where $\tilde{V} = V^{(m)}$ with probability t_m, $m = 1, 2, 3, 4$, and $\sum_{m=1}^{4} t_m = 1$. Note that $p_{X_1X_2|V^{(m)}} = p_{X_1|V^{(m)}} \cdot p_{X_2|V^{(m)}}$ for every $V^{(m)}$, so that $p_{X_1X_2|\tilde{V}} = p_{X_1|\tilde{V}} \cdot p_{X_2|\tilde{V}}$ as required by (6.4).

We next show that $V^{(m)}$ corresponds to Mode m in Figure 6.6. Using $\tilde{V}$ rather than V does not change $p_{X_1X_2Y}$, so that the sum-rate bound $I(X_1X_2;Y)$ is not changed. If $\tilde{V} = V^{(2)}$ then $I(X_1;Y|X_2\tilde{V}) = 0$ and $I(X_2;Y|X_1\tilde{V}) = 0$. Similarly, if $\tilde{V} = V^{(3)}$ then $I(X_1;Y|X_2\tilde{V}) = 0$ and $I(X_2;Y|X_1\tilde{V}) = I(X_2;Y|X_1V)$, and if $\tilde{V} = V^{(4)}$ then $I(X_1;Y|X_2\tilde{V}) = I(X_1;Y|X_2V)$ and $I(X_2;Y|X_1\tilde{V}) = 0$. Thus, the bounds (6.1) and (6.2) with V replaced by $\tilde{V}$ define a rectangle whose maximum sum-rate point is at

$$(R_1, R_2) = ((t_1 + t_4)I_1, (t_1 + t_3)I_2) \tag{6.17}$$

where $I_1 = I(X_1;Y|X_2V)$ and $I_2 = I(X_2;Y|X_1V)$. One can make the rate pair in (6.17) any rate point inside the pentagon of Figure 6.5 by choosing the "time-sharing" probabilities t_m, $m = 1, 2, 3, 4$ appropriately. Because the sum-rate bound (6.3) is superfluous for these rate points, one can achieve them with Mode 1 in Figure 6.6.

Chapter 7

Summary and Concluding Remarks

We summarize the contributions of this work and then conclude with some remarks.

Chapter 2:

- We demonstrated that d-separation implies conditional independence in functional dependence graphs with *arbitrary* random variables.
- The proof of this result showed that d-separation is equivalent to making the functional dependence graph a subgraph of the graph $\mathcal{G}^*$ in Figure 2.4.
- A weaker condition than d-separation was introduced and shown to establish conditional independence in functional dependence graphs. This condition was called fd-separation for "functional dependence" separation.

Chapter 3:

- Information-theoretic quantities for dealing with causality were

introduced: *causally conditioned uncertainty* and *causally conditioned directed information*. The definitions were based on earlier definitions given by Marko [37] and Massey [36].

- Bounds, chain rules and stationarity properties of the newly defined quantities were derived.

Chapter 4:

- An explicit characterization of the capacity region of the two-way channel in terms of causally conditioned directed information was derived. The derivation was based on Shannon's coding technique [17].
- It was shown how to simplify the characterization of the capacity region for the common-output two-way channel. A byproduct of the simplification is that simpler coding distributions than for the general two-way channel may be used.
- Concatenated adaptive codewords were used to describe and generalize the coding technique and rate region of Han [22].

Chapter 5:

- An explicit characterization of the capacity region of the multiple-access channel with feedback (MAC-FB) in terms of causally conditioned directed information was derived.
- A generalization of the Cover-Leung rate region [29] was derived.
- First examples of enlargements of the Cover-Leung rate region for discrete MAC-FBs were given. The channels used were noisy binary adder channels.

Chapter 6:

- Feedback strategies for a class of two-user MAC-FBs were designed.
- The strategies were shown to approach all rate points in the capacity region of the class of MAC-FBs. Moreover, it was shown that the strategies do not require time sharing.

Concluding Remarks

The notion that information possesses direction seems natural in most communication scenarios. This dissertation has demonstrated that the directed information definitions of Marko and Massey not only have direction, but in fact lead one to the precise quantities required for determining the capacity region of at least two multi-user channels. There are certainly other channels for which directed information is the quantity one is looking for. For example, the results of Chapters 4 and 5 can be generalized to noisy feedback channels of the sort dealt with in [32]. Whether directed information is *the* quantity for other channels with feedback, or whether one must introduce quantities other than causally conditioned uncertainty, remains to be seen.

The methods we have introduced in Chapters 3, 4 and 5 could be described as a "brute-force" attack. This is especially apparent when one attempts to use the derived capacity expressions for finding *outer* bounds on the capacity regions of the two-way channel and the multiple-access channel with feedback. We certainly have not come closer to finding a *single-letter* characterization [76, pages 29, 35, 39] of the capacity regions of these channels, if this is at all possible. The calculation of rate points on the boundary of the capacity region of the two-way channel and the multiple-access channel with feedback thus remain open problems.

There are several other problems related to this work which are interesting to pursue.

- In Chapter 2, the d-separation and fd-separation criteria prove conditional independence, i.e., that the information $I(A;B|C)$ is exactly zero. In cases where d-separation or fd-separation do not hold, it would be useful to have a graphical technique that can instead guarantee a result such as $I(A;B|C) < \epsilon$.

- In Chapter 3, several properties of directed information were established. A property not dealt with there is the *convexity* of directed information. More knowledge about convexity could lead to simple techniques for optimizing directed information rates. For example, a modification of the iterative algorithm of Arimoto and Blahut [48, page 366] might exist that can find the best Lth inner bound rates of Lemmas 4.2, 5.2 and 5.4.

- The structure of the strategies in Chapter 6 suggests a new random coding approach for multiple-access channels with feedback: perhaps one can use data compression techniques to obtain rate points outside the Cover-Leung region. Of course, it would be even better if simple *strategies* achieving such rate points could be found.

- We have dealt only with channels for which a *regular*, *high rate* and *noiseless* feedback link is available. The analysis of the capacity region and the design of feedback strategies for channels with *irregular* and *low rate* feedback is especially practically relevant.

Bibliography

[1] C. Shannon, "The zero error capacity of a noisy channel," *IRE Trans. Inform. Theory*, vol. 2, pp. 221–238, September 1956, Reprinted in *Claude Elwood Shannon: Collected Papers*, pp. 221-238, (N.J.A. Sloane and A.D. Wyner, eds.) Piscataway: IEEE Press, 1993.

[2] R.W. Lucky, "A survey of the communication theory literature: 1968-1973," *IEEE Trans. Inform. Theory*, vol. 19, pp. 725–739, November 1973.

[3] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and October 1948, Reprinted in *Claude Elwood Shannon: Collected Papers*, pp. 5-83, (N.J.A. Sloane and A.D. Wyner, eds.) Piscataway: IEEE Press, 1993.

[4] J.L. Massey, "The role of feedback in random-accessing and multi--user communications," Kreditantrag, ETH Zürich, 1996.

[5] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Inform. Theory*, vol. 9, no. 3, pp. 136–143, July 1963.

[6] J.P.M. Schalkwijk and T. Kailath, "A coding scheme for additive white noise channels with feedback – Part I: no bandwidth constraint," *IEEE Trans. Inform. Theory*, vol. 12, no. 2, pp. 172–182, April 1966.

[7] J.P.M. Schalkwijk, "A coding scheme for additive white noise channels with feedback – Part II: band-limited signals," *IEEE Trans. Inform. Theory*, vol. 12, no. 2, pp. 183–189, April 1966.

[8] R.G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.

[9] K.S. Zigangirov, "Data transmission over a gaussian channel with feedback," *Problemy Peredachi Informatsii*, vol. 3, no. 2, pp. 98–101, 1967.

[10] A.D. Wyner, "On the Schalkwijk-Kailath coding scheme with a peak energy constraint," *IEEE Trans. Inform. Theory*, vol. 14, no. 1, pp. 129–134, Jan. 1968.

[11] S. Butman, "A general formulation of linear feedback communication systems with solutions," *IEEE Trans. Inform. Theory*, vol. 15, no. 3, pp. 392–400, May 1969.

[12] T.M. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Trans. Inform. Theory*, vol. 35, no. 1, pp. 37–43, Jan. 1989.

[13] A. Dembo, "On gaussian feedback capacity," *IEEE Trans. Inform. Theory*, vol. 35, no. 5, pp. 1072–1089, Sept. 1989.

[14] L.H. Ozarow, "Random coding for additive gaussian channels with feedback," *IEEE Trans. Inform. Theory*, vol. 36, no. 1, pp. 17–22, Jan. 1990.

[15] F. Alajaji, "Feedback does not increase the capacity of discrete channels with additive noise," *IEEE Trans. Inform. Theory*, vol. 41, no. 2, pp. 546–549, March 1995.

[16] A.J. Viterbi, "Information theory in the sixties," *IEEE Trans. Inform. Theory*, vol. 19, pp. 257–262, May 1973.

[17] C. Shannon, "Two-way communication channels," in *Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability*, J. Neyman, Ed., Berkeley, CA, 1961, vol. 1, pp. 611–644, Univ. Calif. Press, Reprinted in *Claude Elwood Shannon: Collected Papers*, pp. 351-384, (N.J.A. Sloane and A.D. Wyner, eds.) Piscataway: IEEE Press, 1993.

[18] E.C. van der Meulen, "A survey of multi-way channels in information theory: 1961-1976," *IEEE Trans. Inform. Theory*, vol. 23, no. 1, pp. 1–37, Jan. 1977.

[19] G. Dueck, "The capacity region of the two-way channel can exceed the inner bound," *Inform. Contr.*, vol. 40, pp. 258–266, March 1979.

[20] J.P.M. Schalkwijk, "The binary multiplying channnel – a coding scheme that operates beyond shannon's inner bound region," *IEEE Trans. Inform. Theory*, vol. 28, no. 1, pp. 107–110, Jan. 1982.

[21] J.P.M. Schalkwijk, "On an extension of an achievable rate region for the binary multiplying channnel," *IEEE Trans. Inform. Theory*, vol. 29, no. 3, pp. 445–448, May 1983.

[22] T.S. Han, "A general coding scheme for the two-way channel," *IEEE Trans. Inform. Theory*, vol. 30, no. 1, pp. 35–44, Jan. 1984.

[23] Z. Zhang, T. Berger, and J.P.M. Schalkwijk, "New outer bounds to capacity regions of two-way channels," *IEEE Trans. Inform. Theory*, vol. 32, no. 3, pp. 383–386, May 1986.

[24] A.P. Hekstra and F.M.J. Willems, "Dependence balance bounds for single-output two-way channels," *IEEE Trans. Inform. Theory*, vol. 35, no. 1, pp. 44–53, Jan. 1989.

[25] R. Ahlswede, "Multi-way communication channels," in *Proc. 2nd Int. Symp. Inform. Theory*, Tsahkadsor, Armenian SSR, 1971, pp. 23–52, Publishing house of the Hungarian Academy of Sciences, 1973.

[26] E.C. van der Meulen, "The discrete memoryless channel with two senders and one receiver," in *Proc. 2nd Int. Symp. Inform. Theory*, Tsahkadsor, Armenian S.S.R., 1971, pp. 103–135, Publishing House of the Hungarian Academy of Sciences, 1973.

[27] H. Liao, "A coding theorem for multiple access communications," in *Proceedings of the 1972 IEEE Int. Symp. on Inform. Theory*, Asilomar, 1972.

[28] N.T. Gaarder and J.K. Wolf, "The capacity region of a multiple-access discrete memoryless channel can increase with feedback," *IEEE Trans. Inform. Theory*, vol. 21, no. 1, pp. 100–102, Jan. 1975.

[29] T.M. Cover and C.S.K. Leung, "An achievable rate region for the multiple-access channel with feedback," *IEEE Trans. Inform. Theory*, vol. 27, no. 3, pp. 292–298, May 1981.

[30] F.J. Willems, "The feedback capacity region of a class of discrete memoryless multiple access channels," *IEEE Trans. Inform. Theory*, vol. 28, no. 1, pp. 93–95, Jan. 1982.

[31] L.H. Ozarow, "The capacity of the white gaussian multiple access channel with feedback," *IEEE Trans. Inform. Theory*, vol. 30, no. 4, pp. 623–629, July 1984.

[32] A.B. Carleial, "Multiple-access channels with different generalized feedback signals," *IEEE Trans. Inform. Theory*, vol. 28, no. 6, pp. 841–850, Nov. 1982.

[33] J.A. Thomas, "Feeback can at most double gaussian multiple access channel capacity," *IEEE Trans. Inform. Theory*, vol. 33, no. 5, pp. 711–716, Sept. 1987.

[34] S. Pombra and T.M. Cover, "Non white gaussian multiple access channels with feedback," *IEEE Trans. Inform. Theory*, vol. 40, no. 3, pp. 885–892, May 1994.

[35] E. Ordentlich, "On the factor-of-two bound for Gaussian multiple-access channels with feedback," *IEEE Trans. Inform. Theory*, vol. 42, no. 6, pp. 2231–2235, Nov. 1996.

[36] J.L. Massey, "Causality, feedback and directed information," in *Proceedings of the 1990 IEEE Int. Symp. on Inform. Theory and Its Appl.*, Hawaii, U.S.A., Nov. 27-30 1990, pp. 303–305.

[37] H. Marko, "The bidirectional communication theory – a generalization of information theory," *IEEE Trans. Comm.*, vol. 21, no. 12, pp. 1345–1351, Dec. 1973.

[38] F.J. Willems, "On multiple access channels with feedback," *IEEE Trans. Inform. Theory*, vol. 30, no. 6, pp. 842–845, Nov. 1984.

[39] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, Calif., 1988.

[40] J. Pearl, "Fusion, propagation, and structuring in belief networks," *Artificial Intelligence*, vol. 29, pp. 241–288, 1986, Reprinted in *Great Papers in Computer Science*, pp. 586-627, P. Laplante, Ed., Elsevier, 1996.

[41] S.L. Lauritzen, A.P. Dawid, B.N. Larsen, and H.-G.Leimer, "Independence properties of directed markov fields," *Networks*, vol. 20, pp. 491–505, 1990.

[42] J.L. Massey, "Causal interpretations of random variables," *Problemy Peredachi Informatsii*, vol. 32, no. 1, pp. 112–116, July 1996.

[43] D. Geiger, T. Verma, and J. Pearl, "Identifying independence in bayesian networks," *Networks*, vol. 20, pp. 507–534, 1990.

[44] J. Pearl and R. Dechter, "Identifying independencies in causal graphs with feedback," in *Uncertainty in Artificial Intelligence, Proceedings of the Twelfth Conference*, E. Horvitz and F. Jensen, Eds., San Francisco, Calif., 1996, pp. 420–426, Morgan Kaufmann.

[45] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction and Search*, Springer, New York, 1993.

[46] P. Spirtes, "Directed cyclic graphical representations of feedback models," in *Uncertainty in Artificial Intelligence, Proceedings of the Eleventh Conference*, P. Besnar and S. Hanks, Eds., San Francisco, Calif., 1995, pp. 491–498, Morgan Kaufmann.

[47] J.T.A. Koster, "Markov properties of nonrecursive causal models," *The Annals of Statistics*, vol. 24, no. 5, pp. 2148–2177, Oct. 1996.

[48] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[49] C. Harpes, G. Kramer, J.L. Massey, "A generalization of linear cryptanalysis and the applicability of matsui's piling-up lemma," in *Advances in Cryptology – Eurocrypt '95, Lecture Notes in Computer Science No. 921*, L.C. Guillou and J.-L. Quisquater, Ed. 1995, pp. 24–38, Springer.

[50] M. Matsui, "Linear cryptanalysis method for des cipher," in *Advances in Cryptology – Eurocrypt '93, Lecture Notes in Computer Science No. 765*, T. Helleseth, Ed. 1994, pp. 386–397, Springer.

[51] R.M. Fano, *The Transmission of Information*, The MIT Press and Wiley, New York, 1961.

[52] J.L. Massey, "Applied digital information theory," Course Notes, ETH Zürich, 1994/95.

[53] I.N. Bronshtein and K.A. Semendyayev, *Handbook of Mathematics*, Springer, Berlin, 3 edition, 1997.

[54] R.E. Blahut, *Principles and Practice of Information Theory*, Addison-Wesley, Reading, Massachusetts, 1987.

[55] R.G. Gallager, *Discrete Stochastic Processes*, Kluwer, Boston, 1996.

[56] S. Lin and D.J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*, Prentice Hall, New Jersey, 1983.

[57] D. Slepian and J.K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. 19, pp. 471–480, July 1973.

[58] H.B. Meeuwissen, "Coding strategies for two-way communication," Research Report, Institute of Continuing Education, Eindhoven University of Technology, 1995.

[59] J.L. Massey, "Information theory of many user communications," Course Notes, ETH Zürich, 1985.

[60] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels*, Academic Press, New York, 1981.

[61] F.J. Willems and E.C. van der Meulen, "Partial feedback for the discrete memoryless multiple access channel," *IEEE Trans. Inform. Theory*, vol. 29, no. 2, pp. 287–290, March 1983.

[62] R.G. Gallager, "A perspective on multiaccess channels," *IEEE Trans. Inform. Theory*, vol. 31, pp. 124–142, March 1985.

[63] J.M. Wozencraft and I.M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965.

[64] A.J. Vinck, "Constructive superposition coding for the binary erasure multiple access channel," in *Proceedings of the Fourth Symp. on Inform. Theory in the Benelux*, E.C. van der Meulen, Ed., Haasrode, Belgium, May 26-27 1983, pp. 179–188.

[65] A.J. Vinck, "On the multiple access channel," in *Proceedings of the Second Joint Swedish-Soviet Int. Workshop on Inform. Theory*, Gränna, Sweden, April 14-19 1985, pp. 24–29.

[66] A.J. Vinck, W.L.M. Hoeks, and K.A. Post, "On the capacity of the two-user M-ary multiple-access channel with feedback," *IEEE Trans. Inform. Theory*, vol. 31, no. 4, pp. 540–543, July 1985.

[67] J.P.M. Schalkwijk, "A class of simple and optimal strategies for block coding on the binary symmetric channel with noiseless feedback," *IEEE Trans. Inform. Theory*, vol. 17, no. 3, pp. 283–287, May 1971.

[68] G. Kramer, "A sequential strategy for the two-user noiseless binary adder channel with feedback," in *Proceedings of the 1997 IEEE Int. Symp. on Inform. Theory*, Ulm, Germany, June 29 - July 4 1997, p. 131.

[69] K.S. Zigangirov, "Information transmission over a binary symmetric channel with noiseless feedback (random transmission time)," *Problemy Peredachi Informatsii*, vol. 4, no. 3, pp. 38–47, 1968.

[70] M.V. Burnashev and K.S. Zigangirov, "One problem of observation control," *Problemy Peredachi Informatsii*, vol. 11, no. 3, pp. 44–52, July-Sept. 1975.

[71] E.J. Weldon, "An upper bound for the probability of a word error for block codes used with the memoryless binary symmetric channel with noiseless information feedback," Technical Memorandum 62-5314-4, Bell Telephone Laboratories, August 1962.

[72] J.M. Ooi and G.W. Wornell, "Fast iterative coding for feedback channels," in *Proceedings of the 1997 IEEE Int. Symp. on Inform. Theory*, Ulm, Germany, June 29 - July 4 1997, p. 133.

[73] N. Abramson, *Information Theory and Coding*, McGraw-Hill, New York, 1963.

[74] O. Lalive d'Epinay and M. Giger, "Strategien für Kanäle mit Rückkopplung," Diploma Thesis, Signal and Inform. Proc. Lab., Abteilung IIIB, ETH Zürich, Wintersemester 1996/97.

[75] P. Vanroose, "Code construction for the noiseless binary switching multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 34, no. 5, pp. 1100–1106, Sept. 1988.

[76] T.M. Cover and B. Gopinath, Eds., *Open Problems in Communication and Computation*, Springer, New York, 1987.

Curriculum Vitae

1970	Born in Winnipeg, Manitoba, Canada on April 8.
1976-87	Elementary, Junior High and High School in Winnipeg.
1987-91	Bachelor of Science in Electrical Engineering at the University of Manitoba, Winnipeg.
1991-92	Master of Science in Electrical Engineering at the University of Manitoba.
1992-96	Post-Diploma in Information Technology at the ETH Zürich, Switzerland.
1996-98	Doctor of Technical Sciences at the ETH Zürich.