On Source Coding for Networks

Thesis by
Michael Fleming

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California
2004
(Defended May 10, 2004)

To God

and

my family

Acknowledgments

I would like to express my immense gratitude to my advisor, Professor Michelle Effros, whose guidance and mentoring made this work possible.

I would like to thank the members of my candidacy and defense committees, Professors Jehoshua Bruck, Michelle Effros, Gary Lorden, Babak Hassibi, Steven Low, and Robert McEliece, who read my thesis manuscript and gave me valuable feedback on my work.

My labmates Diego Dugatkin, Hanying Feng, Qian Zhao, Wei-Hsin Gu, and especially Sidharth Jaggi, helped me with technical details and provided much-appreciated company during long writing sessions. I am grateful also to Jason Colwell and Professor Sze Tan for their advice on optimization and estimation theory.

I would like to thank my family and friends for their constant support throughout my time at Caltech.

Finally, I am grateful to the agencies that funded my research. This work was supported by the F.W.W. Rhodes Memorial Scholarship and a Redshaw Award, the William H. Pickering Fellowship, NSF Grant Nos. MIP-9501977, CCR-9909026, and CCR-0220039, the Lee Center for Advanced Networking at Caltech, the Intel Technology for Education 2000 program, and a grant from the Charles Lee Powell Foundation.

Abstract

In this thesis, I examine both applied and theoretical issues in network source coding.

The applied results focus on the construction of locally rate-distortion-optimal vector quantizers for networks. I extend an existing vector quantizer design algorithm for arbitrary network topologies [1] to allow for the use of side information at the decoder and for the presence of channel errors. I show how to implement the algorithm and use it to design codes for several different systems. The implementation treats both fixed-rate and variable-rate quantizer design and includes a discussion of convergence and complexity. Experimental results for several different systems demonstrate in practice some of the potential performance benefits (in terms of rate, distortion, and functionality) of incorporating a network's topology into the design of its data compression system.

The theoretical work covers several topics. Firstly, for a system with some side information known at both the encoder and the decoder, and some known only at the decoder, I derive the rate-distortion function and evaluate it for binary symmetric and Gaussian sources. I then apply the results for binary sources in evaluating the binary symmetric rate-distortion function for a system where the presence of side information at the decoder is unreliable. Previously, only upper and lower bounds were known for that problem. Secondly, I address

with an example the question of whether feedback from a decoder to an encoder ever enlarges the achievable rate region for lossless network source coding of memoryless sources. Thirdly, I show how cutset methods can yield quick and simple rate-distortion converses for any source coding network. Finally, I present rate-distortion results for two different broadcast source coding systems.

Contents

Acknowledgments **iv**

Abstract **vi**

1 Introduction **1**

2 Background **10**

2.1 Jointly Typical Sequences 10

2.2 Existing Rate-Distortion Results 11

2.3 Vector Quantization 25

3 Network Vector Quantizer Design **27**

3.1 Introduction 27

3.2 Network Description 30

3.3 Locally Optimal NVQ Design 37

3.4 Implementation 45

3.5 Experimental Results 57

3.6 Summary 69

4 Rate-Distortion with Mixed Side Information **71**

4.1 Introduction 71

4.2 $R(D)$ for the Mixed Side Information System 73

4.3 Joint Gaussian Sources 80

4.4 Joint Binary Sources 82

4.5 Heegard and Berger's System 88

4.6 Summary 93

5 Network Source Coding Results **95**

5.1 Feedback in Lossless Coding 95

5.2 Source Coding Converses Via Cutsets 101

5.3 Broadcast Source Coding 109

5.4 Summary 124

6 Conclusions **126**

A The Satellite Weather Image Data Set **130**

B The Binary MSI Example **132**

C Differentiability of $K(a_u)$ **136**

D The Binary HB Example **139**

E Proofs of Lemmas and Theorems **142**

Bibliography **160**

List of Figures

1.1 (a) A point-to-point network. (b) A side information network. (c) A two-user MA network. (d) A two-receiver MR network. (e) A two-channel MD network. (f) A two-receiver BC network. 5

1.2 The relationship between the systems. 6

1.3 (a) The mixed side information system considered in Chapter 4. (b) The system of Heegard and Berger [2] and Kaspi [3]. (c) The three-encoder, three-decoder system considered in Chapter 5. 8

2.1 (a) The conditional rate-distortion network [4]. (b) The network of Berger and Yeung [5]. (c) The network of Kaspi and Berger [6]. 14

3.1 (a) A point-to-point code. (b) A two-receiver BC code. (c) A two-user MA code. (d) A WZ code with side information Z_1. (e) A two-channel MD code. The notation is $X_{t,S}$ for a source and $\hat{X}_{t,S,r}$ for a reproduction; here t is the transmitter, S the set of receivers, and r the reproducer. 32

3.2 A general three-node network. 33

3.3 The 2AWZ network. 44

3.4 A discrete Markov constraint example. (a) The distribution P, uniform over an ellipse. (b,c,d) $\hat{P}$ for $|\mathcal{K}_{2,1}| = |\mathcal{K}_{3,1}| = K$ when: (b) $K = 2^1$, (c) $K = 2^3$, (d) $K = 2^5$. 52

3.5 Optimal encoding at node 2. (a) The estimated performance for a given index set $i_{2,*}$ can be found by summing the performance in two linked 2AWZ subsystems. (b),(c) The two subsystems. 56

3.6 (a) WZVQ performance as a function of side information rate $\log_2(|\mathcal{K}_Z|)/n$ for jointly Gaussian source and side information ($\rho = 0.375$). (b) Various coding performances for the 2AWZ system on satellite data. 59

3.7 (a) Comparison of fixed- and variable-rate coding performances for the 2A system. (b) WZ code performance as a function of source and side information correlation and the number of cosets used in decoder initialization. 61

3.8 Efficiency of network source coding vs. independent coding. (a) Overall performance as a function of correlation. (b),(c) Weighted sum distortion at each node as a function of the Lagrangian parameters (shown from two different angles). 63

3.9 (a) MDVQ performance on the satellite data as a function of the number of descriptions per vector and the channel failure probability. (b) Fixed-rate and entropy-constrained MDVQ performance on Gaussian data compared to the D-R bound. 65

3.10 (a) The network messages and side informations of a six-node ring network. (b) Design time for a ring network. 67

4.1 (a) The conditional rate-distortion system. (b) The Wyner-Ziv system. (c) The MSI system. (d) Heegard and Berger's system. 73

4.2 A distribution achieving $R_{X|YZ}(p, D)$. Here $a = 1 - Dl_{11}$ and N is Gaussian noise, independent of (X, Y, Z), with mean zero and variance $D/(1 - Dl_{11})$. . 82

4.3 Joint distribution of (X, Y, Z) for binary MSI example. 82

4.4 Joint distribution of (W, X, Z) for $|\mathcal{W}| \leq 3$. 83

4.5 The value (top) and the form (bottom) of the optimal solution for different values of p_0 and q_0 when $D = 0.1$. 89

4.6 Numerical results for Heegard and Berger's system, $q_0 = 0.1$, $D_1 = 0.05$. . . . 94

5.1 A bipartite graph considered for the feedback problem. 96

5.2 The lossless coding system used in the proof of Theorem 18. 99

5.3 A general network. The line is an example of a cutset boundary separating the M nodes of the network into two sets, A and B. 103

5.4 The simplified network for the cutset bounds. 105

5.5 The system used for the three-set extension of the cutset approach. 106

5.6 (a) A two-user MA network. (b) A three-node network. 108

5.7 A k-ary broadcast tree. 110

5.8 A broadcast system with three receivers. 118

5.9 A three-encoder, three-decoder lossless coding system. 124

A.1 Sample images from the GOES-8 weather satellite. From left to right: visible spectrum, infrared 2, infrared 5. 131

E.1 A graphical representation of the minimization. (a) The polygon defined by the two planes $n_{12} + n_{13} = I_{12,13}$ and $n_{13} + n_{23} = I_{13,23}$. (b) The polygon defined by all three planes when $0 \leq I_{12,23} \leq I_{13,23} - I_{12,13}$. (c) The polygon defined by all three planes when $I_{13,23} - I_{12,13} \leq I_{12,23} \leq I_{12,13} + I_{13,23}$. 158

List of Tables

1.1 Progress chart for network source coding. 6

3.1 Application of the Markov constraint to a pair of Gaussian sources. 51

3.2 Total number of codewords for various systems. 68

4.1 Possible decoding functions for each symbol, together with their expected distortion contribution. 85

4.2 A possible decoding function f when $|\mathcal{W}| = 3$. 85

A.1 Data source assignments for the NVQ experiments 131

Chapter 1

Introduction

The amount of data transferred over electronic communication networks has increased dramatically in the last three decades. As a result, the motivation for developing efficient source codes for network data compression has never been higher. Yet, despite the growing number of network applications, there are very few source codes in common use that take advantage of the topology of the network in which they operate. Indeed, the vast majority of data compression codes are developed without any consideration whatsoever of the structure of the network; the network is treated as a collection of independent point-to-point links, and a separate code is designed for each link. This "independent" approach to source code design for networks does not achieve the goal of using the network links in the most efficient manner. Although point-to-point codes remove the redundancy in each individual source, they are unable to remove the redundancy between different sources, and are therefore inefficient when applied to statistically dependent sources.

Redundancy between data sources is observed in a growing number of network applications. Video conferencing, sensor networks, and distributed computing all generate multiple,

highly correlated data streams. These applications require a global, "network" approach to code design, one that exploits inter-source dependencies. This observation begs the question of why such an approach is rarely taken in practice. One reason is an incomplete understanding of the magnitude of the potential benefits; rate-distortion results are known for only a few simple networks. Another, perhaps more important reason is a lack of understanding of how to convert existing theory into good practical codes. Indeed, even the definition of "goodness" in networks is not immediately clear. Several factors must be balanced: rates, distortions, design complexity, run-time encoding and decoding complexity, and robustness to changes or failures in the network.

The first part of this thesis investigates the global, "network" approach to source code design using vector quantization (VQ). An algorithm for VQ design for arbitrary networks is outlined in [1], Chapter 2 extends this algorithm and presents the details of its implementation along with experimental results, thereby yielding the first practical data compression codes designed for a general network setting. The definition of goodness adopted in this design is rate-distortion performance; I extend the necessary conditions for rate-distortion optimality of network encoders and decoders from [1] to allow for the presence of both side information and channel errors, and use them as the basis of an iterative design algorithm functionally equivalent to the generalized-Lloyd algorithm. I show that convergence of the algorithm to a local optimum is guaranteed for some systems; for others, approximations required for practical implementation remove this guarantee, although I do observe convergence in all of my experimental work. When necessary, I show how to modify the algorithm to ensure convergence for all systems; however, the modification trades away some rate-distortion performance to gain this guarantee of convergence. The design equations treat

both fixed- and variable-rate quantizer design and take into account the presence of channel errors. However, the implementation of variable-rate design is currently limited due to the lack of lossless entropy codes for general networks. The range of systems for which optimal variable-rate quantizers can be designed will expand as the field of lossless network source coding develops. I design VQs for several different systems and evaluate their performance compared to both an independent coding approach and to rate-distortion bounds. The experimental results demonstrate that, for networks with correlated sources, incorporating the network's topology into the design of its data compression system yields significant increases in performance with respect to an independent coding approach. This work also appears in [7, 8, 9].

Development of more efficient practical codes is supported by a more thorough understanding of source coding theory, which is the focus of the second part of this thesis. Table 1.1 summarizes our current knowledge of source coding theory for the basic network classes illustrated in Figure 1.1. In this table, a citation in one of the table cells indicates that we know the region of achievable rates (for lossless coding) or the rate-distortion function (for lossy coding). Partial knowledge of the rate-distortion function via an achievability (ach) or converse (con) result is as indicated. An 'x' denotes that the area remains largely open. The various systems mentioned are introduced briefly below.

- **Point-to-point [10]:** A point-to-point network, shown in Figure 1.1(a), is the simplest communication network. A single encoder transmits information to a single decoder.

- **Side information[1] (SI) [11, 12, 13]:** A SI network, shown in Figure 1.1(b), is a

[1]My use of the term "side information system" refers specifically to the system with side information

point-to-point communication network in which side information is available at the decoder. In the context of lossy coding, I refer to this system also as the Wyner-Ziv (WZ) system.

- **Multiple Access (MA) [11]:** In an MA network, shown in Figure 1.1(c), two or more senders transmit information to a single receiver.

- **Multi-Resolution (MR) [14, 15]:** MR codes, shown in Figure 1.1(d) generate an embedded source description for two or more receivers. Receiver i receives only the first fraction f_i of the description, where $f_1 < f_2 < \ldots$.

- **Multiple Description (MD) [16, 17, 18]:** An MD code can be used for point-to-point communication over multiple, unreliable communication channels (or over a lossy, packet-based channel in which lost packets cannot be retransmitted). Each channel's source description may be lost, and the decoder reproduces the source by combining all received descriptions. In Figure 1.1(e), we model a two-channel system and represent the different decoding scenarios with three separate decoders.

- **Broadcast (BC) [19, 20]:** In a BC network, shown in Figure 1.1(f), a single sender describes a collection of sources to two or more receivers. A different message can be transmitted to each possible subset of the receivers.

The relationships between the various systems are shown in Figure 1.2. Point-to-point networks are special cases of both MA and BC networks. SI networks are considered as special

available only at the decoder. When side information is available to both encoder and decoder, I use the term "conditional side information system".

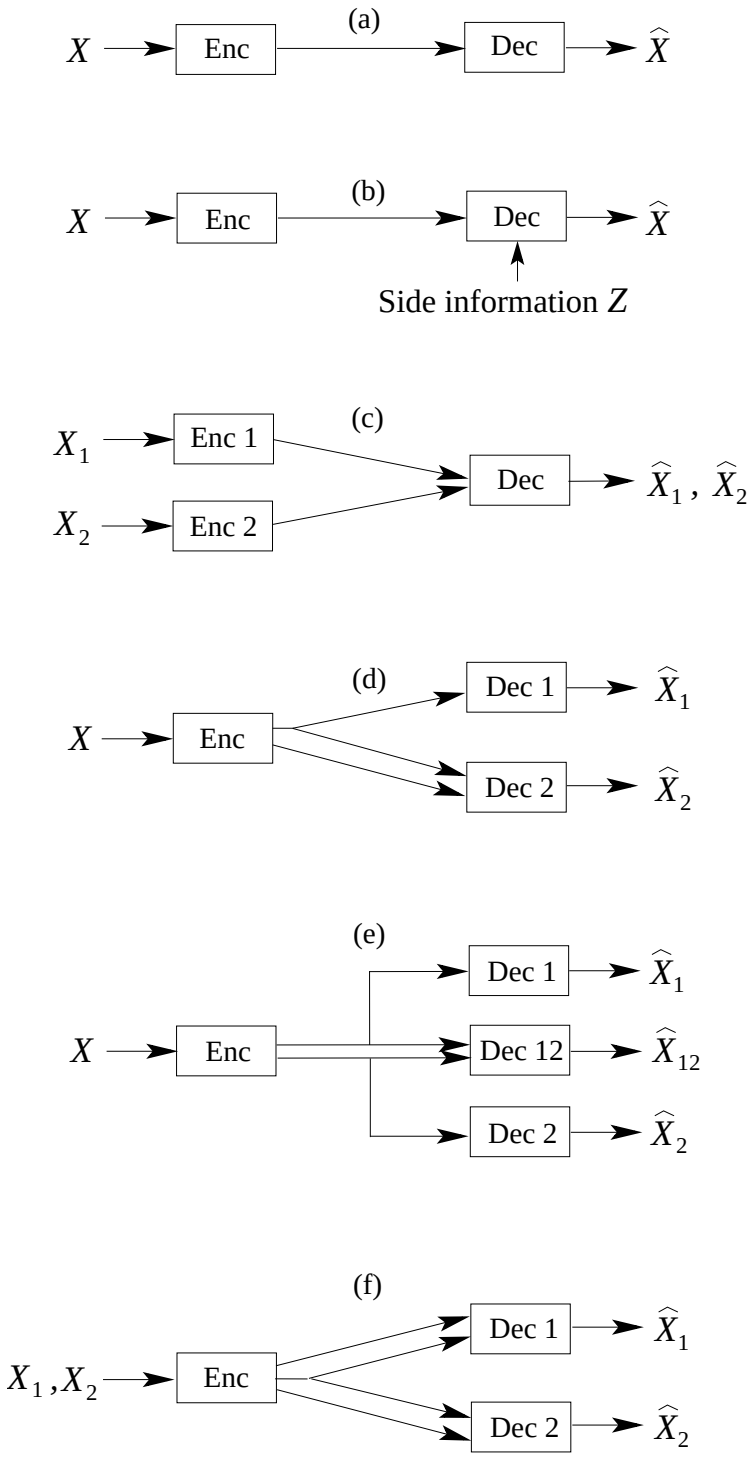


Figure 1.1: (a) A point-to-point network. (b) A side information network. (c) A two-user MA network. (d) A two-receiver MR network. (e) A two-channel MD network. (f) A two-receiver BC network.

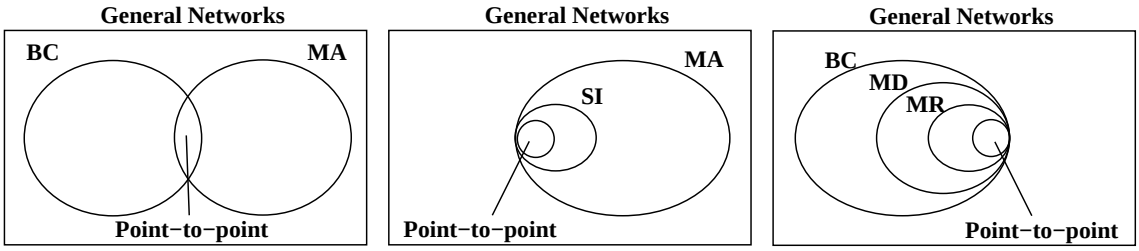


Figure 1.2: The relationship between the systems.

Network	Lossless theory	Lossy theory
Point-to-point	Shannon [10]	Shannon [10]
Side-information	Slepian & Wolf [11]	Wyner & Ziv [12, 13]
Multiple Access	Slepian & Wolf [11]	Tung & Berger Ach/Con:[21, 22]
Multi-Resolution	N/A	Rimoldi [23]
Multiple Description	N/A	Various Ach:[18, 24, 25] Con:[26]
2-receiver Broadcast	Gray & Wyner [19]	Gray & Wyner [19]
M-receiver Broadcast	x	x
General network	Han & Kobayashi Ach/Con:[27]	x

Table 1.1: Progress chart for network source coding.

cases of MA networks for which one source (the side information) has no rate constraint. MR and MD networks are both special cases of the M-receiver BC network.

Consider now the specific example of an environmental remote sensing network with several sensors, each of which takes measurements and transmits them to a central base station, which also makes its own measurements. In encoding its transmission to the base station, each sensor can consider the measurements taken by the base station as side information available to the base station's decoder. If the system uses multi-hop transmissions, then measurements relayed by a sensor act as side information available both to that sensor's encoder and the base station's decoder. Motivated by this framework, I begin the second part of the thesis in Chapter 4 by deriving rate-distortion results for two systems using side information. First, for the system shown in Figure 1.3(a) with some side information known at both the encoder and the decoder, and some known only at the decoder, I derive the rate-distortion function and evaluate it for binary symmetric and Gaussian sources. I then apply the results for the binary source to a second network, shown in Figure 1.3(b), which models point-to-point communication when the presence of side information at the decoder is unreliable [2, 3]. I demonstrate how to evaluate the binary rate-distortion function for that network, closing the gap between previous bounds [2, 28] on its value. The form of the binary rate-distortion function for this second system exhibits an interesting behavior akin to successive refinement, but with side information available to the refining decoder. This work also appears in [29, 30]

Chapter 5 covers three further topics in network source coding theory. The first is a question arising out of a difference between lossless and lossy MA source coding. In lossy coding, the Wyner-Ziv result [12, 13] demonstrates that, in general, a higher rate is required

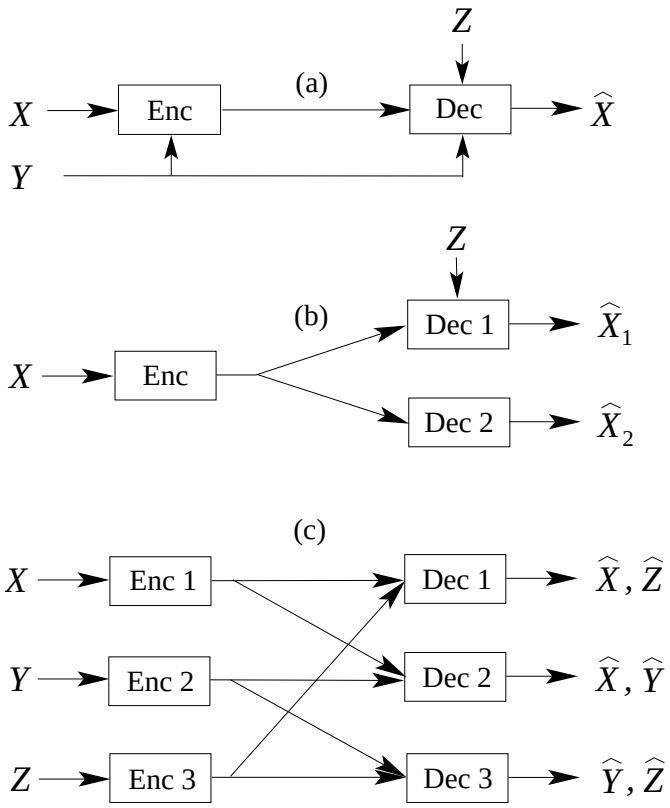


Figure 1.3: (a) The mixed side information system considered in Chapter 4. (b) The system of Heegard and Berger [2] and Kaspi [3]. (c) The three-encoder, three-decoder system considered in Chapter 5.

when an encoder does not have access to all of the messages or side informations available to the decoder. However, the Slepian-Wolf result [11] implies that this result does not carry over to the lossless case. This begs the question of whether feedback from a decoder to an encoder can ever reduce the rate total rate required in lossless coding. I answer this question by means of a detailed example which shows that, as the alphabet size of the sources grows without bound, feedback of a limited rate can reduce by an arbitrary amount the total rate required by the encoders. The second topic is the development of simple rate-distortion converses for networks. Network converses are often difficult to find, but I show that a method based on cutsets yields simple converses for any lossy source coding network. This approach has the advantage of easy applicability but the drawback that the converses are, in many cases, fairly loose. Finally, motivated by the lack of results in general broadcast coding, I look at the rate tradeoff between the three encoders in the three-source three-receiver lossless network of Figure 1.3(c) and apply the outcome to the derivation of an achievability result for three-receiver lossy broadcast source coding. I also derive a rate-distortion result for a tree-structured broadcast coding system.

A summary and discussion of the main contributions concludes the thesis in Chapter 6.

Throughout the thesis, I adopt the notation and definitions from [31] for the basic information theoretic quantities (entropies, mutual informations, etc.). Since the focus is on source coding, all communication channels are assumed to be noiseless unless specified otherwise. The practical work considers strictly stationary information sources; the theoretical work assumes further that the sources are also independent and identically distributed (i.i.d.).

Chapter 2

Background

This chapter defines typical sequences, summarizes several existing results in network rate-distortion theory, and summarizes past work in VQ coding. For the rate-distortion results, I present the definition of the achievable rate-distortion region in detail only for the point-to-point network; a similar definition applies for each of the other networks.

2.1 Jointly Typical Sequences

The rate-distortion proofs in this thesis make use of strongly jointly typical sequences as defined below.

Let $\{X_k\}_{k=1}^K$, with $X_k \in \mathcal{X}_k \ \ \forall k$, denote a finite collection of discrete random variables with some fixed joint distribution $p(x_1, x_2, \ldots, x_K)$. Let S denote an ordered subset of the indices $\{1, 2, \ldots, K\}$ and let $X_S = (X_k : k \in S)$. Denote n independent samples of $X_S \in \mathcal{X}_S$ by $X_S^n = (X_{S,1}, \ldots, X_{S,n})$. For any S and any $a(S) \in \mathcal{X}(S)$, let

$$N(a_S|x_S^n) = \sum_{i=1}^{n} 1_{x_{S,i}=a_S},$$

where $x_{S,i}$ is the ith element of x_S^n and 1_E is the indicator function for event E.

Definition The set $A_\epsilon^{*(n)}$ of ϵ-strongly typical n-sequences is the set of all sequences $(x_1^n, x_2^n, \ldots, x_K^n)$ satisfying for all $S \subseteq \{X_1, X_2, \ldots, X_K\}$ the following two conditions:

1. For all $a_S \in \mathcal{X}_S$ with $p(a_S) > 0$,

$$\left|\frac{1}{n}N(a_S|x_S^n) - p(a_S)\right| \;<\; \frac{\epsilon}{|\mathcal{X}_S|}.$$

2. For all $a_S \in \mathcal{X}_S$ with $p(a_S) = 0$, $N(a_S|x_S^n) = 0$.

The following two lemmas describe useful properties of the strongly typical set.

Lemma 1 *[31, Lemma 13.6.1] Let $X_{S,1}, \ldots, X_{S,n}$ be drawn i.i.d. with distribution $p(x_S)$. Then*

$$\Pr(A_\epsilon^{*(n)}) \to 1 \quad as \quad n \to \infty.$$

Lemma 2 *[31, Lemma 13.6.2] Let $Y_1, \ldots, Y_n$ be drawn i.i.d. with distribution $p(y)$. For $x^n \in A_\epsilon^{*(n)}$, the probability that $(x^n, Y^n) \in A_\epsilon^{*(n)}$ is bounded by*

$$2^{-n(I(X;Y)+\epsilon_1)} \;\leq\; \Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon_1)},$$

where $\epsilon_1 \to 0$ as $\epsilon \to 0$ and $n \to \infty$.

2.2 Existing Rate-Distortion Results

2.2.1 The Point-to-Point Network

Let $X \in \mathcal{X}$ be an independent and identically distributed (i.i.d.) source taking values in source alphabet $\mathcal{X}$ and distributed according to $p(x)$. Let $\hat{\mathcal{X}}$ be a reproduction alphabet,

and let $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ be a measure of the distortion between symbols from the two alphabets.

A $(2^{nR}, n, \Delta_n)$ point-to-point rate-distortion code for source X under distortion measure d is defined by encoder and decoder functions (α^n, β^n) such that

$$\begin{aligned} \alpha^n &: \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\} \\ \beta^n &: \{1, 2, \ldots, 2^{nR}\} \to \hat{\mathcal{X}}^n \\ \Delta_n &= \frac{1}{n} \sum_{i=1}^{n} Ed(X_i, \hat{X}_i), \end{aligned}$$

where $\hat{X}_i$ is the ith component of $\hat{X}^n = \beta^n(\alpha^n(X^n))$ and the expectation is with respect to the source distribution. A rate-distortion pair (R, D) is achievable if there exists a sequence of $(2^{nR}, n, \Delta_n)$ rate-distortion codes (α^n, β^n) with $\lim_{n\to\infty} \Delta_n \leq D$. The rate-distortion region is defined as

$$\mathcal{R} = \overline{\{(R, D) : (R, D) \text{ is achievable}\}},$$

where the overbar denotes set closure. The rate-distortion function is defined as

$$R_X(D) = \inf_D \{R : (R, D) \in \mathcal{R}\}.$$

The following theorem gives an information-theoretic characterization of the rate-distortion function.

Theorem 1 *[10, Section 28]*

$$R_X(D) = \inf_{\hat{X} \in \mathcal{M}_X(D)} I(X; \hat{X}),$$

where $\mathcal{M}_X(D)$ is the closure of the set of all random variables $\hat{X}$ described by a test channel $p(\hat{x}|x)$ such that $Ed(X, \hat{X}) \leq D$.

The infimum above has been evaluated for a variety of sources, including binary and Gaussian sources (see, for example, [31]). For i.i.d. sources, it can be evaluated numerically via a globally-optimal iterative descent algorithm [32, 33].

2.2.2 The Conditional Rate-Distortion Network

When side information $Y \in \mathcal{Y}$ is available to both encoder and decoder, as in Figure 2.1(a), the rate-distortion function is called the conditional rate-distortion function and is given by the following theorem.

Theorem 2 *[4, Pg. 8]*

$$R_{X|Y}(D) = \inf_{\hat{X} \in \mathcal{M}_{X|Y}(D)} I(X; \hat{X}|Y),$$

where $\mathcal{M}_{X|Y}(D)$ is the closure of the set of all random variables $\hat{X}$ described by a test channel $p(\hat{x}|x, y)$ such that $Ed(X, \hat{X}) \leq D$.

The infimum above has been evaluated for jointly Gaussian sources [13].

2.2.3 The Wyner-Ziv Network

When side information $Z \in \mathcal{Z}$ is available to only the decoder and not to the encoder, as in Figure 1.1(b), the rate-distortion function is called the Wyner-Ziv (WZ) rate-distortion function. The following theorem gives its form for both discrete and continuous sources, but for continuous sources the achievability part of the theorem is proven only for distortion measures satisfying the following two conditions:

1. For all $\hat{x} \in \hat{\mathcal{X}}$, $Ed(X, \hat{x}) < \infty$.

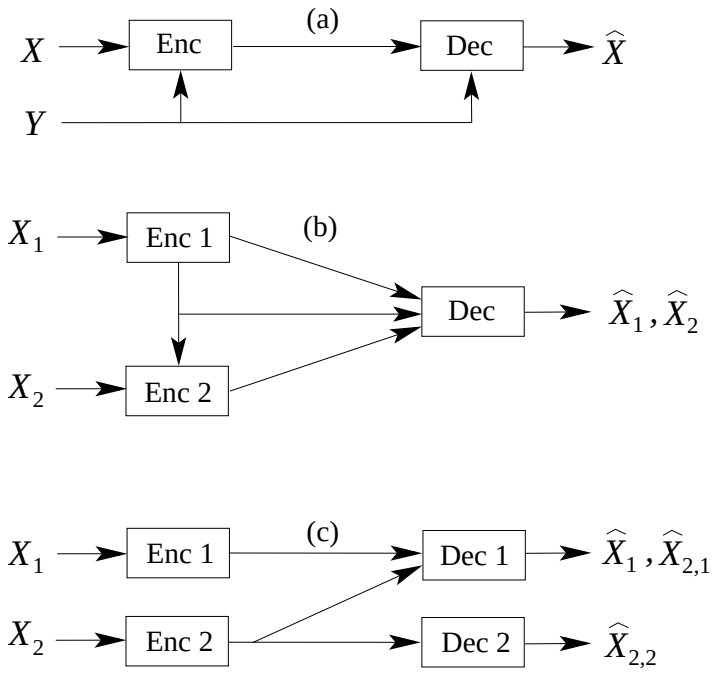


Figure 2.1: (a) The conditional rate-distortion network [4]. (b) The network of Berger and Yeung [5]. (c) The network of Kaspi and Berger [6].

2. For all random variables $\hat{X}$ such that $0 < Ed(X, \hat{X}) < \infty$ and all $\epsilon > 0$, there exists a finite subset $\{\hat{x}_1, \ldots, \hat{x}_N\} \subseteq \hat{\mathcal{X}}$, and a quantizer $f_Q : \hat{\mathcal{X}} \to \{\hat{X}_i\}$ such that $Ed(X, f_Q(\hat{X})) \leq (1 + \epsilon)Ed(X, \hat{X})$.

Condition 2 is a smoothness constraint used in generalizing the WZ rate-distortion proof from discrete to continuous alphabets [13]. Wyner notes that it is not especially restrictive, showing that when $\mathcal{X} = \hat{\mathcal{X}} = \mathbb{R}$ it holds for all r-th power distortion measures, $d(x, \hat{x}) = |x - \hat{x}|^r$ with $r > 0$.

Theorem 3 *[12, Theorem 1], [13, Theorems 2.1,2.2]*

$$\begin{aligned} R_{X|\{Z\}}(D) &= \inf_{W \in \mathcal{M}_{X|\{Z\}}(D)} I(X; W|Z) \\ &= \min_{W \in \mathcal{M}_{X|\{Z\}}(D)} I(X; W) - I(W; Z), \end{aligned}$$

where $\mathcal{M}_{X|\{Z\}}(D)$ *is the closure of the set of all random variables* W *such that* $W \to X \to Z$ *forms a Markov chain and there exists a function* $f : \mathcal{W} \times \mathcal{Z} \to \hat{\mathcal{X}}$ *for which* $Ed(X, f(W, Z)) \leq D$.

The set notation around the Z in the descriptor $R_{X|\{Z\}}(D)$ denotes that the side information is available only at the decoder. Roughly speaking, the Markov chain condition in the statement of Theorem 3 reflects the restriction that the description chosen by the encoder must be based on X alone since the encoder does not have direct access to Z. The WZ rate-distortion function has been evaluated for both binary symmetric [12] and jointly Gaussian [13] sources.

2.2.4 The Multiple-Access Network

In the two-user MA network of Figure 1.1(c), dependent sources X_1 and X_2 are described by two separate encoders to a single decoder. Encoder 1 uses rate R_1 and encoder 2 uses rate R_2. The decoder makes reproductions $\hat{X}_1$ and $\hat{X}_2$ satisfying the distortion constraints $Ed(X_i, \hat{X}_i) \leq D_i$, $i = 1, 2$.

Let $\mathcal{R}_{MA}(D_1, D_2)$ be the closure of the set of all achievable rate vectors for distortions (D_1, D_2). A complete single-letter characterization of $\mathcal{R}_{MA}(D_1, D_2)$ is not available, but achievability and converse results exist. Berger and Tung [34, 21, 22] give the following results. Define $\mathcal{R}_{ach}(D_1, D_2)$ to be the closure of the set of all rate pairs (R_1, R_2) for which there exist auxiliary random variables W_1 and W_2 such that

1.

$$\begin{aligned} R_1 &\geq I(X_1, X_2; W_1|W_2) \\ R_2 &\geq I(X_1, X_2; W_2|W_1) \\ R_1 + R_2 &\geq I(X_1, X_2; W_1, W_2), \end{aligned}$$

2. There exist functions $f_1(W_1, W_2)$ and $f_2(W_1, W_2)$ satisfying

$$Ed(X_i, f_i(W_1, W_2)) \leq D_i, \quad i = 1, 2.$$

3. $W_1 \to X_1 \to X_2 \to W_2$ forms a Markov chain.

Define $\mathcal{R}_{con}(D_1, D_2)$ similarly, but replace the Markov requirement in condition 1 with the requirements that $W_1 \to X_1 \to X_2$ and $X_1 \to X_2 \to W_2$ form Markov chains.

Theorem 4 *[21, Theorem 4.1,5.1], [22, Theorem 6.1,6.2]*

$$\mathcal{R}_{ach}(D_1, D_2) \subseteq \mathcal{R}_{MA}(D_1, D_2) \subseteq \mathcal{R}_{con}(D_1, D_2)$$

Berger and Tung provide the M-user extension of the above and evaluate the two-user region for joint Gaussian sources in [34, 21]. Oohama derives further results for joint Gaussian sources in [35, 36].

Berger and Yeung give a matching achievability and converse for the case $D_1 = 0$ in [37]. In that case, let $\mathcal{R}^*(D_2)$ be the closure of the set of all rate pairs (R_1, R_2) for which there exists an auxiliary random variables W_2 such that

1.

$$\begin{aligned} R_1 &\geq H(X_1|W_2) \\ R_2 &\geq I(X_2; W_2|X) \\ R_1 + R_2 &\geq H(X) + I(X_2; W_2|X). \end{aligned}$$

2. A function $f_2(X, W_2)$ exists satisfying $Ed(X_2, f_2(X, W_2)) \leq D_2$.

3. $X_1 \to X_2 \to W_2$ forms a Markov chain.

4. $|\mathcal{W}_2| \leq |\mathcal{X}_2| + 2$.

Theorem 5 *[37, Theorem 1]*

$$\mathcal{R}_{MA}(0, D_2) = \mathcal{R}^*(D_2)$$

The above region is determined almost completely for jointly symmetric binary sources [37].

Kaspi and Berger [6] look at how inter-encoder communication affects the rate-distortion region by adding a common link of rate R_0 from encoder 1 to encoder 2 and the decoder, as shown in Figure 2.1(b). Their result is the following. Let $\mathcal{R}_{KB}(D_1, D_2)$ be the closure of the set of all achievable rate triples for distortions (D_1, D_2). Let $\mathcal{R}_{ach}$ be the closure of the set of all rate triples for which there exist auxiliary random variables U_1, U_2, and W such that

1.

$$\begin{aligned} R_0 &\geq I(X_2; W|X_1) \\ R_1 &\geq I(X_1; U_1|U_2, W) \\ R_2 &\geq I(X_2; U_2|U_1, W) \\ R_0 + R_1 + R_2 &\geq I(X_1, X_2; U_1, U_2, W). \end{aligned}$$

2. There exist functions $f_1(U_1, U_2, W)$ and $f_2(U_1, U_2, W)$ satisfying $Ed(X_i, f_i(U_1, U_2, W)) \leq D_i$, $i = 1, 2$.

3. $U_1 \to (X_1, W) \to (X_2, W) \to U_2$ and $X_1 \to X_2 \to W$ form Markov chains.

4. $|\mathcal{U}| \leq |\mathcal{X}_1||\mathcal{X}_2| + 6|\mathcal{X}_1| + 5$, $|\mathcal{U}_2| \leq |\mathcal{X}_2|^2 + 6|\mathcal{X}_2| + 5$, $|\mathcal{W}| \leq |\mathcal{X}_2| + 6$.

Theorem 6 *[6, Theorems 2.1 - 2.5]*

$$\mathcal{R}_{ach}(D_1, D_2) \subseteq \mathcal{R}_{KB}(D_1, D_2),$$

with equality in the following (not necessarily exhaustive) cases:

(1) $D_1 = 0$. *(2)* $D_2 = 0$. *(3)* $R_2 = 0$. *(4)* $R_0 > H(X_2|X_1)$.

2.2.5 Multiple Access with Encoder Breakdown

Consider the case when $D_1 = 0$, but there exists a chance that the encoder for X_1 may break down. This can be modeled by the system of Figure 2.1(c) using two decoders. Decoder 1 corresponds to the case when the encoder of X_1 does not break down; decoder corresponds to the case when it does. Here X_1 is described at rate R_1 to decoder 1 and X_2 is described at common rate R_2 to both decoders. X_1 must be reproduced losslessly at decoder 1 and X_2 must be reproduced to distortion D_{21} at decoder 1 and distortion D_{22} at decoder 2. Berger and Yeung [5] give the following characterization of the rate-distortion region. Let $\mathcal{R}_{MAEB}(D_{21}, D_{22})$ be the closure of the set of all achievable rate vectors for distortions (D_{21}, D_{22}). Let $\mathcal{R}^*(D_{21}, D_{22})$ be the closure of the set of all rate pairs (R_1, R_2) for which there exist auxiliary random variables U and V such that

1.
$$\begin{aligned} R_1 &\geq H(X_1|U) \\ R_2 &\geq I(X_2;U) + I(X_2;V|X_1,U). \end{aligned}$$

2. There exist functions $\hat{X}_{21} = f_1(X_1, U, V)$ and $\hat{X}_{22} = f_2(U)$ satisfying

$$\begin{aligned} Ed(X_2, f_1(X,U,V)) &\leq D_{21} \\ Ed(X_2, f_2(U)) &\leq D_{22}. \end{aligned}$$

3. $X_1 \to X_2 \to (U, V)$ forms a Markov chain.

4. $|\mathcal{U}| \leq |\mathcal{X}_2| + 3$, $|\mathcal{V}| \leq |\mathcal{X}_2| + 3$.

Theorem 7 *[5, Theorem 1]*

$$\mathcal{R}_{MAEB}(D_{21}, D_{22}) = \mathcal{R}^*(D_{21}, D_{22})$$

When $D_{21} = 0$, Berger and Yeung evaluate this region for jointly symmetric binary sources [5].

2.2.6 The Multi-Resolution Network

Figure 1.1(d) shows two-receiver MR coding. A single source X is described to two decoders. Both decoders receive a common description at rate R_1, and decoder 2 receives an additional private description at rate R_2. Decoders 1 and 2 make reproductions $\hat{X}_1$ and $\hat{X}_2$ at distortions D_1 and $D_2 \leq D_1$, respectively. The rate-distortion region is given by Rimoldi [23]. Let $\mathcal{R}_{MR}(D_1, D_2)$ be the closure of the set of all achievable rate vectors for distortions (D_1, D_2), and let $\mathcal{R}^*(D_1, D_2)$ be the closure of the set of all rates for which there exist $\hat{X}_1$ and $\hat{X}_2$ such that

$$\begin{aligned} R_1 &\geq I(X; \hat{X}_1) \\ R_2 &\geq I(X; \hat{X}_2|\hat{X}_1) \\ Ed(X, \hat{X}_1) &\leq D_1 \\ Ed(X, \hat{X}_2) &\leq D_2. \end{aligned}$$

Theorem 8 *[23, Theorem 1]*

$$\mathcal{R}_{MR}(D_1, D_2) = \mathcal{R}^*(D_1, D_2).$$

An M-receiver result is also given in [23]. The rate distortion region has been evaluated for several sources; for more details see the discussion on successive refinement in Section 2.2.10.

2.2.7 The Multiple Description Network

Figure 1.1(e) shows two-user MD coding. A single source is described on two different channels at rates R_1 and R_2, respectively. Decoder 1 receives only the channel 1 description and makes reproduction $\hat{X}_1$; decoder 2 receives only the channel 2 description and

makes reproduction $\hat{X}_2$; decoder 12 receives both descriptions and makes reproduction $\hat{X}_{12}$. Let $\mathcal{R}_{MD}(D_1, D_2, D_{12})$ be the closure of the set of all achievable rate vectors for distortions (D_1, D_2, D_{12}). An exact single letter characterization of $\mathcal{R}_{MD}(D_1, D_2, D_{12})$ remains unknown, but achievability and converse results exist. El Gamal and Cover [18] give the following achievability result. Let $\mathcal{R}_{ach}(D_1, D_2, D_{12})$ be the closure of the set of all rates (R_1, R_2) for which there exist $\hat{X}_1$, $\hat{X}_2$, and $\hat{X}_{12}$ such that

$$\begin{aligned} R_1 &\geq I(X; \hat{X}_1) \\ R_2 &\geq I(X; \hat{X}_2) \\ R_1 + R_2 &\geq I(X; \hat{X}_{12}, \hat{X}_1, \hat{X}_2) + I(\hat{X}_1; \hat{X}_2) \\ Ed(X, \hat{X}_t) &\leq D_t, \quad t \in \{1, 2, 12\}. \end{aligned}$$

Theorem 9 *[18, Theorem 1]*

$$\mathcal{R}_{ach}(D_1, D_2, D_{12}) \subseteq \mathcal{R}_{MD}(D_1, D_2, D_{12}).$$

Ahlswede shows that this achievability result is tight for the special case of multiple description coding in which $R_1 + R_2 = R_X(D_{12})$ [38], and Ozarow shows that it is tight for Gaussian sources [17]. However, Zhang and Berger show it to be loose in general [24]. They provide both a counterexample to its tightness and the following new achievable region. Redefine $\mathcal{R}_{ach}(D_1, D_2, D_{12})$ to be the closure of the set of all rates (R_1, R_2) for which there exist auxiliary random variables $\hat{X}_0$, $\hat{X}_1$, $\hat{X}_2$ such that

1.

$$\begin{aligned} R_i &\geq I(X; \hat{X}_0, \hat{X}_i), \quad i = 1, 2 \\ R_1 + R_2 &\geq 2I(X; \hat{X}_0) + I(\hat{X}_1; \hat{X}_2 | \hat{X}_0) + I(X; \hat{X}_1, \hat{X}_2 | \hat{X}_0). \end{aligned}$$

2. There exist functions $f_i(\hat{X}_0, \hat{X}_i)$, $i = 1, 2$, and $f_{12}(\hat{X}_0, \hat{X}_1, \hat{X}_2)$ satisfying

$$\begin{aligned} Ed(X, f_i(\hat{X}_0, \hat{X}_i)) &\leq D_i, \qquad i = 1, 2 \\ Ed(X, f_{12}(\hat{X}_0, \hat{X}_1, \hat{X}_2)) &\leq D_{12}. \end{aligned}$$

Theorem 10 *[24, Theorem 1]*

$$\mathcal{R}_{ach}(D_1, D_2, D_{12}) \subseteq \mathcal{R}_{MD}(D_1, D_2, D_{12}),$$

A generalization of this result to M descriptions is given in [25].

The following converse for MD is provided by Sher and Feder [26]. Let $\mathcal{R}_c(D_1, D_2, D_{12})$ be the closure of the set of all rates (R_1, R_2) for which there exist $\hat{X}_1$, $\hat{X}_2$, and $\hat{X}_{12}$ such that

$$\begin{aligned} R_1 &\geq I(X; \hat{X}_1) \\ R_2 &\geq I(X; \hat{X}_2) \\ R_1 + R_2 &\geq I(X; \hat{X}_{12}|\hat{X}_1, \hat{X}_2) + I(X; \hat{X}_1) + I(X; \hat{X}_2) \\ Ed(X, \hat{X}_t) &\leq D_t, \quad t \in \{1, 2, 12\}. \end{aligned}$$

Theorem 11 *[26, Theorem 1]*

$$\mathcal{R}_{MD}(D_1, D_2, D_{12}) \subseteq \mathcal{R}_c(D_1, D_2, D_{12}).$$

2.2.8 The Network with Unreliable Side Information

Consider a Wyner-Ziv system in which the presence of side information at the decoder is unreliable. The network of Figure 1.3(b) models such a system using two decoders. Decoder 1 corresponds to the case when the side information is available; decoder 2 corresponds to the case when it is absent. This network is studied by both Heegard and Berger [2] and

Kaspi [3]. They give different but equivalent characterizations of the rate-distortion region; I state here Heegard and Berger's result. Let $\mathcal{R}_{USI}(D_1, D_2)$ be the closure of the set of all achievable rate vectors for distortions (D_1, D_2), and let $\mathcal{R}^*(D_1, D_2)$ be the closure of the set of all rates R for which there exist auxiliary random variables U and V such that

1.

$$R \geq I(X;U) + I(X;V|U,Z).$$

2. There exist functions $f_1(U,V,Z)$ and $f_2(U)$ satisfying

$$\begin{aligned} Ed(X, f_1(U,V,Z)) &\leq D_1 \\ Ed(X, f_2(U)) &\leq D_2. \end{aligned}$$

3. $X \to Z \to (U,V)$ forms a Markov chain.

4. $|\mathcal{U}| \leq |\mathcal{X}| + 2$ and $|\mathcal{V}| \leq (|\mathcal{X}| + 1)^2$.

Theorem 12 *[2, Theorem 1]*

$$\mathcal{R}_{USI}(D_1, D_2) = \mathcal{R}^*(D_1, D_2)$$

Both papers give additional results. Kaspi gives the rate-distortion region when the side information is also available to the encoder. Heegard and Berger generalize their achievability result to the case of several decoders, each with different side information. They also evaluate their region for jointly Gaussian sources and provide an upper bound for binary symmetric sources. That upper bound is tightened by Kerpez [28], who also provides a loose lower bound. I close the gap between these bounds in Chapter 4.

2.2.9 The Two-Receiver Broadcast Network

Figure 1.1(f) shows a two-receiver BC network. Source X_1 must be described to decoder 1, and source X_2 must be described to decoder 2. There are three channels for the descriptions; a common channel to both decoders of rate R_{12}, and two private channels, one to each of the two decoders, of rates R_1 and R_2, respectively. The rate-distortion region is given by Gray and Wyner [19]. Let $\mathcal{R}_{BC}(D_1, D_2)$ be the closure of the set of all achievable rate vectors for distortions (D_1, D_2), and let $\mathcal{R}^*(D_1, D_2)$ be the closure of the set of all rate pairs (R_1, R_2) for which an auxiliary random variable W exists satisfying

$$\begin{aligned} R_{12} &\geq I(X_1, X_2; W) \\ R_1 &\geq R_{X_1|W}(D_1) \\ R_2 &\geq R_{X_2|W}(D_2). \end{aligned}$$

Theorem 13 *[19, Theorem 8]*

$$\mathcal{R}_{BC}(D_1, D_2) = \mathcal{R}^*(D_1, D_2)$$

I generalize the achievability result to the three receiver case in Chapter 5. I also derive tight results for a special case of M-receiver broadcast coding.

2.2.10 Rate Loss and Successive Refinement

Zamir [39] defines the rate loss $L(D)$ for WZ coding as the difference between the WZ and the conditional rate-distortion functions; thus $L(D) = R_{X|\{Z\}}(D) - R_{X|Z}(D)$. The rate loss describes the difference in achievable rate when the side information is available to just the

decoder compared to when it is available to both the encoder and the decoder. It is always non-negative.

In [39], Zamir shows that for a continuous source and the r-th power distortion measure, the rate loss is bounded by a constant which depends on the source alphabet and the distortion measure, but not on the source distribution. For example, for continuous alphabet sources and squared-error distortion, $L(D) \leq \frac{1}{2}$ for all D. This bound shows that the penalty paid for Y not being available at the encoder cannot be arbitrarily large.

The concept of rate loss can be applied to other coding scenarios such as MR coding [40]. For MR codes, rate loss is found to be zero for successively refinable sources, defined below.

Definition [14] [15, Definitions 1 and 2] A source X is said to be *successively refinable* under a distortion measure if, for that source and distortion measure, successive refinement from distortion D_1 to distortion D_2 is achievable for every $D_1 \geq D_2$. *Successive refinement* from distortion D_1 to $D_2 \leq D_1$ is said to be achievable if there exists a sequence of encoding and decoding functions

$$\begin{aligned}
\alpha_1^n &: \mathcal{X}^n \to \{1, \ldots, 2^{nR_1}\} \\
\alpha_1^n &: \mathcal{X}^n \to \{1, \ldots, 2^{n(R_2-R_1)}\} \\
\beta_1^n &: \{1, \ldots, 2^{nR_1}\} \to \hat{\mathcal{X}}^n \\
\beta_2^n &: \{1, \ldots, 2^{nR_1}\} \times \{1, \ldots, 2^{n(R_2-R_1)}\} \to \hat{\mathcal{X}}^n
\end{aligned}$$

such that for $\hat{X}_1^n = \beta_1^n(\alpha_1^n(X^n))$ and $\hat{X}_2^n = \beta_2^n(\alpha_1^n(X^n), \alpha_2^n(X^n))$,

$$\begin{aligned}
\limsup_{n\to\infty} Ed(X^n, \hat{X}_1^n) &\leq D_X(R_1) \\
\limsup_{n\to\infty} Ed(X^n, \hat{X}_2^n) &\leq D_X(R_2),
\end{aligned}$$

where $D_X(R)$ is the distortion-rate function for the source.

Examples of successively refinable sources include Gaussian sources under squared-error distortion, arbitrary discrete distributions under Hamming distortion, and Laplacian sources under absolute error [15]. An example of a problem that is not successively refinable is that of the Gerrish source [41, 15] (which has $|\mathcal{X}| = 3$ and $p(x) = (\frac{1-p}{2}, p, \frac{1-p}{2})$), although the maximal excess rate $R_2 - R_X(D_2)$ when $R_1 = R_X(D_1)$ is very small [42].

Rate loss results for MR codes [40, 43] bound the difference in total rate $\sum_{i=1}^{k} R_i$ used to achieve distortion D_i in resolution i and the corresponding rate $R_X(D_i)$ for an optimal point-to-point code. They give source-independent bounds similar to those of Zamir for the WZ system and show that only a small penalty is paid in using a single multi-resolution source description in place of a family of optimal point-to-point (single-resolution) descriptions.

2.3 Vector Quantization

Past work on VQ design typically takes one of two approaches. Either the codebook is first initialized in some way and then trained using an iterative descent algorithm ("unconstrained" design), or a lattice or other structure is imposed on the codebook ("constrained" design). The work in this thesis considers unconstrained VQ design. Prior work in this area is summarized below.

Globally optimal scalar quantizer (SQ) design can be done in polynomial-time for both fixed-rate coding [44, 45, 46, 47] and variable-rate coding [48]. However, globally optimal VQ design is NP-hard even for fixed-rate VQ and only two codewords [49]. It is convenient, therefore, to consider locally optimal design via iterative descent techniques. An iterative

descent algorithm to design locally rate-distortion-optimal fixed-rate VQs for the simplest network, the point-to-point network, appears in [50], and its extension to variable-rate coding via the inclusion of an entropy constraint appears in [51]. Fixed-rate VQ design for transmission over a noisy channel is considered in [52, 53].

The approaches of [50] and [51] have been generalized for application to MR, MD, and BC networks. Examples of MR coding include tree-structured VQ [54, 55], locally optimal fixed-rate MRSQ [56], variable-rate MRSQ [57], and locally optimal variable-rate MRSQ and MRVQ [58, 59]. Examples of MD coding include locally optimal two-description fixed-rate MDVQ [60], locally optimal two-description fixed-rate [61] and entropy-constrained MDSQ [62].

There are significantly fewer VQ design algorithms for MA systems. Two schemes for two-user MA coding appear in [63], but they differ from the previous works mentioned in that they are not optimized explicitly for rate-distortion performance.

Building on a part of this thesis first published in [7], fixed and variable-rate BC coding is covered in [20, 64]. That work is further generalized in [1], which presents an algorithm for VQ design in a general network with multiple encoders and multiple decoders.

Subsequent to the publication of the VQ design algorithm from this thesis [8, 9], an algorithm has been developed that achieves globally near-optimal VQ design for many of the systems studied here [65]. That algorithm relies directly on the optimal encoder and decoder definitions first described in [1] and generalized in Chapter 3 to allow for the use of side information at the decoders and for the presence of channel errors.

Chapter 3

Network Vector Quantizer Design

3.1 Introduction

Good code design algorithms are a necessary precursor to widespread use of network data compression algorithms. This chapter treats VQ design for networks. The choice of VQs (which include SQs as a special case) is motivated by their practicality, generality, and close relationship to theory.

As mentioned in Chapter 1, practical coding for networks can be approached in one of two ways. In the independent approach, a separate code is designed for each communication link. In the network approach, the network topology is taken into account. In this chapter I focus solely on the network approach and develop an iterative algorithm for the design of network VQs (NVQs) for any network topology. The resulting algorithm generates locally, but not necessarily globally, rate-distortion-optimal NVQs for some systems. For other systems, approximations required for practical implementation remove this guarantee, although I do observe convergence in all of my experimental work.

My contribution to VQ design comprises two parts. In the first, I solve the problem of VQ design for M-channel multiple description coding [7]. That work is generalized by Zhao and Effros for BC coding in [20, 64], and, building on [7, 20, 64], Effros presents a design algorithm for general network VQ in [1]. In the second part, I extend the design equations in [1] to allow both for the use of side information at the decoders and for the presence of channel errors. Then, I demonstrate in detail how to implement the algorithm, and I discuss its convergence and complexity. Finally, I present experimental results for several types of network VQs, demonstrating their rate-distortion improvements over independent VQs and comparing their performance to rate-distortion bounds. Since the framework of [1] and its extension presented here subsume that of my first work [7], I present everything in the newer framework.

Following the entropy-constrained coding tradition (see, for example, [51, 59, 66, 67]), I describe lossy code design as quantization followed by entropy coding. The only loss of generality associated with the entropy-constrained approach is the restriction to solutions lying on the lower convex hull of achievable entropies and distortions. I here focus exclusively on the quantizer design,[1] considering entropy codes only insofar as their index description rates affect quantizer optimization.

While the entropy codes of [51, 59] are lossless codes, entropy coding for many network systems requires the use of codes with asymptotically negligible (but non-zero) error prob-

[1]The topic of entropy code design for network systems is a rich field deserving separate attention; see, for example, [19, 64, 68, 69, 70, 71, 72, 73, 74].

abilities [68, 74, 75], called *near-lossless* codes.[2] The use of near-lossless entropy codes is assumed where necessary. As in [51, 59, 66, 67], I approximate entropy code performance using the asymptotically optimal values – reporting rates as entropies and assuming zero error probability. This approach is consistent with past work. It is also convenient since tight bounds on the non-asymptotic performance are not currently available and the current high level of interest in entropy coding for network systems promises rapid changes in this important area. The extension of my approach to include entropy code optimization and account for true (possibly non-zero) error probabilities in the iterative design is trivial, and I give the optimization equations in their most general form to allow for this extension.

Although my algorithm allows fixed- and variable-rate quantizer design for arbitrary networks, potential optimality in variable-rate design is currently limited to a select group of systems. It requires the availability of either optimal theoretical entropy constraints or optimal practical entropy codes. Optimal entropy constraints are available for MR, MD, WZ, and two-user MA systems. Also, some points on the boundary of the achievable region for two-user BC coding are achievable using practical codes [64]. For multi-user MA systems, the achievable rate region is known but optimal theoretical entropy constraints are not easily derived, nor have optimal practical near-lossless codes been created. For multi-receiver BC and general networks (e.g., the general three-node network of Figure 3.2), even the asymptotically optimal near-lossless rates are unknown. For such networks, I must design variable-rate quantizers using rates that are known to be achievable in place of the unknown optimal rates. The resulting quantizers are necessarily suboptimal.

[2]For instance, achieving the Slepian-Wolf rate bounds in a multiple access system requires the use of near-lossless codes. Lossless codes cannot achieve the bounds for all sources.

Scalability and complexity are important considerations for any network algorithm. The scalability of my NVQ implementation depends on the interconnectivity of the network. For an M-node network in which the in-degree of each node is constant as M grows, design complexity increases linearly in M, and code design for large networks is feasible. If, however, the in-degree of each node increases with M, then the design complexity increases exponentially in M and my approach is useful for small networks only. Once design is complete, the run-time complexity of my algorithm need not be prohibitive for any size of network. Optimal decoding can be implemented easily using table lookup. Optimal encoding is more complicated, but if encoding complexity is critical, then it can be greatly reduced by approximating each encoder with a hierarchical structure of tables following the approach of [76].

The chapter is organized as follows. I develop a framework for network description in Section 3.2. The optimal design equations for an NVQ are presented in Section 3.3, and their implementation discussed in Section 3.4. Section 3.5 presents experimental results for specific network design examples, and I draw conclusions in Section 3.6.

3.2 Network Description

This section develops a framework for describing network components and defines the meaning of optimality for network source codes. Due to the complexity of a general network, this discussion requires a significant amount of notation; I simplify where possible.

Consider a dimension-n code for an M-node network. In the most general case, every node communicates with every other node, and a message may be intended for any subset of nodes. Let $X_{t,S}^n \in \mathcal{X}_{t,S}^n$ denote the source to be described by node t to the nodes in set

$S \subseteq \mathcal{M} = \{1, \ldots, M\}$. For example, $X^n_{1,\{2,3\}}$ is the source described by node 1 to nodes 2 and 3. If S contains only one index, I write $X^n_{t,\{r\}} = X^n_{t,r}$. Let $\hat{X}^n_{t,S,r} \in \hat{\mathcal{X}}^n_{t,S,r}$ denote the reproduction of source $X^n_{t,S}$ at node $r \in S$. Thus $\hat{X}^n_{1,\{2,3\},2}$ is node 2's reproduction of source $X^n_{1,\{2,3\}}$. Reproductions $\hat{X}^n_{1,\{2,3\},2}$ and $\hat{X}^n_{1,\{2,3\},3}$ can differ since nodes 2 and 3 jointly decode the description of $X^n_{1,\{2,3\}}$ with different source descriptions. The source and reproduction alphabets can be continuous or discrete, and typically $\hat{\mathcal{X}}^n_{t,S,r} = \mathcal{X}^n_{t,S}$.

For each node $t \in \mathcal{M}$, let $\mathcal{S}(t)$ denote a collection of sets such that for each $S \in \mathcal{S}(t)$, there exists a source to be described by node t to precisely the members of $S \subseteq \mathcal{M}$. Then $X^n_{t,*} = (X^n_{t,S})_{S \in \mathcal{S}(t)}$ gives the collection of sources described by node t. Similarly, for each $r \in \mathcal{M}$, let $\mathcal{T}(r) = \{(t, S) \in \mathcal{S} : r \in S\}$ be the set of source descriptions received by node r, where $\mathcal{S} = \{(t, S) : t \in \mathcal{M}, S \in \mathcal{S}(t)\}$ is the set of sources in the network. Then $\hat{X}^n_{*,r} = (\hat{X}^n_{t,S,r})_{(t,S) \in \mathcal{T}(r)}$ gives the collection of reconstructions at node r. Finally, let $\mathcal{T} = \{(t, S, r) : r \in \mathcal{M}, (t, S) \in \mathcal{T}(r)\}$ denote the set of all transmitter-message-receiver triples.

For each node $r \in \mathcal{M}$, denote the side information available at node r by $Z^n_r \in \mathcal{Z}^n_r$. Alphabet $\mathcal{Z}_r$ can be continuous or discrete.

Figure 3.1 recasts some of the network examples of Chapter 1 into this notation.

Figure 3.2 shows the example of a general three-node network. Each node transmits a total of three different source descriptions. Node 1, for instance, encodes a source intended for node 2 only, a source intended for node 3 only, and a source intended for both. These are denoted by $X^n_{1,2}$, $X^n_{1,3}$, and $X^n_{1,\{2,3\}}$, respectively, giving $X^n_{1,*} = (X^n_{1,2}, X^n_{1,3}, X^n_{1,\{2,3\}})$. Each node in the network receives and decodes four source descriptions. The collection of reproductions at node 1 is $\hat{X}^n_{*,1} = (\hat{X}^n_{2,1,1}, \hat{X}^n_{2,\{1,3\},1}, \hat{X}^n_{3,1,1}, \hat{X}^n_{3,\{1,2\},1})$; their descriptions are jointly decoded with the help of side information Z^n_1. The total number of reproductions is greater

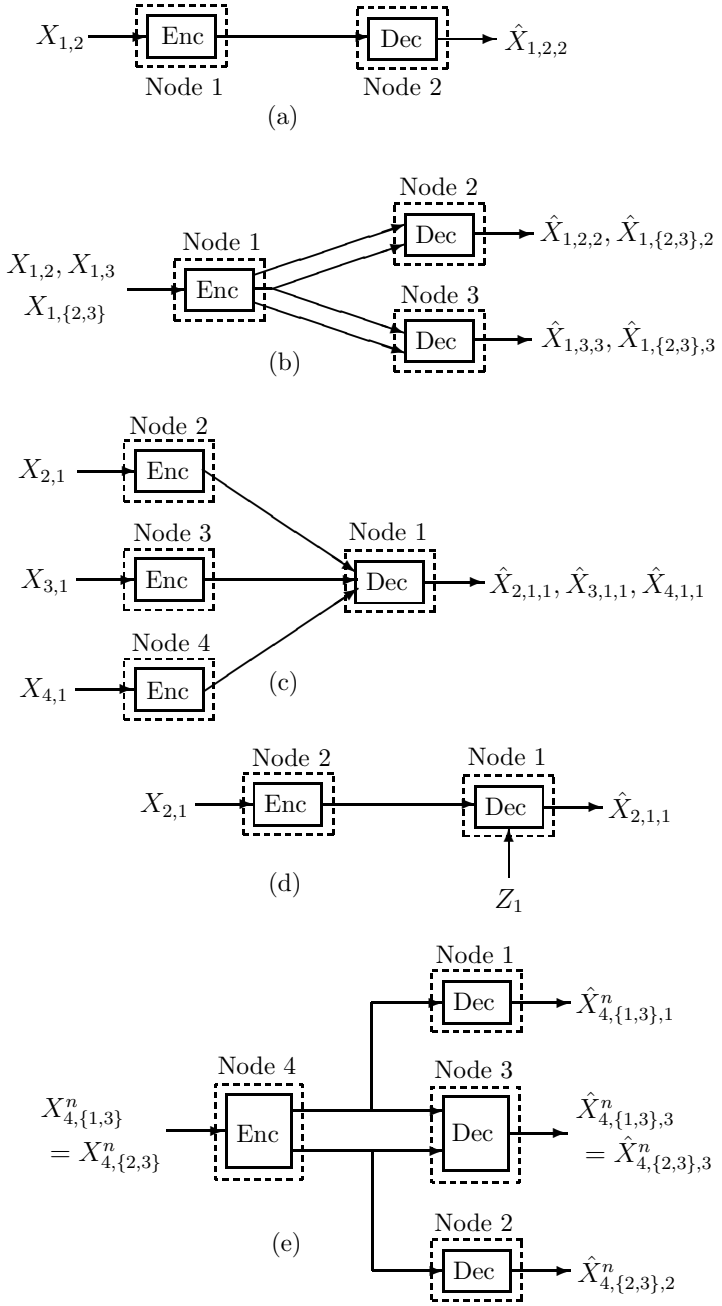


Figure 3.1: (a) A point-to-point code. (b) A two-receiver BC code. (c) A two-user MA code. (d) A WZ code with side information Z_1. (e) A two-channel MD code. The notation is $X_{t,S}$ for a source and $\hat{X}_{t,S,r}$ for a reproduction; here t is the transmitter, S the set of receivers, and r the reproducer.

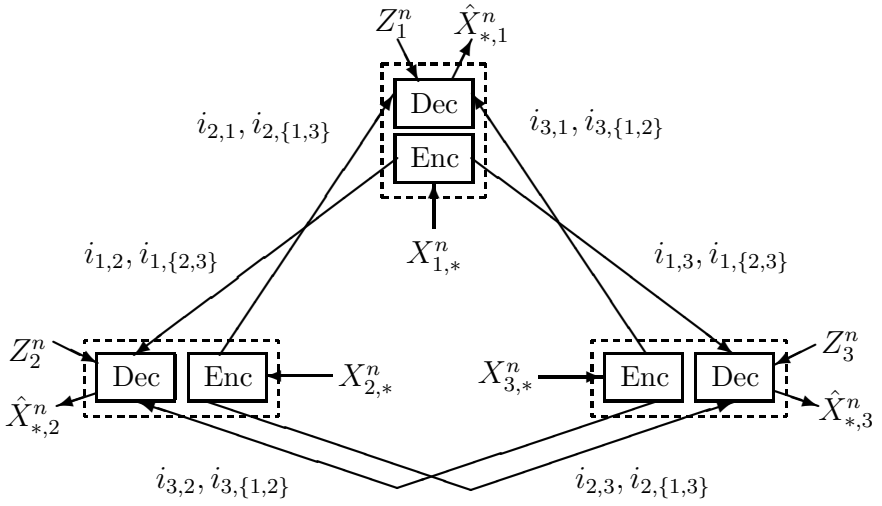


Figure 3.2: A general three-node network.

than the number of sources since some sources are reproduced at more than one node.

A network encoder comprises two parts: a quantizer encoder, followed by an entropy encoder. A network decoder comprises two complementary parts: an entropy decoder followed by a quantizer decoder. For variable-rate NVQ design, the network's entropy coders may be lossless or near-lossless, and following [51] and practical implementations employing arithmetic codes, I allow the entropy coders to operate at a higher dimension than the quantizers. For the case of fixed-rate NVQ design, the entropy coders are simply lossless codes operating at a fixed rate.

For any vector $X_{t,*}^n$ of source n-vectors, the quantizer encoder at node t, given by $\alpha_t : \mathcal{X}_{t,*}^n \to \mathcal{I}_{t,*}$, maps $X_{t,*}^n$ to a collection of indices $i_{t,*}$ in the index set $\mathcal{I}_{t,*}$. In theory, $\mathcal{I}_{t,*}$ may be finite or countably infinite; in practice a finite $\mathcal{I}_{t,*}$ is assumed. Here $i_{t,*} = (i_{t,S})_{S \in \mathcal{S}(t)}$, and for each $S \in \mathcal{S}(t)$, $i_{t,S} \in \mathcal{I}_{t,S}$. The collection of indices $i_{t,*}$ is mapped by the fixed- or variable-rate entropy encoder at node t to a concatenated string of binary descriptions $c_{t,*} \in \mathcal{C}_{t,*}$. The channel conveys each individual description $c_{t,S}$ to precisely the receivers

$r \in S$.

For any $r \in \mathcal{M}$, the entropy decoder at node r receives the codewords $c_{*,r}$ and side information Z_r^n and outputs index reconstructions $\hat{i}_{*,r} \in \mathcal{I}_{*,r}$. Except in a few special cases (e.g., when a coding error occurs), these are identical to the corresponding transmitted indices. Denote the quantizer decoder at node r by $\beta_r : \mathcal{I}_{*,r} \times \mathcal{Z}_r^n \to \hat{\mathcal{X}}_{*,r}^n$. It maps indices $\hat{i}_{*,r} \in \mathcal{I}_{*,r}$ and side information Z_r^n to a collection of reproduction vectors $\hat{X}_{*,r}^n$ such that $\hat{X}_{t,S,r}^n \in \hat{\mathcal{X}}_{t,S,r}^n$ for each $(t,S) \in \mathcal{T}(r)$. Let $\beta_{t,S,r}^n(\hat{i}_{*,r}, Z_r^n)$ denote the reproduction of $X_{t,S}^n$ made by receiver r. Then $\beta_r(\hat{i}_{*,r}, Z_r^n) = \hat{X}_{*,r}^n$ implies that $\beta_{t,S,r}(\hat{i}_{*,r}, Z_r^n) = \hat{X}_{t,S,r}^n$ for each $(t,S) \in \mathcal{T}(r)$. Note $\beta_{t,S,r}$ depends on $\hat{i}_{*,r}$ rather than simply $\hat{i}_{t,S}$ since $\hat{i}_{*,r}$ is jointly decoded.

Associate two mappings with each entropy code. The first, $\ell_t : \mathcal{I}_{t,*} \to [0,\infty)$, is the rate used to describe $i_{t,*}$. In practice, $\ell_t(i_{t,*})$ is the length of the entropy code's corresponding codewords $c_{t,*}$; for entropy-constrained design, $\ell_t(i_{t,*})$ is a function of the entropy bound [51]. The rate used to describe a particular $i_{t,S}$ is given by $\ell_{t,S} : \mathcal{I}_{t,*} \to [0,\infty)$; it depends on all of the indices from node t because the mapping is done jointly.

The second mapping, given by $f_r : \mathcal{I}_{*,r} \times \mathcal{Z}_r^n \to \mathcal{I}_{*,r}$, maps indices $i_{*,r}$ transmitted to node r, together with side information Z_r^n, to the indices $\hat{i}_{*,r}$ received after entropy decoding. Let $\alpha_{t,S}(x_{t,*}^n)$ denote the component of α_t that produces codeword $i_{t,S}$. Then $\hat{i}_{*,r} = f_r(i_{*,r}, z_r^n)$, where $i_{*,r} = \left(\alpha_{t',S'}(X_{t',*}^n)\right)_{(t',S') \in \mathcal{T}(r)}$. Typically, $f_r(i_{*,r}, z_r^n) = i_{*,r}$. Exceptions are caused by coding errors and a few special cases discussed in Section 3.4.

I restrict the joint encoding of the entropy codes to ensure unique decodability. Every entropy decoder must be able to uniquely decode each of its codewords using only the other codewords and the side information available to it. For example, in the MD system of Figure 3.1(e), the entropy encoder at node 4 encodes indices $i_{4,\{1,3\}}$ and $i_{4,\{2,3\}}$ so that

they can be individually decoded at nodes 1 and 2. This requires that the coding be done independently. However, the restriction is not so severe that all entropy codings in all systems need be done independently. In the system of Figure 3.1(b), for example, a conditional entropy code for $i_{1,2}$ given $i_{1,\{2,3,4\}}$ can be used since the two indices are jointly decoded at node 1. The restriction that the entropy codes be uniquely decodable does not imply that the encoders are one-to-one mappings; different source symbols may be given the same description if the decoder has other information (side information or messages from other sources) that allows it to distinguish them.

The performance of a network source code $Q^n = (\{\alpha_t\}_{t\in\mathcal{M}}, \{\beta_r\}_{r\in\mathcal{M}}, \{\ell_t\}_{t\in\mathcal{M}}, \{f_r\}_{r\in\mathcal{M}})$ is measured in rate and distortion. In particular, for each $(t, S, r) \in \mathcal{T}$, let $d_{t,S,r} : \mathcal{X}_{t,S} \times \hat{\mathcal{X}}_{t,S,r} \to [0, \infty)$ be a nonnegative distortion measure between the alphabets $\mathcal{X}_{t,S}$ and $\hat{\mathcal{X}}_{t,S,r}$. Let the distortion between vectors of symbols be additive, so that for any $n > 1$,

$$d^n_{t,S,r}(x^n_{t,S}, \hat{x}^n_{t,S,r}) = \sum_{k=1}^{n} d_{t,S,r}(x_{t,S}(k), \hat{x}_{t,S,r}(k)).$$

Here $x_{t,S}(k)$ and $\hat{x}_{t,S,r}(k)$ denote the kth symbols in vectors $x^n_{t,S}$ and $\hat{x}^n_{t,S,r}$, respectively.[3] Although not required for the validity of the results here, for simplicity of notation I assume that the distortion measures are identical and omit the subscripts. I also omit the superscript since it is clear from the arguments whether d is operating on a scalar or a vector.

Denote by $\mathcal{Q}^{\mathrm{fr},n}$ and $\mathcal{Q}^{\mathrm{vr},n}$ the classes of n-dimensional fixed- and variable-rate NVQs, respectively. Let $x^n_{*,*} = (x^n_{1,*}, x^n_{2,*}, \ldots, x^n_{M,*})$ denote a particular value for the collection of random source vectors $X^n_{*,*} = (X^n_{1,*}, X^n_{2,*}, \ldots, X^n_{M,*})$. Similarly, let $z^n_* = (z^n_1, z^n_2, \ldots, z^n_M)$ denote a particular value for the side information $Z^n_* = (Z^n_1, Z^n_2, \ldots, Z^n_M)$. The (instantaneous)

[3]This notation for the kth element differs from that introduced in Chapter 2 so as to avoid too many subscripts. It will be used in this chapter only.

rate and distortion vectors associated with coding source vector $x^n_{*,*}$ with code $Q^n \in \mathcal{Q}^{(\mathrm{fr}|\mathrm{vr}),n}$ given side information z^n_* are, respectively,[4]

$$\begin{aligned}
\mathbf{r}(x^n_{*,*}, Q^n) &= \left(r_{t,S}(x^n_{t,*}, Q^n)\right)_{(t,S)\in\mathcal{S}} = \left(\ell_{t,S}(\alpha_t(x^n_{t,*}))\right)_{(t,S)\in\mathcal{S}} \\
\mathbf{d}(x^n_{*,*}, z^n_*, Q^n) &= \left(d(x^n_{t,S}, \hat{x}^n_{t,S,r})\right)_{(t,S,r)\in\mathcal{T}} = \left(d\left(x^n_{t,S}, \beta_{t,S,r}(\hat{i}_{*,r}, z^n_r)\right)\right)_{(t,S,r)\in\mathcal{T}}.
\end{aligned}$$

Assume that the source and side information vectors together form a strictly stationary[5] ergodic random process with source distribution P. Let E denote the expectation with respect to P. The expected rate and distortion in describing n symbols from P with code Q^n are $\mathbf{R}(P, Q^n) = (R_{t,S}(P, Q^n))_{(t,S)\in\mathcal{S}}$ and $\mathbf{D}(P, Q^n) = (D_{t,S,r}(P, Q^n))_{(t,S,r)\in\mathcal{T}}$, where

$$\begin{aligned}
R_{t,S}(P, Q^n) &= Er_{t,S}\left(X^n_{t,*}, Q^n\right) = E\ell_{t,S}\left(\alpha_t(X^n_{t,*})\right) \\
D_{t,S,r}(P, Q^n) &= Ed\left(X^n_{t,S}, \hat{X}^n_{t,S,r}\right) = Ed\left(X^n_{t,S}, \beta_{t,S,r}(\hat{I}_{*,r}, Z^n_r)\right).
\end{aligned}$$

By [1, Lemmas 1,2,3] and the associated discussion, optimal NVQ performance is achieved by minimizing the weighted operational rate-distortion functional

$$j^{(\mathrm{fr}|\mathrm{vr}),n}(P, \mathbf{a}, \mathbf{b}) = \inf_{Q^n\in\mathcal{Q}^{(\mathrm{fr}|\mathrm{vr}),n}} \sum_{(t,S)\in\mathcal{S}} \frac{1}{n}\left[a_{t,S}R_{t,S}(P, Q^n) + \sum_{r\in S} b_{t,S,r}D_{t,S,r}(P, Q^n)\right]. \quad (3.1)$$

The weighted operational rate-distortion functionals may be viewed as Lagrangians for minimizing a weighted sum of distortions subject to a collection of constraints on the corresponding rates. They can also be viewed as Lagrangians for minimizing a weighted sum of rates subject to a collection of constraints on the corresponding distortions. The weights $\mathbf{a}$ and $\mathbf{b}$ embody the code designer's priorities on the rates and distortions. They are constrained to

[4]The superscript (fr|vr) implies that the given result applies in parallel for fixed- and variable-rate.

[5]The condition of strict stationarity could be replaced by a condition of asymptotic mean stationarity in the results that follow. Strict stationarity is used for simplicity.

be non-negative, so that higher rates and distortions yield a higher Lagrangian cost. Code design depends on the *relative* values of these weights, and hence without loss of generality I set

$$\sum_{(t,S)\in\mathcal{S}} \left[a_{t,S} + \sum_{r\in S} b_{t,S,r} \right] = 1.$$

In practice, $\mathbf{a}$ and $\mathbf{b}$ cannot easily be chosen to guarantee specific rates or distortions: they reflect trade-offs over the entire network. In a typical code design scenario, the goal is to minimize a weighted sum of distortions subject to the constraint $\mathbf{R}(P, Q^n) = \mathbf{R}^*$. In this case, I set $\mathbf{b}$ according to the distortion weights and adopt a gradient descent approach to find appropriate values for $\mathbf{a}$. Denote by $Q^n(\mathbf{a}, \mathbf{b})$ the quantizer produced by the algorithm (described in the following section) when the Lagrangian constants are $(\mathbf{a}, \mathbf{b})$. The gradient descent minimizes the absolute rate difference $\chi(\mathbf{a}) = |\mathbf{R}(P, Q^n(\mathbf{a}, \mathbf{b})) - \mathbf{R}^*|^2$ as a function of $\mathbf{a}$.

I call a fixed- or variable-rate network source code Q^n *optimal* if it achieves a point on $j^{\mathrm{fr},n}(P, \mathbf{a}, \mathbf{b})$ or $j^{\mathrm{vr},n}(P, \mathbf{a}, \mathbf{b})$. Section 3.3 considers locally optimal NVQ design.

3.3 Locally Optimal NVQ Design

The goal in NVQ design is to find a code Q^n that optimizes the weighted cost in (3.1). Following the strategy of [50] and [51], I consider below the necessary conditions for optimality of $\{\alpha_t\}$ and $\{\beta_r\}$ when the other system components are fixed. Using these conditions, I design NVQs through an iterative descent technique functionally equivalent to the generalized Lloyd algorithm. Throughout the discussion, I compare the NVQ conditions with those for the point-to-point system of Figure 3.1(a) to highlight the changes involved in moving from

independent to network design.

The Lagrangian cost for a given code Q^n is

$$j^n(P, \mathbf{a}, \mathbf{b}, Q^n) = \sum_{(t,S)\in\mathcal{S}} \frac{1}{n} \left[a_{t,S} R_{t,S}(P, Q^n) + \sum_{r\in S} b_{t,S,r} D_{t,S,r}(P, Q^n) \right]. \tag{3.2}$$

An algorithm to design a code minimizing this cost is as follows [1].

Initialize the system components $\{\alpha_t\}$, $\{\beta_r\}$, lengths $\{\ell_t\}$, and mappings $\{f_r\}$.

Repeat

Optimize each α_t and β_r in turn, holding every other component fixed.

Update the coding rates $\{\ell_t\}$ and mappings $\{f_r\}$, holding all $\{\alpha_t\}$ and $\{\beta_r\}$ fixed.

Until the code's cost function j^n converges.

Provided that each optimization reduces the cost functional j^n, which is bounded below by zero, the algorithm will converge. In practice, I make approximations to simplify the optimizations and cannot always guarantee a reduction of the cost function (see Section 3.4 for more details). However, except when close to a minimum, I do observe a consistent reduction in j^n in my experiments.

Decoder optimization is simple, even in the most general case. However, encoder optimization is not. Messages produced by an encoder are jointly decoded with messages from other encoders and with side information. However, each encoder knows neither the input to the other encoders nor the side information exactly, so it must operate based on the *expected* behavior of these other quantities. The expectation complicates the design process.

Now consider the component optimizations in detail, beginning with the decoders.

Optimal Decoders

Choose some $R \in \mathcal{M}$, and consider necessary conditions for the optimality of β_R when all encoders $\{\alpha_t\}_{t\in\mathcal{M}}$, all other decoders $\{\beta_r\}_{r\in\mathcal{M}\cap\{R\}^c}$, all length functions $\{\ell_t\}_{t\in\mathcal{M}}$, and all mappings $\{f_r\}_{r\in\mathcal{M}}$ are fixed. The optimal decoder $\beta_R^\star = (\beta_{t,S,R}^\star)_{(t,S)\in\mathcal{T}(R)}$ for index vector $\hat{i}_{*,R} = (\hat{i}_{t,S})_{(t,S)\in\mathcal{T}(R)}$ and side information z_R^n satisfies

$$\beta_{t,S,R}^\star(\hat{i}_{*,R}, z_R^n) = \arg\min_{\hat{x}^n\in\hat{\mathcal{X}}_{t,S,R}^n} E\left[d(X_{t,S}^n,\hat{x}^n)\middle|\, Z_R^n = z_R^n, f_R\left(\left(\alpha_{t',S'}(X_{t',*}^n)\right)_{(t',S')\in\mathcal{T}(R)}, z_R^n\right) = \hat{i}_{*,R}\right]. \quad (3.3)$$

The expectation is with respect to the source distribution P.

The optimal decoder for the point-to-point system, shown in Figure 3.1(a), satisfies

$$\beta_{t,R,R}^\star(\hat{i}_{t,R}) = \arg\min_{\hat{x}^n\in\hat{\mathcal{X}}_{t,R,R}^n} E\left[d(X_{t,R}^n,\hat{x}^n)\middle|\, f_R(\alpha_{t,R}(X_{t,R}^n)) = \hat{i}_{t,R}\right], \quad (3.4)$$

where I have relabeled node 1 as t and node 2 as R. In the point-to-point case, the optimal reproduction for $\hat{i}_{t,R}$ is the vector $\hat{x}^n \in \hat{\mathcal{X}}_{t,R,R}^n$ that minimizes the expected distortion in the Voronoi cell indexed by $\hat{i}_{t,R}$. This Voronoi cell contains all source vectors $X_{t,R}^n$ such that $f_R(\alpha_{t,R}(X_{t,R}^n)) = \hat{i}_{t,R}$. In the network case, the equation takes the same form, but with the Voronoi cell now indexed by a *collection* of indices $\hat{i}_{*,R}$ *and* the side information z_R^n.

In general, the optimal network decoder depends on the full distribution P rather than merely the distribution of the message under consideration. This dependence arises from the joint nature of the decoding process.

The optimal decoder can be extended to allow for channel coding errors[6]. The distribution of channel coding errors is assumed to be independent of the sources or side information given

[6]Incorporating the stochastic effects of channel coding errors into quantizer design allows control of the sensitivity of the source code to channel errors. It also allows for quantizer design in the case of a joint

the transmitted indices. I describe the effect of channel errors on the indices received by decoder r by the random mapping $G_r : \mathcal{I}_{*,r} \to \mathcal{I}_{*,r}$. Indices $i_{*,r}$ transmitted by the encoders are transformed into $G_r(i_{*,r})$ by the channel, and decoded as $\hat{i}_{*,r} = f(G_r(i_{*,r}), z_r^n)$ by the entropy code. The optimal decoder becomes

$$\begin{aligned}\beta^\star_{t,S,R}(\hat{i}_{*,R}, z_R^n) &= \arg \min_{\hat{x}^n \in \hat{\mathcal{X}}^n_{t,S,R}} E\left[d(X^n_{t,S}, \hat{x}^n) \middle| Z_R^n = z_R^n, f(G_R(I_{*,R}), z_R^n) = \hat{i}_{*,R} \right] \\ &= \arg \min_{\hat{x}^n \in \hat{\mathcal{X}}^n_{t,S,R}} \sum_{i_{*,R} \in \mathcal{I}_{*,R}} \Big(E\left[d(X^n_{t,S}, \hat{x}^n) \middle| Z_R^n = z_R^n, I_{*,r} = i_{*,R} \right] \cdot \\ &\qquad \Pr(I_{*,R} = i_{*,R} | Z_R^n = z_R^n, f(G_R(I_{*,R}), z_R^n) = \hat{i}_{*,R}) \Big) .\end{aligned}$$

Optimal Encoders

Now choose some $T \in \mathcal{M}$ and consider necessary conditions for the optimality of α_T when $\{\alpha_t\}_{t \in \mathcal{M} \cap \{T\}^c}$, $\{\beta_r\}_{r \in \mathcal{M}}$, $\{\ell_t\}_{t \in \mathcal{M}}$, $\{f_r\}_{r \in \mathcal{M}}$ are fixed. The optimal encoder $\alpha_T^\star$ satisfies

$$\begin{aligned}\alpha^\star_T(x^n_{T,*}) = \arg \min_{i_{T,*} \in \mathcal{I}_{T,*}} \Bigg[&\sum_{S \in \mathcal{S}(T)} a_{T,S} \ell_{T,S}(i_{T,*}) + \\ &\sum_{r \in S' : S' \in \mathcal{S}(T)} \sum_{(t,S) \in \mathcal{T}(r)} b_{t,S,r} E\left[d(X^n_{t,S}, \beta_{t,S,r}(f_r(I_{*,r}, Z_r^n), Z_r^n)) \middle| X^n_{T,*} = x^n_{T,*}, I_{T,*} = i_{T,*} \right] \Bigg] . \quad (3.5)\end{aligned}$$

Compare this to the equation for optimizing the encoder of the point-to-point system

$$\alpha^\star_T(x^n_{T,r}) = \arg \min_{i_{T,r} \in \mathcal{I}_{T,r}} \left[a_{T,r} \ell_{T,r}(i_{T,r}) + b_{T,r,r} d\left(X^n_{T,r}, \beta_{T,r,r}\left(f(i_{T,r}) \right) \right) \right] . \quad (3.6)$$

In the point-to-point case (3.6), the encoder's choice of index $i_{T,r}$ affects only one reproduction $\hat{X}_{T,r,r}$ at only one node r. In the network case (3.5), the indices chosen by encoder

source-channel code. Since the source-channel separation theorem does not hold for network coding (see for example [31, pp. 448-9]), joint source-channel codes are required for optimal performance in some networks.

$\alpha_T^\star$ have a much more widespread impact. As expected, the indices $\alpha_T(x_{T,*}^n)$ affect the reproductions for $X_{T,*}^n$, but they also affect some reproductions for $X_{t,*}^n$ ($t \neq T$) because each decoder β_r *jointly* maps its set of received indices to the corresponding vector of reproductions. Thus $i_{T,*}$ affects *all* reproductions at any node r to which T transmits a message. The minimization considers the weighted distortion over *all* of these reproductions.

The other major difference between the point-to-point and network equations is the expectation in the distortion term. The encoder in the point-to-point case knows $\hat{X}_{T,r,r}$ exactly for any possible choice of $i_{T,r}$. This is not true in the network case. For example, suppose encoder α_T transmits to node r. It does not know any of the indices received by r from other nodes, nor the side information available to r. These unknowns are jointly decoded with the message(s) from α_T to produce the reproductions at r, and hence α_T cannot completely determine the reproductions knowing only its own choice of indices. Encoder α_T must take a conditional expectation over the unknown quantities, conditioned on all of the information it does know, to determine its best choice of indices. In (3.5), the use of capitalization for $I_{*,r} = (I_{t',S'})_{(t',S')\in\mathcal{T}(r)}$ denotes the fact that for any $t' \neq T$, $i_{t',S'}$ is unknown to α_T and must be treated as a random variable. The expectation is taken over the conditional distribution on $X_{t,S}^n$, $I_{*,r}$, and the side information Z_r^n given $X_{T,*}^n = x_{T,*}^n$ and $I_{T,*} = i_{T,*}$. For any $t' \neq T$, the distribution on $I_{t',S'}$ is governed by the corresponding (fixed) encoder $\alpha_{t'}$ together with the conditional distributions on the inputs to that encoder. Evaluating the conditional expectations in the equation for α_T is the primary difficulty in implementing the design algorithm, as discussed in Section 3.4.

The optimal encoder can be extended to allow for channel coding errors. Representing

their effects by a random mapping G_r as before, the optimal encoder becomes

$$\alpha_T^\star(x_{T,*}^n) = \arg \min_{i_{T,*} \in \mathcal{I}_{T,*}} \left[\sum_{S \in \mathcal{S}(T)} a_{T,S} \ell_{T,S}(i_{T,*}) + \sum_{r \in S' : S' \in \mathcal{S}(T)} \sum_{(t,S) \in \mathcal{T}(r)} b_{t,S,r} E \left[d(X_{t,S}^n, \beta_{t,S,r}(f_r(G_r(I_{*,r}), Z_r^n), Z_r^n)) \middle| X_{T,*}^n = x_{T,*}^n, I_{T,*} = i_{T,*} \right] \right].$$

The expectation here is over both the source and the channel error distributions.

Entropy Coding Rates

I now consider how the state of the art in lossless and near-lossless coding affects entropy-constrained design. Networks fall into three categories in this regard.

First are systems for which there exist practical codes achieving arbitrarily close to the entropy bounds and for which we also know the theoretically optimal codeword lengths. For example, in point-to-point coding (Figure 3.1(a)), the entropy bound $R_{1,2} \geq H(I_{1,2})$ can be approximated using either Huffman or arithmetic coding. In addition, the codeword lengths given by $\ell_1(i_{1,2}) = -\log_2 p(i_{1,2})$ yield an expected rate equal to $H(I_{1,2})$ and satisfy Kraft's inequality. For systems in this category (including MR, MD, and 2-receiver BC systems), I follow [51] and design entropy-constrained NVQs using the theoretically optimal lengths.

Second are systems for which we cannot assign theoretical optimal codeword lengths, but we can still design practical lossless codes with rates close to the entropy bounds. This category includes WZ and 2A systems. Slepian and Wolf [11] give the achievable rate region for near-lossless coding in a two-access (2A) system (the generalization to M-encoder MA systems appears in [77]). However, these bounds alone are insufficient to determine the optimal codeword lengths. Generalizations of the point-to-point solution can yield lengths achieving

points such as $(R_{2,1}, R_{3,1}) = (H(I_{2,1}), H(I_{3,1}|I_{2,1}))$ on the boundary of the achievable rate region. However, there is no 2A-equivalent of Kraft's inequality with which to prove that there could exist uniquely decodable codes with those lengths. I turn to practical codes, such as the 2A code in [78], and use their codeword lengths in entropy-constrained design.

Third are systems for which we cannot assign theoretically optimal codeword lengths and lack techniques for designing optimal codes. For example, in lossless M-receiver BC coding, even the optimal performance is unknown when $M > 2$. However, we can assign theoretical lengths and even design practical codes to achieve rates unobtainable by an independent approach. For systems in this category (including the three-node network of Figure 3.2), I use the best known achievable rates, practical or theoretical, in the entropy constraint. Improved entropy constraints for these systems will likely become available as the field of lossless network coding develops.

Assuming that the near-lossless entropy codes achieve their asymptotic error probability of zero and that the distortion measure cannot be infinite, then there is no need to consider the increase in distortion resulting from entropy coding errors. For practical codes, which have a small but non-zero probability of error, the distortion increase should be taken into account in the near-lossless code optimization. The distortion caused by an error can be calculated using the training set and the fixed encoders and decoders. The near-lossless design algorithm presented in [78] minimizes a weighted sum of rate and probability of error, but it can easily be altered to weight each error by the expected distortion it generates as opposed to simply the error probability. This then ensures that the entropy code optimization is conducted with the same priorities as the quantizer design and will never result in an increased Lagrangian cost.

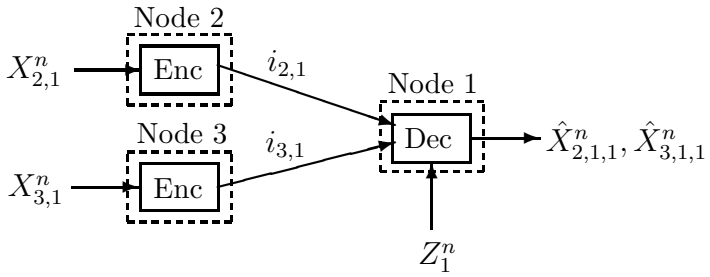


Figure 3.3: The 2AWZ network.

Network Scalability

Scalability is a key issue for some network source coding applications. If a code is trained for a particular network, but that network is then altered by adding or deleting a node, how much code redesign is required to accommodate the change?

Starting with the WZ system of Figure 3.1(d), consider adding a third node that describes a new source $X_{3,1}$ to the decoder at node 1. This creates a 2A system with side information at the decoder (a 2AWZ system), as depicted in Figure 3.3. There are at least two options for updating the code. In the first, the encoder at node 2 stays the same, the decoder retains its previous codebook for jointly decoding $(\hat{i}_{2,1}, Z_1)$, and a new conditional codebook for decoding $\hat{i}_{3,1}$, conditioned on $(\hat{i}_{2,1}, Z_1)$, is trained. This greedy approach keeps the previous system intact and simply adds new components. However, the correlation between $X_{3,1}$ and $X_{2,1}$ is exploited only in the decoding of $X_{3,1}$ and not $X_{2,1}$. Optimally exploiting the correlation so as to minimize the Lagrangian cost (3.2) requires a second, global approach, in which the original WZ codebook is extended to jointly decode all three inputs $(\hat{i}_{2,1}, \hat{i}_{3,1}, Z_1)$. For this, the whole system must be retrained.

Now consider deleting node 2 from the 2-access system of Figure 3.3, so that the decoder no longer receives $\hat{i}_{2,1}$. A new decoder can be formed from the existing one by simply

averaging over the various codewords for different values of $\hat{i}_{2,1}$, for each fixed $(\hat{i}_{3,1}, Z_1)$. Provided that the existing decoder cells are convex with respect to $\hat{i}_{2,1}$, this is a good strategy. Otherwise, global retraining is necessary.

These cases exemplify canonical network alterations: a new code can be formed with little cost by making local, greedily-designed component additions or subtractions, but aiming for optimal performance requires retraining the entire network at a greater cost.

3.4 Implementation

This section considers the implementation of the design algorithm. I focus on practical evaluation of the terms in the optimality conditions (3.3,3.5) from the previous section. I also discuss the use of side information at the decoders.

In practice, I optimize my codes with respect to a training data set. A key assumption I make is that the empirical joint distribution of the training set is close to the true joint source distribution.

Several experimental results in Section 3.5 assume that the entropy codes achieve their asymptotic bound of error probability zero. This does not imply that $f_r(i_{*,r}, z_r^n) = i_{*,r}$. Non-identity mappings must be used to deal with empty cells that arise during training. In designing a point-to-point VQ, there may be training iterations in which no training vectors are mapped to a particular Voronoi cell because its codeword is not the nearest neighbor of any of the training vectors. In entropy-constrained VQ (ECVQ), such cells are removed from consideration by associating with their index an infinite rate. Thus, an encoder designed to minimize $aD + bR$ for some $b > 0$ never uses that index. The same empty cell phenomenon

occurs in network coding. However, in an MA system, a decoder cell is jointly indexed by two or more encoder indices. Rates are not associated with individual cells, but with each separate index. Even if index pair (i, j) corresponds to an empty cell, this does not mean that either i or j individually should be assigned infinite rate (describing index i requires infinite rate only if *all* cells $(i, \cdot)$ become empty). Since we cannot in general remove cells from consideration by altering their rate, and since the encoders work independently, it is possible that an empty cell may inadvertently be indexed. This must be avoided, since to save rate we usually do not require that the entropy code preserve the indices of empty cells. In practice, we begin by assuming that no cells are empty, and as cells do become empty we merge each of them with a full neighboring cell (when side information z^n is present, we can choose a different merging for each z^n). The cell merging is incorporated into the entropy code to allow a saving in rate. Any reference to an empty cell is redirected to the non-empty neighbor, and this redirection is made known to the encoders through the mappings $\{f_r\}$. As in ECVQ, no cells that become empty are ever filled again; training vectors mapped to an empty cell indexed by (i_2, i_3) are always redirected to the appropriate non-empty neighbor $f_r(i_2, i_3, z^n)$. Thus, $f_r(i_2, i_3, z^n)$ fills two roles: handling both empty cells (in a "once empty always empty" manner) and near-lossless coding errors.

The terms in the optimality condition for a network decoder (3.3) are no more difficult to evaluate than those for the point-to-point decoder (3.4). For the point-to-point decoder, optimization trains each codeword using the set of training vectors falling into that codeword's Voronoi cell. For example, when the distortion measure is squared-error, evaluating the expectation in (3.4) places each codeword at the mean value of its associated training vectors. Network decoder optimization (3.3) requires no change in approach.

The difficulties in NVQ design arise in the optimization of the network encoders. In point-to-point VQ, encoder optimization is implemented through a nearest neighbor search: the encoder chooses the index $i_{T,r}$ that minimizes the Lagrangian in (3.6). In NVQ design, I again search over all possible encoder indices, but computing the Lagrangian in (3.5) may require the evaluation of a conditional expectation. I divide network encoders into two types, one for which the conditional expectation is necessary and one for which it is not. Each encoder's type is determined by the nature of the decoders it transmits to. Call any decoder that uses side information, or that receives messages from two or more encoders, a *joint* decoder. Call all other decoders *individual* decoders.

A *Type I* encoder transmits messages only to individual decoders. Type I encoders know exactly the reproductions associated with each possible index choice, and hence no conditional expectation is necessary. Optimization of Type I encoders is done via a straightforward modified nearest-neighbor search in the same way as point-to-point encoder optimization.

A *Type II* encoder transmits to one or more joint decoders. Since it lacks some of the information used by the joint decoders, a Type II encoder cannot determine the decoders' reproductions. Its optimization therefore requires a conditional expectation over the unknown messages or side information at each joint decoder. The discussion that follows illustrates the implementation of Type II encoders using two examples. The first is a 2A system with side information at the decoder (the 2AWZ system). The second is a three-node network, which adds the additional complication of having Type II encoders that transmit to more than one joint decoder.

In addition to the implementation of Type II encoders, the 2AWZ example addresses the use of side information at a network decoder and the initialization of the components of a

network system. The discussion generalizes from 2AWZ to arbitrary networks.

3.4.1 Two-User Multiple Access with Side Information (2AWZ)

This section discusses Type II encoder implementation, the use of side information, and initialization methods for the 2AWZ network shown in Figure 3.3. The two encoders $\alpha_2 : \mathcal{X}_{2,1}^n \to \mathcal{I}_{2,1}$ and $\alpha_3 : \mathcal{X}_{3,1}^n \to \mathcal{I}_{3,1}$ operate on sources $X_{2,1}^n$ and $X_{3,1}^n$ respectively to produce indices $i_{2,1}$ and $i_{3,1}$. The decoder $\beta_1 : \mathcal{I}_{2,1} \times \mathcal{I}_{3,1} \times {\mathcal{Z}_1}^n \to \hat{\mathcal{X}}_{2,1,1}^n \times \hat{\mathcal{X}}_{3,1,1}^n$ jointly decodes the corresponding received indices $(\hat{i}_{2,1,1}, \hat{i}_{3,1,1}) = f_1(i_{2,1}, i_{3,1}, z_1^n)$ using side information z_1^n. The decoder reproductions are denoted individually as $\beta_{2,1,1}(\hat{i}_{2,1,1}, \hat{i}_{3,1,1}, z_1^n) = \hat{x}_{2,1,1}^n$ and $\beta_{3,1,1}(\hat{i}_{2,1,1}, \hat{i}_{3,1,1}, z_1^n) = \hat{x}_{3,1,1}^n$. I assume that $(X_{2,1}^n, X_{3,1}^n, Z_1^n)$ are dependent random variables, and, for notational convenience, I use Z_1^n and Z^n interchangeably.

For now, assume that $\mathcal{Z}^n$ is discrete and that $|\mathcal{Z}^n|$ is small, so that the total number of possible decoder codewords, $|\mathcal{I}_{2,1}||\mathcal{I}_{3,1}||\mathcal{Z}^n|$, is significantly smaller than the size of the training set. Later, I discuss how to work with a large or continuous $\mathcal{Z}^n$.

Encoder Implementation

Considering α_2 and rewriting the expectations from (3.5) in terms of sums over the sets $\mathcal{I}_{3,1}$ and $\mathcal{Z}^n$ gives

$$\begin{aligned}\alpha_2^\star(x_{2,1}^n) \;=\; \arg\min_{i_{2,1}} \Bigg[& a_{2,1}\,|\ell_2(i_{2,1})| + \sum_{i_{3,1}} \sum_{z^n} \Bigg(b_{2,1,1} d\left(x_{2,1}^n, \beta_{2,1,1}\left(f_1(i_{2,1}, i_{3,1}, z^n), z^n\right)\right) \\ & + b_{3,1,1} E\left[d(X_{3,1}^n, \beta_{3,1,1}(f_1(i_{2,1}, i_{3,1}, z^n), z^n)) \middle| X_{2,1}^n = x_{2,1}^n, \alpha_3(X_{3,1}^n) = i_{3,1}, Z^n = z^n \right] \Bigg) \\ & \cdot \Pr\left(\alpha_3(X_{3,1}^n) = i_{3,1}, Z^n = z^n \middle| X_{2,1}^n = x_{2,1}^n\right) \Bigg]. \end{aligned} \tag{3.7}$$

An analytical model for the source distribution is generally unavailable, so I estimate $\Pr(\alpha_3(X_{3,1}^n) = i_{3,1}, Z^n = z^n | x_{2,1}^n)$ and the expectation over $X_{3,1}^n$ using the training data. Since the alphabet $\mathcal{X}_{2,1}^n$ is in general very large (e.g., $|\mathcal{X}_{2,1}^4| = 256^4$ for an 8-bit greyscale image and VQ dimension 4), the number of conditional probability terms to be estimated is very large. Given a limited training set size, estimation techniques of possible interest include algorithms using kernels, histograms, and combinations of the two [79]. I here use histograms due to their low computational complexity. I partition $\mathcal{X}_{2,1}^n$ into a finite number of bins and estimate conditional distributions over $\mathcal{I}_{3,1} \times \mathcal{Z}^n$ for each bin. Denote by $\delta_2 : \mathcal{X}_{2,1}^n \to \mathcal{K}_{2,1} = \{1, \ldots, |\mathcal{K}_{2,1}|\}$ the function that maps a sample $x_{2,1}^n$ to the index $k_{2,1}$ of its corresponding histogram bin.

Denote by $\Gamma = \{(x_{2,1}^n, x_{3,1}^n, z^n)\}$ the *list*[7] of training vectors, and define

$$\begin{aligned} \Gamma(k_{2,1}, i_{3,1}, z^n) &= \{(x_{2,1}'^n, x_{3,1}'^n, z'^n) \in \Gamma : \delta_2(x_{2,1}'^n) = k_{2,1}, \alpha_3(x_{3,1}'^n) = i_{3,1}, z'^n = z^n\} \\ \Gamma(k_{2,1}) &= \{(x_{2,1}'^n, x_{3,1}'^n, z'^n) \in \Gamma : \delta_2(x_{2,1}'^n) = k_{2,1}\}. \end{aligned}$$

I estimate $\Pr(\alpha_3(X_{3,1}^n) = i_{3,1}, Z^n = z^n | x_{2,1}^n)$ in (3.7) by replacing the condition on $x_{2,1}^n$ with a condition on $\delta_2(x_{2,1}^n)$, giving

$$\Pr(\alpha_3(X_{3,1}^n) = i_{3,1}, Z^n = z^n | x_{2,1}^n) \approx \Pr(\alpha_3(X_{3,1}^n) = i_{3,1}, Z^n = z^n | \delta_2(x_{2,1}^n)) = \frac{\left|\Gamma(\delta_2(x_{2,1}^n), i_{3,1}, z^n)\right|}{\left|\Gamma(\delta_2(x_{2,1}^n))\right|},$$

which I evaluate from the training data using the current (fixed) α_3. The expectation over $X_{3,1}^n$ is evaluated using the known mappings (from the previous optimization[8]) for all of the training vectors.

[7] Γ is defined as a list rather than a set, because any training vector that appears multiple times in Γ should be counted multiple times in any list size or sum calculation.

[8] All components except α_2 are held fixed from the previous optimization.

Convergence

In the above discussion, I make approximations that allow a significant reduction in the number of conditional distributions to be estimated. These approximations represent a deviation from the optimal encoder as specified by the design equations, and, as a result, convergence of the algorithm is no longer guaranteed (although it is observed in practice for training sets considered). I now show that by altering our cost function, convergence can be guaranteed at the cost of some performance degradation.

Let $K_{2,1} = \delta_2(X_{2,1}^n)$ and $K_{3,1} = \delta_3(X_{3,1}^n)$. Suppose $X_{2,1}^n \rightarrow K_{2,1} \rightarrow K_{3,1} \rightarrow X_{3,1}^n$, and $(X_{2,1}^n, X_{3,1}^n) \rightarrow (K_{2,1}, K_{3,1}) \rightarrow Z^n$ form Markov chains. Then the approximation of conditioning on $k_{2,1} = \delta_2(x_{2,1}^n)$ and $k_{3,1} = \delta_3(x_{3,1}^n)$ becomes exact, and I can implement the optimal encoder exactly. These Markov properties do not hold in general, but by building a probability model in which they are forced to hold, I get a design algorithm that is guaranteed to converge. For any source distribution $P(x_{2,1}^n, x_{3,1}^n, z^n, k_{2,1}, k_{3,1})$, there exists a corresponding distribution $\hat{P}(x_{2,1}^n, x_{3,1}^n, z^n, k_{2,1}, k_{3,1})$ that satisfies the Markov properties, where

$$\hat{P}(x_{2,1}^n, x_{3,1}^n, z^n, k_{2,1}, k_{3,1}) = P(x_{2,1}^n)P(k_{2,1}|x_{2,1}^n)P(k_{3,1}|k_{2,1})P(x_{3,1}^n|k_{3,1})P(z^n|k_{2,1}, k_{3,1}).$$

Define a new cost function $\hat{j}^n(\hat{P}, \mathbf{a}, \mathbf{b}, Q^n)$ that differs from $j^n(P, \mathbf{a}, \mathbf{b}, Q^n)$ in (3.2) in that expectations are taken with respect to $\hat{P}$ rather than P.[9] Thus $\hat{j}^n$ gives the expected system performance with respect to $\hat{P}$, where $\hat{P}$ has the properties we desire. Both the optimal encoders and the optimal decoder for $\hat{j}^n$ can be implemented exactly (in a computationally

[9]In (3.2), expectations are taken over a distribution of the form $P(x_{2,1}^n, x_{3,1}^n, z^n)$. This can easily be extended to the form $P(x_{2,1}^n, x_{3,1}^n, z^n, k_{2,1}, k_{3,1})$, since $k_{2,1}$ and $k_{3,1}$ are deterministic functions of $x_{2,1}^n$ and $x_{3,1}^n$.

$x_{2,1} \backslash x_{3,1}$	$x_{3,1} \leq 0$	$x_{3,1} > 0$
$x_{2,1} > 0$	$(2-\eta)P(x_{2,1})P(x_{3,1})$	$\eta P(x_{2,1})P(x_{3,1})$
$x_{2,1} \leq 0$	$\eta P(x_{2,1})P(x_{3,1})$	$(2-\eta)P(x_{2,1})P(x_{3,1})$

Table 3.1: Application of the Markov constraint to a pair of Gaussian sources.

feasible manner), and hence convergence is guaranteed. However, the code is now optimized with respect to $\hat{P}$ rather than the true distribution P, and it does not perform as well in practice as a code designed with the non-convergent algorithm on P, as shown by the experimental results in Sections 3.5.1 and 3.5.2.

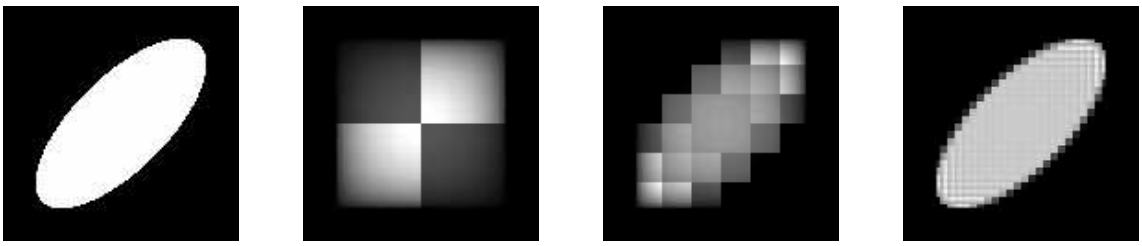
I give two examples to build intuition of how enforcement of the Markov property alters the joint distribution. For simplicity, I omit the side information.

Let vector $(X_{2,1}, X_{3,1})$ be Gaussian with mean 0, variance 1, and correlation ρ, i.e.,

$$P(x_{2,1}, x_{3,1}) = \frac{1}{2\pi(1-\rho^2)} \exp\left(-\frac{x_{2,1}^2 + 2\rho x_{2,1} x_{3,1} + x_{3,1}^2}{2(1-\rho^2)}\right),$$

giving marginals $P(x_{2,1}) = (1/\sqrt{2\pi})\exp(-x_{2,1}^2/2)$ and $P(x_{3,1}) = (1/\sqrt{2\pi})\exp(-x_{3,1}^2/2)$. Consider scalar quantization ($n = 1$) and allocate one bit to each of $k_{2,1}$ and $k_{3,1}$. Choose $\delta_2(x) = \delta_3(x) = 1(x > 0)$ and $\eta = 4\int_0^\infty \int_0^\infty P(x_{2,1}, x_{3,1}) dx_{2,1} dx_{3,1}$, where $1(\cdot)$ is the indicator function. Then $\hat{P}$ for the four possible values of $(k_{2,1}, k_{3,1})$ is given by Table 3.1. It is a weighted product of the marginals of P. The weights, which reflect the correlation, are such that the integral over each quadrant is the same for $\hat{P}$ and P. For independent sources, $\rho = 0$, $\eta = 1$, and hence $P = \hat{P}$. For highly correlated sources with $\rho \approx 1$, $\eta \approx 2$ and $\hat{P}$ smears the positive correlation over the first and third quadrants. The second and fourth quadrants have little or no probability mass, consistent with the original distribution.

In general, $\hat{P}$ is a weighted product of the marginals of P in which the weighting can



(a) (b) (c) (d)

Figure 3.4: A discrete Markov constraint example. (a) The distribution P, uniform over an ellipse. (b,c,d) $\hat{P}$ for $|\mathcal{K}_{2,1}| = |\mathcal{K}_{3,1}| = K$ when: (b) $K = 2^1$, (c) $K = 2^3$, (d) $K = 2^5$.

be different for different values of $(K_{2,1}, K_{3,1})$. As $|\mathcal{K}_{2,1}|$ and $|\mathcal{K}_{3,1}|$ grow, $\hat{P}$ more closely resembles P. Figure 3.4 illustrates this point for a discrete distribution on a square grid with $|\mathcal{X}_{2,1}| = |\mathcal{X}_{3,1}| = 2^7$ ($n = 1$ as before). The original uniform distribution (Figure 3.4(a)) is approximated to greater accuracy (Figure 3.4(b)-(d)) by increasing $|\mathcal{K}_{2,1}|$ and $|\mathcal{K}_{3,1}|$.

For my experiments I use 8-bit greyscale images at dimension 4, with $|\mathcal{X}_{2,1}^4| = |\mathcal{X}_{3,1}^4| = 2^{32}$, and $|\mathcal{K}_{2,1}| = |\mathcal{K}_{3,1}| \approx 2^9$. Thus $\hat{P}$ will only coarsely model the source correlation, but this appears to be sufficient for the sources and rates used in the experiments.

Initialization

The first step in implementing generalized Lloyd design is to initialize the system encoders and decoders. Since iterative descent design can at best give only a locally optimal solution, initialization can have a significant impact on final performance[10] When theory suggests a

[10]A variety of annealing techniques have been applied to traditional VQ design (e.g., [80, 81]) in an attempt to address the local optimality problem. These techniques can be generalized to NVQ design. While several authors have conjectured that these techniques yield global optima, this conjecture remains unproven. Recent work on both point-to-point a restricted class of network VQ problems demonstrates the existence of polynomial-time approximation algorithms that guarantee fixed-rate codes with performance

useful structure for a codebook, this can be made use of in initialization. Here I outline two initialization methods. One is based on point-to-point coding methods and is suitable for weakly correlated sources and side information; the other is based on a binning structure (mirroring the binning structure used to prove the Slepian-Wolf theorem for lossless coding) as described in [71, 82] and is suitable for strongly correlated sources and side information. For both approaches I initialize the entropy codes with the codeword lengths used for fixed-rate coding, and with identity mappings $\{f_r\}$. Consequently, I equate $\hat{i}_{*,r}$ with $i_{*,r}$ in the following discussion on quantizer initialization.

In the point-to-point method, I design a codebook with cells convex with respect to each of $i_{2,1}$ and $i_{3,1}$. I begin by designing a point-to-point VQ for each of $X_{2,1}^n$ and $X_{3,1}^n$. The network encoders α_2 and α_3 are initialized as the corresponding point-to-point encoders. If there were no side information, I could construct the joint decoder by simply taking the cross product of the two point-to-point codebooks. With side information, $|\mathcal{Z}^n|$ initial codewords must be specified for each $(i_{2,1}, i_{3,1})$ pair. Using the point-to-point encoders, partition the training set into lists $\Gamma(i_{2,1}, i_{3,1}, z^n)$, where

$$\Gamma(i_{2,1}, i_{3,1}, z^n) = \{(x_{2,1}'^n, x_{3,1}'^n, z'^n) \in \Gamma : \alpha_2(x_{2,1}'^n) = i_{2,1}, \alpha_3(x_{3,1}'^n) = i_{3,1}, z'^n = z^n\}.$$

Initialize the decoder β_1 by setting each $\beta_{2,1,1}(i_{2,1}, i_{3,1}, z^n)$ and $\beta_{3,1,1}(i_{2,1}, i_{3,1}, z^n)$ to be the centroids (with respect to the appropriate distortion measures) of the codewords from the list $\Gamma(i_{2,1}, i_{3,1}, z^n)$.

For sources highly correlated with the side information at the decoders, I use a binning structure. The resulting codebook has cells that are non-convex for a given $i_{2,1}$ and $i_{3,1}$:

within a factor $(1+\epsilon)$ of the optimal (see [65] and the references therein). The NVQ results given there postdate this work.

non-contiguous regions of one source alphabet can be quantized to the same index, and the decoder relies on the other source and the side information to distinguish the correct region. The set of non-contiguous regions assigned to a particular index is called a *coset*. For the WZ system, in which source $x_{3,1}^n$ does not appear, I create a binning structure with $2^{r_c} \leq |\mathcal{Z}^n|$ cosets starting with a quantizer that maps each z^n into one of 2^{r_c} different values and a lattice-based codebook for source $x_{2,1}^n$ at rate $R_{2,1}$. Translate the lattice by small amounts (less than or equal to half the distance between adjacent lattice points) in 2^{r_c} different directions; the images of a lattice point under the different translations are allocated to different cosets. Each individual coset consists of a translated copy of the original lattice[11]. For an MA code, the base lattice is formed from a cross product of lattices for each of the individual sources.

While the design algorithm can produce optimized VQs with a binning structure, I have not observed it to do so experimentally if a binning structure was not used in initialization.

Detailed Side Information

The previous discussion assumes $|\mathcal{Z}^n|$ is small. When $|\mathcal{Z}^n|$ is large or $\mathcal{Z}$ is continuous, the number of decoder codewords required prohibits a practical implementation as above. I solve this problem by coarsely quantizing the side information before it is given to the decoder. The quantized side information is denoted by an index $k_Z \in \mathcal{K}_Z = \{1, 2, \ldots, |\mathcal{K}_Z|\}$, and I redefine β_1 so that $\beta_1 : \mathcal{I}_{2,1} \times \mathcal{I}_{3,1} \times \mathcal{K}_Z \rightarrow \hat{\mathcal{X}}_{2,1,1}^n \times \hat{\mathcal{X}}_{3,1,1}^n$ instead of $\beta_1 : \mathcal{I}_{2,1} \times \mathcal{I}_{3,1} \times \mathcal{Z}^n \rightarrow$

[11] An alternative to obtaining the cosets by translation is to encode the source using the original lattice, then partition the training vectors mapped to a particular lattice point into 2^{r_c} sets using their quantized z^n values and initialize as described in the low correlation method.

$\hat{\mathcal{X}}^n_{2,1,1} \times \hat{\mathcal{X}}^n_{3,1,1}$. I allow a different quantization of $\mathcal{Z}^n$ for each pair of received indices $(\hat{i}_{2,1}, \hat{i}_{3,1})$, and denote by $\delta_Z : \mathcal{I}_{2,1} \times \mathcal{I}_{3,1} \times \mathcal{Z}^n \to \mathcal{K}_Z$ the quantizer encoder that determines k_Z given received index pair $(\hat{i}_{2,1}, \hat{i}_{3,1})$ and side information z^n. The quantizer codewords are denoted $\{\phi_Z(\hat{i}_{2,1}, \hat{i}_{3,1}, k_Z)\}_{k_Z=1}^{K_Z}$, and are initialized by clustering on the list of training vectors

$$\Gamma(i_{2,1}, i_{3,1}) = \{(x'^n_{2,1}, x'^n_{3,1}, z'^n) : \alpha_2(x'^n_{2,1}) = i_{2,1}, \alpha_3(x'^n_{3,1}) = i_{3,1}, z'^n \in \mathcal{Z}^n\}.$$

I create $|\mathcal{K}_Z|$ clusters from this list, number the clusters from 1 to K_Z, and set the decoder codewords $\beta_{2,1,1}(\hat{i}_{2,1}, \hat{i}_{3,1}, k_Z)$ and $\beta_{3,1,1}(\hat{i}_{2,1}, \hat{i}_{3,1}, k_Z)$ as the centroids of the appropriate cluster.

Quantizing the side information is a practical rather than an optimal strategy. However, assuming that the joint source-side information distribution is reasonably smooth, then provided the quantization is of a significantly higher rate than that used for the source messages, essentially all of the correlation between the source messages and side information is captured. The experimental results of Section 3.5.1 suggest that on practical data sets, $|\mathcal{K}_Z|$ (and hence the number of decoder codewords) can be kept small while paying little or no penalty in rate-distortion performance.

3.4.2 A General Three-Node Network

This section uses the example of a general three-node network to discuss the implementation of a Type II encoder that transmits to more than one other node.

Consider the implementation of the design conditions for encoder 2 for the three-node network shown in Figure 3.2. Encoder 2 participates in two 2AWZ subsystems: it cooperates with encoder 3 to send information to decoder 1 and with encoder 1 to send information to decoder 3. Since the indices chosen by encoder 2 affect reproductions at both nodes 1 and 3,

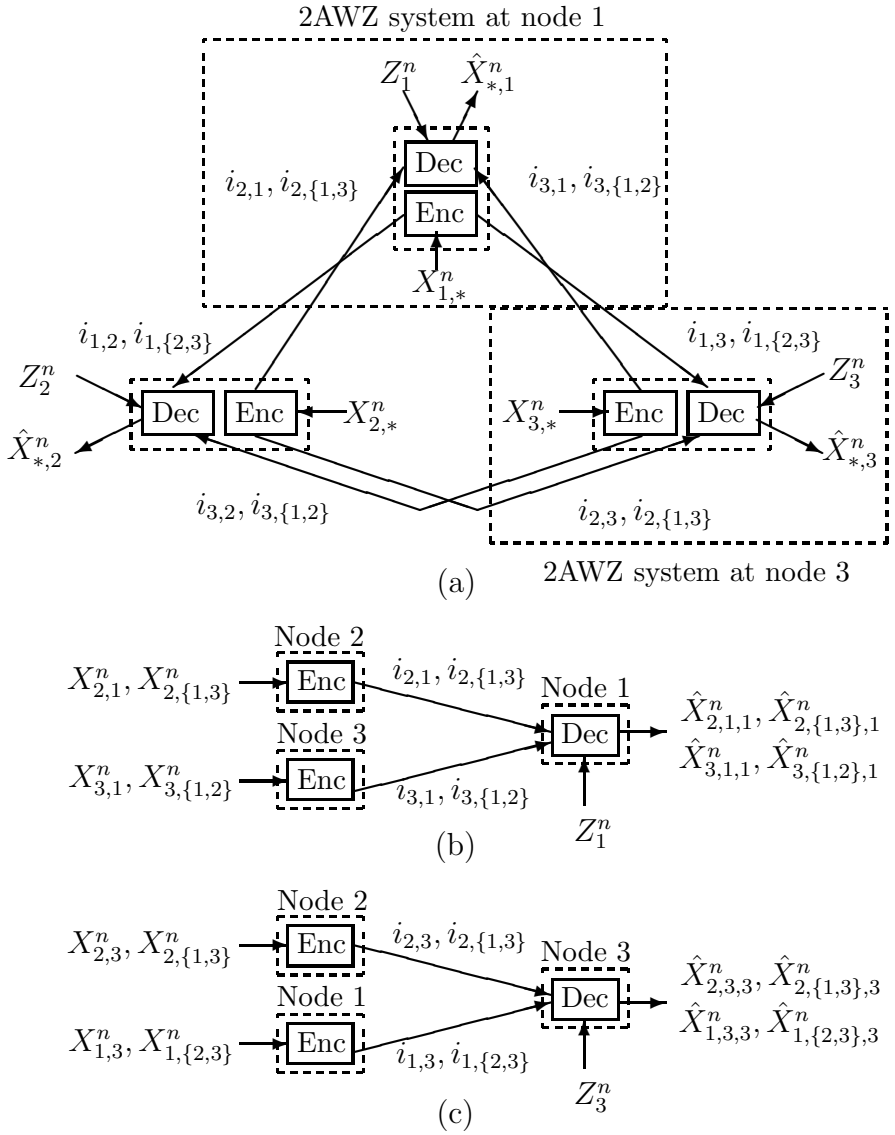


Figure 3.5: Optimal encoding at node 2. (a) The estimated performance for a given index set $i_{2,*}$ can be found by summing the performance in two linked 2AWZ subsystems. (b),(c) The two subsystems.

as shown in Figure 3.5, the distortion terms for both decoders must be evaluated in a single minimization. From the design equations, $\alpha_2^\star$ is given by

$$\begin{aligned}\alpha_2^\star(x_{2,*}^n) = \arg \min_{i_{2,*} \in \mathcal{I}_{2,*}} \Bigg(& \sum_{S \in \mathcal{S}(2)} a_{2,S} \ell_{2,S}(i_{2,*}) \\ & + \sum_{(t,S) \in \mathcal{T}(1)} b_{t,S,1} E\left[d(X_{t,S}^n, \beta_{t,S,1}(\hat{I}_{*,1}, Z_1^n)) \middle| X_{2,*}^n = x_{2,*}^n, I_{2,*} = i_{2,*} \right] \\ & + \sum_{(t,S) \in \mathcal{T}(3)} b_{t,S,3} E\left[d(X_{t,S}^n, \beta_{t,S,3}(\hat{I}_{*,3}, Z_3^n)) \middle| X_{2,*}^n = x_{2,*}^n, I_{2,*} = i_{2,*} \right] \Bigg), \end{aligned} \tag{3.8}$$

where

$$\begin{aligned} \mathcal{S}(2) &= \{\{1,3\},1,3\}, \\ \mathcal{T}(1) &= \{(2,\{1,3\}),(2,1),(3,\{1,2\}),(3,1)\}, \\ \mathcal{T}(3) &= \{(1,\{2,3\}),(1,3),(2,\{1,3\}),(2,3)\}. \\ \hat{I}_{*,1} &= f_1\left((\alpha_{t',S'}(X_{t',*}^n)_{(t',S') \in \mathcal{T}(1)}, Z_1^n\right) \\ \hat{I}_{*,3} &= f_3\left((\alpha_{t',S'}(X_{t',*}^n)_{(t',S') \in \mathcal{T}(3)}, Z_3^n\right). \end{aligned}$$

All terms in (3.8) are of similar form to those for the 2AWZ system, and the same approximation methods can be used to evaluate the conditional expectations. In general, any Type II encoder sending to more than one other node can be designed using the same approach for sending to only one other node; there are just more distortion and rate terms to evaluate.

3.5 Experimental Results

In this section I build NVQs for different network systems and present experimental results. I discuss three systems in detail: the 2AWZ and general three-node networks described in

Section 3.4, and the MD network introduced in Section 3.1. A closely related discussion of broadcast VQ can be found in [9].

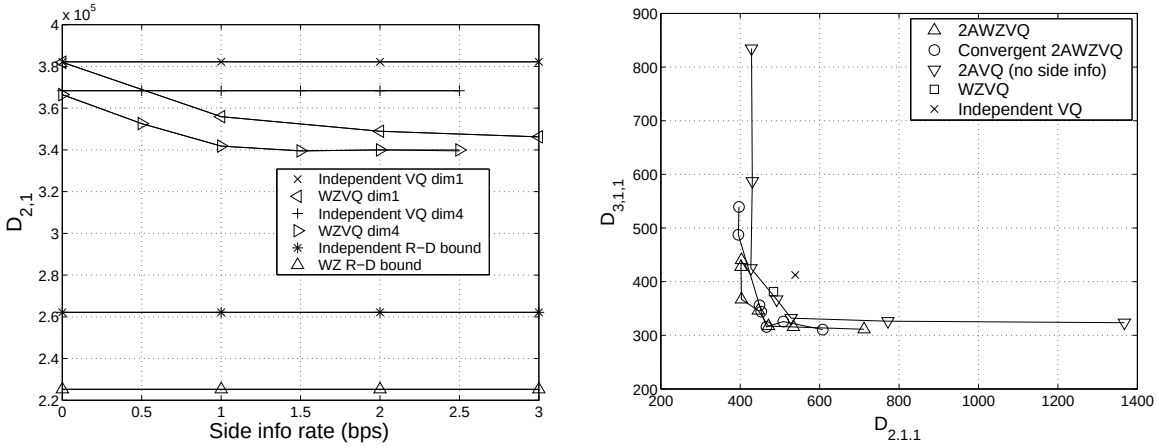
For each of the three systems I give a brief introduction, a discussion of entropy codeword length selection for entropy-constrained coding, and experimental results. The experiments compare the performance of NVQs to that of independent VQs and to available rate-distortion bounds. Additionally, I examine the scalability of network codes using an MD network and a ring network as examples. All experiments use the squared-error (MSE) distortion measure.

3.5.1 The 2AWZ Network

The previous section treats this system in detail, and I therefore skip its introduction.

There are currently no practical codes for the 2AWZ system, nor provably optimal theoretical codeword lengths. However, there are practical codes for the 2A system [78]. I perform both 2AWZ and 2A experiments, the former at fixed-rate only, the later at both fixed- and variable-rate. The variable-rate 2A design uses the near-lossless codes from [78], and the non-zero contribution from coding errors is included in the reported distortion.

I conduct four experiments. The first studies the use and impact of side information in a WZ system, removing one of the users of the 2AWZ system for simplicity of result presentation. The second compares network to independent coding performance on the full 2AWZ system. The third compares fixed- and variable rate coding performance on the 2A system. The fourth investigates the benefit of initializing the decoder to have a binning structure, again using a WZ system for result simplicity.



(a) (b)

Figure 3.6: (a) WZVQ performance as a function of side information rate $\log_2(|\mathcal{K}_Z|)/n$ for jointly Gaussian source and side information ($\rho = 0.375$). (b) Various coding performances for the 2AWZ system on satellite data.

For the first (WZ) experiment, I generate i.i.d. jointly Gaussian data with correlation $\rho = 0.375$ between the source and the side information. I quantize the side information as discussed in Section 3.4. I use fixed-rate codes of rate 1 bit per sample (bps) and vary the vector dimension n and the number $|\mathcal{K}_Z|$ of values used to quantize the side information. For good practical code performance I want $|\mathcal{K}_Z|$ high to make full use of the correlation between the message and the side information. However, I must limit $|\mathcal{K}_Z|$ to limit the number of decoder codewords and hence the design complexity. I compare the performance of WZVQs and independent VQs to two rate-distortion bounds; one is the WZ R-D bound, the other is the point-to-point R-D bound (which uses no side information).

The distortion for the different codes is shown in Figure 3.6(a). The R-D bounds are independent of $|\mathcal{K}_Z|$ and are plotted for comparative purposes; they show the optimal theoretically achievable performance for any vector dimension n, and, in the WZ R-D case,

arbitrarily high $|\mathcal{K}_Z|$. The results show that the use of side information in the WZ codes improves performance by approximately 0.4 dB and bridges 20% of the gap in distortion between independent coding and the WZ R-D bound at dimension 4 and correlation 0.375. The gain is even higher at dimension 1. For both dimensions it is of comparable size to the difference in the two R-D bounds, suggesting that the WZ codes are making efficient use of the side information. The results also show that almost all of the benefit of side information is captured with a low value of $|\mathcal{K}_Z|$, validating quantization of the side information as a method for reducing design complexity.

For the second experiment, I train and test various fixed-rate codes for the full 2AWZ system. I use satellite weather images for the data set[12]. All codes use vector dimension 4, rate 0.75 bps, and Lagrangian weights $a_{2,1} = a_{3,1} = 0$, $b_{2,1,1} + b_{3,1,1} = 1$. I use $|\mathcal{K}_Z| = 16$ different values to quantize the side information.

Figure 3.6(b) shows a plot of distortions $D_{2,1,1}$ and $D_{3,1,1}$ for the various coding techniques. For any choice of Lagrangian weights, independent code design yields the same code and hence contributes a single point to the graph. For network code design, the Lagrangian weights trade off the importance of the two reproductions and yield different codes. I display the performance of network codes both with (2AWZVQ) and without (2AVQ) the use of side information. For the 2AWZVQ codes, I include results obtained using the convergent as well as the non-convergent algorithm. I also show the performance achieved by using two separately decoded WZVQs, one for each source. The 2AWZVQs show a gain of at least 1.17dB in each reproduction over the independent VQs. This gain arises from both the joint decoding of messages and the use of side information; the distortions achieved by the 2AVQs

[12]This data set is described in detail in Appendix A.

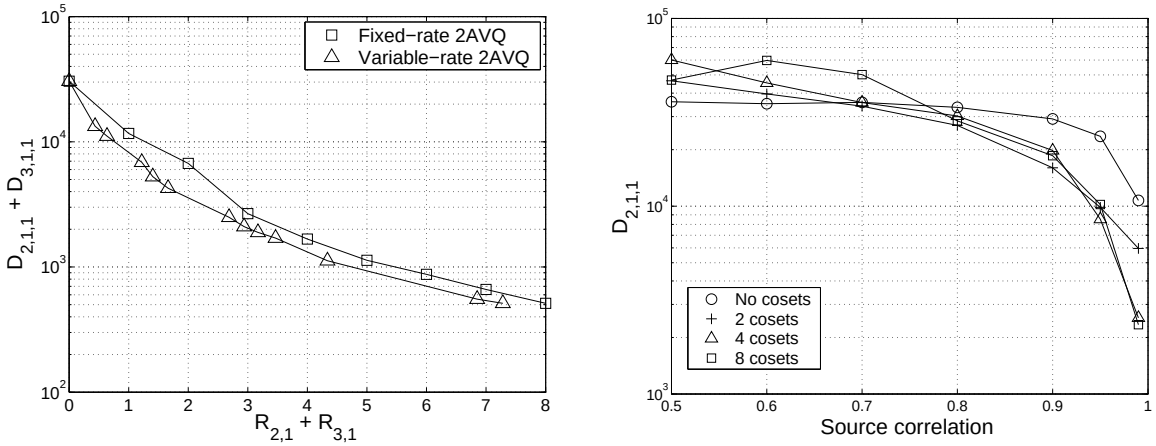


Figure 3.7: (a) Comparison of fixed- and variable-rate coding performances for the 2A system. (b) WZ code performance as a function of source and side information correlation and the number of cosets used in decoder initialization.

and by the WZVQs suggest that both contributions are significant. The results obtained by the convergent and non-convergent algorithms are similar on this data set.

The third experiment compares the performance of fixed- and variable-rate codes for the 2A system. I again use satellite weather images for the data and vector dimension 4. I set the Lagrangian weights $b_{2,1,1} = b_{3,1,1} = 1$ and $a_{2,1} = a_{3,1} = a$, where a is varied to produce codes targeting different points on the convex hull of the achievable rate-distortion region. For variable-rate design I use the codeword lengths and mappings of real near-lossless 2A codes. Figure 3.7(a) shows the sum $D_{2,1,1} + D_{3,1,1}$ of the two distortions as a function of the total rate $R_{2,1} + R_{3,1}$ for several fixed- and variable-rate codes. The variable-rate codes consistently outperform their fixed-rate counterparts by 1.2 dB.

The fourth experiment uses WZ codes to investigate the benefits of initializing with a binning structure when the source and side information are highly correlated. The source

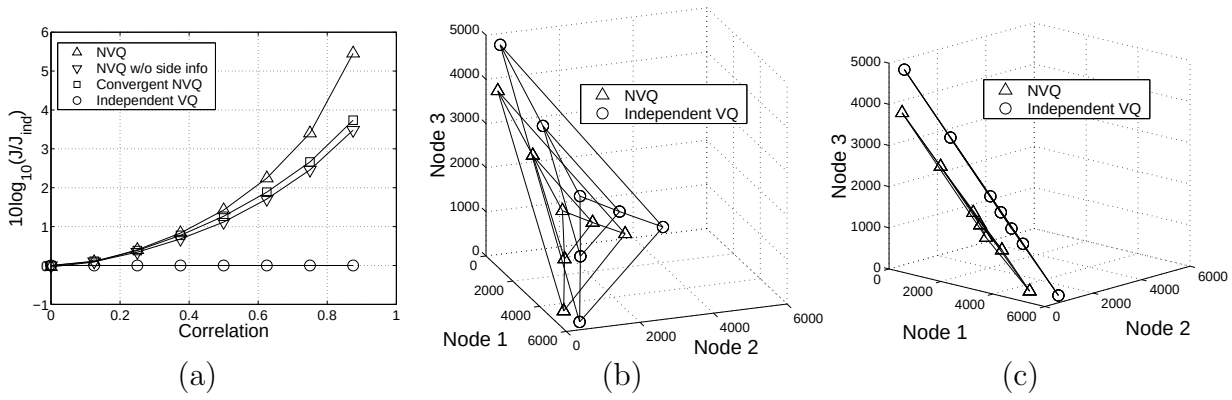
and side information are i.i.d. jointly Gaussian, with mean 0, variance 1, and correlation ρ. I use fixed-rate codes of dimension 1, $|\mathcal{K}_Z| = 64$ different values to quantize the side information, and initialize the decoder with 2^{r_c} cosets, $r_c \in \{0, 1, 2, 3\}$. Figure 3.7 shows the performance obtained by different codes as a function of ρ. For low correlations, such as $\rho = 0.5$, a binning structure significantly hampers performance, and the training optimization removes the binning structure as best it can. The final performance is similar to that of the code with no binning. For high correlation, performance is significantly improved using a binning-structured decoder. The desired number of cosets increases with the correlation.

3.5.2 A General Three-Node Network

Section 3.4 includes a detailed introduction of general three-node networks. Optimal entropy codes and coding bounds are currently unavailable for the general three-node network. I perform all of the experiments using fixed-rate codes.

I conduct two experiments for the general three-node network. The first shows the efficiency of NVQs as a function of inter-source correlation; the second compares the trade-off in performance at each node as a function of the Lagrangian weights controlling the optimization. All experiments show fixed-rate coding results at vector dimension 4 and rate 0.5 bps for each of the nine sources. The side information used by the decoder at each node consists of the three sources to be encoded at that node and is quantized to $|\mathcal{K}_Z| = 16$ levels.

The performance gain of NVQs over independent VQs is a function of inter-source correlation. Figure 3.8(a) shows a plot of the Lagrangian cost (3.2) in dB as a function of correlation for i.i.d. jointly Gaussian sources. For each sample, the correlation between any



(a) (b) (c)

Figure 3.8: Efficiency of network source coding vs. independent coding. (a) Overall performance as a function of correlation. (b),(c) Weighted sum distortion at each node as a function of the Lagrangian parameters (shown from two different angles).

two sources has a constant value ρ. Equal weighting is given to each reproduction. As the correlation between sources increases, the performance gain of NVQs over independent VQs increases significantly, exceeding 2 dB for $\rho \geq 0.6$. Also, unlike the 2AWZ experiments, for this system and data set the alterations required to ensure convergent design do impair performance.

Figures 3.8(b,c) show the weighted sum of the distortions at each of the three nodes using the satellite data set. Varying the Lagrangian weights $\{b_{t,S,r}\}$ traces out the surface shown from two different angles in the figures. (I constrain the Lagrangian weights to keep all weights corresponding to reproductions at the same node equal.) The surface corresponding to the NVQs lies approximately 1 dB closer to the origin than that of independently VQs, indicating an average of 1 dB improvement over the set of source reproductions at each node.

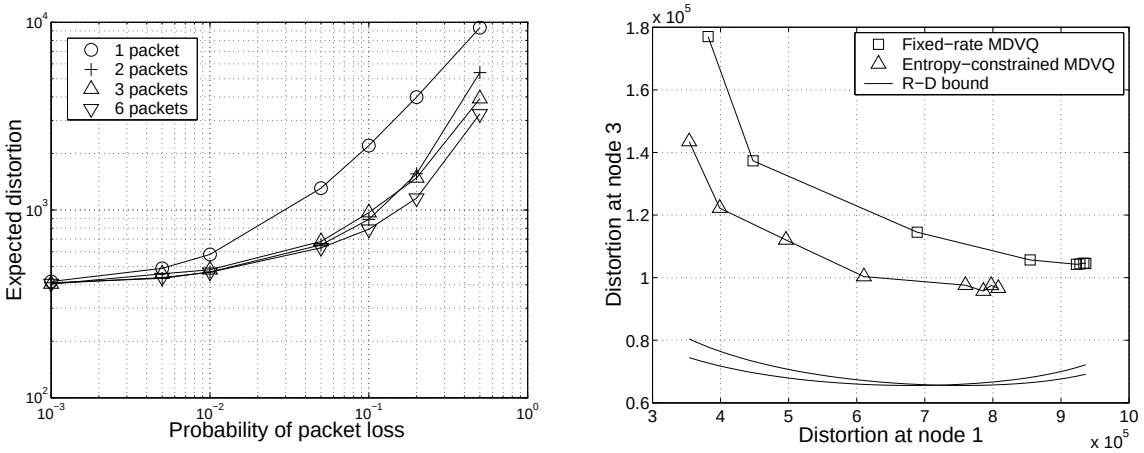
3.5.3 A Multiple Description System

An MD system transmitting descriptions over K unreliable channels can give rise to $2^K - 1$ non-trivial sets of received descriptions. Figure 3.1(e) casts the system into a network model by treating the decoder for each of the $2^K - 1$ non-trivial sets as a separate node in the network. I label the encoder with index $M = 2^K$ and each decoder with the integer representation of a binary vector $\mathbf{e} = (e_1, \ldots, e_K)$, where $e_k = 1$ if channel k is operational and 0 otherwise. The K channel descriptions now correspond to K network messages; network message k is received by all decoders that have $e_k = 1$. Although some decoders receive two or more descriptions, each outputs only one reproduction since all descriptions are of the same original source. The system contains only one encoder and no side information, so the encoder is of Type I and can be implemented without approximations.

Since the K descriptions must be individually decodable (for each description there is some decoder that receives that and only that description) the optimal coding bound for description k is the entropy $H(I_k)$ of the index used for message k. This bound can be approximated in practice by entropy codes (e.g., arithmetic codes), and I associate with each index i_k the optimal average length, $-\log_2 \Pr(I_k = i_k)$.

Given the probability of each $\mathbf{e} \in \{0, 1\}^K$, the expected distortion of our code is minimized by setting the weight on each reproduction's distortion to be the probability of receiving exactly the set of descriptions used to make that reproduction. I then adjust the relative sizes of the weights on the description rates to achieve our desired code rates.

Two experiments demonstrate the performance of MD codes. In the first, I use the satellite weather data set to train and test fixed-rate MDVQs. Each code uses a different



(a) (b)

Figure 3.9: (a) MDVQ performance on the satellite data as a function of the number of descriptions per vector and the channel failure probability. (b) Fixed-rate and entropy-constrained MDVQ performance on Gaussian data compared to the D-R bound.

number of descriptions to encode each four-dimensional data vector, but all have the same total encoded bit-rate of 6 bits per vector. The MDVQs considered use: one six-bit description, two three-bit descriptions, three two-bit descriptions, and six one-bit descriptions, respectively. For each code, I transmit different descriptions of the same vector on different channels, and I assume that the different channels all have equal failure rates. Figure 3.9(a) shows the expected reproduction distortion as a function of channel failure probability and the number of descriptions used. Moving from a single description (as in traditional coding) to two descriptions greatly slows the degradation in performance as a function of channel failure probability. Using more than two descriptions yields even better performance.

The second experiment compares fixed-rate and entropy-constrained two-description, 4-dimensional MDVQ performance to the distortion-rate bound using i.i.d. Gaussian data. I choose a rate of 1 bps per description and design codes for different probabilities of channel

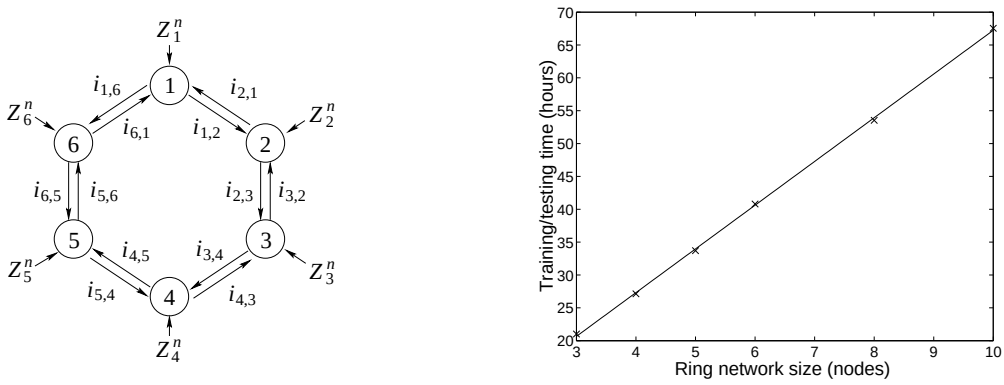
failure. Each code is characterized by the distortions it achieves at the three decoders. However, for all of the codes designed in this experiment I found the distortion at node 2 to be almost constant, so Figure 3.9(b) simply plots the distortion at node 3 against that at node 1. I also plot the distortion-rate bound for the code rate used in the experiment. The bound depends on node 2's distortion, which varied very slightly over the results; the two lines defining the bound correspond to the smallest and largest values of node 2's observed distortion. The results demonstrate the reduction in distortion achieved by variable-rate compared to fixed-rate coding. This reduction varies from 0.5 dB for low channel failure probability to 1.3 dB for high channel failure probability.

3.5.4 Network Scalability

The complexity of network code design depends on the following factors.

The number of codewords Assigning every codeword a weight equal to the number of encoders that access that codeword, design complexity is linear in the sum weight of all network codewords. A node's in-degree is the number of codewords transmitting to that node. Network design is linear in the number of nodes M when the in-degree of each node is kept constant and exponential in M when in-degree grows linearly with M.

The size of the training set Code design complexity is linear in the size of the training set, which must be large enough to ensure that each encoder's histogram estimation of the data's joint distribution is accurate. The number of histogram bins required by an encoder transmitting to node r is linear in the number of codewords at node r, hence the number of training vectors needed is linear in the number of codewords at each node. In



(a) (b)

Figure 3.10: (a) The network messages and side informations of a six-node ring network. (b) Design time for a ring network.

addition, the size of each training set vector is linear in the number of network sources. If the in-degree of each node is constant as M increases, the required training set size will increase linearly, but if the in-degree of each node increases with M, the training set size must increase exponentially.

Fixed- vs. variable-rate design Training complexity is roughly the same for both fixed- and variable-rate design. However, if the network rates must meet specific constraints, then the Lagrangian parameters in variable-rate design must be optimized appropriately. Using a conjugate gradient approach, this increases design complexity approximately by a factor equal to the number of Lagrangian parameters. Symmetry in a network can be exploited to constrain the parameter optimization.

Figure 3.10(a) shows a ring network in which each node communicates with its two neighbors. This is an example of a network in which the in-degree of each node remains constant as M increases; the design time therefore increases linearly with M as evidenced by the experimental design times shown in Figure 3.10(b). The design times are for fixed-rate

Network	Codewords per decoder	Total weight of codewords
MA	2^{2M}	$M2^{2M}$
M-receiver limited BC	16	$16M$
M-receiver full BC	2^{2^M}	$M2^{2^M}$
M-node ring	256	$512M$
General M-node	$2^{(M-1)2^{M-1}+4}$	$M(M-1)2^{(M-1)2^{M-1}+4}$

Table 3.2: Total number of codewords for various systems.

quantizer design on a 1GHz Intel Xeon processor at vector dimension 4, with 4 bits per message and 4 bit side information quantization. Table 3.2 below indicates approximate design complexity for several networks by counting the total weight of network codewords. I assume a vector dimension of 4, with 2 bits per network message and 4 bits for side information quantization. Two types of BC network are considered: a limited one in which there is private information for each individual receiver, but only one common information (for all receivers), and a full one in which every subset of receivers will receive a different common information. The table shows that my design algorithm is not suitable for large MA, fully-connected BC, and fully-connected general networks. However, almost no real networks will be so connected as to have nodes that transmit a separate message to every possible subset of other nodes. Practical networks would be much more likely to follow a model such as the limited BC or the M-node ring, for which the design complexity scales linearly with M and for which my algorithm is appropriate. The exponential increase in design complexity for MA networks is a concern; suboptimal design techniques would need to be adopted for large M, such as dividing the nodes into fixed-sized groups and jointly

decoding each group separately.

Once a network has been designed, encoding and decoding complexity can be made very low if desired. Approximating the optimal encoders using hierarchical coding allows both encoding and decoding to be done via table lookup [76].

3.6 Summary

I extend the algorithm presented in [1] to design locally optimal vector quantizers for general networks. The extensions allow the use of side information at the decoders and allow for design in the presence of channel errors. Both fixed- and variable-rate VQ design are considered. For some network systems, variable-rate VQ design is complicated either by the fact that the theoretically optimal codeword lengths for the entropy code are unknown, or by the absence of good techniques for designing practical codes to approximate the optimal performance. In these cases I optimize relative to the best available bounds on the entropy code's codeword lengths.

I also provide a much-needed discussion of how to implementation the algorithm in practice, a topic considered only sparingly in [1]. I show that the primary difficulty in implementation is the evaluation of conditional expectations required to design the optimal encoders for a network with joint decoders. I provide approximations to reduce the computational complexity of evaluating the expectations and also to reliably estimate the joint statistics of the training data. Making these approximations removes the guarantee of convergence in iterative code design. In practice, however, I do observe convergence, which suggests that the approximations are reasonable. When required, I show how to ensure convergence at

some cost in rate-distortion performance.

My NVQ experiments demonstrate the performance improvements that network-based design yields over independent design. When applied to a satellite weather data set, 2AWZ and three-node network codes both show distortion improvements of more than 1 dB over independent coding. This increase results from the ability of network-designed codes to exploit the redundancy between the different sources in the network; point-to-point design treats every source as independent and thus does not take advantage of this type of redundancy. For networks where sources are highly correlated, such as sensor networks, network-based coding can be significantly more efficient.

Chapter 4

Rate-Distortion with Mixed Side Information

4.1 Introduction

Side information is often available to improve the rate-distortion performance of data compression codes. For example, consider an environmental remote sensing network with several sensors, each of which takes measurements and transmits them to a central base station, which also makes its own measurements. In encoding its transmission to the base station, each sensor can consider the measurements taken by the base station as side information available to the base station's decoder. If the system uses multi-hop transmissions, then measurements relayed by a sensor act as side information available both to that sensor's encoder and the base station's decoder.

Figure 4.1(a) shows the conditional rate-distortion system in which side information is available at both the encoder and decoder. Figure 4.1(b) shows the Wyner-Ziv system [12,

13], in which side information is available only at the decoder. Combining the two types of side information into one system yields the system shown in Figure 4.1(c), which I call the *mixed side information* (MSI) system.

The multi-hop sensor network considered above provides a simple example in which both types of side information are present. Another example comes from the system of Heegard and Berger [2] and shown in Figure 4.1(d), in which the presence of side information at the decoder is unreliable. The system requires two decoders, one for the case when side information is present (decoder 1) and the other for when it is absent (decoder 2). We can approach coding for this system using a two-part source description. The first part is decoded without side information and ensures a minimum reproduction fidelity at both decoders. The second part requires side information Z for its decoding and serves as refinement information at decoder 1. It is not useful to decoder 2. Once the first part is chosen, it can be viewed as side information, known to both the encoder and decoder, for the coding of the second part. Thus the coding of the second part is a mixed side information problem[1].

In this chapter, I consider rate-distortion theory for the MSI system, since solution of examples like the MSI system is a prerequisite of the solution of more general problems. A simple observation allows the derivation of the MSI rate-distortion function directly from the Wyner-Ziv system. I use that result to generalize Zamir's rate loss result and Wyner's Gaussian example from the Wyner-Ziv system. I then solve a new binary example that expands significantly on the corresponding example for the Wyner-Ziv system and apply the

[1]The work in this chapter was initially motivated by a desire to close the gap in the bounds on the binary-source rate-distortion example proposed by Heegard and Berger in [2] and considered further in [28]. In this chapter I provide a solution that does indeed close that gap.

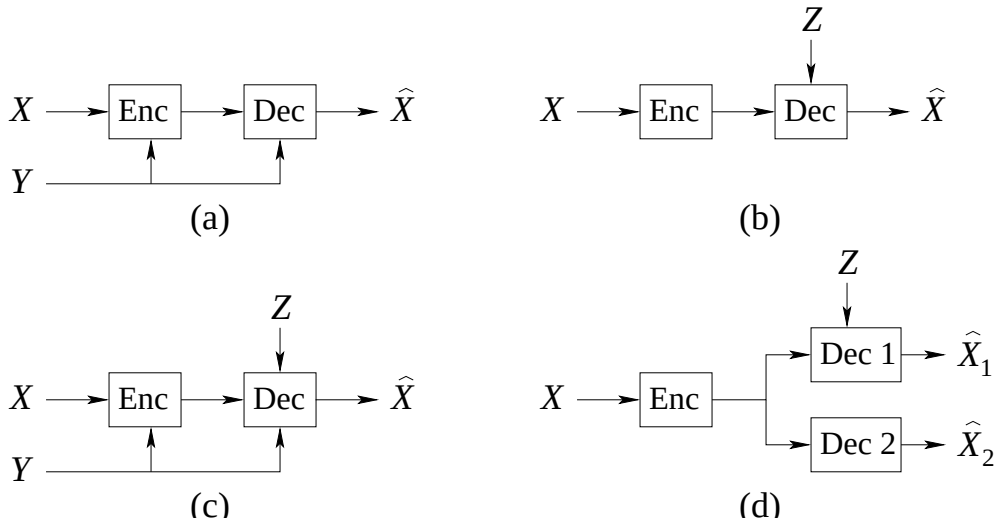


Figure 4.1: (a) The conditional rate-distortion system. (b) The Wyner-Ziv system. (c) The MSI system. (d) Heegard and Berger's system.

result to solve the corresponding and more complicated binary example in the Heegard and Berger system.

4.2 $R(D)$ for the Mixed Side Information System

This section defines notation, derives the rate-distortion function for the MSI system, and bounds the system's rate loss.

Let X, Y, and Z be a triple of random variables with alphabets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, respectively, and with joint distribution $p(x, y, z)$. I assume $I(X; Y, Z) < \infty$. Let $\hat{\mathcal{X}}$ be a reconstruction alphabet and let $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0, \infty)$ be a distortion measure. An (n, M, D) code for the MSI system consists of an encoder α^n and decoder β^n,

$$\begin{aligned} \alpha^n &: \mathcal{X}^n \times \mathcal{Y}^n \to \{1, 2, \ldots, M\} \\ \beta^n &: \{1, 2, \ldots, M\} \times \mathcal{Y}^n \times \mathcal{Z}^n \to \hat{\mathcal{X}}^n, \end{aligned}$$

such that $E\frac{1}{n}d(X_k, \hat{X}_k) \leq D$, where $\hat{X}^n = \beta^n(\alpha^n(X^n, Y^n), Y^n, Z^n)$. A rate R is said to be D-admissible if for every $\epsilon > 0$ there exists for some n an $(n, M, D + \epsilon)$ code with $n^{-1} \log M \leq R + \epsilon$.

The MSI rate-distortion function is defined as

$$R_{X|Y\{Z\}}(p, D) = \inf\{R : R \text{ is } D\text{-admissible}\}.$$

The set notation in the subscript denotes side information available only at the decoder. Following this pattern, the conditional rate-distortion function is $R_{X|Y}(p, D)$ and the Wyner Ziv rate-distortion function is $R_{X|\{Z\}}(p, D)$.

As in [13], I impose the following two conditions on d:

1. For all $\hat{x} \in \hat{\mathcal{X}}$, $Ed(X, \hat{x}) < \infty$.

2. For all random variables $\hat{X}$ such that $0 < Ed(X, \hat{X}) < \infty$, and all $\epsilon > 0$, there exists a finite subset $\{\hat{x}_1, \ldots, \hat{x}_N\} \subseteq \hat{\mathcal{X}}$, and a quantizer $f_Q : \hat{\mathcal{X}} \to \{\hat{X}_i\}$ such that $Ed(X, f_Q(\hat{X})) \leq (1 + \epsilon)Ed(X, \hat{X})$.

Condition 2 is a smoothness constraint used in generalizing the Wyner-Ziv rate-distortion proof from discrete to continuous alphabets [13]. Wyner notes that it is not especially restrictive, showing that when $\mathcal{X} = \mathbb{R}$ it holds for all r-th power distortion measures, $d(x, \hat{x}) = |x - \hat{x}|^r$ with $r > 0$.

Theorem 14 below gives an information-theoretic characterization of $R_{X|Y\{Z\}}(p, D)$.

Theorem 14 *:*

$$\begin{aligned} R_{X|Y\{Z\}}(p, D) &= \inf_{W \in \mathcal{M}_{X|Y\{Z\}}(p,D)} I(X; W|Y, Z) \\ &= \inf_{W \in \mathcal{M}_{X|Y\{Z\}}(p,D)} \left[I(X; W|Y) - I(W; Z|Y)\right], \end{aligned} \tag{4.1}$$

where $\mathcal{M}_{X|Y\{Z\}}(p, D)$ *is the set of all random variables* W *described by a test channel* $\mu(w|x, y)$ *with the property* $W \to (X, Y) \to Z$ *and for which there exists an* $f : \mathcal{W} \times \mathcal{Y} \times \mathcal{Z} \to \hat{\mathcal{X}}$ *such that*

$$\int\int\int\int p(x, y, z)\mu(w|x, y)d(x, f(w, y, z))dwdxdydz \;\leq\; D.$$

If the alphabets $\mathcal{X}$, $\mathcal{Y}$, *and* $\mathcal{Z}$ *are finite, then the infimum becomes a minimum and it suffices to consider in that minimum only those* W *with* $|\mathcal{W}| \leq |\mathcal{X}||\mathcal{Y}| + 1$.

The theorem is proved first for finite-alphabet sources and then modified to apply to discrete and continuous sources.

Proof of Theorem 14 For any pair of finite-alphabet random variables $A \in \mathcal{A}$ and $B \in \mathcal{B}$ with a joint distribution $p(a, b)$ such that $I(A; B) < \infty$, and any distortion measure d' satisfying conditions 1 and 2, the Wyner-Ziv rate-distortion function $R_{A|\{B\}}(p, D)$ is given by [13] as:

$$R_{A|\{B\}}(p, D) = \inf_{W \in \mathcal{M}_{A|\{B\}}(p,D)} [I(A; W) - I(W; B)] \tag{4.2}$$

$$= \inf_{W \in \mathcal{M}_{A|\{B\}}(p,D)} I(A; W|B), \tag{4.3}$$

where $\mathcal{M}_{A|\{B\}}(p, D)$ is the set of all random variables W described by a test channel $\mu(w|a)$ with the property $W \to A \to B$, and for which there exists an $f : \mathcal{W} \times \mathcal{B} \to \hat{\mathcal{A}}$ such that

$$\int\int\int p(a, b)\mu(w|a)d'(a, f(w, b))dwdadb \leq D.$$

Choose $A = (X, Y)$ and $B = (Y, Z)$, and let d' have the form $d'(A, \hat{A}) = d(X, \hat{X})$, where d satisfies conditions 1 and 2 so that d' also does. With these substitutions, $R_{A|\{B\}}(p, D)$ is the rate-distortion function for a system in which Y is both a source and a side information,

i.e., Y is known to both the encoder and the decoder. Under the distortion measure d', which measures only the distortion of X and ignores that of Y, the system is equivalent to the MSI system of Figure 4.1(c) with distortion measure d. Thus

$$\begin{aligned} R_{X|Y\{Z\}}(p,D) &= \inf_{W\in\mathcal{M}_{A|\{B\}}(p,D)} I(A;W|B) \\ &= \inf_{W\in\mathcal{M}_{A|\{B\}}(p,D)} I(X,Y;W|Y,Z) \\ &= \inf_{W\in\mathcal{M}_{A|\{B\}}(p,D)} I(X;W|Y,Z), \end{aligned} \tag{4.4}$$

where $\mathcal{M}_{A|\{B\}}(p,D)$ is the set of all random variables W described by a test channel $\mu(w|a) = \mu(w|x,y)$ with the property $W \to A \to B$ (which is equivalent to $W \to (X,Y) \to Z$), and for which there exists an $f(w,b) = f(w,y,z)$ such that

$$\int\int\int p(a,b)\mu(w|a)d'(a,f(w,a,b))dwdadb \;\leq\; D,$$

which is equivalent to

$$\int\int\int\int p(x,y,z)\mu(w|x,y)d(x,f(w,y,z))dwdxdydz \;\leq\; D.$$

Observe that $\mathcal{M}_{A|\{B\}}(p,D) = \mathcal{M}_{X|Y\{Z\}}(p,D)$, and hence (4.4) can be written

$$R_{X|Y\{Z\}}(p,D) = \inf_{W\in\mathcal{M}_{X|Y\{Z\}}(p,D)} I(X;W|Y,Z).$$

For finite-alphabet sources, the infimum in (4.3) becomes a minimum [12]; propagating this result through the argument above shows that it also applies to the MSI rate-distortion function. The bound on the cardinality of the auxiliary random variable W is derived using the support lemma of Ahlswede and Körner [83, Lemma 3]. The support lemma implies that W needs at least $|\mathcal{X}||\mathcal{Y}| - 1$ letters to preserve $p(x,y|w)$, plus two more to preserve $I(X;W|Y,Z)$ and $Ed(X,f(W,Z))$, giving a total of $|\mathcal{X}||\mathcal{Y}| + 1$.

For discrete and continuous sources, the above proof must be modified. Application of the Wyner-Ziv result as above requires $I(A;B) < \infty$, but for $A = (X,Y)$ and $B = (Y,Z)$ we have $I(A;B) = I(X,Y;Y,Z) > I(Y;Y)$, which may not be finite if Y does not have a finite alphabet. However, the requirement $I(A;B) < \infty$ is used by Wyner to establish his converse; the achievability is still applicable. Thus, I use the same approach as above to prove achievability, and prove the converse directly as shown below, requiring now only that $I(X;Y,Z) < \infty$.

Consider any MSI rate-distortion code (α^n, β^n). Let the ith reproduced symbol be denoted by $\beta_i^n : \{1,\ldots,2^{nR}\} \times \mathcal{Y}^n \times \mathcal{Z}^n \to \hat{\mathcal{X}}$. Let $T = \alpha^n(X^n, Y^n)$ denote the encoded version of X^n when the side information available to both encoder and decoder is Y^n. Let X_i denote the ith component of X, X^{i-1} denote $(X_1,\ldots,X_{i-1})$, and X_{i+1}^n denote $(X_{i+1},\ldots,X_n)$. Then

$$
\begin{aligned}
nR &\overset{(a)}{\geq} H(T) \\
&\geq H(T|Y^n,Z^n) \\
&\geq I(X^n;T|Y^n,Z^n) \\
&\overset{(b)}{=} I(X^n;T,Y^n,Z^n) - I(X^n;Y^n,Z^n) \\
&\overset{(c)}{=} \sum_{i=1}^{n} \left[I(X_i;T,Y^n,Z^n|X_1^{i-1}) - I(X_i;Y_i,Z_i) \right] \\
&= \sum_{i=1}^{n} \left[I(X_i;T,Y^n,Z^n,X_1^{i-1}) - I(X_i;X_1^{i-1}) - I(X_i;Y_i,Z_i) \right] \\
&\overset{(d)}{=} \sum_{i=1}^{n} \left[I(X_i;T,Y^n,Z^n,X_1^{i-1}) - I(X_i;Y_i,Z_i) \right] \\
&\geq \sum_{i=1}^{n} \left[I(X_i;T,Y^n,Z^n) - I(X_i;Y_i,Z_i) \right] \\
&\overset{(e)}{=} \sum_{i=1}^{n} \left[I(X_i;W_i,Y_i,Z_i) - I(X_i;Y_i,Z_i) \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} [I(W_i; X_i|Y_i, Z_i)] \\
&= \sum_{i=1}^{n} [I(W_i; X_i, Z_i|Y_i) - I(W_i; Z_i|Y_i)] \\
&= \sum_{i=1}^{n} [I(W_i; X_i|Y_i) + I(W_i; Z_i|X_i, Y_i) - I(W_i; Z_i|Y_i)] \\
&= \sum_{i=1}^{n} [I(W_i; X_i|Y_i) - I(W_i; Z_i|Y_i)] \\
&\overset{(f)}{\geq} \sum_{i=1}^{n} R_{X|Y\{Z\}}(p, Ed(X_i, \beta_i^n(W_i, Y_i, Z_i))) \\
&\overset{(g)}{\geq} nR_{X|Y\{Z\}}\left(p, E\frac{1}{n}\sum_{i=1}^{n} d(X_i, \beta_i^n(W_i, Y_i, Z_i))\right) \\
&\geq nR_{X|Y\{Z\}}(p, D),
\end{aligned}
$$

where the labeled steps are justified by the following.

(a) T can take at most 2^{nR} distinct values.

(b) The mutual informations here are well-defined due to the assumption $I(X; Y, Z) < \infty$, which implies $I(X^n; Y^n, Z^n) < \infty$.

(c) (X, Y, Z) is i.i.d. and hence $I(X^n; Y^n, Z^n) = \sum_{i=1}^{n} I(X_i; Y_i, Z_i)$

(d) X is i.i.d. and hence $I(X_i; X_1^{i-1}) = 0$.

(e) By define $W_i = (T, Y_1^{i-1}, Y_{i+1}^n, Z_1^{i-1}, Z_{i+1}^n)$. Since (X, Y, Z) are i.i.d. and $T = f(X_i, Y_i)$, W_i does not contain any information about Z_i that is not already in (X_i, Y_i) and $W_i \to (X_i, Y_i) \to Z_i$ forms a Markov chain.

(f) By (4.1) and the fact that, since (X, Y, Z) is i.i.d., $p = p(x_i, y_i, z_i)$ is independent of i. Also, since β_i^n is a function of (T, Y^n, Z^n), it can also be written as a function of (W_i, Y_i, Z_i).

(g) From the convexity of $R_{X_i|Y_i\{Z_i\}}(Ed(X_i, \beta_i^n(W_i, Y_i, Z_i)))$ (this can be shown using the techniques of [31, Lemma 14.9.1]) □

The following theorem decomposes the rate-distortion function over different values of the side information Y. The proof parallels Gray's proof [4] for the discrete conditional rate-distortion function and is omitted.

Theorem 15 *Let $R_{X|y\{Z\}}(p, D)$ denote the rate-distortion function for the MSI system when $Y = y$ is constant. Then*

$$R_{X|Y\{Z\}}(p, D) = \inf_{\{D_y\}\in\mathcal{D}(p,D)} \int_y R_{X|y\{Z\}}(p, D_y)p(y)dy,$$

where

$$\mathcal{D}(p, D) = \left\{ \{D_y, y \in \mathcal{Y}\} : \int_y D_y p(y)dy \leq D \right\}.$$

The minimum on the right hand side is achieved when the D_y are chosen so that the rate-distortion functions $R_{X|y\{Z\}}(p, D_y)$, $y \in \mathcal{Y}$, all have the same slope at their respective distortions D_y.

Theorem 15 makes rigorous the intuition that we can code distinctly for each value of y and that the distortion at different y values should differ so that all rate-distortion curves operate at points of equal slope.

Zamir [39] defines the rate loss for the Wyner-Ziv system as the difference between the Wyner-Ziv and the conditional rate-distortion functions, $L_{X|\{Z\}}(p, D) = R_{X|\{Z\}}(p, D) - R_{X|Z}(p, D)$. Extending this definition, define the rate loss for the MSI system as the difference between the MSI and conditional rate-distortion functions

$$L_{X|Y\{Z\}}(p, D) \stackrel{\triangle}{=} R_{X|Y\{Z\}}(p, D) - R_{X|YZ}(p, D).$$

The rate loss bound derived by Zamir for the Wyner-Ziv system [39] applies unchanged to the mixed side information system via an argument parallel to that in [39]. For a continuous

source and the r-th power distortion measure, the bound is a constant, independent of the source distribution. For example, for continuous alphabet sources and squared-error distortion, $L(p, D) \leq \frac{1}{2}$ for all (p, D). This rate loss bound shows that the penalty paid for Y not being available at the encoder cannot be arbitrarily large. It also provides a way to bound $R_{X|Y\{Z\}}$ when its direct computation is difficult but that of $R_{X|YZ}$ is straightforward.

4.3 Joint Gaussian Sources

Wyner showed in [13] that for a Gaussian source and the squared-error distortion measure, the rate-distortion function $R_{X|\{Z\}}(p, D)$ for the Wyner-Ziv system is equal to the conditional rate-distortion function $R_{X|Z}(p, D)$; the rate-distortion function is the same whether the side information is available at the encoder or not. I generalize this result for the MSI system. Once again, no penalty in rate need be paid even though some side information is not available at the encoder.

The conditional rate-distortion function $R_{X|YZ}(p, D)$ is defined as

$$R_{X|YZ}(p, D) = \inf_{\hat{X} \in \mathcal{M}_{X|YZ}(p,D)} I(X; \hat{X}|Y, Z),$$

where $\mathcal{M}_{X|YZ}(p, D)$ is the set of random variables $\hat{X}$ described by a test channel $\mu(\hat{x}|x, y, z)$ such that $Ed(X, \hat{X}) \leq D$. Returning to the MSI system, let $W \in \mathcal{M}_{X|Y\{Z\}}(p, D)$, let f be the decoding function such that $Ed(X, f(W, Y, Z)) \leq D$, and let $\hat{X} = f(W, Y, Z)$. The data processing inequality and the observation that $W \in \mathcal{M}_{X|Y\{Z\}}(p, D)$ and $\hat{X} = f(W, Y, Z)$ together imply $\hat{X} \in \mathcal{M}_{X|YZ}(p, D)$. Thus

$$I(X; W|Y, Z) \geq I(X; \hat{X}|Y, Z) \geq R_{X|YZ}(p, D).$$

Minimizing with respect to $W \in \mathcal{M}_{X|Y\{Z\}}(p, D)$, we obtain

$$R_{X|Y\{Z\}}(p, D) \geq R_{X|YZ}(p, D),$$

with equality if and only if the $\hat{X}$ achieving the minimum in the definition of $R_{X|YZ}(p, D)$ can be related to the W and f achieving the minimum in the definition of $R_{X|Y\{Z\}}(p, D)$ via $\hat{X} = f(W, Y, Z)$ with $I(X; W|Y, Z) = I(X; \hat{X}|Y, Z)$. This requires $I(X; W|\hat{X}, Y, Z) = 0$.

Now consider a zero-mean, jointly Gaussian random variable (X, Y, Z) with $EX^2 = \sigma_X^2$. Denote by K the covariance matrix of (X, Y, Z), and let $L = (l_{ij}) = K^{-1}$. Given Y and Z, X is Gaussian with conditional mean and variance given by

$$E[X|Y, Z] = -\frac{l_{12}}{l_{11}}Y - \frac{l_{13}}{l_{11}}Z, \qquad Var[X|Y, Z] = \frac{1}{l_{11}}.$$

For this source, I show that $R_{X|YZ}(p, D)$ and $R_{X|Y\{Z\}}(p, D)$ are given by

$$R_{X|YZ}(p, D) \;=\; R_{X|Y\{Z\}}(p, D) = \begin{cases} \frac{1}{2}\log\frac{1}{l_{11}D}, & 0 < D < \frac{1}{l_{11}} \\ 0, & D \geq \frac{1}{l_{11}}. \end{cases}$$

Figure 4.2(a) shows a random variable that achieves $I(X; \hat{X}|Y, Z) = \frac{1}{2}\log 1/(l_{11}D)$. The test channel of Figure 4.2(a) is consistent with an application of Theorem 15; for each $y \in Y$, it is the same channel used by Wyner [13]. The conditional variance is the same for all $y \in \mathcal{Y}$, suggesting that $D_y = D$ for all y; this does indeed yield the optimal result. Redrawing Figure 4.2(a) as shown in Figure 4.2(b), we have

$$\hat{X} = f(W, Y, Z) = W - (1 - a)\left(\frac{l_{12}}{l_{11}}Y + \frac{l_{13}}{l_{11}}Z\right),$$

where $W = a(X + N)$ implies $W \to (X, Y) \to Z$. Together, $\hat{X}$, Y, and Z allow calculation of W. Thus $I(X; W|\hat{X}, Y, Z) = 0$, which permits $R_{X|Y\{Z\}}(p, D) = R_{X|YZ}(p, D)$.

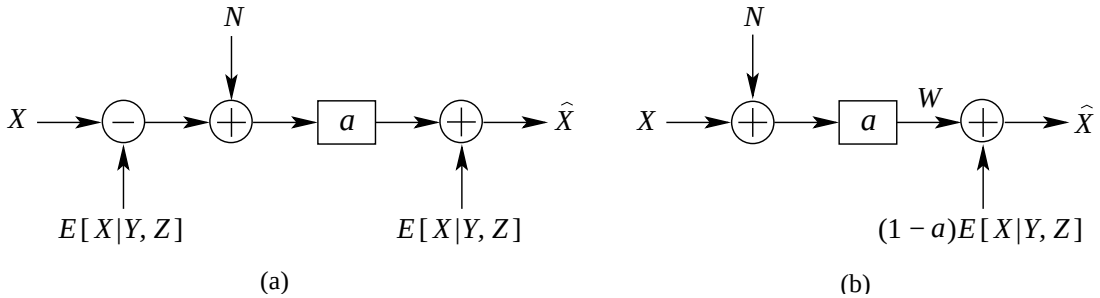


Figure 4.2: A distribution achieving $R_{X|YZ}(p, D)$. Here $a = 1 - Dl_{11}$ and N is Gaussian noise, independent of (X, Y, Z), with mean zero and variance $D/(1 - Dl_{11})$.

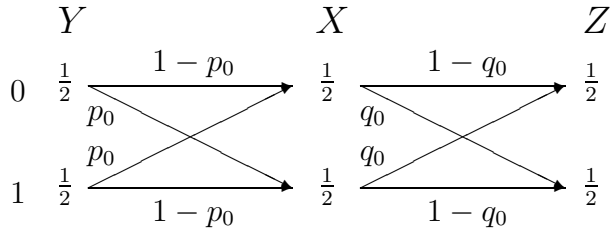


Figure 4.3: Joint distribution of (X, Y, Z) for binary MSI example.

4.4 Joint Binary Sources

Let X, Y, and Z be binary sources, and let Y and Z be related to X via binary symmetric channels, as shown in Figure 4.3. All three variables have marginals of $(\frac{1}{2}, \frac{1}{2})$, and the crossover probabilities of the channels are $p_0 < \frac{1}{2}$ and $q_0 < \frac{1}{2}$ as shown. Denote by $p(x, y, z)$ the joint distribution, and adopt the Hamming distortion measure.

For this problem, Theorem 15 gives

$$R_{X|Y\{Z\}}(D) \;=\; \frac{1}{2}R_{X|Y=0\{Z\}}(D_0) + \frac{1}{2}R_{X|Y=1\{Z\}}(D_1)$$

for some D_0 and D_1 such that $\frac{1}{2}D_0 + \frac{1}{2}D_1 = D$. By symmetry, $D_1 = D_2 = D$, and $R_{X|Y=0\{Z\}}(D_1) = R_{X|Y=1\{Z\}}(D_2)$. Thus $R_{X|Y\{Z\}}(D) = R_{X|Y=0\{Z\}}(D) = R_{X|\{Z\}}(q, D)$, where $q(x, z) = p(x, z|y = 0)$. From here on, I concentrate on finding $R_{X|\{Z\}}(q, D)$. The problem is

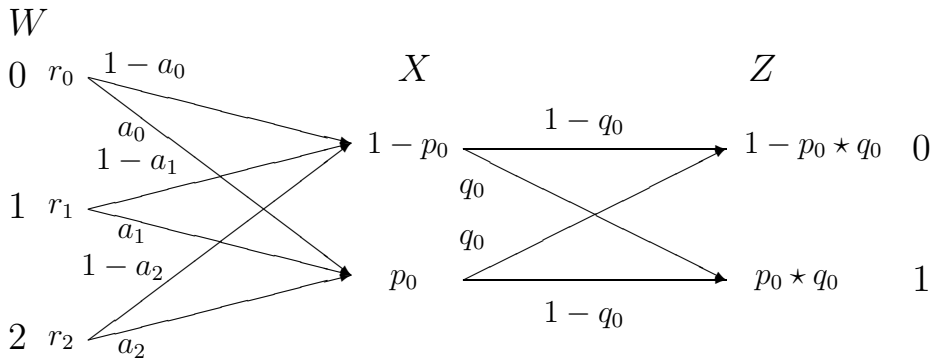


Figure 4.4: Joint distribution of (W, X, Z) for $|\mathcal{W}| \leq 3$.

now similar to the binary symmetric example solved in [12], except that the marginals on X and Z are now skewed: the marginal on X is $(1-p_0, p_0)$, and that on Z is $(1-p_0 \star q_0, p_0 \star q_0)$, where $a \star b \stackrel{\triangle}{=} a(1-b) + b(1-a)$.

By the cardinality bound on the auxiliary random variable [12], the minimum in the definition of $R_{X|\{Z\}}(q, D)$ need consider only those W with $|\mathcal{W}| \leq 3$. The general form of such a W is shown in Figure 4.4. From that figure, we have

$$\begin{aligned} I(X;W) - I(W;Z) &= H(X) - H(X|W) - H(Z) + H(Z|W) \\ &= H(X) - H(Z) + \sum_{i=0}^{2} \Pr(W=w) \left[H(Z|W=w) - H(X|W=w) \right] \\ &= H(p_0) - H(p_0 \star q_0) + \sum_{w=0}^{2} r_w \left[H(a_w \star q_0) - H(a_w) \right], \\ &= -G(p_0) + \sum_{w=0}^{2} r_w G(a_w), \end{aligned} \tag{4.5}$$

where $G(u) \stackrel{\triangle}{=} H(u \star q_0) - H(u)$. The parameters $(a_0, a_1, a_2, r_0, r_1, r_2)$ obey the constraints $0 \leq a_w \leq 1$, $r_w \geq 0$ for all $w \in \{0, 1, 2\}$, and also

$$\begin{aligned} r_0 + r_1 + r_2 &= 1 \\ r_0 a_0 + r_1 a_1 + r_2 a_2 &= p_0. \end{aligned} \tag{4.6}$$

Let $\mathcal{F} = \{f : \mathcal{W} \times \mathcal{Z} \to \{0,1\} : |\mathcal{W}| = 3, |\mathcal{Z}| = 2\}$. Then

$$R_{X|\{Z\}}(q, D) = \min_{f \in \mathcal{F}} R_{X|\{Z\}}(q, D, f), \tag{4.7}$$

where

$$R_{X|\{Z\}}(q, D, f) = \min_{W \in \mathcal{M}_{X|\{Z\}}(q,D,f)} \left[-G(p_0) + \sum_{w=0}^{2} r_w G(a_w) \right], \tag{4.8}$$

and $\mathcal{M}_{X|\{Z\}}(p, D, f)$ is the set of all random variables W with $|\mathcal{W}| = 3$ described by a test channel $\mu(w|x)$ with the property $W \to X \to Z$ and for which

$$\sum_{w=0}^{2} \sum_{x=0}^{1} \sum_{z=0}^{1} q(x, z)\mu(w|x)d(x, f(w, z)) \leq D.$$

To compute the rate-distortion function, I consider each possible decoder $f \in \mathcal{F}$ in turn, and evaluate the minimization over $W \in \mathcal{M}_{X|\{Z\}}(p, D, f)$ for each.

For a particular f and W, the expected distortion of the system is given by

$$\begin{aligned} Ed(X, f(W, Z)) &= \sum_{w=0}^{2} \sum_{z=0}^{1} q(w, z) E\left[d(X, f(W, Z))|W = w, Z = z\right] \\ &= \sum_{w=0}^{2} \sum_{z=0}^{1} r_w q(z|w) \Pr(X \neq f(w, z)). \end{aligned}$$

For each symbol, there are four possible choices for the decoding rule $f(w, \cdot)$. These are shown in Table 4.1, together with $r_w q(z|w) \Pr(X \neq f(w, z))$, their corresponding contribution to the expected distortion. Since $q_0 < \frac{1}{2}$, we have $r_w q_0 < r_w(1 - q_0)$, implying that any decoder with $f(w, 0) = 1$ and $f(w, 1) = 0$ for some w can never be optimal; for any such decoder the expected distortion is always lowered by setting $f(w, 0) = 0$ and $f(w, 1) = 1$. Therefore, we need not further consider decoders with $f(w, 0) = 1$ and $f(w, 1) = 0$ for any w. For all other decoders, Table 4.1 gives the distortion constraint as a function of r_w and a_w. Table 4.2 gives an example; the corresponding distortion constraint is $r_0 a_0 + r_1 a_1 + r_2 q_0 \leq D$.

$f(w,0)$	$f(w,1)$	$r_w q(z\|w) \Pr(X \neq f(w,z))$
0	0	$r_w a_w$
0	1	$r_w q_0$
1	0	$r_w(1-q_0)$
1	1	$r_w(1-a_w)$

Table 4.1: Possible decoding functions for each symbol, together with their expected distortion contribution.

w	$f(w,0)$	$f(w,1)$
0	0	0
1	0	0
2	0	1

Table 4.2: A possible decoding function f when $|\mathcal{W}| = 3$.

Lemma 3 below shows that for any f, the distortion constraint is tight at all points of interest on the $R_{X|\{Z\}}(q, D, f)$ curve. Thus, I can restrict my attention to test channels that meet the distortion constraint with equality.

Lemma 3 *Consider a finite-alphabet Wyner-Ziv system with source X, side information Z, and decoding function f. Let $D_{max} \stackrel{\triangle}{=} \min\{D : R_{X|\{Z\}}(q, D, f) = 0\}$. Then for all $D \leq D_{max}$, the minimum over all test channels $\mu(w|x) \in \mathcal{M}_{X|\{Z\}}(q, D, f)$ in the definition of $R_{X|\{Z\}}(q, D, f)$ can be replaced by a minimum over the subset of test channels in $\mathcal{M}_{X|\{Z\}}(q, D, f)$ for which*

$$\sum_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} \sum_{z \in \mathcal{Z}} q(x, z)\mu(w|x)d(x, f(w, z)) = D.$$

Proof of Lemma 3 For any D_1 and D_2 such that $D_1 \leq D_2$,

$$\mathcal{M}_{X|\{Z\}}(q, D_1, f) \subseteq \mathcal{M}_{X|\{Z\}}(q, D_2, f).$$

Combining this with a timesharing argument, we have that $R_{X|\{Z\}}(q, D, f)$ is convex in D and that $R_{X|\{Z\}}(q, D, f)$ is strictly decreasing for all D such that $R_{X|\{Z\}}(q, D, f) > 0$. Thus, for any $0 < D \leq D_{max}$ and any $\epsilon > 0$, $R_{X|\{Z\}}(q, D, f) < R_{X|\{Z\}}(q, D - \epsilon, f)$, i.e.,

$$\min_{\mu(w|x) \in \mathcal{M}_{X|\{Z\}}(q,D,f)} I(X; W|Z) \quad < \quad \min_{\mu(w|x) \in \mathcal{M}_{X|\{Z\}}(q,D-\epsilon,f)} I(X; W|Z).$$

Since $\mathcal{M}_{X|\{Z\}}(q, D - \epsilon, f) \subseteq \mathcal{M}_{X|\{Z\}}(q, D, f)$, the W attaining the minimum on the left hand side must be in $\mathcal{M}_{X|\{Z\}}(q, D, f) - \mathcal{M}_{X|\{Z\}}(q, D - \epsilon, f)$, and such a W must achieve an expected distortion $D(W)$ satisfying $D - \epsilon < D(W) \leq D$. The result then follows from ϵ arbitrary. □

To compute the rate-distortion function, consider each decoding function f in turn. For the f from Table 4.2, the problem is to minimize

$$g(a_0, a_1, a_2, r_0, r_1, r_2) \stackrel{\triangle}{=} r_0 G(a_0) + r_1 G(a_1) + r_2 G(a_2) \tag{4.9}$$

over all $(a_0, a_1, a_2, r_0, r_1, r_2)$ that satisfy

$$\begin{aligned} r_0 + r_1 + r_2 - 1 &= 0 && (4.10)\\ r_0 a_0 + r_1 a_1 + r_2 a_2 - p_0 &= 0 && (4.11)\\ r_0 a_0 + r_1 a_1 + r_2 q_0 &= D && (4.12)\\ r_w &\geq 0, \quad i \in \{0,1,2\} && (4.13)\\ 0 \leq a_w &\leq 1, \quad i \in \{0,1,2\}. && (4.14) \end{aligned}$$

Since G is convex [12], the function $g(r_0, r_1, r_2, a_0, a_1, a_2)$ is convex in each of its parameters. Thus, there can be only one local extreme value, and, if it exists, it is the global minimum. The three equality constraints (4.10,4.11,4.12) allow a reduction of the number of unsolved parameters from six to three. I reduce the search space further using insights obtained by applying Lagrange multipliers to the optimization; details are provided in Appendix B. The resulting numerical solution for $R_{X|\{Z\}}(q, D, f)$ leaves at most one free parameter, as in the solution of the binary example given by Wyner and Ziv. After finding $R_{X|\{Z\}}(q, D, f)$ for each f, $R_{X|\{Z\}}(q, D)$ is given by (4.7).

Figure 4.5 summarizes the form of the optimal solution for various values of p_0 and q_0 when $D = 0.1$. The results for other values of D are qualitatively the same, but the region for which $R = 0$ grows as D grows.

When both p_0 and q_0 are close to $D = 0.1$, only two symbols are required. Symbol one

conveys, "set $\hat{X} = 0$," and symbol two, "set $\hat{X}$ using the best estimate obtained from the side information." Symbol two costs little rate to describe and gives an expected distortion $Ed(X, \hat{X}) = \min(p_0, q_0) > D$. Symbol one complements symbol two by allowing us to occasionally describe the (skewed) source at a higher quality. As both p_0 and q_0 increase, the distortion constraint becomes tighter, and we soon require a third symbol, "set $\hat{X} = 1$." When both p_0 and q_0 are large, symbol two drops out of use since a reproduction based on the side information has high expected distortion.

Wyner and Ziv's solution for a symmetric marginal on X is a special case of the three-symbol solution in which $r_0 = r_2$.

4.5 Heegard and Berger's System

In [2], Heegard and Berger pose a binary rate-distortion problem for the system of Figure 4.1(d). Choosing X and Z to be symmetric binary sources, they relate the two via a binary symmetric channel of crossover probability q_0 and derive an upper bound on the rate-distortion function. They conjecture that this bound is tight. In [28], Kerpez shows that their bound is loose and provides new upper and lower bounds. In this section, I use the insights gained from the MSI system to show how to compute directly the rate-distortion function for this example, closing the gap between the existing bounds.

The rate-distortion function for the Heegard and Berger (HB) system is

$$R_{HB}(D_1, D_2) = \min_{(U,V) \in \mathcal{M}_{HB}(D_1, D_2)} \left[I(X;U) + I(X;V|U,Z) \right],$$

where $\mathcal{M}_{HB}(D_1, D_2)$ is the set of auxiliary random variables (U, V) such that $(U, V) \rightarrow X \rightarrow Z$ and there exist reproduction functions $\hat{X}_1 = f_1(U, V, Z)$ and $\hat{X}_2 = f_2(U)$ such that $\hat{X}_1$

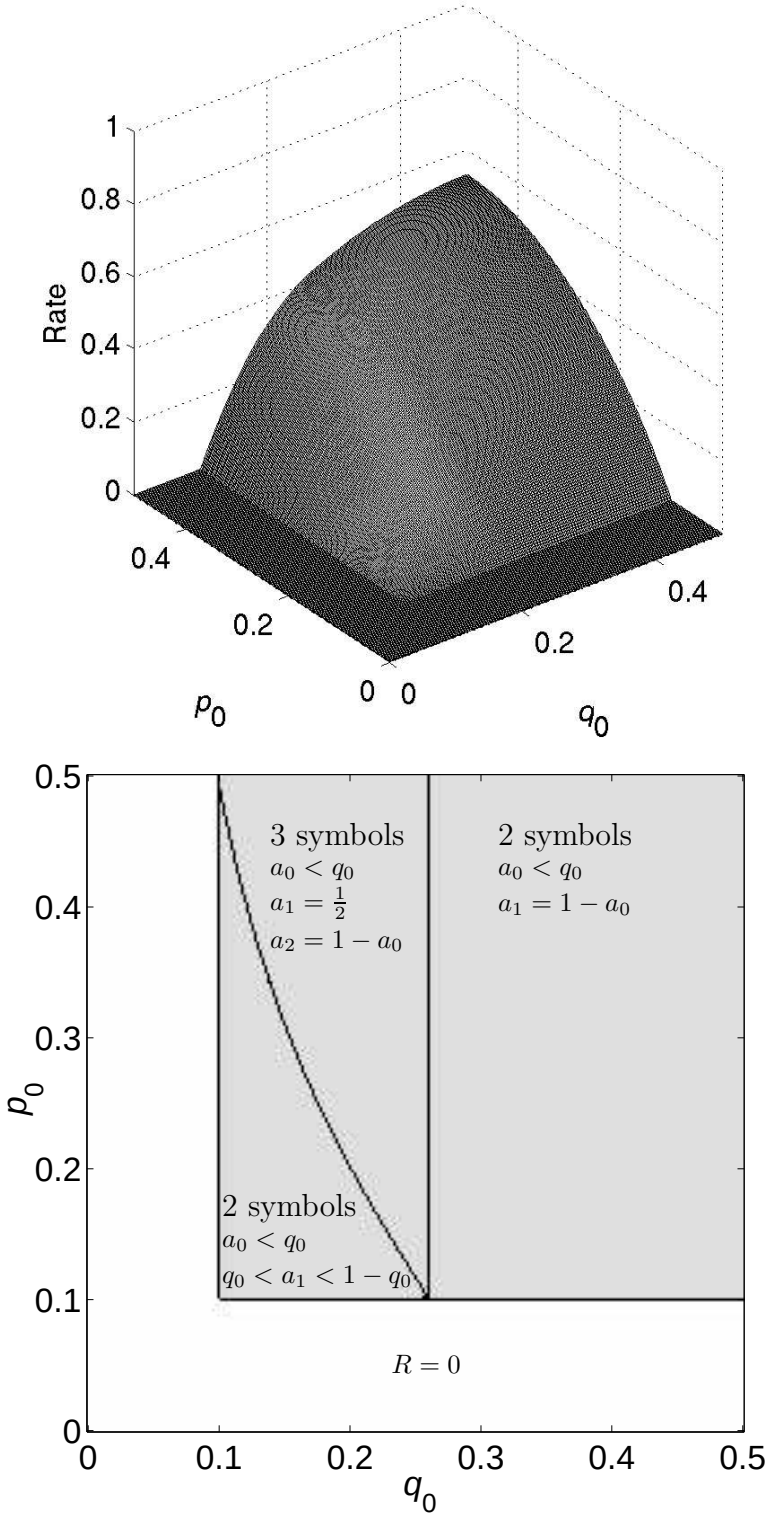


Figure 4.5: The value (top) and the form (bottom) of the optimal solution for different values of p_0 and q_0 when $D = 0.1$.

and $\hat{X}_2$ satisfy $Ed(X, \hat{X}_1) \leq D_1$ and $Ed(X, \hat{X}_2) \leq D_2$, respectively [2]. An alternative form is given by Kaspi in [3].

The condition $(U, V) \rightarrow X \rightarrow Z$ can be rewritten using the following lemma.

Lemma 4 *The condition $(U, V) \rightarrow X \rightarrow Z$ is equivalent to the two conditions $U \rightarrow X \rightarrow Z$ and $V \rightarrow (U, X) \rightarrow Z$.*

Proof of Lemma 4 First, I prove the forward part. Assume that $(U, V) \rightarrow X \rightarrow Z$. Then

$$p(u|x, z) = \int p(u, v|x, z)dv = \int p(u, v|x)dv = p(u|x).$$

Thus $U \rightarrow X \rightarrow Z$. Using this result,

$$\begin{aligned} p(v|u, x, z)p(u|x) &= p(v|u, x, z)p(u|x, z) \\ &= p(u, v|x, z) \\ &= p(u, v|x) \\ &= p(v|u, x)p(u|x), \end{aligned}$$

and hence $V \rightarrow (U, X) \rightarrow Z$.

For the converse, assume $U \rightarrow X \rightarrow Z$ and $V \rightarrow (U, X) \rightarrow Z$. Then

$$p(u, v|x, z) = p(v|u, x, z)p(u|x, z) = p(v|u, x)p(u|x) = p(u, v|x).$$

Thus $(U, V) \rightarrow X \rightarrow Z$. □

For any U, there exists a V and f_1 such that $Ed(X, f_1(U, V, Z)) \leq D_1$ (for instance, $V = X$ and $f_1(U, V, Z) = V$). I can therefore rewrite $R_{HB}(D_1, D_2)$ with the help of Lemma 4 as

$$R_{HB}(D_1, D_2) = \min_{U \in \mathcal{M}_{HB}^U(D_2)} \left[I(X; U) + \min_{V \in \mathcal{M}_{HB}^V(U, D_1)} I(X; V|U, Z) \right],$$

where

$$\begin{aligned}\mathcal{M}_{HB}^{U}(D_2) &= \{U : U \to X \to Z,\ \exists f_2 \text{ s.t. } Ed(X, f_2(U)) \leq D_2\} \\ \mathcal{M}_{HB}^{V}(U, D_1) &= \{V : V \to (U, X) \to Z,\ \exists f_1 \text{ s.t. } Ed(X, f_1(U, V, Z)) \leq D_1\}.\end{aligned}$$

Noting the equivalence of $\mathcal{M}_{HB}^{V}(U, D_1)$ and $\mathcal{M}_{X|U\{Z\}}(p(u, x, z), D_1)$,

$$R_{HB}(D_1, D_2) = \min_{U \in \mathcal{M}_{HB}^{U}(D_2)} \left[I(X;U) + R_{X|U\{Z\}}(p(u, x, z), D_1) \right].$$

For the binary example this yields

$$R_{HB}(D_1, D_2) = \min_{U \in \mathcal{M}_{HB}^{U}(D_2)} \sum_{u \in \mathcal{U}} p(u) \left[1 - H(X|U = u) + R_{X|U=u\{Z\}}(p(u, x, z), D_1) \right].$$

The variable U must achieve the distortion constraint at decoder 2. As shown in [2], its alphabet is bounded according to $|\mathcal{U}| \leq |\mathcal{X}| + 2 = 4$. We can therefore represent U by its marginal probabilities r_u and transition probabilities $a_u = \Pr(X = 1|U = u)$, $u \in \{0, 1, 2, 3\}$. These parameters must satisfy

$$r_0 + r_1 + r_2 + r_3 = 1 \tag{4.15}$$

$$r_0 a_0 + r_1 a_1 + r_2 a_2 + r_3 a_3 = \frac{1}{2} \tag{4.16}$$

$$0 \leq r_u, \quad u \in \{0, 1, 2, 3\} \tag{4.17}$$

$$0 \leq a_u \leq 1, \quad u \in \{0, 1, 2, 3\}. \tag{4.18}$$

For each u, the distribution $p(u, x, z)$ is entirely characterized by that symbol's transition probability a_u and the side information crossover probability q_0. In what follows, I write $R_{X|U=u\{Z\}}(p(u, x, z), D_1)$ in the form $R(a_u, q_0, D_1)$ to make explicit its functional dependence on these parameters. It is the binary MSI rate-distortion function determined in the previous section when $p_0 = a_u$.

Assume that $R(a_u, q_0, D_1)$ is differentiable[2] with respect to a_u, and define $K(a_u) \triangleq R(a_u, q_0, D_1) - H(a_u)$. Finding the optimal U is equivalent to finding the $(a_0, a_1, a_2, a_3, r_0, r_1, r_2, r_3)$ that minimize

$$1 + r_0 K(a_0) + r_1 K(a_1) + r_2 K(a_2) + r_3 K(a_3), \tag{4.19}$$

subject to the constraints (4.15)-(4.18) together with a distortion constraint for decoder 2. Appendix D outlines how to evaluate this minimization and hence determine $R_{HB}(D_1, D_2)$ using a search over only two parameters, matching the complexity required to evaluate the existing bounds by Heegard and Berger and Kerpez.

Evaluating $R_{HB}(D_1, D_2)$, I find a significant region of (q_0, D_1, D_2)-space for which the bounds of Heegard and Berger and Kerpez are loose; an example is shown in Figure 4.6. The rate-distortion function is at some points as much as 0.056 bits per symbol below the minimum of the two prior upper bounds, and at others up to 0.2143 bits per symbol above Kerpez's lower bound[3].

There is one locally minimal solution to the minimization that is always present in the case when none of the inequality constraints is active. That minimum requires two symbols and occurs when $a_0 = D_2$, $a_1 = 1 - D_2$, $r_0 = \frac{1}{2}$, and $r_1 = \frac{1}{2}$, i.e., when U is related to X via a binary symmetric channel with crossover probability D_2. In practice, I find that this is the optimal solution for all q_0, D_1, and D_2 tested, and I conjecture that it is a unique optimal

[2]I show in Appendix C that although $R(a_u, q_0, D_1)$ is not differentiable everywhere with respect to a_u, I can alter it by an insignificant amount so as to smooth it and make it so.

[3]Contrary to a conjecture by Kerpez, his solution is not everywhere better than Heegard and Berger's. I thank Sidharth Jaggi for verifying a counterexample that at $(q_0, D_1, D_2) = (0.1, 0.05, 0.25)$, Kerpez's bound $R_{HB} \leq 0.4116$ is looser than Heegard and Berger's bound $R_{HB} \leq 0.3970$.

solution. However, since $K(a_u)$ is not convex (because $R(a_u, q_0, D_1)$ is not a convex function of a_u), I cannot easily prove the uniqueness of this solution. I can conclude that it at least provides an extremely tight upper bound which can be computed with a search over only one parameter rather than two.

A binary symmetric U achieves the rate-distortion function in the absence of decoder 1 (i.e., U achieves $R_X(D_2)$). The corresponding V achieves $R_{X|U\{Z\}}(p, D_1)$. This situation parallels successive refinement, except that the refinement description now works in cooperation with side information. Interestingly, Heegard and Berger's Gaussian example in [2] exhibits the same pattern. There, too, the variable U is chosen as it would be to achieve $R_X(D_2)$, and V is chosen to provide the necessary refinement. Since both binary and Gaussian sources are successively refinable, this suggests that a two-step approach (choose U so as to achieve $R_X(D_2)$, then choose V so as to achieve $R_{X|U\{Z\}}(p, D_1)$) might achieve the Heegard and Berger rate-distortion function for all successively refinable sources for which $R_X(D_2)$ is achieved by a U generated from the addition of appropriate i.i.d. noise to X.

For general sources, the two-step approach bounds the HB rate-distortion function from above in terms of the traditional and MSI rate-distortion functions. The MSI rate-distortion function is in turn bounded in relation to the conditional rate-distortion function by the rate loss results of Section 4.2.

4.6 Summary

I derive rate-distortion results for a system with some side information known at both the encoder and decoder and some known only at the decoder. Both the rate-distortion func-

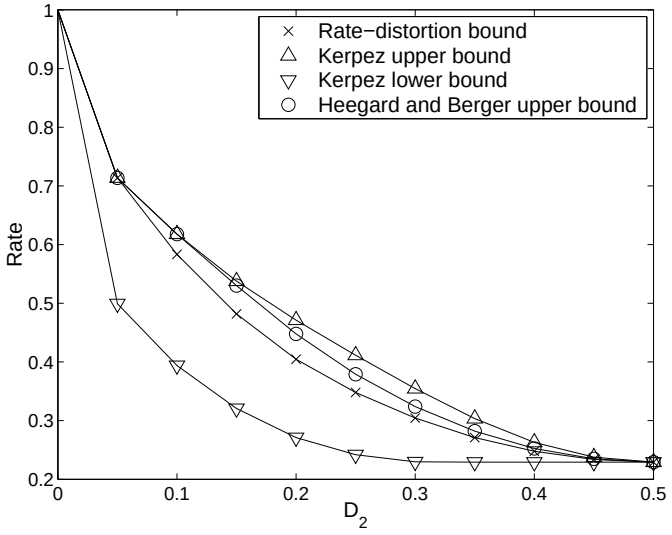


Figure 4.6: Numerical results for Heegard and Berger's system, $q_0 = 0.1$, $D_1 = 0.05$.

tion and the rate loss are direct generalizations of existing results for the conditional rate-distortion and Wyner-Ziv systems. Two rate-distortion examples are studied in depth. The Gaussian example generalizes easily from the Wyner-Ziv case; the binary example is considerably more complicated, but I present an easily computable solution. I use this to help us solve a more difficult binary rate-distortion problem for the system of Heegard and Berger (HB). Comparison of the new binary HB solution and the existing Gaussian HB solution show that they both use a separable, two-step approach to construct auxiliary random variables. That a two-step approach is optimal suggests the existence of a new type of successive refinement for which the second part of the description is decoded together with side information. It also suggests a two-step approach might yield good results for practical coding.

The results in this chapter yield insight into the use of side information in general networks. However, continuing further along the same lines is likely to yield slow progress. Thus, I now change tack and in the next chapter look directly at three different source coding problems for larger networks.

Chapter 5

Network Source Coding Results

This chapter treats three different topics in large network source coding theory. First, I ask and answer the question of whether feedback from a decoder to an encoder can enlarge a set of achievable rates in lossless source coding. Next, I show how to use cutsets to derive simple source coding converses for any network. Finally, I present two new results in broadcast source coding.

5.1 Feedback in Lossless Coding

Consider lossless source coding for a bipartite graph like the example in Figure 5.1. Encoder j, $j = 1, 2, \ldots, J$, sees a source X_j of which it creates a description. This description is made available to some subset of the decoders via noiseless communication channels. Decoder β_k, $k = 1, 2, \ldots, K$, which has access to side information Z_k, reproduces some subset of the sources described to it. For instance, in the example of Figure 5.1, decoder 1 receives descriptions of X_1 and X_3 and reproduces both of them; decoder 2 receives descriptions of

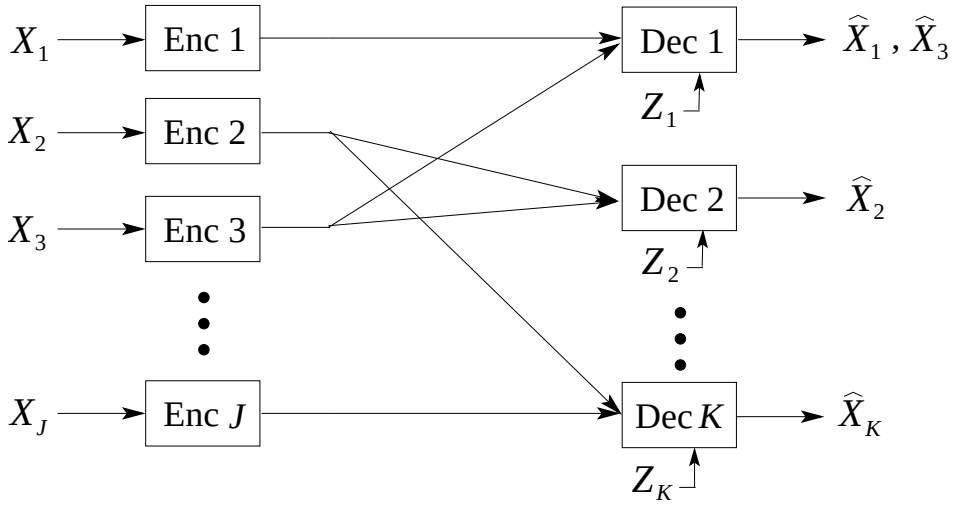


Figure 5.1: A bipartite graph considered for the feedback problem.

X_2 and X_3 but reproduces only X_2. Since the coding is lossless, the achievable rate region is the set of rate vectors $\mathbf{R} = (R_j)_{j\in\{1,\ldots,J\}}$ that allow the reproductions to be made with an arbitrarily low probability of error. This section considers whether feedback from the decoders to the encoders can enlarge the achievable rate region.

For point-to-point networks, it is straightforward to show that feedback does not enlarge the achievable region for either lossless or lossy coding. The information known by the decoder is always a subset of that known by the encoder, and hence there is no information that the decoder can feed back to the encoder that the encoder does not already know. The following theorem, proved in Appendix E, makes this notion concrete for lossy coding. It generalizes the corresponding lossless result [84].

Theorem 16 *For an i.i.d. source X, $R_{FB}(D) = R(D)$, where $R_{FB}(D)$ is the rate-distortion function when feedback is permitted from the decoder to the encoder.*

For the same reason as for the point-to-point case, feedback does not enlarge the rate-distortion region for the the M-receiver BC network: none of the decoders can feed back to

the encoder any information that it does not already know.

The same argument does not apply to WZ or MA systems. For lossy coding, feedback is known to help in the WZ system. There exist sources for which the WZ rate-distortion function is strictly greater than the corresponding conditional rate-distortion function [12]; if the side information at the decoder of the WZ system is fed back to the encoder in its entirety, then the achievable rate-distortion region expands to that of the corresponding conditional rate-distortion system. Since WZ is a special case of MA, it follows that feedback also helps in lossy MA coding. However, in contrast to lossy MA coding, the Slepian-Wolf result [11] implies that lossless MA networks derive no advantage from feedback. This is shown in the case of two-user MA coding in the following theorem, proved in Appendix E. The M-user extension follows from a straightforward generalization.

Theorem 17 *For the two-user multiple access system with i.i.d. sources, the achievable rate region for the case when feedback is permitted from the decoder to each of the encoders is the same as the achievable rate region for the case when no feedback is permitted.*

For general lossless networks, the exact achievable region without feedback is known in the special case when every decoder reproduces every source described to it [85]. For instance, in the example of Figure 5.1, decoder 2 receives descriptions of X_2 and X_3 but reproduces only X_2; to apply the results of [85], decoder 2 would need to reproduce both X_2 and X_3. For such networks, a generalized form of the Slepian-Wolf result holds, and, once again, feedback from the decoders to the encoders does not lower the total rate required.

Since several classes of lossless coding systems derive no rate advantage from feedback, it seems plausible that no lossless source coding network ever does. Theorem 18, however,

shows that feedback can increase the achievable rate region of a lossless source code. It is proved by example in the following section.

Theorem 18 *There exist lossless coding systems in which feedback of finite rate from one of the decoders to one of the encoders is sufficient to reduce the total rate required by the encoders. This reduction in rate can be arbitrarily large as the sizes of the source alphabets increase without bound.*

5.1.1 The Feedback Example

Consider the system shown in Figure 5.2. Sources X and Y are described by encoders α_X and α_Y at rates R_X and R_Y, respectively. The decoder builds a reproduction $\hat{X}$ of X from the two descriptions and the side information Z. The probability of error of an n-dimensional code $(\alpha_X^n, \alpha_Y^n, \beta^n)$ is

$$P_e^{(n)} = \Pr(\beta^n(\alpha_X^n(X^n), \alpha_Y^n(Y^n), Z^n) \neq X^n).$$

A pair of rates (R_X, R_Y) is achievable if for any $\epsilon > 0$ there exists a sequence of codes with rates $(R_X + \epsilon, R_Y + \epsilon)$ such that $P_e^{(n)} \to 0$ as $n \to \infty$. The achievable rate region $\mathcal{R}$ is the closure of the set of all achievable rates. Since Y need not be reproduced at the decoder, the system is not a special case of Slepian and Wolf's setup [11]. Rather, it is an example of source coding with side information as considered by Sgarro [86], but with additional side information Z present at the decoder.

Suppose now that feedback of rate R_{FB} is allowed from β^n to α_Y^n. For the example below it suffices to permit feedback of the form $g^n(Z^n)$, where g is any measurable function of Z^n, and to allow α_Y^n to be a function of both Y^n and the feedback $g^n(Z^n)$. Denote

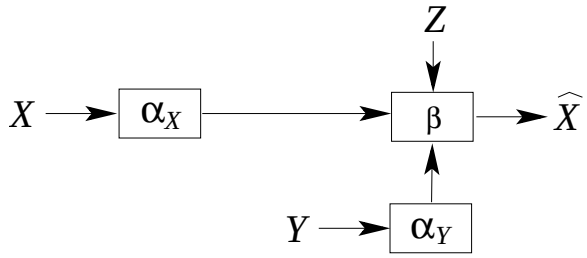


Figure 5.2: The lossless coding system used in the proof of Theorem 18.

the corresponding achievable rate region by $\mathcal{R}_{FB}$. Using an example reminiscent of [87, Example 8], I show the existence of sources (X, Y, Z) such that feedback from β^n to α_Y^n of rate $R_{FB} = 1 + \delta$, where $\delta > 0$ is arbitrarily small, lowers the minimum rate R_Y when $R_X = 0$.

Let X be distributed uniformly on $\mathcal{X} = \{1, \ldots, M\}$, i.e.,

$$p(x) = \begin{cases} \frac{1}{M}, & x \in \{1, \ldots, M\} \\ 0, & \text{otherwise.} \end{cases}$$

The distribution $p(y, z|x)$ is specified in the following way. Let θ be a random variable drawn i.i.d. uniformly on $\{0, 1\}$. When $\theta = 0$, Y is distributed uniformly on $\{1, \ldots, M\}$ and $Z = X$. When $\theta = 1$, $Y = X$ and $Z = M + 1$.

From the above definitions, we can recover X from Y and Z via the function

$$f(y, z) = \begin{cases} y, & Z = M + 1 \\ z, & \text{otherwise.} \end{cases} \tag{5.1}$$

Thus, X can be reproduced by the decoder even when $R_X = 0$. However, the rate R_Y required when $R_X = 0$ differs depending on whether or not feedback is present from the decoder to the encoder of Y.

Theorem 18 follows from the following two lemmas, proved below.

Lemma 5 *When feedback of rate R_{FB} is permitted from β^n to α_Y^n, rates $(R_X, R_Y, R_{FB}) = (0, \frac{1}{2}\log M, 1)$ are achievable for the given sources.*

Lemma 6 *When feedback is not permitted, if $(R_X, R_Y) = (0, R_Y)$ is achievable for the given sources, then $R_Y \geq \log M$.*

Proof of Theorem 18 Combining the two lemmas shows that when $R_X = 0$, the rate R_Y required when feedback is absent is at least twice the rate needed when feedback of rate 1 is allowed between decoder and encoder. The rate difference is $\frac{1}{2}\log M$, which increases without bound as the alphabet size M increases. The required feedback rate remains fixed at $R_{FB} = 1$ for any M. □

Proof of Lemma 5 Define the feedback $g^n(Z^n)$ according to $g^n(Z^n) = (g_1(Z_1), \ldots, g_n(Z_n))$, where $g_i(Z_i) = 1_{\{Z_i = M+1\}} = \theta_i$. The decoder describes θ^n to α_Y^n using rate $R_{FB} = H(\theta_i) = 1$, and α_Y^n describes to the decoder those samples of Y corresponding to time slots when $\theta_i = 1$, using $H(Y) = \log M$ bits per instance. Thus $R_Y = \frac{\log M}{n} E[\sum_{i=1}^n \theta_i] = \frac{1}{2}\log M$. The decoder reconstructs X^n by setting $\hat{X}_i = Y_i$ when $Z_i = M+1$ and $\hat{X}_i = Z_i$ otherwise. This establishes the achievability of rates $(R_X, R_Y, R_{FB}) = (0, \frac{1}{2}\log M, 1)$. □

Proof of Lemma 6 Lemma 6 is established by considering $\hat{X} = f(Y, Z)$ as a function of Y and Z to be calculated by the decoder and applying a functional source coding result. From [87, Theorem 1], the minimum rate R_Y when $R_X = 0$ is given by

$$R_Y = H_G(Y|Z),$$

where G is the characteristic graph of Y, Z, and f. By definition, the vertex set of G is the support set of Y, and two distinct vertices y and y' are connected if there is a z such that

$\Pr(Y = y, Z = z) > 0$, $\Pr(Y = y', Z = z) > 0$, and $f(y, z) \neq f(y', z)$. When $z = M + 1$, then for any $y \neq y'$, $\Pr(Y = y, Z = M + 1) = \Pr(Y = y', Z = M + 1) = \frac{1}{M} > 0$ and $f(y, M + 1) \neq f(y', M + 1)$. Thus, the characteristic graph is fully connected, yielding $H_G(Y|Z) = H(Y|Z)$. For the current problem, $H(Y|Z) = H(Y) = \log M$. Therefore, recovering X with arbitrarily low probability of error when $R_X = 0$ requires $R_Y \geq \log M$. □

The example above essentially forms a functional coding problem out of a more standard lossless coding problem. As such, it illustrates that ordinary lossless coding problems in general networks can share all of the characteristics of functional coding problems, including rate reductions from feedback. Central to the example is the existence of the "helper source" [85] Y; although the decoder receives a description of Y, it is not required to reproduce Y.

The constraint $R_X = 0$ in the above problem reduces the problem to a very specialized case. It is reasonable to ask whether we can observe the same behavior when $R_X > 0$. The answer is yes; an example is obtained simply by replacing the source $X' = (X, V)$, where V is the result of a fair coin toss that is independent of Y and Z. This imposes the requirement $R_{X'} \geq 1$ so as to allow for the description of the coin toss, but leaves the achievable rate for Y unchanged.

5.2 Source Coding Converses Via Cutsets

Finding the exact limits of source code performance is very difficult in all but the simplest of networks. Indeed, the limits of code performance are still unknown even for some three-node networks. However, insight can sometimes be obtained by bounding the achievable rate regions rather than finding them exactly. For instance, determining how the set of achievable

rates scales with network size does not require exact knowledge of the rate region; it suffices to have bounds on both sides that scale in a similar fashion. Also, achieving a performance close to optimal is often enough for practical applications.

A cutset is a partition of the set of network nodes into two or more subsets. A converse for a network is obtained by bounding the rate flowing from one subset to the other for every possible cutset partition of the nodes. For channel coding, a converse result bounding the achievable rate region for a general network is developed using a cutset approach in [31, Theorem 14.10.1], but no parallel result exists for source coding. This section uses cutsets to derive network source coding converses. The cutset approach has the advantage of easy applicability to even the most complex networks, but the drawback that the converses obtained are often quite loose.

I show for two- and three-partition cutsets how to bound network source coding rates, and I demonstrate the resulting converses using a two-access network and a three-node network. Further generalizations using more than three subsets are possible but are not considered here.

5.2.1 Cutset Theorems

Figure 5.3 shows a general network of M nodes. Consider a cutset that partitions the network nodes into two sets, A and $B = A^c$, and denote by $R_{A\to B}$ the total description rate flowing from nodes in A to nodes in B. The set of sources known by nodes in A is X_A, the set of sources known by nodes in B is X_B, and the side information known by nodes in B is Z_B. The set of sources being described by nodes in A to nodes in B is $X_{A\to B}$;

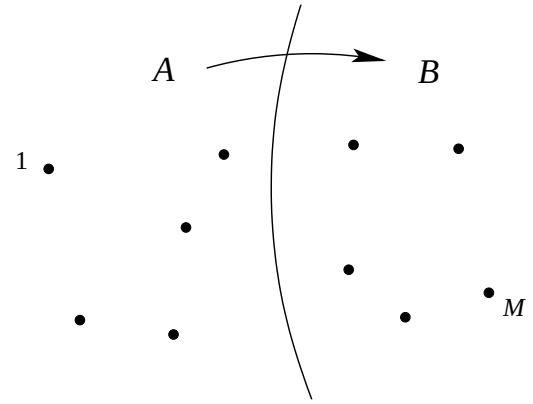


Figure 5.3: A general network. The line is an example of a cutset boundary separating the M nodes of the network into two sets, A and B.

the remaining sources in A are $X_{A\not\to B}$, giving $X_A = (X_{A\to B}, X_{A\not\to B})$. Similar definitions apply to give $X_B = (X_{B\to A}, X_{B\not\to A})$. For simplicity, assume that A and B share no sources or side information in common, although the results can easily be restated to cover that contingency. Assume also that the distortion measure satisfies the two conditions (that there exists an escape symbol with finite expected distortion and that the distortion measure is fairly smooth) imposed in Section 2.2.3.

The theorem below bounds the rate flowing from A to B using an MSI rate-distortion function as defined in Chapter 4. However, since there are multiple components of $X_{A\to B}$ (each source transmitted from A to B is a separate component), each of which may have a separate distortion constraint, Theorem 14 must be extended to the case of several jointly encoded sources. This requires only minor modifications; if we jointly encode a compound source $X = (X_1, \ldots, X_M)$ with distortion constraints $(D_1, \ldots, D_M)$ on each of the individual sources the rate-distortion function takes the same information-theoretic form, but we must introduce a different reproduction function f for each source. Thus, the set $\mathcal{M}_{X|Y\{Z\}}(p, (D_1, \ldots, D_M))$ becomes the set of all W such that for each $i = 1, \ldots, M$ there

exists a reproduction function $f_i(w, y, z)$ meeting distortion constraint D_i.

Since each component of $X_{A\to B}$ may be reproduced at more than one node in B (a source may be described to, say, three nodes in B, each with a different distortion constraint), there may be multiple distortion constraints for the same component. However, it is convenient to work with a single distortion constraint for each source. Define the vector $D_{A\to B}$ to be a collection of distortion constraints, one per component of $X_{A\to B}$, where for each component we choose the weakest constraint imposed by the nodes in B that reproduced that component.

Theorem 19 *Choose any $A \subseteq \{1, 2, \ldots, M\}$ and $B = A^c$, and assume $I(X_{A\to B}; X_B, Z_B) < \infty$. Then any code achieving distortions $D_{A\to B}$ has rates $R_{A\to B}$ satisfying*

$$R_{A\to B} \;\geq\; R_{X_{A\to B}|X_{B\to A}\{X_{B\not\to A}Z_B\}}(D_{A\to B}). \tag{5.2}$$

Proof of Theorem 19 Fix A and $B = A^c$. Perform the following steps, each of which is guaranteed not to increase the minimum rate flowing from A to B.

1. Replace all nodes in A by a single node that has full knowledge of all of the sources known by or described in part to nodes in A. This set of sources is $(X_A, X_{B\to A})$.

2. Replace all nodes in B by a single node that has access to the set of all received messages, sources, and side information known by any node in B.

These two steps are equivalent to allowing unlimited communication between the nodes in A and from B to A while assuming that that communication requires no expenditure in rate. They reduces the network to the system in Figure 5.4, for which the rate-distortion function is $R_{X_{A\to B}|X_{B\to A}\{X_{B\not\to A},Z_B\}}(D_{A\to B})$. Hence

$$R_{A\to B} \;\geq\; R_{X_{A\to B}|X_{B\to A}\{X_{B\not\to A},Z_B\}}(D_{A\to B}).$$

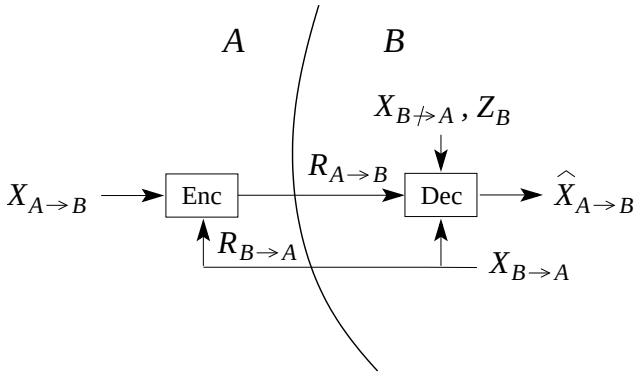


Figure 5.4: The simplified network for the cutset bounds.

□

Remark 1 One of the assumptions of Theorem 19 is that the encoders in A have perfect knowledge of the sources $X_{B\to A}$. This assumption involves two approximations. In general, the encoders receive a description of those sources only to some non-zero distortion. Also, the encoders cannot make full use of those descriptions unless they are prepared to delay coding their own data until the descriptions are fully received.

Remark 2 For lossless coding, the Slepian-Wolf result implies that feedback from B to A need not be taken into account (as shown in the previous section) and thus neither of the approximations identified in Remark 1 are relevant in lossless coding. In this case, (5.2) can be rewritten simply as $R_{A\to B} \geq H(X_{A\to B}|X_B, Z_B)$.

The two-subset partitions considered in Theorem 19 allow easy derivations of simple converse conditions. However, they do not typically yield bounds on the rates of common sources, nor do they make explicit the tradeoff between the rates of common and private sources. Capturing these features requires partitioning the nodes into three or more subsets. I concentrate here on the case of three subsets; bounds obtained for more than three are in general too difficult to compute.

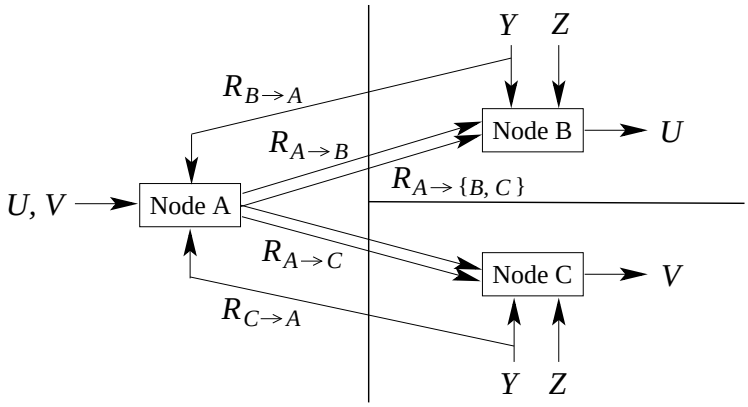


Figure 5.5: The system used for the three-set extension of the cutset approach.

Partitions with two subsets reduced our network to the system shown in Figure 5.4. With three sets, there are several systems we could reduce it to; the one I detail here is the two-receiver broadcast system with side information shown in Figure 5.5. Let

- U denote the collection of sources $(X_{A\to(B,C)}, X_{A\to B})$ transmitted by nodes in A to nodes in B.

- V denote the collection of sources $(X_{A\to(B,C)}, X_{A\to C})$ transmitted by nodes in A to nodes in C.

- Y denote the collection of sources $X_{(B\to A)\cup(C\to A)}$ that are known by either B or C and are fed back to A either in part or in entirety.

- Z denote the collection of sources $X_{(B\not\to A)\cup(C\not\to A)}$ that are known by either B or C but are not fed back to A.

- $D_U = (D_{A\to(B,C)}, D_{A\to B})$, where for each individual source its weakest distortion constraint imposed by any node in B is assumed in the definition of these compound distortion measures. Similarly, let $D_V = (D_{A\to(B,C)}, D_{A\to C})$.

The following theorem, proved in Appendix E, gives the converse result.

Theorem 20 *Let (A, B, C) be any partition of the network nodes into three sets for which the nodes in B do not share information with the nodes in C about the sources transmitted to them from A. Let (U, V, Y, Z) as defined above be drawn i.i.d. with joint distribution $p(u, v, y, z)$, and let $\mathcal{M}$ be the set of all auxiliary random variables W jointly distributed with (U, V, Y, Z) such that $W \to (U, V, Y) \to Z$ forms a Markov chain. Let*

$$\begin{aligned}\mathcal{R}(W, D_U, D_V) = \{(R_{A\to(B,C)}, R_{A\to B}, R_{A\to C}) \quad : \quad & R_{A\to(B,C)} \geq I(U, V; W|Y, Z) \\ & R_{A\to B} \geq R_{U|Y\{Z\}}(D_U) \\ & R_{A\to C} \geq R_{V|Y\{Z\}}(D_V)\}.\end{aligned}$$

The set $\mathcal{R}^*(D_U, D_V)$ of achievable $(R_{A\to(B,C)}, R_{A\to B}, R_{A\to C})$ for distortions (D_U, D_V) satisfies

$$\mathcal{R}^*(D_U, D_V) \subseteq \overline{\bigcup_{W\in\mathcal{M}} \mathcal{R}(W, D_U, D_V)}.$$

5.2.2 Cutset Examples

Consider the two-user MA network of Figure 5.6(a). Theorem 19 gives

$$R_{2,1} \geq \inf_{W\in\mathcal{M}_{2\to\{1,3\}}(D_{2,1,1})} I(X_{2,1}; W_{2,1}|X_{3,1}), \quad R_{3,1} \geq \inf_{W\in\mathcal{M}_{3\to\{1,2\}}(D_{3,1,1})} I(X_{3,1}; W_{3,1}|X_{2,1})$$
$$R_{2,1} + R_{3,1} \geq R_{X_{2,1},X_{3,1}}(D_{2,1,1}, D_{3,1,1}),$$

where $R_{X_{2,1},X_{3,1}}(D_{2,1,1}, D_{3,1,1})$ is the rate-distortion function for source pair $(X_{2,1}, X_{3,1})$. In the lossless case, the bounds simplify to the Slepian-Wolf bounds and are tight. In the lossy case, the bounds are similar to those by Berger and Tung [34, 21], but are not as tight since they do not require that the auxiliary random variables achieving the first two bounds be

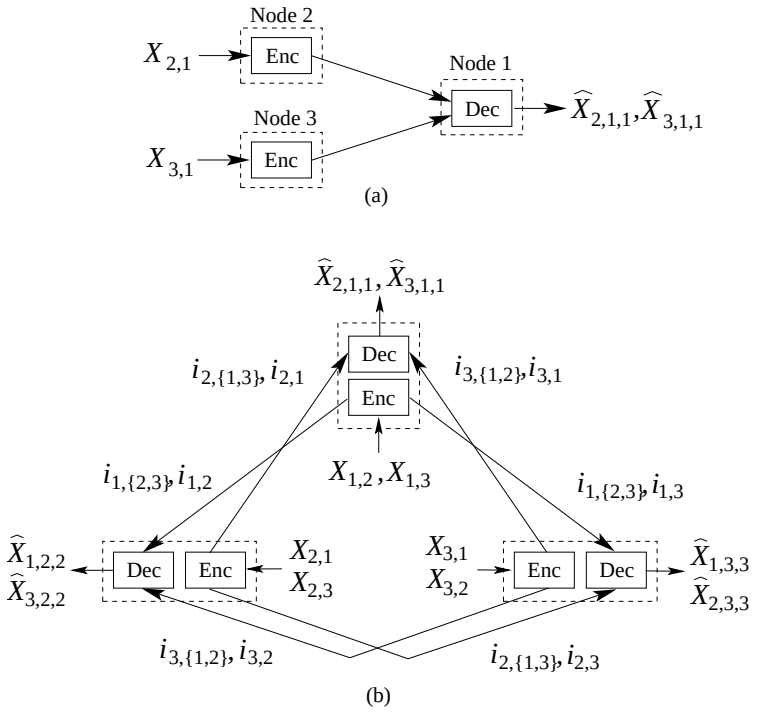


Figure 5.6: (a) A two-user MA network. (b) A three-node network.

the same as those used in achieving the rate-distortion function in the third. They also replace $I(X_{2,1}; W_{2,1}|W_{3,1})$ with the smaller term $I(X_{2,1}; W_{2,1}|X_{3,1})$ in the first bound and the symmetric replacement is made in the second bound.

Figure 5.6(b) shows a three-node network with two sources encoded at each node. Assume that each node has both a common channel and private channel to the other two nodes. Theorem 20 gives the following converse result for the rate leaving node 1. Let $\mathcal{M}$ be the set of all finite-alphabet random variables W such that $W \rightarrow (X_{1,2}, X_{1,3}, X_{2,1}, X_{3,1}) \rightarrow (X_{2,3}, X_{3,2})$ forms a Markov chain.

$$\begin{aligned}\mathcal{R}(W, D_{1,2,2}, D_{1,3,3}) = \{(R_{1,\{2,3\}}, R_{1,2}, R_{1,3}) \quad : \quad & R_{1,\{2,3\}} \geq I(X_{1,2}, X_{1,3}; W|X_{2,1}, X_{3,1}, X_{2,3}, X_{3,2}), \\ & R_{1,2} > R_{X_{1,2}|X_{2,1},X_{3,1}\{X_{2,3},X_{3,2}\}}(D_{1,2,2}), \\ & R_{1,3} > R_{X_{1,3}|X_{2,1},X_{3,1}\{X_{2,3},X_{3,2}\}}(D_{1,3,3})\}.\end{aligned}$$

Then if the rate triple $(R_{1,\{2,3\}}, R_{1,2}, R_{1,3})$ is achievable for distortions $(D_{1,2,2}, D_{1,3,3})$, it must lie within the region

$$\mathcal{R}^*(D_{1,2,2}, D_{1,3,3}) = \overline{\bigcup_{W \in \mathcal{M}} \mathcal{R}(W)}.$$

Similar results are obtained for the rates leaving nodes 2 and 3.

5.3 Broadcast Source Coding

The general broadcast networks defined in Chapter 3 allow a different message to be passed to every subset of the receivers. In practice, such a general framework is impractical because the number of subsets grows exponentially with the number of receivers. More commonly, only some of the subsets are considered, as in the limited broadcast network introduced in Section 3.5, where only one common information is permitted. In this section, I derive tight rate-distortion results for a new practical network model: a tree-based broadcast network. I also look at three-receiver broadcast coding, applying techniques from multiple-description coding and Slepian-Wolf coding to derive an achievability result.

For simplicity, the results below are stated with only one source intended for each receiver. To accommodate common sources intended to be reconstructed at several nodes, each individual source can simply be defined as a compound source containing one or more common sources. For example, in a two receiver case with sources X_1 and X_2, a common source X_{12} is accommodated by setting $X_1 = (X_{12}, X_1')$ and $X_2 = (X_{12}, X_2')$ and redefining the distortion measures appropriately.

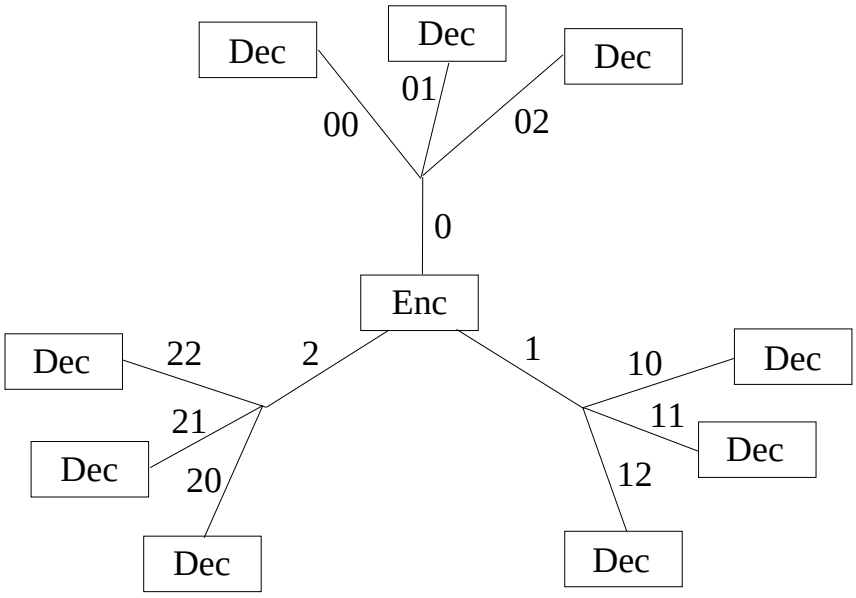


Figure 5.7: A k-ary broadcast tree.

5.3.1 Broadcast Coding for Tree Networks

Consider a wireless transmitter broadcasting to many receivers in various directions and at various distances. In any particular direction, the information received by nodes further away is a subset of the information received by closer nodes.

Figure 5.7 models this situation using a k-ary tree to describe the information dependencies. The receivers are the leaves of the tree. Each node in the tree is labeled according to the description of its path from the encoder. For each $0 \leq \ell \leq L$, $\mathcal{B}_\ell = \{0, 1, \ldots, k-1\}^\ell$ denotes the set of paths from the encoder to the nodes at depth ℓ (the root is considered to be at depth 0 and is described by path $b = \emptyset$). The set of all paths of length less than or equal to k is $\mathcal{B}^k = \cup_{\ell=0}^{k} \mathcal{B}_\ell$, and the set of all paths is $\mathcal{B} = \mathcal{B}^L$. For any node b, let $\text{anc}(b) = \{b' : b' \text{ is a proper prefix of } b\}$ be the set of all ancestors of b. Similarly, let $\text{dec}(b) = \{b' : b \text{ is a prefix of } b'\}$ be the set of all descendants of b, including b itself. Thus

$\mathrm{dec}(b)$ is the subtree anchored by b. The entire tree can be considered as the subtree $\mathrm{dec}(\emptyset)$ of the encoder.

The degraded broadcast channel is modeled as several separate channels as follows. For each subtree $\mathrm{dec}(b)$, $b \in \mathcal{B}$, assume that there exists a common channel at rate R_b to all of the members of that subtree. When b describes a leaf, that subtree contains only one receiver, and the channel is a private channel to receiver b.

There are k^L sources $\{X_b\}_{b\in\mathcal{B}_L}$, where source X_b is intended for the receiver at leaf b. That receiver makes a reconstruction $\hat{X}_b$ of source X_b that satisfies distortion constraint $Ed(X_b, \hat{X}_b) \leq D_b$. The sources are i.i.d. with joint distribution $p(\{x_b\}_{b\in\mathcal{B}_L})$, and the set of all sources is denoted $\mathbf{X} = (X_b)_{b\in\mathcal{B}_L}$.

For each $b \in \mathcal{B}$, let $T_b \in \mathcal{T}_b$ denote the index transmitted on the channel to subtree $\mathrm{dec}(b)$. An encoder and decoders for the system are then defined as

$$\begin{aligned} \alpha^n &: \prod_{b\in\mathcal{B}_L} \{\mathcal{X}_b^n\} \to \prod_{b\in\mathcal{B}_L} \{\mathcal{T}_b\} \\ \beta_b^n &: \prod_{a\in\{\{b\}\ \cup\ \mathrm{anc}(b)\}} \mathcal{T}_a \to \hat{\mathcal{X}}_b^n \quad \forall b \in \mathcal{B}_L. \end{aligned}$$

Given distortion measures $d_b : \mathcal{X}_m \times \hat{\mathcal{X}}_m \to [0, \infty)$, $m = 1, \ldots, M$, the distortions of a code are given by

$$\Delta_b = \frac{1}{n}\sum_{i=1}^{n} d_b(X_{b,i}, \hat{X}_{b,i}),$$

where $\hat{X}_b^n = \beta_b^n(\alpha^n(\mathbf{X}^n))$.

A set of rates $\mathbf{R} = (R_b)_{b\in\mathcal{B}}$ is achievable for distortions $\mathbf{D} = (D_b)_{b\in\mathcal{B}_L}$ if for any $\epsilon, \delta > 0$ there exists an $n \geq 1$ and a code $(\alpha^n, \{\beta_b^n\}_{b\in\mathcal{B}_L})$ with rates not exceeding $\mathbf{R} + (\epsilon, \ldots, \epsilon)$ and distortions not exceeding $\mathbf{D} + (\delta, \ldots, \delta)$. The set of all achievable $\mathbf{R}$ for fixed $\mathbf{D}$ is denoted $\mathcal{R}(\mathbf{D})$.

To describe the rate-distortion region, for each $b \in \mathcal{B}^{L-1}$ associate an auxiliary random variable W_b with the channel to the nodes in subtree $\mathrm{dec}(b)$. Let $W_{\mathrm{anc}(b)}$ denote the set $\{W_{b'}\}_{b' \in \mathrm{anc}(b)}$, and let $\mathcal{R}^*(\mathbf{D})$ be the closure of the set of all rates $\mathbf{R}$ for which there exists a set of random variables $\{W_b\}_{b \in \mathcal{B}^{L-1}}$ satisfying the following conditions.

1. For each $b \in \mathcal{B}^{L-1}$, $R_b \geq I(\mathbf{X}; W_b | W_{\mathrm{anc}(b)})$.

2. For each $b \in \mathcal{B}_L$, $R_b \geq R_{X_b|W_{\mathrm{anc}(b)}}(D_b)$.

Theorem 21 $\mathcal{R}(\mathbf{D}) = \mathcal{R}^*(\mathbf{D})$

Proof of Theorem 21 The theorem is proved in two parts, the converse then the achievability.

To establish the converse, consider an arbitrary code $(\alpha^n, \{\beta_b^n\}_{b \in \mathcal{B}_L})$ of some dimension n with distortions satisfying $\Delta_b \leq D_b$ for $b \in \mathcal{B}_L$. I show that $\mathbf{R} \in \mathcal{R}^*(\mathbf{D})$ by constructing a corresponding set of auxiliary random variables $\{W_b\}_{b \in \mathcal{B}^{L-1}}$.

Let $\{T_b\}_{b \in \mathcal{B}} = \alpha^n(\mathbf{X}^n)$ be the messages produced by the encoder. I bound each of the common rates R_b, $b \in \mathcal{B}^{L-1}$, as follows

$$\begin{aligned}
nR_b &\geq H(T_b) \\
&\overset{(a)}{\geq} H(T_b | T_{\mathrm{anc}(b)}) \\
&\geq I(T_b; \mathbf{X}^n | T_{\mathrm{anc}(b)}) \\
&= H(\mathbf{X}^n | T_{\mathrm{anc}(b)}) - H(\mathbf{X}^n | T_b, T_{\mathrm{anc}(b)}) \\
&= \sum_{i=1}^{n} \left[H(\mathbf{X}_i | \mathbf{X}^{i-1}, T_{\mathrm{anc}(b)}) - H(\mathbf{X}_i | \mathbf{X}^{i-1}, T_b, T_{\mathrm{anc}(b)}) \right] \\
&\overset{(b)}{=} \sum_{i=1}^{n} \left[H(\mathbf{X}_i | W_{\mathrm{anc}(b),i}) - H(\mathbf{X}_i | W_{b,i}, W_{\mathrm{anc}(b),i}) \right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} I(\mathbf{X}_i; W_{b,i}|W_{\mathrm{anc}(b),i}) \\
&\overset{(c)}{=} \sum_{i=1}^{n} I(\mathbf{X}_i; W_{b,i}|W_{\mathrm{anc}(b),i}, Q=i) \\
&= nI(\mathbf{X}_Q; W_{b,Q}|W_{\mathrm{anc}(b),Q}, Q) \\
&\overset{(d)}{=} nI(\mathbf{X}_Q; W_{b,Q}, Q|W_{\mathrm{anc}(b),Q}, Q) \\
&\overset{(e)}{=} nI(\mathbf{X}; W_b|W_{\mathrm{anc}(b)}),
\end{aligned}
$$

where the labeled steps are justified by the following.

(a) By defining $T_{\mathrm{anc}(b)} = \{T_{b'}\}_{b' \in \mathrm{anc}(b)}$.

(b) By defining $W_{b,i} = (\mathbf{X}^{i-1}, T_b)$ for every $b \in \mathcal{B}^{L-1}$ and $W_{\mathrm{anc}(b),i} = \{W_{b',i}\}_{b' \in \mathrm{anc}(b)}$.

(c) By introducing a timesharing variable Q, independent of all of the other random variables, uniformly distributed on $1, \ldots, n$.

(d) $\mathbf{X}_Q$ and Q are independent (the distribution of $\mathbf{X}$ is i.i.d. and does not depend on Q).

(e) By defining $W_b = (W_{b,Q}, Q)$ for every $b \in \mathcal{B}^{L-1}$ and from $\mathbf{X}$ being i.i.d.

I bound each of the private rates $\{R_b\}_{b \in \mathcal{B}_L}$ as follows

$$
\begin{aligned}
nR_b &\overset{(a)}{\geq} H(\hat{X}_b^n|T_{\mathrm{anc}(b)}) \\
&\geq I(\mathbf{X}^n; \hat{X}_b^n|T_{\mathrm{anc}(b)}) \\
&= \sum_{i=1}^{n} I(\mathbf{X}_i; \hat{X}_b^n|\mathbf{X}^{i-1}, T_{\mathrm{anc}(b)}) \\
&\geq \sum_{i=1}^{n} I(\mathbf{X}_i; \hat{X}_{b,i}|\mathbf{X}^{i-1}, T_{\mathrm{anc}(b)}) \\
&\geq \sum_{i=1}^{n} I(X_{b,i}; \hat{X}_{b,i}|\mathbf{X}^{i-1}, T_{\mathrm{anc}(b)}) \\
&= \sum_{i=1}^{n} I(X_{b,i}; \hat{X}_{b,i}|W_{\mathrm{anc}(b),i})
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} I(X_{b,i}; \hat{X}_{b,i} | W_{\text{anc}(b),i}, Q = i) \\
&= nI(X_{b,Q}; \hat{X}_{b,Q} | W_{\text{anc}(b),Q}, Q) \\
&\overset{(b)}{\geq} nR_{X_{b,Q}|W_{\text{anc}(b),Q}Q}(\Delta_b) \\
&= nR_{X_b|W_{\text{anc}(b)}}(\Delta_b) \\
&\overset{(c)}{\geq} nR_{X_b|W_{\text{anc}(b)}}(D_b),
\end{aligned}
$$

where the labeled steps are justified by the following.

(a) There are at most 2^{nR_b} values of $\hat{X}_b^n$ in the range of β_b^n for any given $T_{\text{anc}(b)}$.

(b) From the definition of the conditional rate-distortion function. The average distortion incurred under distribution $p(\mathbf{X}_Q, W_{\text{anc}(b),Q})$ is

$$
\begin{aligned}
Ed(X_{b,Q}, \hat{X}_{b,Q}) &= E[Ed(X_{b,Q}, \hat{X}_{b,Q})|Q] \\
&= \frac{1}{n} \sum_{i=1}^{n} d(X_{b,i}, \hat{X}_{b,i}) \\
&= \Delta_b.
\end{aligned}
$$

(c) $R_{X_b|W_{\text{anc}(b)}}(\cdot)$ is a decreasing function.

Thus, $\mathbf{R} \in \mathcal{R}^*(\mathbf{D})$, and the converse is established. I proceed with the achievability result.

Let $p(\{X_b\}_{b\in\mathcal{B}_L}, \{W_b\}_{b\in\mathcal{B}^{L-1}})$ be given. For each $b \in \mathcal{B}_L$, fix $p(\hat{x}_b|x_b, w_{\text{anc}(b)})$ to be the distribution that achieves the minimum in the definition of $R_{X_b|W_{\text{anc}(b)}}(D_b)$. I show that for these distributions there exists a sequence of codes with asymptotic distortions $\mathbf{D}$ and rates $\mathbf{R}$ such that $\mathbf{R} \in \mathcal{R}^*(\mathbf{D})$.

For each $b \in \mathcal{B}$, let $j_b \in \{1, \ldots, 2^{nR_b}\}$ be the index to be transmitted over the channel to the receivers in subtree dec(b). I create the codebook in the following way. Beginning

from $b = \emptyset$ and working in a breadth-first fashion through all $b \in \mathcal{B}^{L-1}$, for each b and each $j_{\text{anc}(b)} = \{j_{b'}\}_{b' \in \text{anc}(b)}$, draw 2^{nR_b} sequences $W_b^n(j_{\text{anc}(b)}, j_b)$ uniformly with replacement from $A_\epsilon^{*(n)}(W_b|W_{\text{anc}(b)}^n(j_{\text{anc}(b)}))$. Next, for each $b \in \mathcal{B}_L$ and each $j_{\text{anc}(b)}$, draw 2^{nR_b} sequences $\hat{X}_b(j_{\text{anc}(b)}, j_b)$ uniformly with replacement from $A_\epsilon^{*(n)}(\hat{X}_b|W_{\text{anc}(b)}^n(j_{\text{anc}(b)}))$.

The encoder and decoders operate as follows. The encoder receives $\mathbf{X}^n$ and chooses an index set $\{j_b\}_{b \in \mathcal{B}}$ such that for every $b \in \mathcal{B}_L$,

$$(\mathbf{X}^n, W_{\text{anc}(b)}^n(j_{\text{anc}(b)}), \hat{X}_b^n(j_{\text{anc}(b)}, j_b)) \in A_\epsilon^{*(n)}.$$

If there does not exist such an index set, the encoder declares an error. If there exists more than one such set, the encoder chooses among them randomly. In the absence of an error, the encoder transmits each index j_b on the channel to the receivers in subtree $\text{dec}(b)$. Decoder b receives indices $(j_{\text{anc}(b)}, j_b)$ and declares a reproduction $\hat{X}_b^n(j_{\text{anc}(b)}, j_b)$.

The following is an exhaustive list of error events for the code.

1. $E_0 = \{\mathbf{X}^n \notin A_\epsilon^{*(n)}\}$.

2. For each $b \in \mathcal{B}^{L-1}$, $E_b = E_0^c \cap \{\nexists j_b : (\mathbf{X}^n, W_{\text{anc}(b)}^n(j_{\text{anc}(b)}), W_b(j_b)) \in A_\epsilon^{*(n)}\}$.

3. For each $b \in \mathcal{B}_L$,

$$E_{b,1} = (E_0 \cup (\cup_{b' \in \mathcal{B}^{L-1}} E_{b'}))^c \cap \{\nexists j_b : (\mathbf{X}^n, W_{\text{anc}(b)}^n(j_{\text{anc}(b)}), \hat{X}_b^n(j_{\text{anc}(b)}, j_b)) \in A_\epsilon^{*(n)}\}.$$

4. For each $b \in \mathcal{B}_L$,

$$E_{b,2} = (E_0 \cup (\cup_{b' \in \mathcal{B}^{L-1}} E_{b'}) \cup (\cup_{b' \in \mathcal{B}_L} E_{b',1}))^c \cap \{\frac{1}{n} d(X_b^n, \hat{X}_b^n(j_{\text{anc}(b)}, j_b)) \geq D_b + \delta\}.$$

By the union bound, $P_e^n \to 0$ if the error probabilities of all of these events go to zero as $n \to \infty$. Below, I examine each error event individually.

1. By Lemma 1, $\Pr\{E_0\} \to 0$ as $n \to \infty$.

2. By arguments parallel to the achievability of the rate-distortion function [31, Pg. 355-356], for any $b \in \mathcal{B}^{L-1}$, $\Pr\{E_b\} \to 0$ as $n \to \infty$ provided $R_b > I(\mathbf{X}, W_b|W_{\text{anc}(b)}) + (|b|+3)\epsilon$.

3. Again, by arguments parallel to the achievability of the rate-distortion function, for any $b \in \mathcal{B}^{L-1}$, $\Pr\{E_{b,1}\} \to 0$ as $n \to \infty$ provided $R_b > I(X_b, \hat{X}_b|W_{\text{anc}(b)}) + (L+3)\epsilon$.

4. By definition, occurrence of event $E_{b,2}$ implies that $(X_b^n, \hat{X}_b^n(j_{\text{anc}(b)}, j_b)) \in A_\epsilon^{*(n)}$. I bound the distortion of the code using the properties of strong joint typicality

$$\begin{aligned}
\frac{1}{n} d(X_b^n, \hat{X}_b^n(j_{\text{anc}(b)}, j_b)) &= \frac{1}{n} \sum_{x_b \in \mathcal{X}_b, \hat{x}_b \in \hat{\mathcal{X}}_b} d(x_b, \hat{x}_b) N(x_b, \hat{x}_b | X_b^n, \hat{X}_b^n(j_{\text{anc}(b)}, j_b)) \\
&\leq \frac{1}{n} \sum_{x_b \in \mathcal{X}_b, \hat{x}_b \in \hat{\mathcal{X}}_b} d(x_b, \hat{x}_b) \left(np(x_b, \hat{x}_b) + \frac{n\epsilon}{|\mathcal{X}_b||\hat{\mathcal{X}}_b|} \right) \\
&\leq \sum_{x_b \in \mathcal{X}_b, \hat{x}_b \in \hat{\mathcal{X}}_b} d(x_b, \hat{x}_b) p(x_b, \hat{x}_b) + \epsilon d_{max} \sum_{x_b \in \mathcal{X}_b, \hat{x}_b \in \hat{\mathcal{X}}_b} \frac{1}{|\mathcal{X}_b||\hat{\mathcal{X}}_b|} \\
&= E d(X_b, \hat{X}_b) + \epsilon d_{max} \\
&\leq D_b + \epsilon d_{max},
\end{aligned}$$

where I have assumed

$$d_{max} \stackrel{\triangle}{=} \max_b \max_{x_b, \hat{x}_b} d(x_b, \hat{x}_b) < \infty.$$

Thus, for any $b \in \mathcal{B}_L$, $\Pr\{E_{b,2}\} \to 0$ as $n \to \infty$.

I have shown that the probability of error of the sequence of codes goes to zero as $n \to \infty$ provided

$$\text{For all } b \in \mathcal{B}^{L-1}, \quad R_b > I(\mathbf{X}, W_b|W_{\text{anc}(b)}) + (|b|+3)\epsilon,$$

$$\text{For all } b \in \mathcal{B}_L, \quad R_b > I(\mathbf{X}, W_b | W_{\text{anc}(b)}) + (L+3)\epsilon$$
$$= R_{X_b|W_{\text{anc}(b)}}(D_b) + (L+3)\epsilon.$$

Since $\epsilon > 0$ is arbitrary, it follows that there exists a sequence of codes with asymptotic rates $\mathbf{R} \in \mathcal{R}^*(\mathbf{D})$. □

5.3.2 Three-receiver Broadcast Coding

This section derives non-matching achievability and converse results for a three-receiver broadcast network.

The most general three-receiver broadcast network includes a different channel to each subset of the receivers. There is a common channel to all three receivers, a distinct common channel to each different pair of receivers, and a private channel to each receiver, giving seven channels in total. While results for two-receiver broadcast coding [19] show how to construct codes for the private channels and the common channel to all receivers, they do not show how to construct good codes for the common channels to pairs of receivers. For these channels, the descriptions must be chosen to work well with each other in a pairwise fashion. For instance, the description W_{12} sent to receivers 1 and 2 is decoded with W_{13} at node 1 and with W_{23} at node 2, and it must complement well both of those other two descriptions.

Prior theoretical work on broadcast coding includes work by Zhao and Effros [20, 64] that considers lossless broadcast source coding.

For simplicity, I here focus on coding for the three pairwise channels and set the rate for the other channels to zero. This makes the problem one of creating three descriptions such

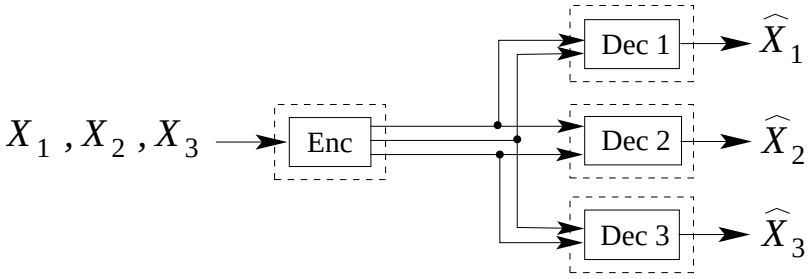


Figure 5.8: A broadcast system with three receivers.

that each different pair will be decoded at one of the decoders.

System Definition

Figure 5.8 shows the broadcast system with common information for each pair of receivers. For the channel to decoders 1 and 2, define a corresponding message index $T_{12} \in \{1, \ldots, 2^{nR_{12}}\}$. Define T_{13} and T_{23} similarly for the other two channels. The encoder and decoders for the system are defined as

$$\begin{aligned}
\alpha^n &: \mathcal{X}_1^n \times \mathcal{X}_2^n \times \mathcal{X}_3^n \to \{1, \ldots, 2^{nR_{12}}\} \times \{1, \ldots, 2^{nR_{13}}\} \times \{1, \ldots, 2^{nR_{23}}\} \\
\beta_1^n &: \{1, \ldots, 2^{nR_{12}}\} \times \{1, \ldots, 2^{nR_{13}}\} \to \hat{\mathcal{X}}_1^n \\
\beta_2^n &: \{1, \ldots, 2^{nR_{12}}\} \times \{1, \ldots, 2^{nR_{23}}\} \to \hat{\mathcal{X}}_2^n \\
\beta_3^n &: \{1, \ldots, 2^{nR_{13}}\} \times \{1, \ldots, 2^{nR_{23}}\} \to \hat{\mathcal{X}}_3^n.
\end{aligned}$$

The distortions of the code are given by

$$\Delta_j = \frac{1}{n} \sum_{i=1}^{n} d(X_{j,i}, \hat{X}_{j,i}) \qquad j = 1, 2, 3.$$

Rates $\mathbf{R} = (R_{12}, R_{13}, R_{23})$ are achievable for distortions $\mathbf{D} = (D_1, D_2, D_3)$ if, for any $\epsilon, \delta > 0$, there exists a code of some dimension having rates not exceeding $\mathbf{R}+(\epsilon, \ldots, \epsilon)$ and distortions not exceeding $\mathbf{D} + (\delta, \ldots, \delta)$. The set of all achievable $\mathbf{R}$ for fixed $\mathbf{D}$ is denoted $\mathcal{R}(\mathbf{D})$.

Results

Let $\mathbf{X} = (X_1, X_2, X_3)$ and let $\mathcal{R}_{con}(\mathbf{D})$ be the convex closure of the set of all rate triples such that there exist auxiliary random variables (W_{12}, W_{13}, W_{23}) satisfying

1.

$$\begin{aligned}
R_{12} &\geq I(\mathbf{X}; W_{12}) - \min\{I(W_{12}; W_{13}), I(W_{12}; W_{23})\} \\
R_{13} &\geq I(\mathbf{X}; W_{13}) - \min\{I(W_{12}; W_{13}), I(W_{13}; W_{23})\} \\
R_{23} &\geq I(\mathbf{X}; W_{23}) - \min\{I(W_{12}; W_{23}), I(W_{13}; W_{23})\} \\
R_{12} + R_{13} &\geq I(\mathbf{X}; W_{12}, W_{13}) \\
R_{12} + R_{23} &\geq I(\mathbf{X}; W_{12}, W_{23}) \\
R_{13} + R_{23} &\geq I(\mathbf{X}; W_{13}, W_{23}) \\
R_{12} + R_{13} + R_{23} &\geq I(\mathbf{X}; W_{12}, W_{13}, W_{23}).
\end{aligned}$$

2. There exist functions f_1, f_2, f_3 satisfying

$$Ed(X_1, f_1(W_{12}, W_{13})) \leq D_1, \ Ed(X_2, f_2(W_{12}, W_{23})) \leq D_2, \ Ed(X_3, f_3(W_{13}, W_{23})) \leq D_3.$$

The following theorem, proved in Appendix E, is the converse result.

Theorem 22 $\mathcal{R}(\mathbf{D}) \subseteq \mathcal{R}_{con}(\mathbf{D})$.

Now let $\mathcal{R}_{ach}(\mathbf{D})$ be the convex closure of the set of all rate triples such that there exist auxiliary random variables (W_{12}, W_{13}, W_{23}) satisfying

1.

$$\begin{aligned}
R_{12} &> I(\mathbf{X}; W_{12}) - \min\{I(W_{12}; W_{13}), I(W_{12}; W_{23})\} \\
R_{13} &> I(\mathbf{X}; W_{13}) - \min\{I(W_{12}; W_{13}), I(W_{13}; W_{23})\} \\
R_{23} &> I(\mathbf{X}; W_{23}) - \min\{I(W_{12}; W_{23}), I(W_{13}; W_{23})\}
\end{aligned}$$

$$\begin{aligned}
R_{12}+R_{13} &> I(\mathbf{X};W_{12},W_{13})+I(W_{12};W_{13}) \\
&\quad -\min\{I(W_{12};W_{13}),I(W_{12};W_{23})+I(W_{13};W_{23})\} \\
R_{12}+R_{23} &> I(\mathbf{X};W_{12},W_{23})+I(W_{12};W_{23}) \\
&\quad -\min\{I(W_{12};W_{23}),I(W_{12};W_{13})+I(W_{13};W_{23})\} \\
R_{13}+R_{23} &> I(\mathbf{X};W_{13},W_{23})+I(W_{13};W_{23}) \\
&\quad -\min\{I(W_{13};W_{23}),I(W_{12};W_{13})+I(W_{12};W_{23})\} \\
R_{13}+R_{13}+R_{23} &> H(W_{12})+H(W_{13})+H(W_{23})-H(W_{12},W_{13},W_{23}|\mathbf{X})-\phi,
\end{aligned}$$

where

$$\begin{aligned}
\phi &= \max_{(n_{12},n_{13},n_{23})\in\mathcal{N}}(n_{12}+n_{13}+n_{23}) \\
\mathcal{N} &= \{(n_{12},n_{13},n_{23}): n_{12}\geq 0, n_{13}\geq 0, n_{23}\geq 0, \\
&\quad n_{12}+n_{13}\leq I(W_{12};W_{13}), n_{12}+n_{23}\leq I(W_{12};W_{23}), n_{13}+n_{23}\leq I(W_{13};W_{23})\}.
\end{aligned}$$

2. There exist functions f_1, f_2, f_3 satisfying

$$Ed(X_1, f_1(W_{12},W_{13}))\leq D_1,\ Ed(X_2, f_2(W_{12},W_{23}))\leq D_2,\ Ed(X_3, f_3(W_{13},W_{23}))\leq D_3.$$

The following theorem, proved in Appendix E, is the achievability result.

Theorem 23 $\mathcal{R}_{ach}(\mathbf{D})\subseteq\mathcal{R}(\mathbf{D})$.

The maximization to obtain ϕ in the definition of the achievability result is evaluated later in this section.

Remark 1 For the case $R_{13}=0$, the achievability result simplifies to that of [24, Thm EGC*], which is known to be loose in general but tight for Gaussian sources. Thus, the achievability result above is also loose in general, but may be tight for Gaussian sources.

Remark 2 For the case in which the auxiliary random variables are symmetric, with identical entropies and with $I(W_{12};W_{13}) = I(W_{12};W_{23}) = I(W_{13};W_{23})$ (as we might expect if the rates are identical and the original sources themselves are symmetric), the converse bounds simplify to

$$\begin{aligned} R_{12} = R_{13} = R_{23} &> I(\mathbf{X};W_{12}) - I(W_{12};W_{13}) \\ R_{12} + R_{13} = R_{12} + R_{23} = R_{13} + R_{23} &> I(\mathbf{X};W_{12},W_{13}) \\ R_{13} + R_{13} + R_{23} &> H(W_{12},W_{13},W_{23}) - I(W_{12},W_{13},W_{23}|\mathbf{X}). \end{aligned}$$

The achievability bounds simplify to

$$\begin{aligned} R_{12} = R_{13} = R_{23} &> I(\mathbf{X};W_{12}) - I(W_{12};W_{13}) \\ R_{12} + R_{13} = R_{12} + R_{23} = R_{13} + R_{23} &> I(\mathbf{X};W_{12},W_{13}) \\ R_{13} + R_{13} + R_{23} &> 3H(W_{12}) - \frac{3}{2}I(W_{12};W_{13}) - H(W_{12},W_{13},W_{23}|\mathbf{X}). \end{aligned}$$

These match in all but total rate.

Outline of Achievability Result

To construct the achievability result, I borrow techniques from both multiple description and Slepian-Wolf coding. Adopting the basic approach of multiple-description achievability results [24, Theorems EGC*, EGC, 1], I introduce an auxiliary random variable for each of the three descriptions, and construct a codebook by drawing sequences from the individual typical sets. Correct operation of the decoders requires that there exists a triple of sequences such that each pair is jointly typical. Rather than determining this number explicitly, I instead choose enough indices to ensure the existence of a triple for which all three

sequences are jointly typical with $\mathbf{X}$ (a stronger condition than required), and then remove the redundancy using Slepian-Wolf coding.

Let the three descriptions for the problem be identified with the random variables $\mathbf{W} = (W_{12}, W_{13}, W_{23})$. The following lemma gives the number of indices required to ensure that a jointly typical triple of sequences exists.

Lemma 7 *[25, Pg 2112-2113] Using $p(x, w_{12}, w_{13}, w_{23})$, draw $2^{nR_{12}}$ sequences $W_{12}^n(k_{12})$ uniformly with replacement from $A_\epsilon^{*(n)}(W_{12})$. Similarly, draw $2^{nR_{13}}$ sequences $W_{13}^n(k_{13})$ with replacement from $A_\epsilon^{*(n)}(W_{13})$, and $2^{nR_{23}}$ sequences $W_{23}^n(k_{23})$ with replacement from $A_\epsilon^{*(n)}(W_{23})$. Suppose $\mathbf{X}^n \in A_\epsilon^{*(n)}$ is given. For any $\delta > 0$,*

$$\Pr\{\exists \mathbf{k} = (k_{12}, k_{13}, k_{23}) : (\mathbf{X}^n, \mathbf{W}^n(\mathbf{k})) \in A_\epsilon^{*(n)}\} > 1 - \delta$$

provided that

$$\begin{aligned}
R_{12} &> I(\mathbf{X}; W_{12}) \\
R_{13} &> I(\mathbf{X}; W_{13}) \\
R_{23} &> I(\mathbf{X}; W_{23}) \\
R_{12} + R_{13} &> I(\mathbf{X}; W_{12}, W_{13}) + I(W_{12}; W_{13}) \\
R_{12} + R_{23} &> I(\mathbf{X}; W_{12}, W_{23}) + I(W_{12}; W_{23}) \\
R_{13} + R_{23} &> I(\mathbf{X}; W_{13}, W_{23}) + I(W_{13}; W_{23}) \\
R_{13} + R_{13} + R_{23} &> H(W_{12}) + H(W_{13}) + H(W_{23}) - H(W_{12}, W_{13}, W_{23}|\mathbf{X}).
\end{aligned}$$

The following section determines how much redundancy can be eliminated by Slepian-Wolf coding in a three-receiver system, giving the solution to the maximization to obtain ϕ in the statement of the achievability result.

Three-receiver Slepian-Wolf Coding

Consider the three-encoder, three-decoder system of Figure 5.9. Encoders 12, 23, and 13 describe sources W_{12}, W_{23}, and W_{13} at rates R_1, R_2, and R_3, respectively. Decoder 1 reproduces W_{12} and W_{13}, decoder 2 reproduces W_{12} and W_{23}, and decoder 3 reproduces W_{13} and W_{23}. The achievable rate region $\mathcal{R}$ for this problem is given by applying a theorem of Csiszár and Körner [85].

Theorem 24 *[85, Theorem 1] The achievable rate region for the system of Figure 5.9 is given by*

$$\begin{aligned}
R_{12} &> \max\{H(W_{12}|W_{13}), H(W_{12}|W_{23})\} \\
R_{13} &> \max\{H(W_{13}|W_{12}), H(W_{13}|W_{23})\} \\
R_{23} &> \max\{H(W_{23}|W_{12}), H(W_{23}|W_{13})\} \\
R_{12}+R_{13} &> H(W_{12}, W_{13}) \\
R_{12}+R_{23} &> H(W_{12}, W_{23}) \\
R_{13}+R_{23} &> H(W_{13}, W_{23}).
\end{aligned}$$

The following theorem, proved in Appendix E, explicitly determines the minimum achievable total rate $R_{min} = \min_{(R_{12},R_{13},R_{23})\in\mathcal{R}}\{R_{12}+R_{13}+R_{23}\}$.

Theorem 25 *Assume, without loss of generality, that $I(W_{12};W_{13}) \le I(W_{13};W_{23})$. Then*

$$R_{min} = \begin{cases} I(W_{12};W_{13}) + I(W_{12};W_{23}), & 0 \le I(W_{12};W_{23}) \le I(W_{13};W_{23}) - I(W_{12};W_{13}) \\ \frac{I(W_{12};W_{13})+I(W_{12};W_{23})+I(W_{13};W_{23})}{2}, & I(W_{13};W_{23}) - I(W_{12};W_{13}) \\ & \quad \le I(W_{12};W_{23}) \le I(W_{12};W_{13}) + I(W_{13};W_{23}) \\ I(W_{12};W_{13}) + I(W_{13};W_{23}), & I(W_{12};W_{23}) \ge I(W_{12};W_{13}) + I(W_{13};W_{23}). \end{cases}$$

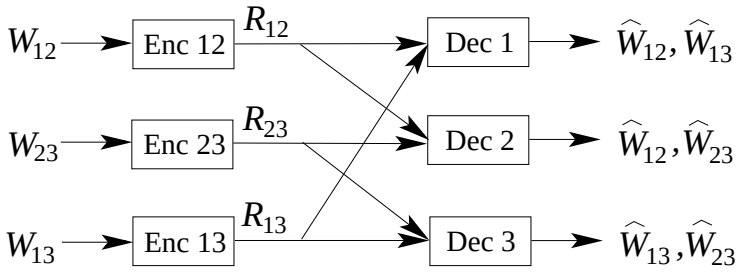


Figure 5.9: A three-encoder, three-decoder lossless coding system.

5.4 Summary

This chapter looks at three specific topics in source coding for general networks.

First, the role of feedback in lossless network source coding is examined. Although many common network systems (e.g., broadcast and multiple access) do not benefit from feedback in the lossless case, I show that there exist some systems for which a limited amount of feedback can enlarge the achievable rate-distortion region by an arbitrary amount.

Second, I present a method to develop converse results bounding the source coding rates in a general network. The approach uses cutsets to partition the network nodes into two sets. It is easy to apply but often yields converses that are quite loose or do not describe the tradeoff between common and private rates. More flexible bounds are obtained by partitioning the network nodes into three or more sets; I present results for a broadcast-like three-set partitioning. Further generalizations to more than three sets are unlikely to be of practical use due to the increasing numbers of auxiliary random variables required for the converse's description. These make its evaluation difficult.

Finally, I look at lossy broadcast coding. I present a non-matching achievability and converse result for a general three-receiver broadcast system. Although the general model for even three receivers proves difficult to analyze, I argue that practical systems do not

resemble the general case in that they send common information only to subsets of receivers that are related, rather than to every possible subset of receivers. Modeling relations between receivers using a tree-structure, I derive matching converse and achievability results for a broadcast network of arbitrary size.

Chapter 6

Conclusions

In this thesis, I examine applied and theoretical issues in network source coding with the aim of designing good network compression systems.

The applied work considers the design of rate-distortion-optimal vector quantizers for general network topologies. I extend the vector quantization design algorithm of [1] to allow for the presence of side information at the decoders and the possibility of channel errors. I then show in detail how to implement the algorithm and thus design vector quantizers for any network. The design considers both fixed- and variable-rate coding, although the applicability of variable-rate network quantizers is limited by our knowledge of lossless network coding. For all but the simplest of lossless coding networks either the optimal achievable rates are unknown, we do not know how to convert the optimal rates into optimal entropy constraints for NVQ design, or we do not have practical entropy codes that approximate the optimal achievable rates.

Experimental results obtained using the algorithm demonstrate that incorporating network topology into code design can yield significant rate-distortion benefits when the network

sources are correlated. On a satellite weather data set, network codes show improvements of more than 1 dB over codes designed independently for each source. For highly correlated Gaussian sources ($\rho \geq 0.6$), improvements of more than 2 dB are possible. These results prove that network source coding is important for applications involving highly correlated sources.

Reducing the complexity of network code design is likely to be a crucial factor in increasing its viability for practical applications. A step in the this direction has already been taken in [65], which uses Chapter 3's extension of the optimality conditions from [1] in a new, low-complexity vector quantization design algorithm. Future work in practical network code design should also consider codes other than VQs; hopefully, the lessons learned here from VQ design will prove useful in adapting other compression techniques (such as wavelet-based coding) from point-to-point systems to networks.

I also investigate several theoretical topics with the aim of extracting insights into the structure and the optimal performance of network data compression codes. Motivated by the availability of side information in applications such as sensor networks, I derive the rate-distortion function for a network with some side information available only at the decoder, and some available to both encoder and decoder. I evaluate this rate-distortion function for the special case of Gaussian and binary sources. Using this result, I solve a similar binary example for the system in which the presence of side information at the decoder is unreliable [2], closing the gap between existing bounds on the solution for that example. The rate-distortion function for the second system involves two auxiliary random variables. Examination of the form of the optimal auxiliary random variables yields insight into the structure of the solution. Both the binary and Gaussian cases exhibit a property akin to

successive refinement, but with side information present at the refinement decoder. Also, in both cases, a two-part coding strategy is optimal, suggesting that the same might be true for all successively refinable sources. I conjecture that a two-part coding strategy is likely to be efficient even for more general sources; rate-loss results for multi-resolution codes show that all sources are nearly successively refinable.

Since most lossless coding systems studied do not benefit from feedback, I ask and answer the question of whether feedback from a decoder to an encoder can enlarge the achievable rate region in lossless coding. A simple example shows that it can and also demonstrates that a limited amount of feedback can yield an arbitrarily large decrease in the rate required by one of the encoders as the alphabet sizes of the sources increase without bound. Although the example considers an extreme case, it illustrates that source coding feedback may be important to consider in network compression system design. Further work in this area might test the magnitude of the benefit available from feedback using real-world examples.

Rate-distortion regions for general networks are often difficult to derive and to evaluate for particular sources. I demonstrate how to use cutsets to derive simple network converses by bounding the information rates flowing between different subsets of a network. The resulting converses are easy to evaluate, but are often quite loose.

Finally, I develop two rate-distortion results for broadcast source coding. Adopting a degraded broadcast model, I look at source coding for a broadcast tree network, and derive its rate-distortion region. I also combine multiple-description and Slepian-Wolf coding techniques to derive an achievability and a converse result for three-receiver broadcast source coding.

Rate-distortion results give the greatest insight into practical code design when the opti-

mal forms of the auxiliary random variables are determined. I detail the form of the auxiliary random variable for the mixed side information system in Chapter 4. However, as shown in that chapter, finding the optimal form is not easy even for simple networks and simple sources. Thus, it seems unlikely that significant progress toward better code design will be accomplished by systematically pursuing precise rate-distortion results for larger and more complicated networks. However, many of today's potential applications do involve large networks. To design practical systems for these applications, we need to begin considering how to apply what we already know to larger networks, even if we do not achieve the theoretically optimal performance by doing so. We might also relax our definition of theoretical optimality to mean simply "scales optimally in the number of network nodes" rather than optimal down to the last bit. The techniques described in this thesis give a few initial steps toward pursuing practically and theoretically viable new angles on rate-distortion theory for large networks.

Appendix A

The Satellite Weather Image Data Set

The satellite weather data set used Chapter 3 was obtained courtesy of NASA and the University of Hawaii. It contains collections of images from three geosynchronous weather satellites. Each satellite records 8-bit greyscale images in frequency bands ranging from infrared to the visible spectrum. For each satellite, I use images from three bands. Each image is cropped to 512×512 pixels. Table A.1 shows the assignment of satellite images to data sources for the WZ, 2AWZ, 2A, three-node, MD, and BC system experiments. Figure A.1 shows sample images. The training and testing sets are non-overlapping and consist of eight and four images per source, respectively.

Satellite Name	Frequency Band	WZ	2AWZ	2A	Three-Node	MD	BC
GMS-5	Visible				$X_{1,\{2,3\}}$		
GMS-5	Infrared 1				$X_{1,2}$		
GMS-5	Infrared 2				$X_{1,3}$		
GOES-8	Visible	Z_1	Z_1	Z_1	$X_{2,\{1,3\}}$	X	$X_{1,\{2,3\}}$
GOES-8	Infrared 2	$X_{2,1}$	$X_{2,1}$	$X_{2,1}$	$X_{2,3}$		$X_{1,2}$
GOES-8	Infrared 5		$X_{3,1}$		$X_{2,1}$		$X_{1,3}$
GOES-10	Visible				$X_{3,\{1,2\}}$		
GOES-10	Infrared 2				$X_{3,1}$		
GOES-10	Infrared 5				$X_{3,2}$		

Table A.1: Data source assignments for the NVQ experiments

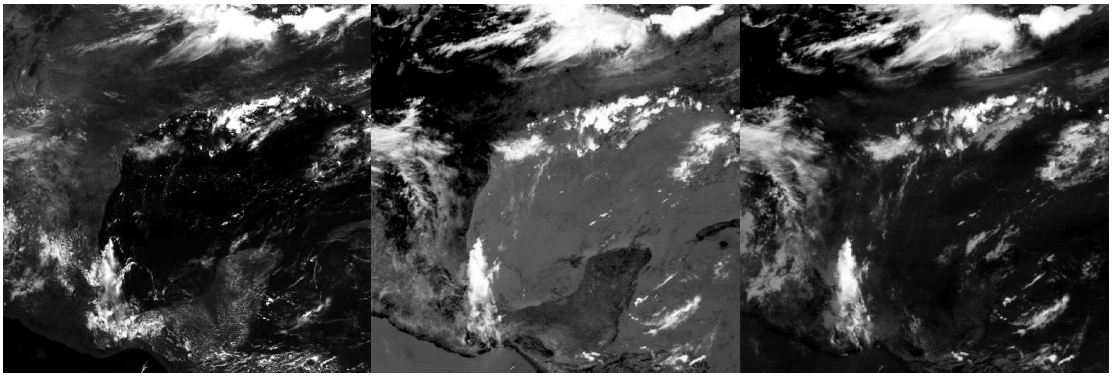


Figure A.1: Sample images from the GOES-8 weather satellite. From left to right: visible spectrum, infrared 2, infrared 5.

Appendix B

The Binary MSI Example

In this appendix I apply Lagrange multipliers to the minimization required to determine $R_{X|\{Z\}}(q, D, f)$. I use the function f described by Table 4.2 to illustrate our method. For this f, I seek to minimize (4.9) subject to the conditions in (4.10)-(4.14). There are six inequality constraints; the application of Lagrange multipliers will depend on the subset that is active. Assume first that none of the constraints is active. I use the objective function and the equality constraints to form the Lagrangian

$$\begin{aligned} J(a_0, a_1, a_2, r_0, r_1, r_2) &= r_0 G(a_0) + r_1 G(a_1) + r_2 G(a_2) + \lambda_1(r_0 + r_1 + r_2 - 1) \\ &\quad + \lambda_2(r_0 a_0 + r_1 a_1 + r_2 a_2 - p_0) + \lambda_3(r_0 a_0 + r_1 a_1 + r_2 q_0 - D), \end{aligned}$$

and obtain the first-order optimality conditions by differentiating $J(a_0, a_1, a_2, r_0, r_1, r_2)$:

$$\begin{aligned} \frac{\partial}{\partial r_0}: \quad G(a_0) &= \lambda_1 + a_0\lambda_2 + \lambda_3 \\ \frac{\partial}{\partial r_1}: \quad G(a_1) &= \lambda_1 + a_1\lambda_2 + \lambda_3 \\ \frac{\partial}{\partial r_2}: \quad G(a_2) &= \lambda_1 + a_2\lambda_2 + q_0\lambda_3 \end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial a_0}: \qquad & G'(a_0) &&= \lambda_2 + \lambda_3 \\
\frac{\partial}{\partial a_1}: \qquad & G'(a_1) &&= \lambda_2 + \lambda_3 \\
\frac{\partial}{\partial a_2}: \qquad & G'(a_2) &&= \lambda_2. \qquad \text{(B.1)}
\end{aligned}$$

These first-order conditions relate $(a_0, a_1, a_2, \lambda_1, \lambda_2, \lambda_3)$; the equality constraints then give (r_0, r_1, r_2) in terms of (a_0, a_1, a_2). For this example, the first order conditions governing a_0 and a_1 are identical. There is a unique optimal value for the minimization, so it must be achieved when $a_0 = a_1$. This is a pattern that is followed for all f: when the decoding rules for two symbols i and j are equal, i.e., when $f(i, z) = f(j, z)\ \forall z \in \{0, 1\}$, then $a_i = a_j$. Symbols i and j can therefore be combined to form a single symbol with transition probability a_i and marginal probability $r_i + r_j$. The solution then has at most two symbols and four parameters, three of which can be fixed by the equality constraints. This reduces the optimization to a search over one free parameter as desired.

The solution for other decoding functions f is found using a similar approach, and the symmetry of the problem reduces the number of decoding functions that need be considered. From Table 4.1 and the resulting discussion, there are at most three decoding rules $f(w, \cdot)$ that need be considered for each symbol. The problem is symmetric in the three symbols, hence it is only the number of symbols using each decoding rule that is important in computing a solution. When all three symbols have the same decoding rule, then the optimal transition probabilities are the same for each and the three symbols can be combined into one. This one-symbol solution has no free parameters. When two symbols have the same decoding rule, the optimal solution uses two symbols and has only one free parameter. Finally, when all three symbols have different decoding rules, the first order conditions have a

solution that we give in terms of a_0 as

$$(a_1, a_2, \lambda_1, \lambda_2, \lambda_3) = \left(\frac{1}{2}, (1-a_0), -q_0 G'(a_0), 0, G'(a_0)\right),$$

where a_0 itself is the solution to $G(a_0) = (a_0 - q_0)G'(a_0)$. I solve for a_0 numerically and obtain r_0, r_1, and r_2 from the equality constraints.

Considering now the case when some of the inequality constraints are made active (i.e., are made into equality constraints), observe the following:

1. Assume the constraint (4.13) on r_w, $w \in \{0, 1, 2\}$ is active; that is, $r_w = 0$. Then symbol $W = w$ is never used and the optimization need consider only solutions with at most two symbols.

2. If all three of the constraints in (4.14) are active, then all parameters are uniquely determined. If two are active, we have only one free parameter.

From the above, either (1) we require only one or two symbols, (2) two or more of the a_w are constrained, or (3) none of the constraints on the r_w and at most one of the constraints on the a_w are active. Since cases (1) and (2) both leave at most one free parameter, I can easily compute the optimal solution for each. The non-trivial solutions arising from case (3) are listed below.

The first is when $a_0 = 0$ and the distortion constraint is $D = r_1 q_0 + r_2(1 - a_2)$. The first-order conditions yield

$$G(a_1) \;=\; H(q_0) + (a_1 + q)G'(a_1) - q_0 G'(a_2). \tag{B.2}$$

I search over a_2 and use (B.2) to find a_1 given a_2.

The second is when $a_0 = 1$ and the distortion constraint is $D = r_1 a_1 + r_2 q_0$. I search over a_2 and obtain a_1 numerically from

$$G(a_1) \quad = \quad H(q_0) - G'(a_2) + a_1 G'(a_1).$$

The above analysis reduces the search for the optimal W to three solution classes:

1. Solutions with only one or two symbols
2. Solutions with $a_1 = \frac{1}{2}$, $a_2 = 1 - a_0$, and a_0 found numerically
3. Solutions where exactly one of the boundary constraints on a_w is active.

Numerical experiments suggest that the best solutions from class 3 never outperform the best solutions from classes 1 and 2.

Appendix C

Differentiability of $K(a_u)$

To apply the first order optimality conditions in the Heegard and Berger problem, I need to ensure that $K(a_u) = R(a_u, q_0, D_1) - H(a_u)$ is differentiable with respect to a_u for $0 < a_u < 1$. Here, $R(a_u, q_0, D_1)$ is the binary MSI rate-distortion function, determined in Section 4.4, when $p_0 = a_u$. Since $H(a_u)$ is differentiable with respect to a_u for $0 < a_u < 1$, it remains to ensure that $R(a_u, q_0, D_1)$ is also differentiable for $0 < a_u < 1$.

As shown in Lemma 8 below, $R(a_u, q_0, D_1)$ is a continuous function of a_u, differentiable at all but a finite number of points. In an arbitrarily small neighborhood around each of these points, the function can be smoothed to make it differentiable; this can be done without changing the functional value by more than an arbitrarily small amount $\epsilon > 0$. I substitute the smoothed (and differentiable) form of $R(a_u, q_0, D_1)$ for the original in the definition of $K(a_u)$. By doing so, I alter the function I am minimizing in the Heegard and Berger problem by at most ϵ; the effect on the derived rate-distortion result of assuming differentiability of $K(a_u)$ is thus negligible.

Lemma 8 *$R(p_0, q_0, D_1)$ is a continuous function of p_0, differentiable at all but a finite number of points.*

Proof of Lemma 8 $R(p_0, q_0, D_1)$ is given by (4.8) as

$$\begin{aligned} R(p_0, q_0, D_1) &= \min_{f \in \mathcal{F}} \min_{\mu(w|x) \in \mathcal{M}(p_0, q_0, D_1, f)} \left(-G(p_0) + \sum_{w=0}^{2} r_w G(a_w) \right) \\ &= -G(p_0) + \min_{\mu(w|x) \in \mathcal{M}_{X|\{Z\}}(p_0, q_0, D_1, f)} \left(\sum_{w=0}^{2} r_w G(a_w) \right), \end{aligned}$$

where $\mathcal{M}_{X|\{Z\}}(p_0, q_0, D_1, f)$ is the set of all test channels describing an auxiliary random variable W with $|\mathcal{W}| \leq 3$ such that $W \to X \to Z$ and $Ed(X, f(W, Z)) \leq D_1$. Differentiability of $G(p_0)$ is shown in [12]; we concentrate on showing that the remaining term is differentiable.

For any p_0, the minimum over all test channels in the definition of $R(p_0, q_0, D_1, f)$ can be obtained by a test channel with $|\mathcal{W}| = 3$. That no more than three symbols are needed is established in [12], and for any solution with fewer than three symbols, there always exists a corresponding three-symbol solution that yields the same minimum value. (For instance, if the two-symbol solution (a_0, a_1, r_0, r_1) is optimal, then so is the three-symbol solution $(a_0', a_1', a_2', r_0', r_1', r_2') = (a_0, a_1, a_1, r_0, \frac{r_1}{2}, \frac{r_1}{2})$.) For each f, partition $\mathcal{M}(p_0, q_0, D_1, f)$ into a set of interior test channels (all $a_w \in (0, 1)$) and sets of different types of boundary test channels (having one or more $a_w \in \{0, 1\}$). For each of these sets, a set of first-order conditions similar to those in (B.1) is derived. For all sets, I obtain the same result as we did for the conditions in (B.1): the first-order conditions uniquely determine the values of a_0, a_1, and a_2. These values are independent of p_0. The value of p_0 affects only how r_0, r_1, and r_2 are determined as functions of a_0, a_1, and a_2 by applying the equality constraints. Moreover,

the functions specifying r_0, r_1, and r_2 are always linear functions of p_0, because the equality constraint (4.6) is a linear function of p_0. Since the objective function $\sum_{w=0}^{2} r_w G(a_w)$ in the minimization is also a linear function of (r_0, r_1, r_2), this implies that for any of the sets, the minimal value of the objective function changes as a linear function of p_0.

To find $\min_{f\in\mathcal{F}} \min_{\mu(w|x)\in\mathcal{M}_{X|\{Z\}}(p_0,q_0,D_1,f)} \left(\sum_{w=0}^{2} r_w G(a_w)\right)$ as a function of p_0, take the minimum of the solutions yielded by the different sets for each f, followed by the minimum over all f. There are a finite number of functions to consider in the minima, and each is linear in p_0. The desired result follows from the observation that the minimum of a finite number M of linear functions is continuous and is differentiable at all but at most $M-1$ points. □

Appendix D

The Binary HB Example

This appendix outlines the results of applying Lagrange multipliers to the minimization of (4.19) subject to the conditions in (4.15)-(4.18) and a distortion constraint.

There are only two possible decoding rules for each symbol: $f(u) = 0$ or $f(u) = 1$. When $f(u) = 0$, then that symbol contributes an expected distortion of $r_u a_u$; when $f(u) = 1$ it contributes an expected distortion of $r_u(1 - a_u)$.

Consider first the case in which none of the inequality constraints is active. The application of Lagrange multipliers yields that $a_u = c_{1,f}$ for all u such that $a_u \leq 1 - a_u$, and $a_u = c_{2,f}$ otherwise. Since symbols with identical transition probabilities can be combined, then for any f the optimal U requires only two symbols. The two-symbol solution has four parameters, (a_0, a_1, r_0, r_1). Three can be determined from the equality constraints, leaving one to search over. The evaluation of $R(a_u, q_0, D_1)$ also involves a search over one free parameter (as shown in the previous section), so that evaluating the optimal U requires a search over two parameters.

Now consider the case when one or more of the inequality constraints are active. The

inequality constraints on r_u are of little interest since setting any particular r_u to zero simply reduces the number of symbols by one. In applying the inequality constraints on the transition probabilities a_u, first note that if $a_i = a_j = 0$, or $a_i = a_j = 1$ then symbols i and j can be combined into a single symbol. Therefore, there are only three cases of boundary solutions: one or more of the transition probabilities is zero, one or more is one, or some are zero and some are one. In all cases the boundary solution can be computed with a search over at most two parameters. I list below the three cases that require numerical solution of one or more parameters.

1. When $a_0 = 0$ and $D = a_1 r_1 + (1 - a_2) r_2$, then obtain a_1 and a_2 numerically in sequence from

$$\begin{aligned} K(a_1) &= a_1 K'(a_1) \\ K(a_2) &= (a_2 - \frac{1}{2}) K'(a_2) + \frac{1}{2} K'(a_1). \end{aligned}$$

2. When $a_0 = 1$ and $D = a_1 r_1 + (1 - a_2) r_2$, then obtain a_1 and a_2 numerically in sequence from

$$\begin{aligned} K(a_2) &= (a_2 - 1) K'(a_2) \\ K(a_1) &= (a_1 - \frac{1}{2}) K'(a_1) - \frac{1}{2} K'(a_2). \end{aligned}$$

3. When $a_0 = 0$, $a_1 = 1$, and $D = r_1 + a_2 r_2 + (1 - a_3) r_3$, then obtain a_2 and a_3 numerically from

$$\begin{aligned} K(a_2) &= a_2 K'(a_2) \\ K(a_3) &= (a_3 - 1) K'(a_3). \end{aligned}$$

Thus, when one or more of the inequality constraints is active, the solution can still be evaluated by searching over a total of two parameters.

Appendix E

Proofs of Lemmas and Theorems

Theorem 16 *For an i.i.d. source X, $R(D) = R_{FB}(D)$, where $R_{FB}(D)$ is the rate-distortion function when feedback is permitted from the decoder to the encoder.*

Proof of Theorem 16 Since feedback cannot increase the required rate $R(D) \geq R_{FB}(D)$, and it suffices to show that $R_{FB}(D) \geq R(D)$.

Consider a general feedback system for the problem, defined as follows. Encoder α receives a source vector $X^n = (X_1, X_2, \ldots, X_n)$. It transmits to the decoder its first symbol $\hat{X}_1$, an arbitrary function of X^n. The decoder β transmits back Y_1, a function of $\hat{X}_1$ and possibly some noise Z_1. Here Z_1 must be independent of X^n since we have assumed that there is no side information available to the decoder. Upon receipt of Y_1, α then transmits $\hat{X}_2$, a function of X^n and Y_1, to β, and β transmits back Y_2, a function of $(\hat{X}_1, \hat{X}_2, Z_1, Z_2)$. In the ith round, $\hat{X}_i$ is a function of X^n and Y^{i-1}, and Y_i is a function of $(\hat{X}^i, Z^i)$.

Under this feedback scheme, consider a code (α, β) of some dimension n that uses feedback

and achieves distortion

$$\frac{1}{n}\sum_{i=1}^{n} d(X_i, \hat{X}_i) = D.$$

Then

$$\begin{aligned}
nR &\geq H(\hat{X}^n) \\
&\geq H(\hat{X}^n) - H(\hat{X}^n|X^n, Y^n, Z^n) \\
&\overset{(a)}{=} H(\hat{X}^n) - H(\hat{X}^n|X^n, Z^n) \\
&= I(\hat{X}^n; X^n, Z^n) \\
&\geq I(\hat{X}^n; X^n) \\
&= H(X^n) - H(X^n|\hat{X}^n) \\
&= \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|X^{i-1}, \hat{X}^n) \\
&\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) \\
&= \sum_{i=1}^{n} I(X_i; \hat{X}_i) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} R(Ed(X_i, \hat{X}_i)) \\
&= n\sum_{i=1}^{n} \frac{1}{n} R(Ed(X_i, \hat{X}_i)) \\
&\overset{(c)}{\geq} nR(\frac{1}{n}\sum_{i=1}^{n} Ed(X_i, \hat{X}_i)) \\
&= nR(D),
\end{aligned}$$

where the labeled steps are justified by the following.

(a) Y^n is a function of X^n and Z^n.

(b) By the definition of the rate-distortion function without feedback.

(c) By the convexity of the rate-distortion function without feedback.

Thus, any code that achieves distortion D with feedback has rate $R > R(D)$, implying $R_{FB}(D) \geq R(D)$ as required. □

Theorem 17 *For the two-user multiple access system with i.i.d. sources, the achievable rate region for the case when feedback is permitted from the decoder to each of the encoders is the same as the achievable rate region for the case when no feedback is permitted.*

Proof of Theorem 17 Let the sources be X_1 and X_2, let the encoders of X_1 and X_2 be α_1 and α_2, respectively, and let the decoder be β. Let the rate of α_1 be R_1 and the rate of α_2 be R_2. The rate region without feedback is given by the result of Slepian and Wolf [11]

$$\begin{aligned} R_1 &\geq H(X_1|X_2) \\ R_2 &\geq H(X_2|X_1) \\ R_1 + R_2 &\geq H(X_1, X_2). \end{aligned}$$

Achievability of these rates with feedback is immediate; I show below the converse. Consider an n-dimensional code $(\alpha_1^n, \alpha_2^n, \beta^n)$ that uses feedback. Denote by T_1 and T_2 the messages produced by α_1 and α_2 respectively. Decoder β^n must recover X_1^n and X_2^n with arbitrarily low probability of error. Assume the most comprehensive feedback possible to α_1: the decoder feeds back X_2^n in its entirety to α_1^n, so that $T_1 = \alpha_1(X_1^n, X_2^n)$. The rate required by encoder 1 is bounded according to

$$\begin{aligned} R_1 &\geq H(T_1) \\ &\geq H(T_1|X_2^n) \\ &\geq I(X_1^n; T_1|X_2^n) + H(T_1|X_1^n, X_2^n) \end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{=} I(X_1^n; T_1 | X_2^n) \\
&= H(X_1^n | X_2^n) - H(X_1^n | T_1, X_2^n) \\
&\overset{(b)}{=} H(X_1^n | X_2^n) - H(X_1^n | T_1, T_2, X_2^n) \\
&\overset{(c)}{\geq} H(X_1^n | X_2^n) - n\epsilon_n \\
&\overset{(d)}{=} H(X_1 | X_2) - n\epsilon_n,
\end{aligned}
$$

where $\epsilon_n \to 0$ as $n \to \infty$ and the labeled steps are justified by the following.

(a) T_1 is a function of X_1^n and X_2^n.

(b) T_2 is a function of X_2^n (and possibly part of T_1 via feedback from β^n to α_2).

(c) By Fano's inequality.

(d) (X_1, X_2) are i.i.d.

By a symmetric argument, $R_2 \geq H(X_2|X_1)$. It remains to show the bound on $R_1 + R_2$. To establish that, observe that separate encoding with feedback can never be more efficient than joint coding with feedback. Joining the two encoders reduces the system to a point-to-point system of rate R with encoder α^n that observes X_1^n and X_2^n. Theorem 16 proves that feedback does not help in such a system, and by Shannon [10], the rate required for the joint system without feedback is $R \geq H(X_1, X_2)$. Thus, the separated encoders with feedback must use rate at least $R_1 + R_2 \geq R \geq H(X_1, X_2)$. □

Theorem 20 *Let (A, B, C) be any partition of the network nodes into three sets for which the nodes in B do not share information with the nodes in C about the sources transmitted to them from A. Let (U, V, Y, Z) be drawn i.i.d. with joint distribution $p(u, v, y, z)$, and let $\mathcal{M}$ be the set of all auxiliary random variables W jointly distributed with (U, V, Y, Z) such*

that $W \to (U,V,Y) \to Z$ *forms a Markov chain. Let*

$$\begin{aligned}\mathcal{R}(W, D_U, D_V) = \{(R_{A\to(B,C)}, R_{A\to B}, R_{A\to C}) \quad : \quad & R_{A\to(B,C)} \geq I(U,V;W|Y,Z) \\ & R_{A\to B} \geq R_{U|Y\{Z\}}(D_U) \\ & R_{A\to C} \geq R_{V|Y\{Z\}}(D_V)\}.\end{aligned}$$

The set $\mathcal{R}^*(D_U, D_V)$ *of achievable* $(R_{A\to(B,C)}, R_{A\to B}, R_{A\to C})$ *for distortions* (D_U, D_V) *satisfies*

$$\mathcal{R}^*(D_U, D_V) \subseteq \overline{\bigcup_{W\in\mathcal{M}} \mathcal{R}(W, D_U, D_V)}.$$

The proof of Theorem 20 uses the lemma below, which follows from the definition of the MSI rate-distortion function.

Lemma 9 *Let* U, V, Y, Z *be i.i.d. sources. Let* $d'((u,v),(\hat{u},\hat{v})) = d(u,\hat{u})$ *be a distortion measure satisfying the two conditions for the MSI rate-distortion function. Then*

$$R'_{UV|Y\{Z\}}(D) \leq R_{U|Y\{Z\}}(D),$$

where $R'_{UV|Y\{Z\}}(D)$ *and* $R_{U|Y\{Z\}}(D)$ *are the MSI rate-distortion functions measured with respect to distortion measures* d' *and* d *respectively.*

Proof of Theorem 20 The following steps reduce the network to the system in Figure 5.5. At each step, the minimal rates required to flow from A to B and C are guaranteed not to increase.

- Combine all of the nodes in A to a single node knowing (U,V,Y)
- Combine all of the nodes in B to a single node knowing (Y,Z).
- Combine all of the nodes in C to a single node knowing (Y,Z).

It remains to show that the bounds apply to the system of Figure 5.5. For simplicity, rewrite $R_{A\to(B,C)} = R_{UV}$, $R_{A\to B} = R_U$, and $R_{A\to C} = R_V$.

Given a code $(\alpha, \beta_1, \beta_2)$ of some dimension n, let $(T_{UV}, T_U, T_V) = \alpha(U^n, V^n, Y^n)$ be the messages produced by the encoder, and let (R_{UV}, R_U, R_V) be the average rates used by the code. Let the ith character of the reproductions be denoted $\hat{U}_i = \beta_{Bi}(T_{UV}, T_U, Y^n, Z^n)$ and $\hat{V}_i = \beta_{Ci}(T_{UV}, T_V, Y^n, Z^n)$, and let the average distortions be denoted

$$\delta_U = \frac{1}{n}\sum_{i=1}^{n} d(U_i, \beta_{Bi}(T_{UV}, T_U, Y^n, Z^n) \le D_U$$
$$\delta_V = \frac{1}{n}\sum_{i=1}^{n} d(V_i, \beta_{Ci}(T_{UV}, T_U, Y^n, Z^n) \le D_V.$$

I show that there exists a random variable W satisfying the conditions of the theorem such that $(R_{UV}, R_U, R_V) \in \mathcal{R}^*(D_U, D_V)$.

I bound R_{UV} as

$$\begin{aligned}
nR_{UV} &\overset{(a)}{\ge} H(T_{UV}) \\
&\ge H(T_{UV}|Y^n, Z^n) \\
&\ge I(T_{UV}; U^n, V^n|Y^n, Z^n) \\
&= H(U^n, V^n|Y^n, Z^n) - H(U^n, V^n|T_{UV}, Y^n, Z^n) \\
&\overset{(b)}{=} \sum_{i=1}^{n} \left[H(U_i, V_i|Y_i, Z_i) - H(U_i, V_i|U^{i-1}, V^{i-1}, T_{UV}, Y^n, Z^n)\right] \\
&\overset{(c)}{=} \sum_{i=1}^{n} \left[H(U_i, V_i|Y_i, Z_i) - H(U_i, V_i|W_i, Y_i, Z_i)\right] \\
&= \sum_{i=1}^{n} I(U_i, V_i; W_i|Y_i, Z_i) \\
&\overset{(d)}{=} \sum_{i=1}^{n} I(U_i, V_i; W_i|Y_i, Z_i, Q = i) \\
&= nI(U_Q, V_Q; W_Q|Y_Q, Z_Q, Q)
\end{aligned}$$

$$\begin{aligned} &\overset{(e)}{=} nI(U_Q, V_Q; W_Q, Q|Y_Q, Z_Q) \\ &\overset{(f)}{=} nI(U, V; W|Y, Z), \end{aligned}$$

where the labeled steps are justified by the following.

(a) $T_{UV} \in \{1, \ldots, 2^{nR_{UV}}\}$.

(b) (U_i, V_i, Y_i, Z_i) are jointly i.i.d.

(c) By defining $W_i = (U^{i-1}, V^{i-1}, Y^{i-1}, Y_{i+1}^n, Z^{i-1}, Z_{i+1}^n, T_{UV})$.

(d) By introducing a timesharing variable Q uniformly distributed on $1, \ldots, n$.

(e) (U, V, Y, Z) are i.i.d. and hence $H(U_Q, V_Q|Y_Q, Z_Q, Q) = H(U_Q, V_Q|Y_Q, Z_Q)$.

(f) By defining $W = (W_Q, Q)$, $U = U_Q$, $V = V_Q$, $Y = Y_Q$, and $Z = Z_Q$. Note that this choice of U, V, Y, Z still leaves these variables with the same distribution as those in the statement of the theorem since $\Pr(U_Q = u, V_Q = v, Y_Q = y, Z_Q = z) = p(u, v, y, z)$ for any Q.

I bound R_U according to

$$\begin{aligned} nR_U &\overset{(a)}{\geq} H(T_U) \\ &\geq H(T_U|T_{UV}, Y^n, Z^n) \\ &\geq I(U^n, V^n; T_U|T_{UV}, Y^n, Z^n) \\ &= \sum_{i=1}^n I(U_i, V_i; T_U|U^{i-1}, V^{i-1}, T_{UV}, Y^n, Z^n) \\ &= \sum_{i=1}^n \left[H(U_i, V_i|U^{i-1}, V^{i-1}, T_{UV}, Y^n, Z^n) - H(U_i, V_i|U^{i-1}, V^{i-1}, T_U, T_{UV}, Y^n, Z^n)\right] \\ &\overset{(b)}{=} \sum_{i=1}^n \left[H(U_i, V_i|W_i, Y_i, Z_i) - H(U_i, V_i|W_i', W_i, Y_i, Z_i)\right] \\ &= \sum_{i=1}^n I(U_i, V_i; W_i'|W_i, Y_i, Z_i) \\ &= \sum_{i=1}^n I(U_i, V_i; W_i'|W_i, Y_i, Z_i, Q = i) \end{aligned}$$

$$
\begin{aligned}
&= nI(U_Q, V_Q; W'_Q | W_Q, Y_Q, Z_Q, Q) \\
&\overset{(c)}{\geq} nR_{U_Q,V_Q|QW_QY_Q\{Z_Q\}}(D_U) \\
&\overset{(d)}{\geq} nR'_{UV|WY\{Z\}}(D_U) \\
&\overset{(e)}{\geq} nR_{U|WY\{Z\}}(D_U),
\end{aligned}
$$

where the labeled steps are justified by the following.

(a) $T_U \in \{1, \ldots, 2^{nR_U}\}$.

(b) By defining $W'_i = T_U$. Note that since Z_i depends only on (U_i, V_i, Y_i) and is conditionally independent of T_U given (U_i, V_i, Y_i), $W'_i \to (U_i, V_i, Y_i) \to Z_i$ forms a Markov chain.

(c) By the definition of the MSI rate-distortion function with distortion measure $d'((u, v), (\hat{u}, \hat{v})) = d_u(u, \hat{u})$. The average distortion incurred at B under distribution $p(U_Q, V_Q, W_Q, Y_Q, Z_Q)$ is

$$
\begin{aligned}
Ed'((u, v), (\hat{u}, \hat{v})) &= Ed(u, \hat{u}) \\
&= E[Ed(u, \hat{u})|Q] \\
&= \frac{1}{n} \sum_{i=1}^{n} d(u_i, \beta_{Bi}(T_{UV}, T_U, Y^n, Z^n)) \\
&= D_U.
\end{aligned}
$$

(d) $R_{UV|WY\{Z\}}(\cdot)$ is a decreasing function.

(e) By Lemma 9.

A symmetric argument gives $R_V \geq R_{V|WY\{Z\}}(D_V)$. Thus, I have shown the existence of a random variable W such that

$$
\begin{aligned}
R_{UV} &\geq I(U, V; W|Y, Z) \\
R_U &\geq R_{U|WY\{Z\}}(D_U) \\
R_V &\geq R_{V|WY\{Z\}}(D_V).
\end{aligned}
$$

□

Theorem 22 $\mathcal{R}(\mathbf{D}) \subseteq \mathcal{R}_{con}(\mathbf{D})$.

Proof of Theorem 22 Given a code $(\alpha, \beta_1, \beta_2, \beta_3)$ of some dimension n with rates $\mathbf{R}$ and distortions satisfying $\Delta_i \leq D_i$ for $i = 1, 2, 3$, I show that there exist auxiliary random variables (W_{12}, W_{13}, W_{23}) such that $\mathbf{R} \in \mathcal{R}_{con}(\mathbf{D})$.

Let $(T_{12}, T_{13}, T_{23}) = \alpha(X_1^n, X_2^n, X_3^n)$ be the messages produced by the encoder. I bound R_{12} as follows.

$$\begin{aligned}
nR_{12} &\geq H(T_{12}) \\
&\geq H(T_{12}|T_{13}) \\
&\geq I(T_{12}; \mathbf{X}^n|T_{13}) \\
&= H(\mathbf{X}^n|T_{13}) - H(\mathbf{X}^n|T_{12}, T_{13}) \\
&= \sum_{i=1}^{n} \left[H(\mathbf{X}_i|\mathbf{X}^{i-1}, T_{13}) - H(\mathbf{X}_i|\mathbf{X}^{i-1}, T_{12}, T_{13})\right] \\
&\overset{(a)}{=} \sum_{i=1}^{n} \left[H(\mathbf{X}_i|W_{13,i}) - H(\mathbf{X}_i|W_{12,i}, W_{13,i})\right] \\
&\overset{(b)}{=} n \sum_{i=1}^{n} I(\mathbf{X}_i; W_{12,i}|W_{13,i}, Q = i) \\
&= nI(\mathbf{X}_Q; W_{12,Q}|W_{13,Q}, Q) \\
&= nI(\mathbf{X}_Q; W_{12,Q}, Q|W_{13,Q}, Q) \\
&\overset{(c)}{=} nI(\mathbf{X}; W_{12}|W_{13}) \\
&= n\left[I(\mathbf{X}, W_{13}; W_{12}) - I(W_{12}; W_{13})\right] \\
&\geq n\left[I(\mathbf{X}; W_{12}) - I(W_{12}; W_{13})\right],
\end{aligned}$$

where the labeled steps are justified by the following.

(a) By defining $W_{12,i} = (\mathbf{X}^{i-1}, T_{12})$ and $W_{13,i} = (\mathbf{X}^{i-1}, T_{13})$.

(b) By introducing a timesharing variable Q uniformly distributed on $1, \ldots, n$.

(c) By defining $W_{12} = (W_{12,Q}, Q)$ and $W_{13} = (W_{13,Q}, Q)$, and from $\mathbf{X}$ being i.i.d.

Repeating the above steps but conditioning on T_{23} instead of T_{13} yields $R_{12} \geq I(\mathbf{X}; W_{12}) - I(W_{12}; W_{23})$. Hence

$$R_{12} \geq I(\mathbf{X}; W_{12}) - \min\{I(W_{12}; W_{13}), I(W_{12}; W_{23})\}.$$

A symmetric result applies for R_{13} and R_{23}.

Each pair of rates can be bounded in the following way.

$$\begin{aligned}
n(R_{12} + R_{13}) &\geq H(T_{12}, T_{13}) \\
&\geq I(T_{12}, T_{13}; \mathbf{X}^n) \\
&= H(\mathbf{X}^n) - H(\mathbf{X}^n | T_{12}, T_{13}) \\
&= \sum_{i=1}^{n} \left[H(\mathbf{X}_i) - H(\mathbf{X}_i | \mathbf{X}^{i-1}, T_{12}, T_{13}) \right] \\
&= \sum_{i=1}^{n} \left[H(\mathbf{X}_i) - H(\mathbf{X}_i | W_{12,i}, W_{13,i}) \right] \\
&= n \sum_{i=1}^{n} I(\mathbf{X}_i; W_{12,i}, W_{13,i} | Q = i) \\
&= n I(\mathbf{X}_Q; W_{12,Q}, W_{13,Q} | Q) \\
&= n I(\mathbf{X}_Q; W_{12,Q}, Q, W_{13,Q}, Q) \\
&= n I(\mathbf{X}; W_{12}, W_{13}).
\end{aligned}$$

The same approach can be used on the triple of rates to yield

$$R_{12} + R_{13} + R_{23} > I(\mathbf{X}; W_{12}, W_{13}, W_{23}).$$

Finally, since each auxiliary random variable contains its corresponding index (e.g., W_{12}

contains T_{12}), the decoding functions can be recast as functions of the auxiliary random variables as required. □

Theorem 23 $\mathcal{R}_a(\mathbf{D}) \subseteq \mathcal{R}(\mathbf{D})$.

Proof of Theorem 23 Draw $2^{nR'_{12}}$ sequences $W_{12}^n(k_{12})$, $k_{12} \in \{1, \ldots, 2^{nR'_{12}}\}$, uniformly with replacement over $A_\epsilon^{*(n)}(W_{12})$. Similarly draw $2^{nR'_{13}}$ sequences $W_{13}^n(k_{13})$ uniformly with replacement over $A_\epsilon^{*(n)}(W_{13})$, and draw $2^{nR'_{23}}$ sequences $W_{23}^n(k_{23})$ uniformly with replacement over $A_\epsilon^{*(n)}(W_{23})$.

Create three random binning functions. Assign each k_{12} to one of $2^{nR_{12}}$ bins via function $g_{12}(k_{12})$, each k_{13} to one of $2^{nR_{13}}$ bins via $g_{13}(k_{13})$, and each k_{23} to one of $2^{nR_{23}}$ bins via $g_{23}(k_{23})$.

The encoder receives $\mathbf{X}^n$. It chooses an index triple (k_{12}, k_{13}, k_{23}) such that

$$(\mathbf{X}^n, W_{12}^n(k_{12}), W_{13}^n(k_{13}), W_{23}^n(k_{23})) \in A_\epsilon^{*(n)}.$$

If there is more than one such triple, it chooses one at random from the set of such triples. If there is no such triple, it declares an error. In the absence of an error, the encoder transmits index set

$$(i_{12}, i_{13}, i_{23}) = (g_{12}(k_{12}), g_{13}(k_{13}), g_{23}(k_{23}))$$

to the decoders.

Decoder 1 receives indices (i_{12}, i_{13}) and maps them to the unique pair (k_{12}, k_{13}) such that $g_{12}(k_{12}) = i_{12}$, $g_{13}(k_{13}) = i_{13}$, and $(W_{12}^n(k_{12}), W_{13}^n(k_{13})) \in A_\epsilon^{*(n)}$. If there is no such unique pair, it declares an error. In the absence of an error, it declares a reproduction

$f_{12}(W_{12}^n(k_{12}), W_{13}^n(k_{13}))$, where f_{12} is the function from the definition of $\mathcal{R}_a$. Decoders 2 and 3 work similarly.

There are various error events at the encoder:

1. $\mathbf{X}^n \notin A_\epsilon^{*(n)}$. By Lemma 1, the probability of this event is small for n sufficiently large.

2. $\mathbf{X}^n \in A_\epsilon^{*(n)}$, but

$$\nexists(k_{12}, k_{13}, k_{23}) \text{ such that } (\mathbf{X}^n, W_{12}^n(k_{12}), W_{13}^n(k_{13}), W_{23}^n(k_{23})) \in A_\epsilon^{*(n)}.$$

By Lemma 7, the probability of this is small provided that

$$\begin{aligned}
R_{12} &> I(\mathbf{X}; W_{12}) \\
R_{13} &> I(\mathbf{X}; W_{13}) \\
R_{23} &> I(\mathbf{X}; W_{23}) \\
R_{12} + R_{13} &> I(\mathbf{X}; W_{12}, W_{13}) + I(W_{12}; W_{13}) \\
R_{12} + R_{23} &> I(\mathbf{X}; W_{12}, W_{23}) + I(W_{12}; W_{23}) \\
R_{13} + R_{23} &> I(\mathbf{X}; W_{13}, W_{23}) + I(W_{13}; W_{23}) \\
R_{12} + R_{13} + R_{23} &> H(W_{12}) + H(W_{13}) + H(W_{23}) - H(W_{12}, W_{13}, W_{23}|\mathbf{X}).
\end{aligned}$$

There are also error events at the decoders. For decoder 1:

1. There exists a $k'_{12} \neq k_{12}$ such that $g(k'_{12}) = g(k_{12}) = i_{12}$ and $(W_{12}^n(k'_{12}), W_{13}^n(k_{13})) \in A_\epsilon^{*(n)}$. Since, by Lemma 2, the probability that a randomly chosen W_{12}^n is jointly typical with $W_{13}^n(k_{13})$ is less than $2^{-n(I(W_{12};W_{13})-3\epsilon)}$,

$$\Pr(\exists k'_{12} : g_{12}(k'_{12}) = i_{12}, (W_{12}^n(k'_{12}), W_{13}^n(k_{13})) \in A_\epsilon^{*(n)}) \leq 2^{n(R'_{12}-R_{12})} 2^{-n(I(W_{12};W_{13})-3\epsilon)}.$$

This can be made arbitrarily small provided n is sufficiently large and $R'_{12} - R_{12} < I(W_{12}; W_{13}) - 3\epsilon$. To prevent the same type of error for k_{13} we require $R'_{13} - R_{13} < I(W_{12}; W_{13}) - 3\epsilon$. Finally, to avoid the case in which there exists a pair (k'_{12}, k'_{13}) with $k'_{12} \neq k_{12}$, $k'_{13} \neq k_{13}$, but $(W^n_{12}(k'_{12}), W^n_{13}(k'_{13})) \in A^{*(n)}_\epsilon)$, we require $R'_{12} - R_{12} + R'_{13} - R_{13} < I(W_{12}; W_{13}) - 3\epsilon$.

2. The code does not meet the distortion requirement, i.e.,

$$Ed(X^n_1, f(W^n_{12}(k_{12}), W^n_{13}(k_{13}))) \geq D_1 + \epsilon.$$

The probability of this event is small since in the absence of the other error events,

$$(X^n_1, W^n_{12}(k_{12}), W^n_{13}(k_{13})) \in A^{*(n)}_\epsilon.$$

By the same argument as [21, Pg. 48], this ensures that their distortion is smaller than $D_1 + \epsilon$.

Combining all of the above rate constraints, the probability of error can be made arbitrarily small provided n is sufficiently large, and

$$\begin{aligned}
R_{12} &> I(\mathbf{X}; W_{12}) - \min\{I(W_{12}; W_{13}), I(W_{12}; W_{23})\} \\
R_{13} &> I(\mathbf{X}; W_{13}) - \min\{I(W_{12}; W_{13}), I(W_{13}; W_{23})\} \\
R_{23} &> I(\mathbf{X}; W_{23}) - \min\{I(W_{12}; W_{23}), I(W_{13}; W_{23})\} \\
R_{12} + R_{13} &> I(\mathbf{X}; W_{12}, W_{13}) + I(W_{12}; W_{13}) \\
&\quad - \min\{I(W_{12}; W_{13}), I(W_{12}; W_{23}) + I(W_{13}; W_{23})\} \\
R_{12} + R_{23} &> I(\mathbf{X}; W_{12}, W_{23}) + I(W_{12}; W_{23}) \\
&\quad - \min\{I(W_{12}; W_{23}), I(W_{12}; W_{13}) + I(W_{13}; W_{23})\}
\end{aligned}$$

$$\begin{aligned} R_{13} + R_{23} &> I(\mathbf{X}; W_{13}, W_{23}) + I(W_{13}; W_{23}) \\ &\quad - \min\{I(W_{13}; W_{23}), I(W_{12}; W_{13}) + I(W_{12}; W_{23})\} \\ R_{13} + R_{13} + R_{23} &> H(W_{12}) + H(W_{13}) + H(W_{23}) - H(W_{12}, W_{13}, W_{23}|\mathbf{X}) - \phi, \end{aligned}$$

where

$$\begin{aligned} \phi &= \max_{(n_{12}, n_{13}, n_{23}) \in \mathcal{N}} (n_{12} + n_{13} + n_{23}) \\ \mathcal{N} &= \{(n_{12}, n_{13}, n_{23}) : n_{12} \geq 0, n_{13} \geq 0, n_{23} \geq 0, \\ &\quad n_{12} + n_{13} \leq I(W_{12}; W_{13}), n_{12} + n_{23} \leq I(W_{12}; W_{23}), n_{13} + n_{23} \leq I(W_{13}; W_{23})\}. \end{aligned}$$

□

Theorem 25 *Assume, without loss of generality, that* $I(W_{12}; W_{13}) \leq I(W_{13}; W_{23})$. *Then*

$$R_{min} = \begin{cases} I(W_{12}; W_{13}) + I(W_{12}; W_{23}), & 0 \leq I(W_{12}; W_{23}) \leq I(W_{13}; W_{23}) - I(W_{12}; W_{13}) \\ \frac{I(W_{12}; W_{13}) + I(W_{12}; W_{23}) + I(W_{13}; W_{23})}{2}, & I(W_{13}; W_{23}) - I(W_{12}; W_{13}) \\ & \quad \leq I(W_{12}; W_{23}) \leq I(W_{13}; W_{23}) + I(W_{12}; W_{13}) \\ I(W_{12}; W_{13}) + I(W_{13}; W_{23}), & I(W_{12}; W_{23}) \geq I(W_{12}; W_{13}) + I(W_{13}; W_{23}). \end{cases}$$

Proof of Theorem 25 The proof is in two parts. I first characterize R_{min} as the solution to a maximization problem and then solve the maximization problem.

The bounds of Theorem 24 can be rewritten as

$$\begin{aligned} R_{12} &> H(W_{12}) - \min\{I(W_{12}; W_{13}), I(W_{12}; W_{23})\} \\ R_{23} &> H(W_{13}) - \min\{I(W_{12}; W_{13}), I(W_{13}; W_{23})\} \\ R_{13} &> H(W_{23}) - \min\{I(W_{12}; W_{23}), I(W_{13}; W_{23})\} \\ R_{12} + R_{13} &> H(W_{12}) + H(W_{23}) - I(W_{12}; W_{23}) \end{aligned}$$

$$\begin{aligned} R_{12} + R_{23} &> H(W_{12}) + H(W_{13}) - I(W_{12}; W_{13}) \\ R_{23} + R_{13} &> H(W_{13}) + H(W_{23}) - I(W_{13}; W_{23}). \end{aligned} \tag{E.1}$$

For any admissible (R_{12}, R_{23}, R_{13}), let $n_{12} = H(W_{12}) - R_{12}$, $n_{13} = H(W_{13}) - R_{23}$, and $n_{23} = H(W_{23}) - R_{13}$, so that

$$R_{12} = H(W_{12}) - n_{12} \qquad R_{23} = H(W_{13}) - n_{13} \qquad R_{13} = H(W_{23}) - n_{23}.$$

Since $(R_{12}, R_{13}, R_{23}) > (H(W_{12}), H(W_{13}), H(W_{23}))$ is clearly achievable (by independent coding), we are interested in the case when n_{12}, n_{13}, and n_{23} are all non-negative. Under this condition, bounds (E.1) above are, line by line, equivalent to

$$\begin{aligned} n_{12} &\leq \min\{I(W_{12}; W_{13}), I(W_{12}; W_{23})\} \\ n_{13} &\leq \min\{I(W_{12}; W_{13}), I(W_{13}; W_{23})\} \\ n_{23} &\leq \min\{I(W_{12}; W_{23}), I(W_{13}; W_{23})\} \\ n_{12} + n_{23} &\leq I(W_{12}; W_{23}) \\ n_{12} + n_{13} &\leq I(W_{12}; W_{13}) \\ n_{13} + n_{23} &\leq I(W_{13}; W_{23}). \end{aligned}$$

The last three of these make the first three redundant. Thus, $(H(W_{12}) - n_{12}, H(W_{13}) - n_{13}, H(W_{23}) - n_{23}) \in \mathcal{R}$ if and only if $(n_{12}, n_{13}, n_{23}) \in \mathcal{N}$, where

$$\begin{aligned} \mathcal{N} = \{&(n_{12}, n_{13}, n_{23}) : n_{12} \geq 0, n_{13} \geq 0, n_{23} \geq 0, \\ &n_{12} + n_{13} \leq I(W_{12}; W_{13}), n_{12} + n_{23} \leq I(W_{12}; W_{23}), n_{13} + n_{23} \leq I(W_{13}; W_{23})\}. \end{aligned}$$

From the definition of R_{min},

R_{min}

$$= \min_{(R_{12},R_{23},R_{13})\in\mathcal{R}} \{R_{12} + R_{23} + R_{13}\}$$

$$= \min_{(H(W_{12})-n_{12},H(W_{13})-n_{13},H(W_{23})-n_{23})\in\mathcal{R}} \{(H(W_{12}) - n_{12} + H(W_{13}) - n_{13} + H(W_{23}) - n_{23})\}$$

$$= \min_{(n_{12},n_{13},n_{23})\in\mathcal{N}} \{(H(W_{12}) - n_{12} + H(W_{13}) - n_{13} + H(W_{23}) - n_{23})\}$$

$$= H(W_{12}) + H(W_{13}) + H(W_{23}) - \max_{(n_{12},n_{13},n_{23})\in\mathcal{N}} \{(n_{12} + n_{13} + n_{23})\}.$$

This is the maximization problem that must be solved.

Now, for convenience, write $I_{12,13} = I(W_{12}; W_{13})$, $I_{12,23} = I(W_{12}; W_{23})$, and $I_{13,23} = I(W_{13}; W_{23})$. Set $\mathcal{N}$ is defined by three inequality constraints, each of which describes a plane in (n_{12}, n_{13}, n_{23})-space. The region defined by the inequalities is a polygon in the positive quadrant with faces corresponding to the given planes; the solution to the maximization is represented by the point in the polygon that has the greatest taxicab distance $d_{taxi}(n_{12}, n_{13}, n_{23}) = n_{12} + n_{13} + n_{23}$ from the origin. Call this point G.

Figure E.1 depicts the polygon formed by the two planes $n_{12}+n_{13} = I_{12,13}$ and $n_{13}+n_{23} = I_{13,23}$. The third plane, not shown, is oriented vertically and is defined by $n_{12} + n_{23} = I_{12,23}$. The intersection of this third plane with the polygon shown in Figure E.1(a) determines the solution of the maximization.

For $I_{12,23} = 0$, the third plane intersects the line $n_{12} = n_{23} = 0$, and the solution is $G = D = (0, I_{12,13}, 0)$. As $I_{12,23}$ grows, the third plane moves out from the n_{13} axis as shown in Figure E.1(b), and solution point $G = (0, I_{12,13}, I_{12,23})$ moves out along line $\overline{DE}$. When $I_{12,23} = I_{13,23} - I_{12,13}$, G reaches E. A further increase in $I_{12,23}$ results in a polygon of the form shown in Figure E.1(c), and solution point G moves out along the line $\overline{EF}$ from E towards F. When $I_{12,23} = I_{13,23} + I_{12,23}$, G reaches F. Further increases in $I_{12,23}$ have no effect on G.

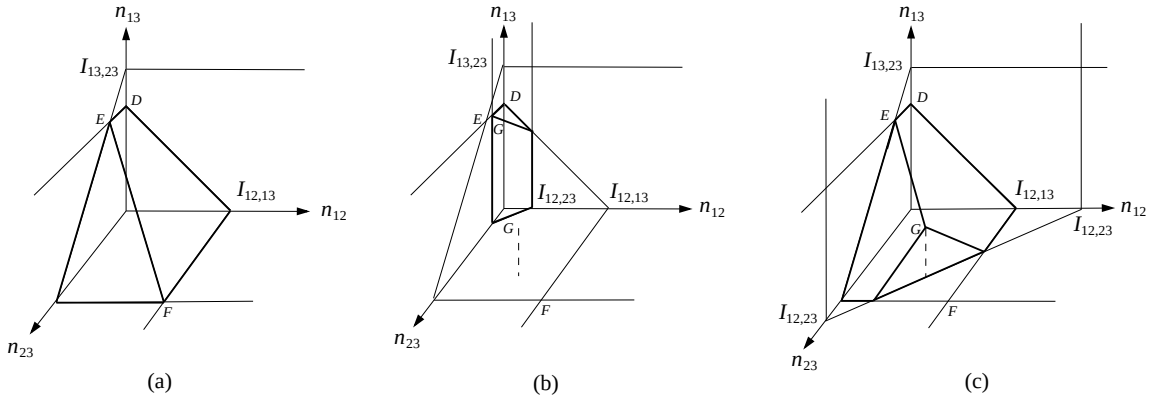


Figure E.1: A graphical representation of the minimization. (a) The polygon defined by the two planes $n_{12} + n_{13} = I_{12,13}$ and $n_{13} + n_{23} = I_{13,23}$. (b) The polygon defined by all three planes when $0 \leq I_{12,23} \leq I_{13,23} - I_{12,13}$. (c) The polygon defined by all three planes when $I_{13,23} - I_{12,13} \leq I_{12,23} \leq I_{12,13} + I_{13,23}$.

Thus, for $0 \leq I_{12,23} \leq I_{13,23} - I_{12,13}$, the optimal solution is $G = (0, I_{12,13}, I_{12,23})$, which is a taxicab distance of $d_{taxi}(G) = I_{12,13} + I_{12,23}$ from the origin.

For $I_{13,23} - I_{12,13} \leq I_{12,23} \leq I_{12,13} + I_{13,23}$, G is the intersection of the line $\overline{EF}$ and the plane $n_{12} + n_{23} = I_{12,23}$. Since $E = (0, I_{12,13}, I_{13,23} - I_{12,13})$ and $F = (I_{12,13}, 0, I_{13,23})$, line $\overline{EF}$ is the set of points

$$\lambda E + (1 - \lambda)F = ((1 - \lambda)I_{12,13}, \lambda I_{12,13}, I_{13,23} - \lambda I_{12,13}) \qquad \forall \lambda \in \mathbb{R}. \tag{E.2}$$

The intersection occurs at the value of λ satisfying

$$(1 - \lambda)I_{12,13} + I_{13,23} - \lambda I_{12,13} \;\; = \;\; I_{12,23},$$

which yields

$$\lambda = \frac{I_{13,23} + I_{12,13} - I_{12,23}}{2I_{12,13}}.$$

Substituting back into (E.2) gives $G = (\frac{I_{12,23}+I_{12,13}-I_{13,23}}{2}, \frac{I_{12,13}+I_{13,23}-I_{12,23}}{2}, \frac{I_{13,23}+I_{12,23}-I_{12,13}}{2})$, which is a taxicab distance of $d_{taxi}(G) = \frac{I_{12,13}+I_{13,23}+I_{12,23}}{2}$ from the origin.

Finally, for any $I_{12,23} \geq I_{12,13} + I_{13,23}$, the optimal solution is $G = F$, which is a taxicab distance of $d_{taxi}(G) = I_{12,13} + I_{13,23}$ from the origin.

Thus,

$$\max_{(n_{12},n_{13},n_{23})\in\mathcal{N}} \{n_{12} + n_{13} + n_{23}\} =$$

$$\begin{cases} I(W_{12};W_{13}) + I(W_{12};W_{23}) & \text{if } 0 \leq I(W_{12};W_{23}) \leq I(W_{13};W_{23}) - I(W_{12};W_{13}) \\ \frac{I(W_{12};W_{13})+I(W_{12};W_{23})+I(W_{13};W_{23})}{2} & \text{if } I(W_{13};W_{23}) - I(W_{12};W_{13}) \\ & \quad \leq I(W_{12};W_{23}) \leq I(W_{12};W_{13}) + I(W_{13};W_{23}) \\ I(W_{12};W_{13}) + I(W_{13};W_{23}) & \text{if } I(W_{12};W_{23}) \geq I(W_{12};W_{13}) + I(W_{13};W_{23}). \end{cases}$$

□

Bibliography

[1] M. Effros. Network source coding. In *2000 Conf. on Info. Sciences and Systems*, Princeton, New Jersey, March 2000. IEEE.

[2] C. Heegard and T. Berger. Rate distortion when side information may be absent. *IEEE Trans. on Information Theory*, 31(6):727–734, November 1985.

[3] A. H. Kaspi. Rate-distortion function when side-information may be present at the decoder. *IEEE Trans. on Information Theory*, 40(6):2031–4, November 1994.

[4] R. M. Gray. Conditional rate-distortion theory. *Stanford Electronic Labs Technical Report*, (6502-2), October 1972.

[5] T. Berger and R. W. Yeung. Multiterminal source encoding with encoder breakdown. *IEEE Trans. on Information Theory*, 35(2):237–244, March 1989.

[6] A. H. Kaspi and T. Berger. Rate-distortion for correlated sources with partially separated encoders. *IEEE Trans. on Information Theory*, 28(6):828–840, November 1982.

[7] M. Fleming and M. Effros. Generalized multiple description vector quantization. In *Proc. of the Data Compression Conf.*, pages 3–12, Snowbird, UT, March 1999. IEEE.

[8] M. Fleming and M. Effros. Network vector quantization. In *Proc. of the Data Compression Conf.*, pages 13–22, Snowbird, UT, March 2001. IEEE.

[9] M. Fleming, Q. Zhao, and M. Effros. Network vector quantization. 2004. To appear in *IEEE Trans. on Information Theory*.

[10] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423,623–656, 1948.

[11] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. on Information Theory*, IT-19(4):471–480, July 1973.

[12] A. D. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Trans. on Information Theory*, 22:1–10, January 1976.

[13] A. D. Wyner. The rate-distortion function for source coding with side information at the decoder–ii: General sources. *Information and Control*, 38:60–80, 1978.

[14] V. Koshelev. Heirarchical coding of discrete sources. *Problemy Peredachi Informatsii*, 17(3):20–33, 1981.

[15] W. H. R. Equitz and T. M. Cover. Successive refinement of information. *IEEE Trans. on Information Theory*, IT-37(2):269–275, March 1991.

[16] J. K. Wolf, A. D. Wyner, and J. Ziv. Source coding for multiple descriptions. *Bell System Technical Journal*, 59:1417–1426, October 1980.

[17] L. Ozarow. On a source coding problem with two channels and three receivers. *Bell System Technical Journal*, 59:446–472, December 1980.

[18] A. El Gamal and T. M. Cover. Achievable rates for multiple descriptions. *IEEE Trans. on Information Theory*, IT-28(6):851–857, November 1982.

[19] R. M. Gray and A. D. Wyner. Source coding for a simple network. *Bell Systems Technical Journal*, 53(9):1681–1721, November 1974.

[20] Q. Zhao and M. Effros. Broadcast system source codes: a new paradigm for data compression. In *Conference Record, Thirty-Third Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, October 1999. IEEE.

[21] S. Y. Tung. *Multiterminal Source Coding.* Ph. D. Dissertation, Cornell University, Ithaca, NY, 1978.

[22] T. Berger. Multiterminal source coding. In *The information theory approach to communications*, pages 172–231, Wien, Austria, 1980. Springer-Verlag.

[23] B. Rimoldi. Successve refinement of information: characterization of the achievable rates. *IEEE Trans. on Information Theory*, 40(1):253–259, January 1994.

[24] Z. Zhang and T. Berger. New results in binary multiple descriptions. *IEEE Trans. on Information Theory*, 33(4):502–521, July 1987.

[25] R. Venkataramani, G. Kramer, and V. Goyal. Multiple description coding with many channels. *IEEE Trans. on Information Theory*, IT-49(9):2106–2114, September 2003.

[26] S. Sher and M. Feder. A converse theorem for the multiple description problem. In *Eighteenth Convention of Electrical and Electronics Engineers in Israel*, pages 1.1.4/1–4. IEEE, March 1995.

[27] T. S. Han and K. Kobayashi. A unified achievable rate region for a general class of multiterminal coding systems. *IEEE Trans. on Information Theory*, IT-26(3):277–288, May 1980.

[28] K. J. Kerpez. The rate-distortion function of a binary symmetric source when side-information may be absent. *IEEE Trans. on Information Theory*, 33(3):448–52, May 1987.

[29] M. Fleming and M. Effros. Rate-distortion with mixed types of side information. In *Proc. of the IEEE International Symposium on Info. Theory*, page 144, Yokohama, Japan, June 2003. IEEE.

[30] M. Fleming and M. Effros. Rate-distortion with mixed types of side information. 2004. Submitted to IEEE Trans. Info. Theory.

[31] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[32] S. Arimoto. An algorithm for calculating the capacity of an arbitrary discrete memoryless channel. *IEEE Trans. on Information Theory*, 18(1):14–20, January 1972.

[33] R.E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. on Information Theory*, 18(4):460–473, July 1972.

[34] T. Berger and S. Y. Tung. Encoding of correlated analog sources. In *Proc. of the 1975 IEEE-USSR Joint Workshop on Information Theory*, pages 7–10, New York, NY, 1976. IEEE.

[35] Y. Oohama. Gaussian multiterminal source coding. *IEEE Trans. on Information Theory*, IT-43(6):1912–1923, November 1997.

[36] Y. Oohama. The rate-distortion function for the quadratic Gaussian CEO problem. *IEEE Trans. on Information Theory*, IT-44(3):1057–1079, May 1998.

[37] T. Berger and R. W. Yeung. Multiterminal source encoding with one distortion criterion. *IEEE Trans. on Information Theory*, 35(2):228–236, March 1989.

[38] R. Ahlswede. The rate-distortion region for multiple descriptions without excess rate. *IEEE Trans. on Information Theory*, 31(6):721–726, November 1985.

[39] R. Zamir. The rate loss in the Wyner-Ziv problem. *IEEE Trans. on Information Theory*, 42:2073–2084, November 1996.

[40] L. Lastras and T. Berger. All sources are nearly successively refinable. *IEEE Trans. on Information Theory*, 47(3):918–926, March 2000.

[41] A. M. Gerrish. *Estimation of information rates.* Ph. D. Dissertation, Yale University, New Haven, CT, 1963.

[42] M. Effros. Distortion-rate bounds for fixed- and variable-rate multi-resolution source codes. *IEEE Trans. on Information Theory*, 45(6):1887–1910, September 1999.

[43] H. Feng and M. Effros. Improved bounds for the rate loss of multiresolution source codes. *IEEE Trans. on Information Theory*, 49(4):809–821, April 2003.

[44] J. D. Bruce. *Optimum Quantization*. Ph.D. Dissertation, M.I.T., Cambridge, MA, 1964.

[45] D. K. Sharma. Design of absolutely optimal quantizers for a wide class of distortion measures. *IEEE Trans. on Information Theory*, 24(6):693–702, November 1978.

[46] X. Wu. *Algorithmic approach to mean-square quantization*. Ph. D. Dissertation, University of Calgary, Calgary, Alberta, Canada, 1988.

[47] X. Wu and K. Zhang. Quantizer monotonicities and globally optimal scalar quantizer design. *IEEE Trans. on Information Theory*, 39(3):1049–1053, May 1993.

[48] D. Muresan and M. Effros. Quantization as histogram segmentation: globally optimal scalar quantizer design in network systems. In *Proc. Data Compression Conference*, pages 302–311, Snowbird, UT, April 2002. IEEE.

[49] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. In *Proceedings of the 10th Annual ACM–SIAM Symposium on Discrete Algorithms (SODA)*, 1999.

[50] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Trans. on Communications*, 28:84–95, January 1980.

[51] P. A. Chou, T. Lookabaugh, and R. M. Gray. Entropy-constrained vector quantization. *IEEE Trans. on Acoustics Speech and Signal Processing*, 37(1):31–42, January 1989.

[52] N. Farvardin. A study of vector quantization for noisy channels. *IEEE Trans. on Information Theory*, IT-36(4):799–809, July 1990.

[53] T. Duman and M. Salehi. Optimal quantization for finite-state channels. In *Proc. of the IEEE International Symposium on Info. Theory*, page 378, Whistler, British Columbia, Canada, September 1995.

[54] A. Buzo, A. H. Gray Jr., R .M. Gray, and J .D. Markel. Speech coding based upon vector quantization. *IEEE Trans. on Information Theory*, 28:562–574, October 1980.

[55] E. A. Riskin and R. M. Gray. A greedy tree growing algorithm for the design of variable rate vector quantizers. *IEEE Trans. on Signal Processing*, 39:2500–2507, November 1991.

[56] H. Brunk and N. Farvardin. Fixed-rate successively refinable scalar quantizers. In *Proc. of the Data Compression Conf.*, pages 250–259, Snowbird, UT, April 1996. IEEE.

[57] H. Jafarkhani, H. Brunk, and N. Farvardin. Entropy-constrained successively refinable scalar quantization. In *Proc. of the Data Compression Conf.*, pages 337–346, Snowbird, UT, March 1997. IEEE.

[58] M. Effros. Practical multi-resolution source coding: TSVQ revisited. In *Proc. of the Data Compression Conf.*, pages 53–62, Snowbird, UT, March 1998. IEEE.

[59] M. Effros and D. Dugatkin. Multi-resolution vector quantization. 2000. Submitted to IEEE Trans. Info. Theory for publication.

[60] V. A. Vaishampayan. Vector quantizer design for diversity systems. In *Proc. of the 25th Annual Conf. on Info. Sciences and Systems*, pages 564–569. IEEE, March 1991.

[61] V. A. Vaishampayan. Design of multiple description scalar quantizers. *IEEE Trans. on Information Theory*, 39(3):821–834, May 1993.

[62] V. A. Vaishampayan and J. Domaszewicz. Design of entropy-constrained multiple description scalar quantizers. *IEEE Trans. on Information Theory*, 40(1):245–250, January 1994.

[63] T. J. Flynn and R. M. Gray. Encoding of correlated observations. *IEEE Trans. on Information Theory*, IT-33(6):773–787, November 1987.

[64] Q. Zhao and M. Effros. Lossless and lossy broadcast system source codes: theoretical limits, optimal design, and empirical performance. In *Proc. Data Compression Conference*, pages 63–72, Snowbird, UT, March 2000. IEEE.

[65] M. Effros and L. Schulman. Rapid near-optimal VQ design with a deterministic data net. In *Proc. of the IEEE International Symposium on Info. Theory*. IEEE.

[66] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Boston, 1992.

[67] F. Kossentini, M. J. T. Smith, and C. F. Barnes. Necessary conditions for the optimality of variable-rate residual vector quantizers. *IEEE Trans. on Information Theory*, 41:1903–14, November 1995.

[68] H. S. Witsenhausen. The zero-error side information problem and chromatic numbers. *IEEE Trans. on Information Theory*, 22:592–593, 1976.

[69] N. Alon and A. Orlitsky. Source coding and graph entropies. *IEEE Trans. on Information Theory*, 42(5):1329–1339, September 1996.

[70] J. Korner and A. Orlitsky. Zero-error information theory. *IEEE Trans. on Information Theory*, 44(6):2207–2228, October 1998.

[71] S. S. Pradhan and K. Ramchandran. Distributed source coding using syndromes (DISCUS): design and construction. In *Proc. Data Compression Conference*, pages 158–167, Snowbird, UT, March 1999. IEEE.

[72] Y. Yan and T. Berger. On instantaneous codes for zero-error coding of two correlated sources. In *Proc. of the IEEE International Symposium on Info. Theory*, page 344, Sorrento, Italy, June 2000.

[73] P. Koulgi, E. Tuncel, S. Regunathan, and K. Rose. Minimum redundancy zero-error source coding with side information. In *Proc. of the IEEE International Symposium on Info. Theory*, page 282, Washington DC, USA, June 2001.

[74] Q. Zhao and M. Effros. Lossless and near-lossless source coding for multiple access networks. *IEEE Trans. on Information Theory*, 49:112–128, January 2003.

[75] Q. Zhao and M. Effros. Optimal code design for lossless and near-lossless source coding in multiple access networks. In *Proc. Data Compression Conference*, pages 263–272, Snowbird, UT, March 2001. IEEE.

[76] H. Jafarkhani and N. Farvardin. Channel-matched heirarchical table-lookup vector quantization. *IEEE Trans. on Information Theory*, 46(3):1121–1125, May 2000.

[77] T. M. Cover. A proof of the data compression theorem of Slepian and Wolf for ergodic sources. *IEEE Trans. on Information Theory*, IT-22:226–228, March 1975.

[78] Q. Zhao and M. Effros. Low complexity code design for lossless and near lossless side information source codes. In *Proc. Data Compression Conference*, Snowbird, UT, March 2003. IEEE.

[79] K. Popat and R. W. Picard. Cluster-based probability model and its application to image and texture processing. *IEEE Trans. on Image Proc.*, 6(2):268–284, February 1997.

[80] K. Zeger and A. Gersho. Globally optimal vector quantization design by stochastic relaxation. *IEEE Trans. on Signal Processing*, 40:310–22, February 1992.

[81] K. Rose, E. Gurewitz, and G. C. Fox. Vector quantization by deterministic annealing. *IEEE Trans. on Information Theory*, 38(4):1249–57, July 1992.

[82] R. Zamir, S. Shamai, and U. Erez. Nested linear/lattice codes for structured multiterminal binning. *IEEE Trans. on Information Theory*, 48:1250–76, June 2002.

[83] R. Ahlswede and J. Korner. Source coding with side information and a converse for degraded broadcast channels. *IEEE Trans. on Information Theory*, 21(6):629–637, November 1975.

[84] S. Jaggi. April 2004. Private communication.

[85] I. Csiszár and J. Körner. Towards a general theory of source networks. *IEEE Trans. on Information Theory*, IT-26(2):155–165, March 1980.

[86] A. Sgarro. Source coding with side information at several decoders. *IEEE Trans. on Information Theory*, 23(2):179–182, March 1977.

[87] A. Orlitsky and R. Roche. Coding for computing. *IEEE Trans. on Information Theory*, 47(3):903–917, March 2001.