Functional Compression: Theory and Applications

by

Vishal D. Doshi

B.S., Electrical and Computer Engineering
B.S., Mathematics
University of Illinois at Urbana-Champaign, 2005

Submitted to the Department of Electrical Engineering and Computer Science
and the Engineering Systems Division
in partial fulfillment of the requirements for the degrees of

Master of Science in Electrical Engineering and Computer Science
and
Master of Science in Technology and Policy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
February 2008

Author ....
Department of Electrical Engineering and Computer Science
and the Engineering Systems Divison
January 7, 2008

Certified by ....
Devavrat Shah
Assistant Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by ....
Muriel Médard
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by ....
Dava Newman
Professor of Aeronautics and Astronomics and Engineering Systems
Director, Technology and Policy Program

Accepted by ....
Terry P. Orlando
Professor of Electrical Engineering
Chairman, Department Graduate Committee

Functional Compression: Theory and Applications

by

Vishal D. Doshi

Submitted to the Department of Electrical Engineering and Computer Science
and the Engineering Systems Division
on January 7, 2008, in partial fulfillment of the
requirements for the degrees of
Master of Science in Electrical Engineering and Computer Science
and
Master of Science in Technology and Policy

Abstract

We consider the problem of functional compression. The objective is to separately compress possibly correlated discrete sources such that an arbitrary deterministic function of those sources can be computed given the compressed data from each source. This is motivated by problems in sensor networks and database privacy. Our architecture gives a quantitative definition of privacy for database statistics. Further, we show that it can provide significant coding gains in sensor networks.

We consider both the lossless and lossy computation of a function. Specifically, we present results of the rate regions for three instances of the problem where there are two sources: 1) lossless computation where one source is available at the decoder, 2) under a special condition, lossless computation where both sources are separately encoded, and 3) lossy computation where one source is available at the decoder.

Wyner and Ziv (1976) considered the third problem for the special case $f(X,Y) = X$ and derived a rate distortion function. Yamamoto (1982) extended this result to a general function. Both of these results are in terms of an auxiliary random variable. Orlitsky and Roche (2001), for the zero distortion case, gave this variable a precise interpretation in terms of the properties of the characteristic graph; this led to a particular coding scheme. We extend that result by providing an achievability scheme that is based on the coloring of the characteristic graph. This suggests a layered architecture where the functional layer controls the coloring scheme, and the data layer uses existing distributed source coding schemes. We extend this graph coloring method to provide algorithms and rates for all three problems.

Thesis Supervisor: Devavrat Shah
Title: Assistant Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Muriel Médard
Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgments

I want to thank my advisors, Devavrat Shah and Muriel Médard. I value the long hours spent with Devavrat wrapping up proofs and nailing down concepts. I appreciate all his guidance on the issues that graduate students face. This thesis would not be complete without Muriel's keen insights into the problem. She kept my eyes on the prize.

I want to thank my father, Devendra, my mother, Daxa, and my brother, Chirag. Their love and support helped me through the not-so-exciting times at MIT.

On a more research-oriented note, I want to thank Dr. Aaron Heller for providing the data used in the simulations in Section 5.1. I want to thank Professor Michelle Effros of Caltech for providing valuable comments on this research.

Many of the results in this thesis have been previously presented. Preliminary results for the theorem in Section 3.1 were presented at the 2006 Asilomar Conference on Signals, Systems and Computers [12]. Preliminary results for the theorems in Section 3.2 were presented at the 2007 Data Compression Conference [13]. Preliminary results for the theorems in Section 3.3 were presented at the 2007 International Symposium on Information Theory [11].

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Generally speaking, data compression considers the compression of a source (sources) and its (their) recovery via a decoding algorithm. Functional compression considers the recovery not of the sources, but of a function of the sources. It is a method for reducing the number of bits required to convey relevant information from disparate sources to a third party. The key contributions of this thesis are to provide meaning to the word "relevant" in this context. We will derive the information theoretic limits for a selection of functional compression problems and give novel algorithms to achieve these rates.

1.1 Motivations and Applications

We are motivated to study this problem mainly by two problems. First, consider a statistical database, such as a medical records database. There are enormous amounts of private data in the database. The database manager wants to release certain statistics, or functions, of the data in order to provide utility to, for example, researchers. Thinking of the data as a bit-string, we provide a way for the database manager to compute a certain function while each person reveals a minimal number of their own bits. Thus, our archticture allows for a minimal loss of privacy, given the need to compute certain statistics.

Next, consider a network of wireless sensors measuring temperature in a building. There are bandwidth and power constraints for each sensor, and the sensors do not communicate with each other, only a central receiver. Further, the receiver wishes only to compute the average temperature in the building. Given the receivers intentions, we want to determine if it possible to compress beyond the traditional distributed data compression rate bounds given by Slepian and Wolf.

We can frame both of the above questions as functional compression problems. The common thread is that of using extra information, the end-user's intentions, either to guarantee privacy or to achieve higher compression rates (thus conserving bandwidth and power).

For the statistical database problem, if one knows the rate required to compute a certain malicious (or privacy-compromising) function, one simply needs to ensure that the database is compressed beyond that minimum rate required. This would ensure that the malicious function is *impossible* to compute.

Conversely, in the sensor network problem, one wishes to compute a function, so one would compress the information to a rate such that it is computable. This rate would be smaller than the best-known compression gains because of the utilization of knowledge of the function.

In general, then, the problem is that of determining the jointly minimal statistics of discrete memoryless sources X and Y in order to compute some fixed deterministic function $f(X, Y)$.

We demonstrate the possible rate gains by example.

Example 1.1. *Consider two sources uniformly and independently producing k-bit integers X and Y; assume $k \geq 2$. We assume independence to bring to focus the compression gains from using knowledge of the function. First suppose $f(X, Y) = (X, Y)$ is to be perfectly reconstructed at the decoder. Then, the rate at which X can encode its information is k bits per symbol (bps); the same holds for Y. Thus the sum rate is $2k$ bits per function-value (bpf).*

Next, suppose $f(X, Y) = X + Y \mod 4$. The value of $f(X, Y)$ depends only upon the final two bits of both X and Y. Thus, at most (and in fact, exactly) 2 bps is the encoding rate, for a sum rate of 4 bpf. Note that the rate gain, $2k - 4$ is unbounded *because we are reducing a possibly huge alphabet to one of size* 4.

Finally, suppose $f(X,Y) = X+Y \mod 4$ *as before, but the decoder is allowed to recover* f *up to some distortion. We consider the Hamming distortion function on* f. *Consider recovering* f *up to a* 1*-bit Hamming distortion. One possible coding scheme would simply encode the single least significant bit for both* X *and* Y. *Then one could recover the least significant bit of* $f(X,Y)$, *thus achieving an encoding rate of* 1 *bps per source or* 2 *bpf.* □

Using knowledge of the decoder's final objective helps achieve better compression rates. Example 1.1 is relatively simple because the function is separable. The function need not always be separable: consider $f(X,Y) = |X - Y|$, for example. In that setting, it is much less obvious how to separately encode X and Y. Therefore, we want a more general framework.

This thesis will provide a general framework that allows us to solve both the problem of finding the best possible rates as well as finding coding schemes that allow for approximations of these rates.

Example 1.1 showcases the three specific scenarios considered in this thesis: side information with zero distortion, distributed compression with zero distortion, and side information with nonzero rate distortion. We proceed by placing our results in their historical context.

1.2 Historical Context

We can categorize compression problems with two sources along three dimensions. First, whether there is one source is locally available at the receiver (call this "side information") or whether both sources are communicating separately (call this "distributed"). Second, whether $f(x,y) = (x,y)$ or is more general. Finally, whether there is zero-distortion or non-zero-distortion. In all cases, the goal is to determine the rates (R_x, R_y) at which X and Y must be encoded ($R_y = \infty$ when Y is side information) in order for the decoder to compute $f(X,Y)$ within distortion $D \geq 0$ with high probability.

Table 1.1: Research progress on zero-distortion source coding problems

Problem types	$f(x,y) = (x,y)$	General $f(x,y)$
Side information	Shannon [23]	Orlitsky and Roche [20] *
Distributed	Slepian and Wolf [24] Pradhan and Ramchandran [21] Coleman et al. [8]	Ahlswede and Körner [1] Körner and Marton [18] *

1.2.1 Zero Distortion

First, consider zero distortion. Shannon [23] considers the side information problem where $f(x,y) = (x,y)$. Slepian and Wolf [24] consider the distributed problem where $f(x,y) = (x,y)$. Many practical and near-optimal coding schemes have been developed for both of the above problems such as DISCUS codes by Pradhan and Ramchandran [21] and source-splitting techniques by Coleman et al. [8]. Thus, the left side of Table 1.1 is fairly complete. We provide the precise theorems in a later section.

Orlitsky and Roche provided a single-letter characterization for the side information problem for a general function $f(x,y)$. Ahlswede and Körner [1] determined the rate region for the distributed problem for $f(x,y) = x$. Körner and Marton [18] considered zero-distortion with both sources separately encoded for the function $f(x,y) = x + y \mod 2$. Nevertheless, there has been little work on a general function $f(x,y)$ in the distributed zero-distortion case. This is summarized in Table 1.1.

The * indicates where this thesis makes contributions. Specifically, for zero distortion, we provide a framework that leads to an optimal modular coding scheme for the side information problem for general functions. We give conditions under which this framework can be extended to the distributed problem for general functions.

Table 1.2: Research progress on nonzero-distortion source coding problems

Problem types	$f(x,y) = (x,y)$	General $f(x,y)$
Side information	Wyner and Ziv [28] Feng et al. [16]	Yamamoto [29] *
Distributed	Berger and Yeung [4] Barros and Servetto [3] Wagner et al. [25]	

1.2.2 Nonzero Distortion

Next, consider nonzero distortion problems. Wyner and Ziv [28] considered the side information problem for $f(x,y) = (x,y)$. The rate region for the case of nonzero-distortion with both sources separately encoded is unknown but bounds have been given by Berger and Yeung [4], Barros and Servetto [3], and Wagner, Tavildar, and Viswanath [25]. Wagner et al. considered a specific distortion function for their results (quadratic). In the context of functional compression, all of these theorems are specific to $f(x,y) = (x,y)$.

Feng, Effros, and Savari [16] solved the side information problem in the case where both sources are one step removed from the encoder and decoder (i.e., the encoder and decoder have noisy information). Yamamoto solved the side information problem for a general function $f(x,y)$. This is summarized in Table 1.2.

Again, the * indicates where this thesis makes a contribution. Specifically, for nonzero distortion, we extend the framework derived for zero distortion and apply it in this more general setting. As indicated above, the distributed setting with nonzero distortion and a general function is quite difficult (even the special case $f(x,y) = (x,y)$ is not completely solved).

1.3 Overview of Results

There are several key issues that we hope to elucidate: the idea that knowing the function in some sense provides information, the information theoretic limits in the

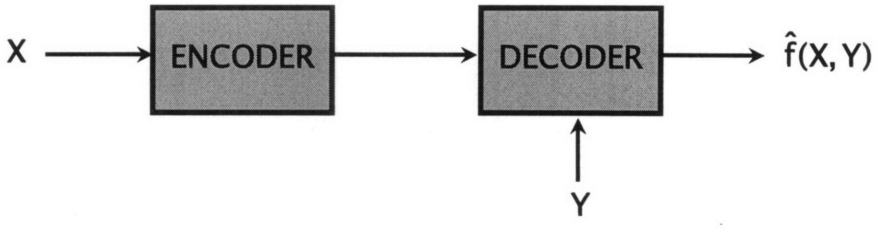

Figure 1-1: The functional compression problem with side information.

above problems, and deriving schemes that are easy to implement. Now, we describe the three problems considered in this thesis along with an overview of the results.

1.3.1 Functional Compression with Side Information

The side information problem is depicted in Figure 1-1. We describe an optimal scheme for encoding X such that $f(X, Y)$ can be computed within expected distortion D at a receiver that knows Y.

The optimal rate for the functional compression with side information problem was given by Orlitsky and Roche in the zero distortion case. While the resulting characterization is complete and tight, it is difficult to calculate even for simple source distributions. For this problem, this thesis provides a new interpretation for that rate through a simple algorithm that can be approximated with available heuristics. Computing the Orlitsky-Roche rate requires optimizing a distribution over an auxiliary random variable W. We provide an interpretation of W that leads to a simple achievability scheme for the Orlitsky-Roche rate that is modular with each module being a well-studied problem. It can be extended to and motivates our functional distributed source coding scheme below.

As mentioned earlier, Yamamoto gave a characterization of the rate distortion function for this problem as an optimization over an auxiliary random variable. We give a new interpretation to Yamamoto's rate distortion function for nonzero distortion. Our formulation of the rate distortion function leads to a coding scheme that

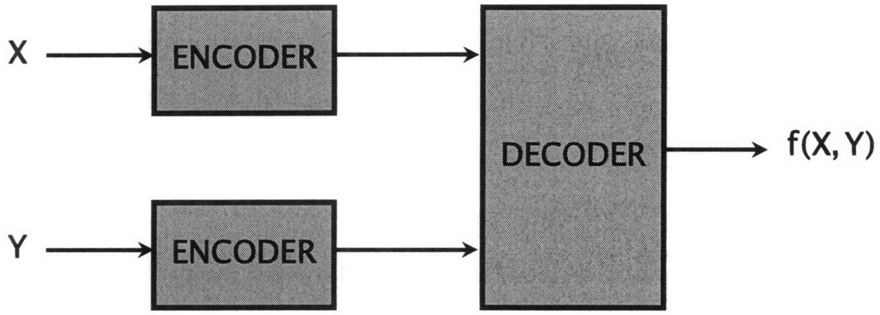

Figure 1-2: The distributed functional compression problem.

extends the coding schemes for the zero distortion case. Further, we give a simple achievability scheme that achieves compression rates that are certainly at least as good as the Slepian-Wolf rates and also at least as good as the zero distortion rate.

For zero distortion, the rate is a special case of the distributed functional compression problem considered next where one source is compressed at entropy-rate, thus allowing for reconstruction at the decoder.

1.3.2 Distributed Functional Compression

The distributed functional compression problem is depicted in Figure 1-2. In this problem, X and Y are separately encoded such that the decoder can compute $f(X, Y)$ with zero distortion and arbitrarily small probability of error. We describe the conditions under which the coding scheme mentioned in the previous section can be extended to the distributed set up. Further, we provide a less general condition depending only upon the probability distribution of the sources under which our scheme is optimal.

Thus, we extend the Slepian-Wolf rate region to a general deterministic function $f(x, y)$. Our rate region is, in general, an inner bound to the general rate region, and we provide conditions under which it is the true rate region.

1.4 Organization

The rest of the thesis is organized as follows. Chapter 2 gives the problem statement and presents the related technical background necessary to understand the main results. Chapter 3 presents those results. The proofs follow in Chapter 4. Chapter 5 gives the implications of the results to the scenarios presented at the start of this chapter. Future research directions and conclusions are given in Chapter 6. Appendix A recalls some information theory fundamentals including the Slepian-Wolf and Wyner-Ziv source coding problems.

Chapter 2

Functional Compression Background

We consider the three problems presented herein within a common framework. The necessary preliminaries for understanding this framework are given in this section. We explain where each problem differs in context. We borrow much of the notation from [9, Chapter 12].

2.1 Problem Setup

We consider two discrete memoryless sources. We assume the sources, $\{X_i\}_{i=1}^{\infty}$ and $\{Y_i\}_{i=1}^{\infty}$, are drawn from finite sets $\mathcal{X}$ and $\mathcal{Y}$ according to a joint distribution $p(x, y)$. Denote by $p(x)$ and $p(y)$ the marginals of $p(x, y)$.

We denote n-sequences of random variables X and Y as $\mathbf{X} = \{X_i\}_{i=k}^{k+n-1}$ and $\mathbf{Y} = \{Y_i\}_{i=k}^{k+n-1}$, respectively, where n and k are clear from context. We generally assume $k = 1$. Because the sequence $(\mathbf{x}, \mathbf{y})$ was drawn i.i.d. according to $p(x, y)$, we can write the probability of any instance of the sequence as $p(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} p(x_i, y_i)$.

The sources encode their messages (at rates $R_x, R_y \in [0, \infty]$) for a common decoder to compute a fixed deterministic function $f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ or $f : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{Z}^n$, its vector extension. (The n will be clear from context.)

For any n, D, R_x, and R_y, we define a *n-distributed functional code* for the joint source (X, Y) and function f as two encoder maps,

$$\begin{aligned} e_x &: \mathcal{X}^n \rightarrow \left\{1, \ldots, 2^{nR_x}\right\}, \\ e_y &: \mathcal{Y}^n \rightarrow \left\{1, \ldots, 2^{nR_y}\right\}, \end{aligned}$$

and a decoder map,

$$r : \left\{1, \ldots, 2^{nR_x}\right\} \times \left\{1, \ldots, 2^{nR_y}\right\} \rightarrow \mathcal{Z}^n.$$

Consider a distortion function, $d : \mathcal{Z} \times \mathcal{Z} \rightarrow [0, \infty)$, with vector extension

$$d(\mathbf{z}_1, \mathbf{z}_2) = \frac{1}{n} \sum_{i=1}^{n} d(z_{1i}, z_{2i}),$$

where $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}^n$. As in [28], we assume that the distortion function satisfies $d(z_1, z_2) = 0$ if and only if $z_1 = z_2$. (Otherwise, one can define the equivalence classes of the function values to make this condition hold.)

The *probability of error* is

$$P_e^n = \Pr[\{(\mathbf{x}, \mathbf{y}) : d(f(\mathbf{x}, \mathbf{y}), r(e_x(\mathbf{x}), e_y(\mathbf{y}))) > D\}].$$

A rate pair, (R_x, R_y), is *achievable* for a distortion D if there exists a sequence of n-distributed functional codes at those rates and distortion level such that $P_e^n \rightarrow 0$ as $n \rightarrow \infty$. The *achievable rate region* is the set closure of the set of all achievable rates.[1] Our most general objective is to find this achievable rate region. If $f(x, y) = (x, y)$, this formulation is the same as the problems considered by Shannon [23] (when $D = 0$ and $R_y = 0$ or $R_y = \infty$), Slepian and Wolf [24] ($D = 0$), etc. Thus, the main difference is in the probability of error definition.

[1] The set closure of a set S is a set closed A such that $S \subseteq A$ and $A \subset T$ for any other closed set T such that $S \subseteq T$.

2.2 Previous Results

We begin by defining a construct useful in formulating all the results.

Definition 2.1. *The characteristic graph $G_x = (V_x, E_x)$ of X with respect to Y, $p(x,y)$, and $f(x,y)$ is defined as follows: $V_x = \mathcal{X}$ and an edge $(x_1, x_2) \in \mathcal{X}^2$ is in E_x if there exists a $y \in \mathcal{Y}$ such that $p(x_1,y)p(x_2,y) > 0$ and $f(x_1,y) \neq f(x_2,y)$.*

Defined thus, G_x is the "confusability graph" from the perspective of the receiver. If $(x_1, x_2) \in E_x$, then the descriptions of x_1 and x_2 must be different to avoid confusion about $f(x,y)$ at the receiver. This was first defined by Shannon when studying the zero error capacity of noisy channels [22]. Witsenhausen [27] used this graph to consider our problem in the case when one source is deterministic, or equivalently, when one encodes X to compute $f(X)$ with 0 distortion. The characteristic graph of Y with respect to X, $p(x,y)$, and $f(x,y)$ is defined analogously and denoted G_y. When notationally convenient and clear, we will drop the subscript.

The importance of using the characteristic graph construct becomes clear when considering independent sets[2] of the graph. By definition of the edge set, knowledge of Y and the independent set uniquely determines $f(x,y)$.

For all vertices in a particular independent set, given Y, f is fixed by definition of the edge set. The characteristic graph thus realizes the intuition behind confusability.

We illustrate this with Example 2.2.

Example 2.2. *To illustrate the idea of confusability and the characteristic graph, consider again the sources from Example 1.1. Both sources are uniformly and independently generating, say, 3-bit integers. The function of interest is $f(X,Y) =$ mod 4. Then, the characteristic graph of X with respect to Y, $p(x,y) = \frac{1}{8}$ and f will be as shown in Figure 2-1. Specifically, we see that there is an edge between x_1 and x_2 if and only if they differ in one or both of their final two bits. This is because, given any y, $f(x_1,y) = f(x_2,y)$ if and only if x_1 and x_2 have the same final two bits.* □

[2] A subset of vertices of a graph G is an independent set if no two nodes in the subset are adjacent to each other in G. With the characteristic graph, independent sets form equivalence classes.

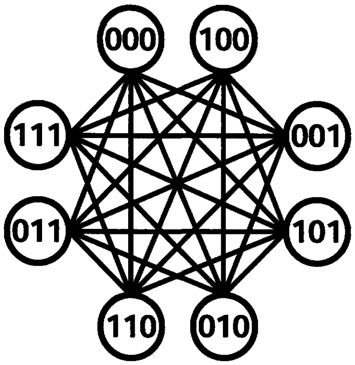

Figure 2-1: Example of a characteristic graph.

Next, we define graph entropy, which we use later to more generally derive the rate required for communication for problems such as in Example 2.2.

Definition 2.3. *Given a graph $G = (V, E)$ and a distribution on the vertices V, Körner [17] defines the graph entropy as:*

$$H_G(X) = \min_{X \in W \in \Gamma(G)} I(W; X), \tag{2.1}$$

where $\Gamma(G)$ is the set of all independent sets of G.

The notation $X \in W \in \Gamma(G)$ means that we are minimizing over all distributions $p(w, x)$ such that $p(w, x) > 0$ implies $x \in w$ where w is an independent set of the graph G. We now demonstrate how this can be used to solve problems like that given in Example 2.2.

Example 2.4. *Consider again the scenario in Example 1.1 as presented in Example 2.2. For the graph in Figure 2-1, the maximally independent sets are the sets with the same final two bits. Thus, to minimize $I(X; W) = H(X) - H(X|W)$ or to maximize $H(X|W)$, we must have $p(w)$ uniform over the* 4 *maximally independent sets. This is because $H(X|w) = 1$ for all maximally independent w. Therefore, we get $H_G(X) = 3 - 1 = 2$.* □

As alluded to earlier and demonstrated in the example, Witsenhausen [27] showed that the graph entropy is the minimum rate at which a single source can be encoded such that a function of that source can be computed with zero distortion. For this result, the vertices of the graph are the support of the random variable X and the edge set is such that x and x' have an edge when $f(x) \neq f(x')$. In other words, it is our characteristic graph from above when there is no randomness in Y.

Orlitsky and Roche [20] defined an extension of Körner's graph entropy, the *conditional graph entropy.*

Definition 2.5. *The conditional graph entropy is*

$$H_G(X|Y) = \min_{\substack{X \in W \in \Gamma(G) \\ W - X - Y}} I(W; X|Y). \tag{2.2}$$

The additional constraint that $W - X - Y$ forms a Markov chain formally states the intuition that W should not contain any information about Y that is not available through X. If X and Y are independent, $H_G(X|Y) = H_G(X)$.

Theorem 2.6 (Orlitisky-Roche Theorem, 2001 [20]). *When G is the characteristic graph of X with respect to Y, $p(x,y)$, and $f(x,y)$, $R_x \geq H_G(X|Y)$ is the rate region for reliable computation of the function $f(X,Y)$ with zero distortion and arbitrarily small probability of error when Y is available as side information.*[3]

A natural extension of this problem is the functional compression with side information problem for nonzero distortion. Yamamoto gives a full characterization of the rate-distortion function for the side information functional compression problem [29] as a generalization of the Wyner-Ziv side-information rate-distortion function [28]. Specifically, Yamamoto gives the rate distortion function as follows.

Theorem 2.7 (Yamamoto Theorem, 1982). *The rate distortion function for the functional compression with side information problem is*

$$R(D) = \min_{p \in \mathcal{P}(D)} I(W; X|Y)$$

[3] The knowledge of Y at the decoder can arise either through side information or an independent description of Y at rate $H(Y)$.

where $\mathcal{P}(D)$ is the collection of all distributions on W given X such that there exists a $g : \mathcal{W} \times \mathcal{Y} \to \mathcal{Z}$ satisfying $E[d(f(\mathbf{X}, \mathbf{Y}), g(\mathbf{W}, \mathbf{Y}))] \leq D$.

This is a natural extension of the Wyner-Ziv result, Theorem A.4. Additionally, the variable $W \in \Gamma(G)$ in the definition of the Orlitsky-Roche rate (Definition 2.2) can be seen as an interpretation (for the zero-distortion case), of Yamamoto's auxiliary variable, W, as a variable over the independent sets of G. Namely, when the distortion $D = 0$, the distributions on W given X for which there exists a reconstruction function g place nonzero probability only if w describes an independent set of G and $x \in w$.

2.3 Graph Entropies

Our results depend on the use of more graph tools, which we now describe. Alon and Orlitsky denoted [2] the OR-power graph of G as $G^n = (V_n, E_n)$ where $V_n = V^n$ and two vertices $(\mathbf{x}_1, \mathbf{x}_2) \in E_n \subseteq V_n \times V_n$ if any component $(x_{1i}, x_{2i}) \in E$. This can be interpreted as an encoding of a block of source observations.

A vertex coloring of a graph is any function $c : V \to \mathbb{N}$ of a graph $G = (V, E)$ such that $(x_1, x_2) \in E$ implies $c(x_1) \neq c(x_2)$. The entropy of a coloring is the entropy of the induced distribution on colors $p(c(x)) = p(c^{-1}(c(x)))$ where $c^{-1}(x) = \{\bar{x} : c(\bar{x}) = c(x)\}$ and is called a *color class.*

Definition 2.8. *Let $\mathcal{A} \subset \mathcal{X} \times \mathcal{Y}$ be a high probability set (i.e., be any subset such that $p(\mathcal{A}) \geq 1 - \varepsilon$). Define $\hat{p}(x, y) = p(x, y)/p(\mathcal{A})$. In other words, $\hat{p}$ is the distribution over (x, y) conditioned on $(x, y) \in \mathcal{A}$. Denote the characteristic graph of X with respect to Y, $\hat{p}$, and f as $\hat{G}_x = (\hat{V}_x, \hat{E}_x)$ and and the characteristic graph of Y with respect to X, $\hat{p}$, and f as $\hat{G}_y = (\hat{V}_y, \hat{E}_y)$. Note that $\hat{E}_x \subseteq E_x$ and $\hat{E}_y \subseteq E_y$. Finally, we say c_x and c_y are ε-colorings of G_x and G_y if they are valid colorings of $\hat{G}_x$ and $\hat{G}_y$ defined with respect to some high probability set $\mathcal{A}$.*

Alon and Orlitsky [2] defined the *chromatic entropy* of a graph G.

Definition 2.9.

$$H_G^\chi(X) = \min_{c \text{ is an } \varepsilon\text{-coloring of } G} H(c(X)).$$

Well-known typicality results (e.g., [9]) imply that there exists a high probability set for which the graph vertices are roughly equiprobable. Thus, the chromatic entropy is a representation of the chromatic number of high probability subgraphs of the characteristic graph. We define a natural extension, the *conditional chromatic entropy*, as follows.

Definition 2.10.

$$H_G^\chi(X|Y) = \min_{c \text{ is an } \varepsilon\text{-coloring of } G} H(c(X)|Y).$$

The above optimizations are minima and not infima because there are only finitely many subgraphs of any fixed G, and thus only finitely many ε-colorings regardless of ε. Later, in order to use typicality results, we allow a block length n and look at a G^n, in order to drive error to zero. Thus, as $n \to \infty$, we will look at the infimum over all n.

These optimizations are NP-hard [6], but they can be approximated using existing heuristics [5, 19]. With these definitions and results, we can now discuss our formal results.

Chapter 3

Main Results

In this chapter, we formally state our results for the problems described earlier. The proofs are provided in Chapter 4.

3.1 Functional Compression with Side Information ($D = 0$)

We begin by describing this zero distortion problem for a single source. For example, consider the case where the desired function depends only on X, and there is no side information at the decoder. Witsenhausen [27] tells us the optimal rate is the graph entropy $H_G(X)$ defined earlier in Definition 2.3 where G is the characteristic graph of X with respect to the function $f(X)$. As stated earlier, the chromatic entropy is a representation of the chromatic number of a high probability subgraph of the characteristic graph. Körner proved [17] that in an appropriate limit, the chromatic entropy approaches the graph entropy:

Theorem 3.1 (Körner Theorem, 1973).

$$\lim_{n \to \infty} \frac{1}{n} H_{G^n}^{\chi}(\mathbf{X}) = H_G(X). \tag{3.1}$$

The implications of this result are that we can compute a function of a discrete memoryless source with vanishing probability of error by first coloring a sufficiently large power graph of the characteristic graph of the source with respect to the function, and then, encoding the colors using any code that achieves the entropy bound on the colored source. The previous approach for achieving rates close to the bound $H_G(X)$ was to optimize with respect to a distribution over W as in the definition of $H_G(X)$. This theorem allows us to move the optimization from finding the optimal distribution to finding the optimal colorings. Thus, our solution modularizes the coding by first creating a graph coloring problem (for which heuristics exist), and then transmitting the colors using any existing entropy-rate code. Moreover, we can extend this technique to the functional side information case.

Now recall the problem described by Figure 1-1. For the problem with two sources, Orlitsky and Roche proved the optimal rate for the zero distortion functional compression problem with side information is $H_G(X|Y)$. Recall from Definition 2.5, $H_G(X|Y)$ is also achieved by optimizing a distribution over W. Thus, our goal is to extend Körner's Theorem to conditional chromatic and conditional graph entropies. We do this in the following theorem.

Theorem 3.2.

$$\lim_{n\to\infty} \frac{1}{n} H^{\chi}_{G^n}(\mathbf{X}|\mathbf{Y}) = H_G(X|Y)$$

This theorem extends the previous result to the conditional case. In other words, in order to encode a memoryless source, we first color a graph G^n for sufficiently large n. Then, we encode each source symbol with its corresponding vertex's color. Finally, we use a Slepian-Wolf code on the sequence of colors achieving a rate arbitrarily close to $H(c(\mathbf{X})|\mathbf{Y})$. This allows for computation of the function at the decoder, and the overall code will then have rate arbitrarily close to the Orlitsky-Roche bound $H_G(X|Y)$ by Theorem 3.2. Thus, it is asymptotically optimal. Figure 3-1 illustrates our coding scheme.

The Orlitsky-Roche achievability proof uses random coding arguments and proves the existence of an optimal coding scheme, but does not specify it precisely. Our coding scheme allows the use of heuristics available for finding good colorings as well

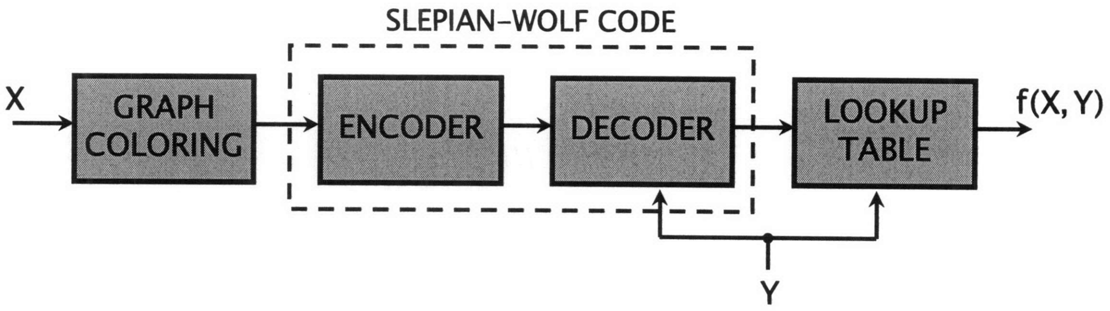

Figure 3-1: Source coding scheme for the zero distortion functional compression problem with side information.

as the use of optimal source codes that achieve the conditional entropy. Finding the minimum entropy colorings required to achieve the bound is NP-hard [6], but the simplest colorings will (weakly)[1] improve over the bound $H(X|Y)$ that arises when trying to recover X completely at the receiver by the Data Processing Inequality. This solution gives the corner points of the achievable rate region for the distributed functional compression problem, considered next.

3.2 Distributed Functional Compression ($D = 0$)

In this section, we prove rate bounds for the distributed functional compression problem. The derived rate region is always achievable and sometimes tight. The region directly evolves from the coloring arguments discussed in the above section.

Recall the problem shown in Figure 1-2. Our goal is to provide an achievability scheme that extends the modular scheme given in Figure 3-1 for the functional side information problem. In other words, we provide a scheme that first precodes the data (coloring) and then uses existing coding schemes on the precoded data. Here, the natural choice for existing coding schemes would be codes that achieve rates close to the Slepian-Wolf rate region boundary.

The Slepian-Wolf Theorem (see Theorem A.3) states that in order to recover

[1] For all functions $f(\mathbf{X}, \mathbf{Y})$, $H(c(\mathbf{X})|\mathbf{Y}) \leq H(\mathbf{X}|\mathbf{Y})$. For any non-injective function $f(\mathbf{X}, \mathbf{Y})$ and large enough block length n, $H(c(\mathbf{X})|\mathbf{Y}) < H(\mathbf{X}|\mathbf{Y})$.

a joint source (X, Y) at a receiver, it is both necessary and sufficient to encode separately sources X and Y at rates (R_x, R_y) where

$$\begin{aligned} R_x &\geq H(X|Y) \\ R_y &\geq H(Y|X) \\ R_x + R_y &\geq H(X, Y). \end{aligned}$$

Let this region be denoted $\mathcal{R}(X, Y)$.

Also, for any n, and functions g_x and g_y defined on $\mathcal{X}^n$ and $\mathcal{Y}^n$ respectively, denote by $\mathcal{R}^n(g_x, g_y)$ the Slepian-Wolf region for the induced variables $g_x(\mathbf{X})$ and $g_y(\mathbf{Y})$ normalized by the block length. In other words, $\mathcal{R}^n(g_x, g_y)$ is the set of all (R_x, R_y) such that:

$$\begin{aligned} R_x &\geq \frac{1}{n} H(g_x(\mathbf{X})|g_y(\mathbf{Y})), \\ R_y &\geq \frac{1}{n} H(g_y(\mathbf{Y})|g_x(\mathbf{X})), \\ R_x + R_y &\geq \frac{1}{n} H(g_x(\mathbf{X}), g_y(\mathbf{Y})). \end{aligned}$$

If Y is sent at rate $H(Y)$, it can be faithfully recovered at the receiver. Thus, the rate for X is then $H_{G_x}(X|Y)$ as given by Orlitsky and Roche. And the rate for Y when $R_x \geq H(X)$ is $H_{G_y}(Y|X)$. Therefore, we know the corner points for the rate region for the distributed functional compression problem.

Our goal is to determine the region as well as give a scheme analogous to the one given in Figure 3-1 that achieves all rates in the given region. We proceed with the following philosophy: color both X and Y and encode the colored sequences using codes achieving the Slepian-Wolf bounds. We want to characterize when this approach is valid. In other words, we want to find the conditions under which colorings of the characteristic graphs are sufficient to determine $f(x, y)$ for the zero distortion problem.

3.2.1 Zigzag Condition

A condition which is necessary and sufficient for a coloring scheme as presented to give a legitimate code is given below.

Condition 3.3. *For any n, consider ε-colorings c_x and c_y of G_x^n and G_y^n with associated probability distribution $\hat{p}$. The colorings c_x and c_y and the source distribution $p(x,y)$ are said to satisfy the condition if for all colors $(\gamma,\sigma) \in c_x(\mathcal{X}^n) \times c_y(\mathcal{Y}^n)$, and all $(\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2) \in c_x^{-1}(\gamma) \times c_y^{-1}(\sigma)$ such that $\hat{p}(\mathbf{x}_1,\mathbf{y}_1)\hat{p}(\mathbf{x}_2,\mathbf{y}_2) > 0$, $f(\mathbf{x}_1,\mathbf{y}_1) = f(\mathbf{x}_2,\mathbf{y}_2)$.*

Condition 3.3 presupposes an encoding scheme. We next present a condition that does not.

Condition 3.4 (Zigzag Condition). *A discrete memoryless source $\{(X_i,Y_i)\}_{i\in\mathbb{N}}$ with distribution $p(x,y)$ satisfies the Zigzag Condition if for any ε and some n, $(\mathbf{x}_1,\mathbf{y}_1)$, $(\mathbf{x}_2,\mathbf{y}_2) \in T_\varepsilon^n$, there exists some $(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \in T_\varepsilon^n$ such that $(\tilde{\mathbf{x}},\mathbf{y}_i),(\mathbf{x}_i,\tilde{\mathbf{y}}) \in T_{\frac{\varepsilon}{2}}^n$ for each $i \in \{1,2\}$, and $(\tilde{x}_j,\tilde{y}_j) = (x_{ij},y_{(3-i)j})$ for some $i \in \{1,2\}$ for each j.*

Under Condition 3.3, deciding whether to use our encoding scheme would require the encoding scheme to already exist because it relies on c_x and c_y. This is problematic. On the other hand, the Zigzag Condition depends only on the source distribution. Further, it implies Condition 3.3. The Zigzag Condition is weaker, but easier to check: if not satisfied, there still may be colorings which satisfy Condition 3.3 and are thus valid and possibly optimal codes.

However, if the Zigzag Condition is satisfied, so is Condition 3.3 and the encoding scheme is optimal. The Zigzag Condition is sufficient to ensure that any coloring of the characteristic graphs will be sufficient to determine the function, and the communication will be at a necessary rate. We want to understand the Zigzag Condition in terms of what sources must satisfy it.

Consider some $(\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2) \in T_\varepsilon^n$. Then, we must have some $(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \in T_\varepsilon^n$ such that $(\tilde{\mathbf{x}},\mathbf{y}_1),(\tilde{\mathbf{x}},\mathbf{y}_2),(\mathbf{x}_1,\tilde{\mathbf{y}}),(\mathbf{x}_2,\tilde{\mathbf{y}}) \in T_{\frac{\varepsilon}{2}}^n$. This is represented in the Figure 3-2. If a solid line is between two values, then the pair is in T_ε^n. If a dashed line is between the

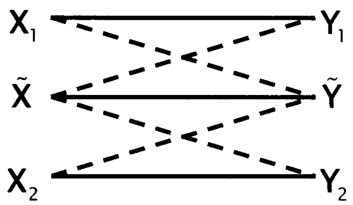

Figure 3-2: An illustration of the Zigzag Condition.

two values, then the paier is in $T^n_{\frac{\varepsilon}{2}}$. Thus, the Zigzag Condition forces many source values to be typical, which suggests this condition is, in fact, strong.

Next, consider a source that does not satisfy the Zigzag Condition. One way we could force it to satisfy the Condition would be to make previously atypical source values typical. However, this would necessarily increase the rate and we have no guarantee of optimality.

3.2.2 Rate Region

Next, define $\mathcal{S}^\varepsilon = \bigcup_{n=1}^{\infty} \bigcup_{(c_x^n, c_y^n)} \mathcal{R}^n(c_x^n, c_y^n)$ where for all n, c_x^n and c_y^n are ε-colorings of G_x^n and G_y^n. Let $\mathcal{S}$ be such that $\mathcal{S} \subseteq \mathcal{S}^\varepsilon$ for all $\varepsilon > 0$, and for any other $\mathcal{A} \subseteq \mathcal{S}^\varepsilon$ for all $\varepsilon > 0$, $\mathcal{A} \subseteq \mathcal{S}$. Thus, $\mathcal{S}$ is the largest set that is a subset of $\mathcal{S}^\varepsilon$ for all $\varepsilon > 0$. Let $\bar{\mathcal{S}}$ be the set closure of $\mathcal{S}$.[2] Finally, let $\mathcal{R}$ be the rate region for the distributed functional compression problem. We can now state the rate region in the notation just given.

Theorem 3.5. *For any $\varepsilon > 0$, $\mathcal{S}^\varepsilon$ is an inner bound to the rate region, and thus, $\mathcal{S}$ is an inner bound to the rate region. In other words, $\mathcal{R} \subseteq \mathcal{S}$. Moreover, under the Zigzag Condition, the rate region for the distributed functional source coding problem is $\mathcal{R} = \bar{\mathcal{S}}$.*

Thus, under the Zigzag Condition, any colorings of high probability subgraphs of G_x^n and G_y^n will allow for computation of the function. Furthermore, no other encodings can achieve better rates. Thus, we have decoupled the encoding for functional compression such that we first color our graph, and then we apply a Slepian-Wolf

[2] The set closure $\bar{\mathcal{S}}$ is the smallest closed set that contains $\mathcal{S}$.

code on the colors, an extension of Theorem 3.2 to the distributed case. Further, while the above characterization is not single letter, *any* coloring (except when the coloring is injective) necessarily does better than the Slepian-Wolf rates, by the Data Processing Inequality (cf. [9]).

Also, we know that the Orlitsky-Roche bound is consistent with our region at $R_y = H(Y)$. To see this, note that if $R_y = H(Y)$, this is the same as saying $c_y(\mathbf{y}) = \mathbf{y}$ for all $\mathbf{y}$ typical with some $\mathbf{x}$. Thus, the rate R_x must be $R_x \geq \frac{1}{n} H(c_x(\mathbf{X})|\mathbf{Y})$ which is minimized at $H_{G_x}(X|Y)$ by Theorem 3.2. Next, we derive a characterization of the minimum joint rate, $R_x + R_y$ in terms of graph entropies.

Corollary 3.6. *Under the Zigzag Condition, if there is a unique point which achieves the minimum joint rate, it must be* $R_x + R_y = H_{G_x}(X) + H_{G_y}(Y)$.

Thus, both encoders can use only $p(x)$ and $p(y)$, respectively, when encoding, and encode at rates $H(c_x(\mathbf{x}))$ and $H(c_y(\mathbf{y}))$, respectively, to achieve the optimal joint rate. Next, we consider the case where there jointly optimal rate is not unique. We want to determine how far $R_x + R_y = H_{G_x}(X) + H_{G_y}(Y)$ is from the optimum.

Theorem 3.7. *Let* $I_{G_x}(X;Y) = H_{G_x}(X) - H_{G_x}(X|Y)$ *be the graph information of* X *with respect to* Y. *Let* $I_{G_y}(Y;X) = H_{G_y}(Y) - H_{G_y}(Y|X)$ *be the graph information of* Y *with respect to* X. *Let* $R_{xy} = R_x + R_y$ *represent the value of the minimum joint rate. Then, under the Zigzag Condition:*

$$\left[H_{G_x}(X) + H_{G_y}(Y)\right] - R_{xy} \leq \min\left\{I_{G_x}(X;Y), I_{G_y}(Y;X)\right\}.$$

Thus, for the case in Corollary 3.6, the mutual information of the minimum entropy colorings of G_x^n and G_y^n goes to zero as $n \to \infty$:

$$\lim_{n\to\infty} \frac{1}{n} I(c(\mathbf{X}); c(\mathbf{Y})) = 0.$$

Further, Theorem 3.7 implies that if f is very simple in the sense that the independent sets of G_x are large, then $H_{G_x}(X)$ and $H_{G_x}(X|Y)$ are close, and $I_{G_x}(X;Y)$ is close to zero. Therefore, coloring followed by fixed block length compression (using $p(x)$, not

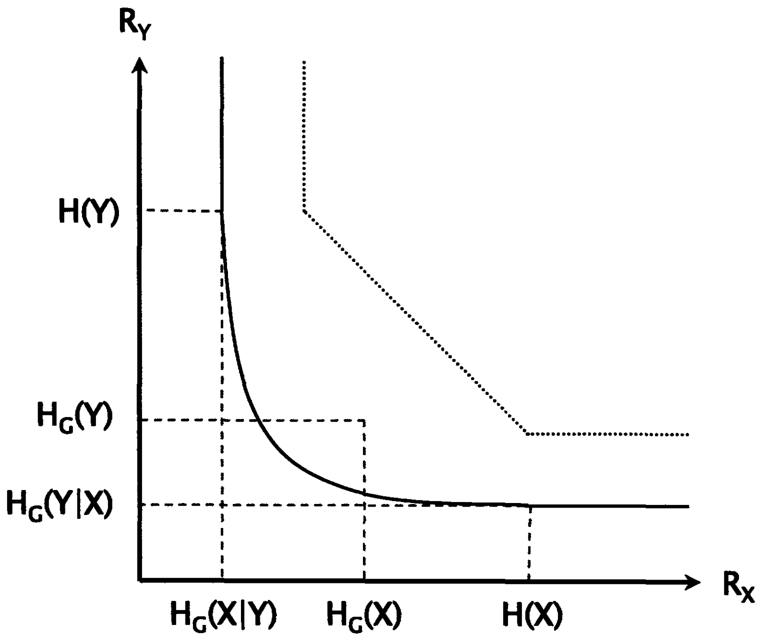

Figure 3-3: An example rate region for the zero distortion distributed functional compression problem.

$p(x, y)$) is not too far from optimal. (Similarly, for G_y.) Another case when the right hand side of Theorem 3.7 is small is when X and Y have small mutual information. In fact, if X and Y are independent, the right hand side is zero and Corollary 3.6 applies.

The region given in Theorem 3.5 has several interesting properties. First, it is convex by time-sharing arguments for any two points in the region. Second, when there is a unique minimal joint rate R_x+R_y, we can give a single-letter characterization for it (Corollary 3.6). When it is not unique, we have given a simple bound on performance.

Figure 3-3 presents a possible rate region for the case where the minimal rate is not unique. (For ease of reading, we drop the subscripts for G_x and G_y and write G for both.)

The "corners" of this rate region are $(H_{G_x}(X|Y), H(Y))$ and $(H(X), H_{G_y}(Y|X))$, the Orlitsky-Roche points, which can be achieved with graph coloring, in the limit sense, as described earlier. For any rate $R_x \in (H_{G_x}(X|Y), H(X))$, the joint rate required is less than or equal to the joint rate required by a time-sharing of the Orlitsky-Roche scheme. The inner region denoted by the dotted line is the Slepian-Wolf rate region.

The other point we characterize is the minimum joint rate point (when unique) given as $(H_{G_x}(X), H_{G_y}(Y))$. Thus, we have given a single-letter characterization for three points in the region.

3.3 Functional Compression with Side Information $(D > 0)$

We now consider the functional rate distortion problem; we give a new characterization of the rate distortion function given by Yamamoto. We also give an upper bound on that rate distortion function which leads to an achievability scheme that mirrors those given in the functional side information problem.

Recall the Yamamoto rate distortion function (Theorem 2.7). According to the Orlitsky-Roche result (Theorem 2.6), when $D = 0$, any distribution over independent sets of the characteristic graph (with the Markov chain $W - X - Y$ imposed) is in $\mathcal{P}(0)$. Any distribution in $\mathcal{P}(0)$ can be thought of as a distribution over independent sets of the characteristic graph.

We show that finding a suitable reconstruction function, $\hat{f}$, is equivalent to finding the g on $\mathcal{W} \times \mathcal{Y}$ from Theorem 2.7. Specifically, for any m, let $\mathcal{F}_m(D)$ denote the set of all functions $\hat{f}_m : \mathcal{X}^m \times \mathcal{Y}^m \to \mathcal{Z}^m$ such that

$$\lim_{m\to\infty} E[d(f(\mathbf{X}, \mathbf{Y}), \hat{f}_m(\mathbf{X}, \mathbf{Y}))] \leq D.$$

To prove this, we will eventually consider blocks of m-vectors; thus, the functions in the expectation above will be on $\mathcal{X}^{mn} \times \mathcal{Y}^{mn}$. Let $\mathcal{F}(D) = \bigcup_{m\in\mathbb{N}} \mathcal{F}_m(D)$. Let $G(\hat{f})$ denote the characteristic graph of $\mathbf{X}$ with respect to $\mathbf{Y}$, $p(\mathbf{x}, \mathbf{y})$, and $\hat{f}$ for any $\hat{f} \in \mathcal{F}(D)$. For each m and all functions $\hat{f} \in \mathcal{F}(D)$, denote for brevity the normalized graph entropy $\frac{1}{m}H_{G(\hat{f})}(\mathbf{X}|\mathbf{Y})$ as $H_{G(\hat{f})}(X|Y)$.

Theorem 3.8.

$$R(D) = \inf_{\hat{f}\in\mathcal{F}(D)} H_{G(\hat{f})}(X|Y)$$

Note that $G(\hat{f})$ must be a subgraph of the characteristic graph G^m (for appropriate m) with respect to f. Because G^m is finite, there are only finitely many subgraphs. Thus, for any fixed error ε and associated block length m, this is a finite optimization. This theorem implies that once the suitable reconstruction function $\hat{f}$ is found, the functional side information bound (and achievability scheme) using the graph $G(\hat{f})$ is optimal in the limit.

This characterization is mostly illustrative. Indeed, $\mathcal{F}(D)$ is an (uncountably) infinite set, but the set of graphs associated with these functions is countably infinite. Moreover, any allowable graph dictates an ordinal function, but it has no meaning in terms of distortion. Given the ordinal function $\hat{f}$, choosing the cardinal values that minimize expected distortion is a tractable optimization problem. This shows that if one could find an approximation function $\hat{f}$, the compression rate may improve (even when $\hat{f}$ is not optimal).

The problem of finding an appropriate function $\hat{f}$ is equivalent to finding a new graph whose edges are a subset of the edges of the characteristic graph. This motivates Corollary 3.9 where we use the a graph parameterized by D to look at a subset of $\mathcal{F}(D)$. The resulting bound is not tight, but it provides a practical tool for tackling a very difficult problem.

Define the D-characteristic graph of X with respect to Y, $p(x, y)$, and $f(x, y)$, as having verticies $V = \mathcal{X}$ and the pair (x_1, x_2) is an edge if there exists some $y \in \mathcal{Y}$ such that $p(x_1, y)p(x_2, y) > 0$ and $d(f(x_1, y), f(x_2, y)) > D$. Denote this graph as G^D. Because $d(z_1, z_2) = 0$ if and only if $z_1 = z_2$, the 0-characteristic graph is the characteristic graph, i.e. $G^0 = G$.

Corollary 3.9. *The rate $H_{G^D}(X|Y)$ is achievable.*

Constructing this graph is not computationally difficult when the number of vertices is small. Given the graph, we have a set of equivalence classes for $\hat{f}$. One can then optimize $\hat{f}$ by choosing those values for the equivalence classes that minimize distortion. However, any legal values (values that lead to the graph G^D) will necessarily still have distortion within D. Indeed, this construction guarantees that not only is expected distortion less than or equal to D, but actual distortion is always less than or equal to D. There are many possible improvements to be made here.

Theorem 3.2 and the corresponding achievability scheme, Corollary 3.9, gives a simple coding scheme that may potentially lead to large compression gains.

3.4 Possible Extensions

In all of the above problems, our achievability schemes are modular, providing a separation between the computation of the function and the correlation between the sources.

The computation module is a graph coloring module. The specific problem of interest for our scheme is NP-hard [6], but there is ample literature providing near-optimal graph coloring heuristics for special graphs or heuristics that work well in certain cases [5, 19].

The correlation module is a standard entropy coding scheme such as Slepian-Wolf coding. There are many practical algorithms with near-optimal performance for these codes. For example DISCUS codes [21] and source-splitting techniques [8].

Given the separation, the problem of functional compression becomes more tractable. While the overall problem may still be NP-hard, one can combine the results from each module to provide heuristics that are good for the specific task at hand.

We note that all results consider only two sources X and Y. Extending the correlation coding to a more general scenario of N sources is a solved problem [9, p. 415]. Thus, it seems likely that given a suitable extension of the graph coloring technique and an extension of the condition required for distributed compression (the Zigzag Condition), a full region for N sources would resolve. However, we focus on the two-source scenario because, as with Slepian-Wolf, we have gained many insights into the problem. We leave the extension to future work.

Chapter 4

Proofs and Ancillary Results

In this chapter, we provide full proofs of all our previously-stated results.

4.1 Functional Compression with Side Information

We recall Theorem 3.2:

$$\lim_{n\to\infty} \frac{1}{n} H^{\chi}_{G^n}(\mathbf{X}|\mathbf{Y}) = H_G(X|Y). \tag{4.1}$$

To prove this, we borrow proof techniques from Körner [17], and Orlitsky and Roche [20]. We first state some more typicality results. We use the notion of ε-strong typicality (see Definition A.2).

Lemma 4.1. *Suppose $(\mathbf{X}, \mathbf{Y})$ is a sequence of n random variables drawn independently and according to the joint distribution $p(x, y)$, which is the marginal of $p(w, x, y)$. Let an n-sequence $\mathbf{W}$ be drawn independently according to its marginal, $p(w)$. Suppose the joint distribution $p(w, x, y)$ forms a Markov Chain, $W - X - Y$. Then, for all $\varepsilon > 0$, there is a $\varepsilon_1 = k \cdot \varepsilon$, where k depends only on the distribution $p(w, x, y)$, such that for sufficiently large n,*

1. *$P[\mathbf{X} \notin T^n_\varepsilon] < \varepsilon_1$, $P[\mathbf{Y} \notin T^n_\varepsilon] < \varepsilon_1$, and $P[(\mathbf{X}, \mathbf{Y}) \notin T^n_\varepsilon] < \varepsilon_1$,*
2. *For all $\mathbf{x} \in T^n_\varepsilon$, $P[(\mathbf{x}, \mathbf{W}) \in T^n_\varepsilon] \geq 2^{-n(I(W;X)+\varepsilon_1)}$.*

3. For all $\mathbf{y} \in T_\varepsilon^n$*,* $P[(\mathbf{y}, \mathbf{W}) \in T_\varepsilon^n] \leq 2^{-n(I(W;X)-\varepsilon_1)}$*.*

4. For all $(\mathbf{w}, \mathbf{x}) \in T_\varepsilon^n$*,*

$$P[(\mathbf{w}, \mathbf{Y}) \in T_\varepsilon^n | (\mathbf{x}, \mathbf{Y}) \in T_\varepsilon^n] \geq 1 - \varepsilon_1.$$

Part 1 follows from [9, Lemma 13.6.1], parts 2 and 3 follow from [9, Lemma 13.6.2], and part 4 follows from [9, Lemma 14.8.1].

4.1.1 Lower Bound

Consider any n and the corresponding OR-product graph G^n. Let c be an ε-coloring of G^n that achieves $H_{G^n}^{\chi}(\mathbf{X}|\mathbf{Y})$, i.e.

$$H_{G^n}^{\chi}(\mathbf{X}|\mathbf{Y}) = H(c(\mathbf{X})|Y).$$

As stated earlier, a minimum entropy coloring exists because the set of all colorings on a graph with a finite number of verticies (here, fewer than $|\mathcal{X}|^n$) is finite. With this coloring, we prove that there is a scheme that encodes at rate $\frac{1}{n}H_{G^n}^{\chi}(\mathbf{X}|\mathbf{Y})$ such that the decoder can compute $f(\mathbf{X}, \mathbf{Y})$ with small probability of error. Proving that would establish achievability of the rate $\frac{1}{n}H_{G^n}^{\chi}(\mathbf{X}|\mathbf{Y})$. Theorem 2.6 proves that no achievable rate can be below $H_G(X|Y)$. This will give us the following lemma.

Lemma 4.2.

$$\liminf_{n\to\infty} \frac{1}{n} H_{G^n}^{\chi}(\mathbf{X}|\mathbf{Y}) \geq H_G(X|Y).$$

Proof. For any $n > 0$, let c, as above, denote a coloring on G^n that achieves $H_{G^n}^{\chi}(\mathbf{X}|\mathbf{Y})$, for G the characteristic graph of X with respect to Y, $p(x, y)$, and $f(x, y)$. Let $\Sigma = \{c(\mathbf{x}) : \mathbf{x} \in \mathcal{X}^n\}$, the set of all colors. For every color $\sigma \in \Sigma$ and $\mathbf{y} \in \mathcal{Y}^n$, let $g(\sigma, \mathbf{y}) = f(\bar{\mathbf{x}}, \mathbf{y})$ where $\bar{\mathbf{x}} \in c^{-1}(\sigma) = \{\mathbf{x} : c(\mathbf{x}) = \sigma\}$ with $p(\bar{\mathbf{x}}, \mathbf{y}) > 0$. If no such $\bar{\mathbf{x}}$ exists, g is undefined.

Consider $(\mathbf{x}, \mathbf{y})$ as an instance of the source. Thus, $p(\mathbf{x}, \mathbf{y}) > 0$. Suppose $c(\mathbf{x}) = \sigma$ and $\mathbf{y}$ are available at the decoder where c is defined as above. Then, there is a decoding error when $g(\sigma, \mathbf{y}) \neq f(\mathbf{x}, \mathbf{y})$. This is true only if there exists some $\bar{\mathbf{x}} \in \mathcal{X}^n$ such that $c(\bar{\mathbf{x}}) = \sigma$, $p(\bar{\mathbf{x}}, \mathbf{y}) > 0$, and $f(\bar{\mathbf{x}}, \mathbf{y}) \neq f(\mathbf{x}, \mathbf{y})$. However, because $c(\bar{\mathbf{x}}) = c(\mathbf{x})$, it is true that $(\mathbf{x}, \bar{\mathbf{x}}) \notin E$, where E is the edge set of G^n. Therefore, for all $\bar{\mathbf{y}} \in \mathcal{Y}^n$ with $p(\mathbf{x}, \bar{\mathbf{y}})p(\bar{\mathbf{x}}, \bar{\mathbf{y}}) > 0$, it must be true that $f(\mathbf{x}, \bar{\mathbf{y}}) = f(\bar{\mathbf{x}}, \bar{\mathbf{y}})$. This means $f(\bar{\mathbf{x}}, \mathbf{y}) = f(\mathbf{x}, \mathbf{y})$ for all $\bar{\mathbf{x}} \in c^{-1}(\sigma)$ with $p(\bar{\mathbf{x}}, \mathbf{y}) > 0$. Thus, $g(\sigma, \mathbf{y}) = f(\mathbf{x}, \mathbf{y})$. This shows that the side information along with just the color is sufficient to determine f.

It remains to be seen how to make $c(\mathbf{x})$ and $\mathbf{y}$ available at the decoder. Recall that if a source is encoded at a rate equal to its entropy, it can be recovered to arbitrarily small probability of error at the decoder. Thus, having $\mathbf{Y}$ available at the decoder as side information is the same as encoding $\mathbf{Y}$ at rate greater than $H(Y)$. Recall the Slepian-Wolf Theorem [24] on sources C and Y states that if $R_y > H(Y)$, an encoding with rate $R_c > H(C|Y)$ suffices to recover (C, Y) at the decoder.

We consider our source as $(c(\mathbf{X}), \mathbf{Y})$. Thus, an encoding of rate at least $H(c(\mathbf{X})|\mathbf{Y})$ suffices to recover the functions with arbitrary probability of error. Encoders (and corresponding decoders) exist by the Slepian-Wolf Theorem. Let $\tilde{e} : \Sigma^m \to \{1, \ldots, 2^{mR_c}\}$ be such an encoding with $\tilde{r} : \{1, \ldots, 2^{mR_c}\} \times \mathcal{Y}^{nm} \to \Sigma^m \times \mathcal{Y}^{nm}$ its corresponding decoder. Thus, the idea here is to first color n-blocks of the source. Then, one encodes m-blocks of the colors. The overall rate will be $\frac{1}{m}H(c(\mathbf{X})|\mathbf{Y}))$.

Formally, fix some $n \in \mathbb{N}$. Suppose $\varepsilon > 0$. With an encoder as above, let m be such that

$$\Pr\left[\tilde{r}(\sigma, \mathbf{Y}) \neq (\sigma, \mathbf{Y})\right] < \varepsilon. \tag{4.2}$$

To show achievability, we need to prove that there exists an encoder $e : \mathcal{X}^{nm} \to \{1, \ldots, 2^{mH(c(\mathbf{X})|\mathbf{Y})}\}$ and a decoder $r : \{1, \ldots, 2^{mH(c(\mathbf{X})|Y)}\} \times \mathcal{Y}^{nm} \to \mathcal{Z}^{nm}$ such that the probability of error is also small:

$$\Pr\left[r(e(\mathbf{X}), \mathbf{Y}) \neq f(\mathbf{X}, \mathbf{Y})\right] < \varepsilon. \tag{4.3}$$

To prove this, define our encoder as

$$e(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \tilde{e}(c(\mathbf{x}_1), \ldots, c(\mathbf{x}_m)).$$

Then define the decoder as

$$r(e(\mathbf{x}_1, \ldots, \mathbf{x}_m), (\mathbf{y}_1, \ldots, \mathbf{y}_m)) = (g(c(\mathbf{x}_1), \mathbf{y}_1), \ldots, g(c(\mathbf{x}_m), \mathbf{y}_m)),$$

when $\tilde{r}$ correctly recovers the pair $(c(\mathbf{x}_i), \mathbf{y}_i)$ and is undefined otherwise.

The probability that $\tilde{r}$ fails is less than ε by (4.2). If $\tilde{r}$ does not fail, then, as described earlier, the function will be correctly recovered. Thus, we have shown (4.3).

Therefore, for any n, the rate $\frac{1}{n}H^{\chi}_{G^n}(\mathbf{X}|\mathbf{Y})$ is achievable. Thus,

$$\liminf_{n\to\infty} \frac{1}{n} H^{\chi}_{G^n}(\mathbf{X}|\mathbf{Y}) \geq H_G(X|Y),$$

completing our proof of the lower bound. □

4.1.2 Upper Bound

Next, we prove that the encoding rate required to recover $c(\mathbf{X})$ given $\mathbf{Y}$ is at most $H_G(X|Y)$:

Lemma 4.3.

$$\limsup_{n\to\infty} \frac{1}{n} H^{\chi}_{G^n}(\mathbf{X}|\mathbf{Y}) \leq H_G(X|Y).$$

Proof. Suppose $\varepsilon > 0$, $\delta > 0$. Suppose n is (sufficiently large) such that: (1) Lemma 4.1 applies with some $\varepsilon_1 < 1$, (2) $2^{-n\delta} < \varepsilon_1$, and (3) $n > 2 + \frac{3}{2\varepsilon_1}$.

Let $p(w, x, y)$ be the distribution that achieves the $H_G(X|Y)$ with the Markov property $W - X - Y$. (This is guaranteed to exist by Theorem 2.6.) Denote by $p(w)$, $p(x)$, and $p(y)$ the marginal distributions. For any integer M, define an *M-system* $(\mathbf{W}_1, \ldots, \mathbf{W}_M)$ where each $\mathbf{W}_i$ is drawn independently with distribution $p(\mathbf{w}) = \prod_{i=1}^{n} p(w_i)$.

Our encoding scheme will declare an error if $(\mathbf{x},\mathbf{y}) \notin T_\varepsilon^n$. This means that the encoder will code over ε_1 colorings of the characteristic graphs. By construction, this error happens with probability less than ε_1. Henceforth assume that $(\mathbf{x},\mathbf{y}) \in T_\varepsilon^n$.

Next, our encoder will declare an error when there is no i such that $(\mathbf{W}_i,\mathbf{x}) \in T_\varepsilon^n$. This occurs with probability

$$\begin{aligned}
\Pr[(\mathbf{W}_i,\mathbf{x}) \notin T_\varepsilon^n \forall i] &\overset{(a)}{\leq} \prod_{i=1}^{M} \Pr[(\mathbf{W}_i,\mathbf{x}) \notin T_\varepsilon^n] \\
&\overset{(b)}{=} (1 - \Pr[(\mathbf{W},\mathbf{x}) \in T_\varepsilon^n])^M \\
&\overset{(c)}{\leq} \left(1 - 2^{-n(I(W;X)+\varepsilon_1)}\right)^M \\
&\overset{(d)}{\leq} 2^{-M \cdot 2^{-n(I(W;X)+\varepsilon_1)}}
\end{aligned}$$

where (a) and (b) follow because the $\mathbf{W}_i$ are independent and identically distributed, (c) follows from Lemma 4.1 part 2, and (d) follows because for $\alpha \in [0,1]$, $(1-\alpha)^n \leq 2^{-n\alpha}$. Assuming $M > 2^{n(I(W;X)+\varepsilon_1+\delta)}$,

$$\Pr[(\mathbf{W}_i,\mathbf{x}) \notin T_\varepsilon^n \forall i] \leq 2^{-\delta n} < \varepsilon_1,$$

because n is large enough such that the final inequality holds. Henceforth, fix an M-system $(\mathbf{W}_1,\ldots,\mathbf{W}_M)$ for some $M > 2^{n(I(W;X)+\varepsilon_1+\delta)}$. Further, assume there is some i such that $(\mathbf{W}_i,\mathbf{x}) \in T_\varepsilon^n$.

For each $\mathbf{x}$, let the smallest (or any) such i be denoted as $\hat{c}(\mathbf{x})$. Note that $\hat{c}$ is an ε_1-coloring of the graph G^n. For each $\mathbf{y}$, define:

$$\begin{aligned}
S(\mathbf{y}) &= \{\hat{c}(\mathbf{x}) : (\mathbf{x},\mathbf{y}) \in T_\varepsilon^n\}, \\
Z(\mathbf{y}) &= \{\mathbf{W}_i : (\mathbf{W}_i,\mathbf{y}) \in T_\varepsilon^n\}, \\
s(\mathbf{y}) &= |S(\mathbf{y})|, \\
z(\mathbf{y}) &= |Z(\mathbf{y})|.
\end{aligned}$$

Then, $s(\mathbf{y}) = \sum_{i=1}^{M} \mathbf{1}_{i \in S(\mathbf{y})}$, because our coloring scheme $\hat{c}$ is simply an assignment of the indices of the M-system. Thus, we know

$$E[s(\mathbf{Y})] = \sum_{i=1}^{M} \Pr[i \in S(\mathbf{Y})].$$

Similarly, we get $z(\mathbf{y}) = \sum_{i=1}^{M} \mathbf{1}_{\mathbf{W}_i \in Z(\mathbf{y})}$. Thus,

$$\begin{aligned} E[z(\mathbf{Y})] &= \sum_{i=1}^{M} \Pr[\mathbf{W}_i \in Z(\mathbf{Y})] \\ &\geq \sum_{i=1}^{M} \Pr[\mathbf{W}_i \in Z(\mathbf{Y}) \text{ and } i \in S(\mathbf{Y})] \\ &= \sum_{i=1}^{M} \Pr[i \in S(\mathbf{Y})] P[\mathbf{W}_i \in Z(\mathbf{Y}) | i \in S(\mathbf{Y})] \end{aligned}$$

We know that if $i \in S(\mathbf{Y})$, there is some $\mathbf{x}$ such that $\hat{c}(\mathbf{x}) = i$ and $(\mathbf{x}, \mathbf{Y}) \in T_\varepsilon^n$. For each such $\mathbf{x}$, we must have (by definition of our coloring), $(\mathbf{W}_i, \mathbf{x}) \in T_\varepsilon^n$. For each such $\mathbf{x}$ where $(\mathbf{W}_i, \mathbf{x}) \in T_\varepsilon^n$,

$$\begin{aligned} \Pr[\mathbf{W}_i \in Z(\mathbf{Y}) | i \in S(\mathbf{Y})] &= \Pr[(\mathbf{W}_i, \mathbf{Y}) \in T_\varepsilon^n | (\mathbf{x}, \mathbf{Y}) \in T_\varepsilon^n] \\ &\geq 1 - \varepsilon_1 \end{aligned}$$

by Lemma 4.1 part 4. Thus, we have $E[z(\mathbf{Y})] \geq (1 - \varepsilon_1) E[s(\mathbf{Y})]$. This, along with Jensen's inequality, imply

$$\begin{aligned} E[\log s(\mathbf{Y})] &\leq \log E[s(\mathbf{Y})] \\ &\leq \log E\left[\frac{z(\mathbf{Y})}{1 - \varepsilon_1}\right]. \end{aligned}$$

Finally, (using a Taylor series expansion) we know $\log \frac{1}{1-\varepsilon_1} \leq \varepsilon_2 = 2\varepsilon_1 + \frac{1}{2}$ when $0 < \varepsilon_1 < 1$. Thus,

$$E[\log s(\mathbf{Y})] \leq \log E[z(\mathbf{Y})] + \varepsilon_2, \tag{4.4}$$

We compute

$$\begin{aligned}E[z(\mathbf{Y})] &= \sum_{i=1}^{M} P[(\mathbf{W}_i, \mathbf{Y}) \in T_\varepsilon^n] \\ &= M \cdot P[(\mathbf{W}_i, \mathbf{Y}) \in T_\varepsilon^n]\end{aligned}$$

because the $\mathbf{W}_i$ are i.i.d. Therefore,

$$E[z(\mathbf{Y})] \leq M \cdot 2^{-n(I(W;Y)-\varepsilon_1)} \tag{4.5}$$

by Lemma 4.1 part 3.

By the definition of $S(\mathbf{y})$, we know that determining $\hat{c}$ given $\mathbf{Y} = \mathbf{y}$ requires at most $\log s(\mathbf{y})$ bits. Therefore, we have

$$H(\hat{c}(\mathbf{X})|\mathbf{Y}) \leq E[\log s(\mathbf{Y})]. \tag{4.6}$$

Putting it all together, we have

$$\begin{aligned}H_{G^n}^{\mathcal{X}}(\mathbf{X}|\mathbf{Y}) &\overset{(a)}{\leq} H(\hat{c}(\mathbf{X})|\mathbf{Y}) \\ &\overset{(b)}{\leq} E[\log s(\mathbf{Y})] \\ &\overset{(c)}{\leq} \log E[z(\mathbf{Y})] + \varepsilon_2 \\ &\overset{(d)}{\leq} \log\left(M \cdot 2^{-n(I(W;Y)-\varepsilon_1)}\right) + \varepsilon_2 \\ &\overset{(e)}{=} \log\left(2^{n(I(W;X)-I(W;Y)+2\varepsilon_1+\delta)} + 1\right) + \varepsilon_2 \\ &\overset{(f)}{\leq} n(I(W;X) - I(W;Y) + 2\varepsilon_1 + \delta) + 1 + \varepsilon_2\end{aligned}$$

where (a) follows by definition of the conditional chromatic entropy, (b) follows from inequality (4.6), (c) follows from inequality (4.4), (d) follows from inequality (4.5), (e) follows by setting $M = \lceil 2^{n(I(W;X)+\varepsilon_1+\delta)} \rceil$, and (f) follows because $\log(\alpha + 1) \leq \log(\alpha) + 1$ for $\alpha \geq 1$.

For Markov chains $W - X - Y$,

$$I(W;X) - I(W;Y) = I(W;X|Y).$$

Thus, for our optimal distribution $p(w,x,y)$, we have

$$H^{\mathbf{X}}_{G^n}(\mathbf{X}|\mathbf{Y}) \leq n(H_G(X|Y) + 2\varepsilon_1 + \delta) + 1 + \varepsilon_2$$

Because $n > 2 + \frac{3}{2\varepsilon_1}$, $\frac{1+\varepsilon_2}{n} < \varepsilon_1$. Thus, $\frac{1}{n}H^{\mathbf{X}}_{G^n}(\mathbf{X}|\mathbf{Y}) \leq H_G(X|Y) + 3\varepsilon_1 + \delta$. This completes the proof for the upper bound:

$$\limsup_{n\to\infty} \frac{1}{n} H^{\mathbf{X}}_{G^n}(\mathbf{X}|\mathbf{Y}) \leq H_G(X|Y).$$

□

The lower and upper bounds, Lemmas 4.2 and 4.3, combine to give Theorem 3.2:

$$\lim_{n\to\infty} \frac{1}{n} H^{\mathbf{X}}_{G^n}(\mathbf{X}|\mathbf{Y}) = H_G(X|Y).$$

□

4.2 Distributed Functional Compression

Recall Theorem 3.5 states that the achievable rate region for the distributed functional compression problem, under the Zigzag Condition (Condition 3.4), is the set closure of the set of all rates that can be realized via graph coloring.

We prove this by first showing that if the colors are available at the decoder, the decoder can successfully compute the function. This proves achievability. Next, we show that all valid encodings are ε-colorings of the characteristic graphs (and their powers). This establishes the converse.

4.2.1 Achievability

We first prove the achievability of all rates in the region given in the theorem statement.

Lemma 4.4. *For sufficiently large n and ε-colorings c_x and c_y of G_x^n and G_y^n, respectively, there exists*

$$\hat{f} : c_x(\mathcal{X}^n) \times c_y(\mathcal{Y}^n) \to \mathcal{Z}^n$$

such that $\hat{f}(c_x(\mathbf{x}), c_y(\mathbf{y})) = f(\mathbf{x}, \mathbf{y})$ for all $(\mathbf{x}, \mathbf{y}) \in T_\varepsilon^n$.

Proof. Suppose $(\mathbf{x}, \mathbf{y}) \in T_\varepsilon^n$, and we have colorings c_x and c_y. We proceed by constructing $\hat{f}$. For any two colors $\gamma \in c_x(\mathcal{X}^n)$ and $\sigma \in c_y(\mathcal{Y}^n)$, let $\hat{\mathbf{x}} \in c_x^{-1}(\gamma)$ and $\hat{\mathbf{y}} \in c_y^{-1}(\sigma)$ be any (say the first) pair such that $p(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in T_\varepsilon^n$. Define $\hat{f}(\gamma, \sigma) = f(\hat{\mathbf{x}}, \hat{\mathbf{y}})$. There must be such a pair because certainly $(\mathbf{x}, \mathbf{y})$ qualifies.

To show this function is well-defined on elements in the support, suppose $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ are both in T_ε^n. Suppose further that $c_x(\mathbf{x}_1) = c_x(\mathbf{x}_2)$ and $c_y(\mathbf{y}_1) = c_y(\mathbf{y}_2)$. Then, we know that there is no edge $(\mathbf{x}_1, \mathbf{x}_2)$ in the high probability subgraph of G_x^n or $(\mathbf{y}_1, \mathbf{y}_2)$ in the edge set of the high probability subgraph of G_y^n by definition of graph coloring.

By the Zigzag Condition, there exists some $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ such that $(\tilde{\mathbf{x}}, \mathbf{y}_1)$, $(\tilde{\mathbf{x}}, \mathbf{y}_2)$, $(\mathbf{x}_1, \tilde{\mathbf{y}})$, $(\mathbf{x}_2, \tilde{\mathbf{y}}) \in T_{\frac{\varepsilon}{2}}^n$. We claim that there is no edge between $(\mathbf{x}_i, \tilde{\mathbf{x}})$ or $(\mathbf{y}_i, \tilde{\mathbf{y}})$ for either i. We prove this for $(\mathbf{x}_1, \tilde{\mathbf{x}})$ with the other cases following naturally. Suppose there were an edge. Thus, there is some $\hat{\mathbf{y}}$ such that $f(\mathbf{x}_1, \hat{\mathbf{y}}) \neq f(\tilde{\mathbf{x}}, \hat{\mathbf{y}})$. This implies that $f(x_{1j}, \hat{y}_j) \neq f(\tilde{x}_j, \hat{y}_j)$ for some j. Define $\tilde{\mathbf{y}}'$ as $\tilde{\mathbf{y}}$ in every component but the j-th, where it is $\hat{y}_j$.

We know that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$|\nu_{(\mathbf{x}_1, \tilde{\mathbf{y}})}(x, y) - p(x, y)| \leq \frac{\varepsilon}{2|\mathcal{X}||\mathcal{Y}|}$$

by Definition A.2 of $\frac{\varepsilon}{2}$-typicality. Therefore,

$$|\nu_{(\mathbf{x}_1, \tilde{\mathbf{y}}')}(x, y) - p(x, y)| \leq \frac{\varepsilon}{2|\mathcal{X}||\mathcal{Y}|}$$

for all (x, y) such that $y \neq \hat{y}_j$ and $y \neq \tilde{y}_j$.

Next, we can choose n large enough such that $n > \frac{2|\mathcal{X}||\mathcal{Y}|}{\varepsilon}$. Then, for $y = \hat{y}_j$ or $y = \tilde{y}_j$, the empirical frequency changes by at most $\frac{1}{n}$. Thus, for all (x, y) (including $y = \hat{y}_j$ and $y = \tilde{y}_j$), we have

$$|\nu_{(\mathbf{x}_1, \tilde{\mathbf{y}}')}(x, y) - p(x, y)| \leq \frac{\varepsilon}{2|\mathcal{X}||\mathcal{Y}|} + \frac{1}{n} \leq \frac{\varepsilon}{|\mathcal{X}||\mathcal{Y}|}$$

Thus, $\tilde{\mathbf{y}}'$ is ε-typical with both $\mathbf{x}_1$ and $\mathbf{x}_2$. By construction, $f(\mathbf{x}_1, \tilde{\mathbf{y}}') \neq f(\mathbf{x}_2, \tilde{\mathbf{y}}')$. Therefore, there must be an edge in the high probability subgraph between $(\mathbf{x}_1, \mathbf{x}_2)$, an impossibility. Thus, there is no edge $(\mathbf{x}_1, \tilde{\mathbf{x}})$. The others follow similarly.

Thus, by definition of the graph,

$$f(\mathbf{x}_1, \mathbf{y}_1) = f(\tilde{\mathbf{x}}, \mathbf{y}_1) = f(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = f(\mathbf{x}_2, \tilde{\mathbf{y}}) = f(\mathbf{x}_2, \mathbf{y}_2).$$

Therefore, our function $\hat{f}$ is well-defined and has the desired property. □

Then, Lemma 4.4 implies that we can successfully compute our function f given colors of the characteristic graphs. Thus, if the decoder is given colors, it can look up f based on its table of $\hat{f}$. The question is now of faithfully (with probability of error less than ε) transmitting these colors to the receiver. However, when we consider the colors as sources, we know the achievable rates.

Lemma 4.5. *For any n, ε-colorings c_x and c_y of G_x^n and G_y^n, respectively, the achievable rate region for joint source $(c_x(\mathbf{X}), c_y(\mathbf{Y}))$ is the set of all rates, (R_x^c, R_y^c), satisfying:*

$$\begin{aligned} R_x^c &\geq H(c_x(\mathbf{X})|c_y(\mathbf{Y})), \\ R_y^c &\geq H(c_y(\mathbf{Y})|c_x(\mathbf{X})), \\ R_x^c + R_y^c &\geq H(c_x(\mathbf{X}), c_y(\mathbf{Y})). \end{aligned}$$

Proof. This follows directly from the Slepian-Wolf Theorem [24] for the separate encoding of correlated sources. □

Suppose the probability of decoder error for the decoder guaranteed in Lemma 4.5 is less than $\frac{\varepsilon}{2}$. Then the total error in the coding scheme of first coloring G_x^n and G_y^n, and then encoding those colors to be faithfully decoded at the decoder is upper bounded by the sum of the errors in each stage. Thus, Lemmas 4.4 and 4.5 together to show that the probability that the decoder errs is less than ε for any ε provided large enough n (and block size m on the colors).

Finally, in light of the fact that n source symbols are encoded for each color, the achievable rate region for the problem under the Zigzag Condition is the set of all rates (R_x, R_y) such that

$$\begin{aligned} R_x &\geq \frac{1}{n} H(c_x^n(\mathbf{X}) | c_y^n(\mathbf{Y})), \\ R_y &\geq \frac{1}{n} H(c_y^n(\mathbf{Y}) | c_x^n(\mathbf{X})), \\ R_x + R_y &\geq \frac{1}{n} H(c_x^n(\mathbf{X}), c_y^n(\mathbf{Y})). \end{aligned}$$

where c_x^n and c_y^n are achievable ε-colorings (for any $\varepsilon > 0$). Thus every $(R_x, R_y) \in \mathcal{S}^\varepsilon$ is achievable for all $\varepsilon > 0$. Therefore, every $(R_x, R_y) \in \mathcal{S}$ is achievable.

4.2.2 Converse

Next, we prove that any distributed functional source code with small probability of error induces a coloring.

Suppose $\varepsilon > 0$. Define for all (n, ε),

$$\mathcal{F}_\varepsilon^n = \{\hat{f} : \Pr[\hat{f}(\mathbf{X}, \mathbf{Y}) \neq f(\mathbf{X}, \mathbf{Y})] < \varepsilon\}.$$

This is the set of all functions that equal f to within ε probability of error. (Note that all achievable distributed functional source codes are in $\mathcal{F}_\varepsilon^n$ for large enough n.)

Lemma 4.6. *Consider some function $g : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$. Any distributed functional code that reconstructs g with zero-error (with respect to a distribution $p(x, y)$) induces colorings on the characteristic graphs of X and Y with respect to g, $p(x, y)$, and Y*

and X, respectively.

Proof. Suppose we have encoders e_x and e_y, decoder d, and characteristic graphs G_x^n and G_y^n. Then by definitions, a zero-error reconstruction implies that for any $(\mathbf{x}_1, \mathbf{y}_2)$, $(\mathbf{x}_2, \mathbf{y}_2)$ such that if $p(\mathbf{x}_1, \mathbf{y}_1) > 0$, $p(\mathbf{x}_2, \mathbf{y}_2) > 0$, $e_x(\mathbf{x}_1) = e_x(\mathbf{x}_2)$, and $e_y(\mathbf{y}_1) = e_y(\mathbf{y}_2)$, then

$$f(\mathbf{x}_1, \mathbf{y}_1) = f(\mathbf{x}_2, \mathbf{y}_2) = r(e_x(\mathbf{x}_1), e_y(\mathbf{y}_1)). \tag{4.7}$$

We now show that e_x and e_y are valid colorings of G_x^n and G_y^n. We demonstrate the argument for X. The argument for Y is analogous. We proceed by contradiction. If it were not true, then there must be some edge with both vertices with the same color. In other words, there must exist $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y})$ such that $p(\mathbf{x}_1, \mathbf{y})p(\mathbf{x}_2, \mathbf{y}) > 0$, $e_x(\mathbf{x}_1) = e_x(\mathbf{x}_2)$, and $f(\mathbf{x}_1, \mathbf{y}) \neq f(\mathbf{x}_2, \mathbf{y})$. This is impossible (by taking $\mathbf{y}_1 = \mathbf{y}_2 = \mathbf{y}$ in equation (4.7)). Hence, we have induced colorings of the characteristic graphs. □

We now show that any achievable distributed functional code also induces an ε-coloring of the characteristic graphs.

Lemma 4.7. *All achievable distributed functional codes induce ε-colorings of the characteristic graphs.*

Proof. Let $g(\mathbf{x}, \mathbf{y}) = r(e_x(\mathbf{x}), e_y(\mathbf{y})) \in \mathcal{F}_\varepsilon^n$ be such a code. Then, we know that a zero-error reconstruction (with respect to p) of g induces colorings, e_x and e_y, of the characteristic graphs with respect to g and p by Lemma 4.6. Let the set of all $(\mathbf{x}, \mathbf{y})$ such that $g(\mathbf{x}, \mathbf{y}) \neq f(\mathbf{x}, \mathbf{y})$ be denoted as $\mathcal{C}$. Then because $g \in \mathcal{F}_\varepsilon^n$, we know that $\Pr[\mathcal{C}] < \varepsilon$. Therefore, the functions e_x and e_y restricted to $\mathcal{C}$ are ε-colorings of G_x and G_y (by definition). □

Thus, the Lemma 4.6 and Lemma 4.7 establish Theorem 3.5 in full.

4.2.3 Minimal Joint Rate

Recall Corollary 3.6 states that under the zigzag condition, when there is a unique point achieving the minimum joint rate, it must be $R_x + R_y = H_{G_x}(X) + H_{G_y}(Y)$.

Proof. First, we recall the rate pair $(H_{G_x}(X), H_{G_y}(Y))$ can be achieved via graph colorings. This is true by the achievability result of Theorem 3.5 along with Theorem 3.1, which states that graph colorings can achieve each of $H_{G_x}(X)$ and $H_{G_y}(Y)$. In the achievability proof above, we showed that, under the zigzag condition, any coloring scheme will lead to achievable rates. Therefore, $(H_{G_x}(X), H_{G_y}(Y))$ is in the rate region. (Note, that we have not yet used the uniqueness of the minimum.)

Suppose (R_x, R_y) achieves the minimum joint rate. By Theorem 3.5, this must be in some Slepian-Wolf region for the colors. Because it is a minimum, we must have $R_x + R_y = \frac{1}{n}H(c_x^n(\mathbf{X}), c_y^n(\mathbf{Y}))$. This can be achieved with $R_x = \frac{1}{n}H(c_x^n(\mathbf{X}))$ and $R_y = \frac{1}{n}H(c_y^n(\mathbf{Y})|c_x^n(\mathbf{X}))$ or with $R_x = \frac{1}{n}H(c_x^n(\mathbf{X})|c_y^n(\mathbf{Y}))$ and $R_y = \frac{1}{n}H(c_y^n(\mathbf{Y}))$.

By assumption, there is only one such point, thus we must have $\frac{1}{n}I(c_x^n(\mathbf{X}); c_y^n(\mathbf{Y})) \to 0$ as $n \to \infty$. Thus, the minimal rate is $\frac{1}{n}H(c_x^n(\mathbf{X})) + \frac{1}{n}H(c_y^n(\mathbf{Y})) \to R_x + R_y$ as $n \to \infty$. We know for all n, $H_{G_x}(X) + H_{G_y}(Y) \leq \frac{1}{n}H(c_x^n(\mathbf{X})) + \frac{1}{n}H(c_y^n(\mathbf{Y}))$ by Theorem 3.1.

Therefore, we must have that the minimum achievable joint rate is $H_{G_x}(X) + H_{G_y}(Y)$. □

This corollary implies that minimum entropy colorings have decreasing mutual information as n increases. Thus, the closer we are to the optimum via graph coloring, the less complicated our Slepian-Wolf codes must be. In the limit, because mutual information is zero, each source only needs to code to entropy. Thus, the Slepian-Wolf codes are unnecessary when achieving the minimal joint rate. (Nevertheless, finding the minimum entropy colorings is, again, NP-hard.)

Next in Theorem 3.7, we consider the case when the minimum is not uniquely achievable.

Proof. The joint rate must always satisfy:

$$\begin{aligned}
R_x + R_y =& \frac{1}{n} H(c_x^n(\mathbf{X}), c_y^n(\mathbf{Y})) \\
=& \frac{1}{n} H(c_x^n(\mathbf{X})) + \frac{1}{n} H(c_y^n(\mathbf{Y}) | c_x^n(\mathbf{X})) \\
\geq& H_{G_x}(X) + \frac{1}{n} H(c_y^n(\mathbf{Y}) | \mathbf{X}) \\
\geq& H_{G_x}(X) + H_{G_y}(Y|X)
\end{aligned}$$

The first inequality follows from the Data Processing Inequality on the Markov chain $c_y^n(\mathbf{Y}) - \mathbf{X} - c_x^n(\mathbf{X})$, and the second follows by definition of the conditional graph entropy. Similarly, we get:

$$\begin{aligned}
R_x + R_y =& \frac{1}{n} H(c_x^n(\mathbf{X}), c_y^n(\mathbf{Y})) \\
=& \frac{1}{n} H(c_x^n(\mathbf{X}) | c_y^n(\mathbf{Y})) + \frac{1}{n} H(c_y^n(\mathbf{Y})) \\
\geq& H_{G_x}(X|Y) + H_{G_y}(Y)
\end{aligned}$$

Thus, the difference between the optimal rate $(R_x + R_y)$, and the rate given in Corollary 3.6 is bounded by the following two inequalities:

$$\begin{aligned}
\left[H_{G_x}(X) + H_{G_y}(Y)\right] - [R_x + R_y] \leq H_{G_x}(X) - H_{G_x}(X|Y) \\
\left[H_{G_x}(X) + H_{G_y}(Y)\right] - [R_x + R_y] \leq H_{G_y}(Y) - H_{G_y}(Y|X)
\end{aligned}$$

□

4.3 Functional Rate Distortion

In this section, we prove Theorem 3.8 and Corollary 3.9 for the functional rate distortion problem.

We restate Theorem 3.8 for completeness:

$$R(D) = \min_{\hat{f} \in \mathcal{F}(D)} H_{G(\hat{f})}(X|Y)$$

Proof. We prove that the given characterization is valid by first showing the rate $H_{G(\hat{f})}(X|Y)$ is achievable for any $\hat{f} \in \mathcal{F}(D)$, and next showing that every achievability scheme must be in $\mathcal{F}(D)$.

By Orlitsky and Roche, we know that the rate $H_{G(\hat{f})}(X|Y)$ is sufficient to determine the function $\hat{f}(\mathbf{X}, \mathbf{Y})$ at the receiver. By definition,

$$\lim_{n \to \infty} E[d(f(\mathbf{X}, \mathbf{Y}), \hat{f}(\mathbf{X}, \mathbf{Y}))] \leq D.$$

Thus, the rate $H_{G(\hat{f})}(X|Y)$ is achievable.

Next, suppose we have any achievable rate R, with corresponding *sequence* of encoding and decoding functions e_1^n and e_2^n respectively. Then the function $\hat{f}(\cdot, \cdot) = e_2^n(e_1^n(\cdot), \cdot)$ is a function $\hat{f} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathcal{Z}^n$ with the property (by achievability) that $\lim_{n \to \infty} E[d(f(\mathbf{X}, \mathbf{Y}), \hat{f}(\mathbf{X}, \mathbf{Y}))] \leq D$ (again because as $n \to \infty$, ε is driven to 0). Thus, $\hat{f} \in \mathcal{F}(D)$, completing the proof of Theorem 3.8. □

Next we prove Corollary 3.9, which states that $H_{G^D}(X|Y)$ is an achievable rate. We show this by demonstrating that any distribution on (W, X, Y) satisfying $W - X - Y$ and $X \in W \in \Gamma(G^D)$ also satisfies the Yamamoto requirement (i.e. is also in $\mathcal{P}(D)$).

Proof. Suppose $p(w, x, y)$ is such that $p(w, x, y) = p(w|x)p(x, y)$, or $W - X - Y$ is a Markov chain. Further suppose that $X \in W \in \Gamma(G^D)$. Then define $g(w, y) = f(x^*, y)$ where x^* is any (say, the first) $x \in w$ with $p(x^*, y) > 0$. This is well-defined because the nonexistence of x such that $p(x, y) > 0$ is a zero probability event, and $x \in w$ occurs with probability one by assumption.

Further, because w is an independent set, for any $x_1, x_2 \in w$, one must have $(x_1, x_2) \notin E^D$, the edge set of G^D. By definition of G^D, this means that for all $y \in \mathcal{Y}$ such that $p(x_1, y)p(x_2, y) > 0$, it must be the case that $d(f(x_1, y), f(x_2, y)) \leq D$.

Therefore,

$$E[d(f(X,Y),g(W,Y))] = E[d(f(X,Y),f(X^*,Y))] \leq D$$

because both $X \in W$ and $X^* \in W$ are probability 1 events.

We have shown that for a given distribution achieving the conditional graph entropy, there is a function g on $\mathcal{W} \times \mathcal{Y}$ that has expected distortion less than D. In other words, any distribution satisfying $W - X - Y$ and $X \in W \in \Gamma(G^D)$ is also in $\mathcal{P}(D)$. Further, any such distribution can be associated with a coding scheme, by Orlitsky and Roche's work [20], that achieves the rate $I(W;X|Y)$. When the distribution is chosen such that $I(W;X|Y)$ is minimized, this is by definition equal to $H_{G^D}(X|Y)$. Thus, the rate $H_{G^D}(X|Y)$ is achievable, proving Corollary 3.9 and providing a single-letter upper bound for $R(D)$. □

Chapter 5

Applications

In this chapter, we present some applications of the work presented in the previous chapters.

5.1 Blue Force Tracking

In this section, we consider a sensor network scenario in which there are several sources communicating with some central receiver. This receiver wishes to learn some function of the sources.

Specifically, we consider Blue Force Tracking, which is a GPS system used by the U.S. Armed Forces to track friendly and enemy movements. Sometimes the nodes in the system communicate with each other, and sometimes they communicate with some central receiver, such as a UAV, which is the case considered here.

We present preliminary experimental results for the algorithm given for the distributed functional compression. We obtained tracking data from SRI International.[1] This data represents GPS location data. It includes information on various mobiles, including latitude and longitude coordinates. We ignored the other information (e.g., altitude) for the purpose of this simulation. See Figure 5-1. The blue represents the trajectory through time of one vehicle, and the red curve represents the trajectory of

[1] We thank Dr. Aaron Heller for providing the data, available at: `http://www.ai.sri.com/ajh/isat`.

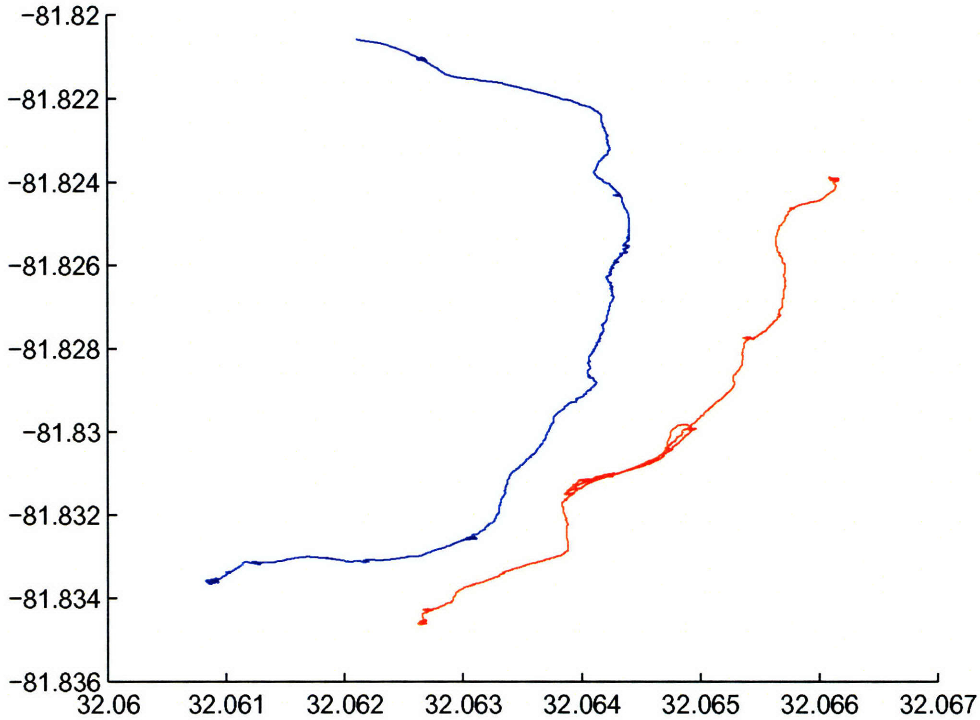

Figure 5-1: Blue Force Tracking Data.

the other.

We focused on these two mobiles, our sources. We assume that our sources are the positional differences (i.e. Δ-encoding), $\mathbf{X}_1$ and $\mathbf{X}_2$, where each is actually a pair, $(\Delta X_{1,\text{LAT}}, \Delta X_{1,\text{LON}})$ and $(\Delta X_{2,\text{LAT}}, \Delta X_{2,\text{LON}})$, of the latitude and longitude data. The use of Δ-encoding assumes that the positional differences form a Markov chain, a common assumption. Our goal is to test the hypothesis that there can be significant encoding gains with even very simple coloring schemes when a function and not the full sources need to be recovered. We consider three relative proximity functions[2] for

[2] We would have liked to use a true proximity function, but then we could not form a valid comparison because our uncolored rate would be in terms of Δ-encoding, but our coloring would necessarily have to be in terms of an encoding of the true position. Therefore, we examine functions that measure how far two mobiles moved towards or away from each other relative to their previous distance, a kind of distance of positional differences.

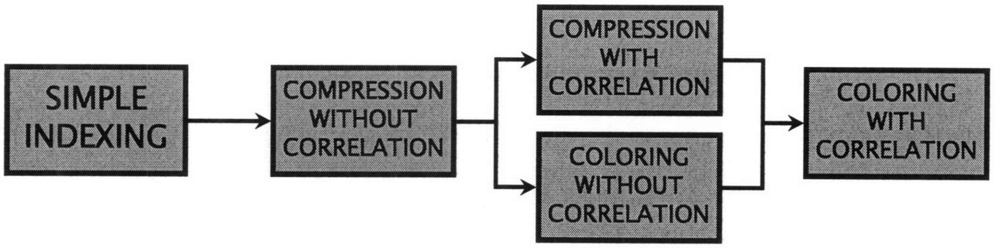
SIMPLE INDEXING
COMPRESSION WITHOUT CORRELATION
COMPRESSION WITH CORRELATION
COLORING WITHOUT CORRELATION
COLORING WITH CORRELATION

Figure 5-2: Levels of compression.

our analysis:

$$f_{\text{LAT}}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{1}_{|\Delta X_{1,\text{LAT}} - \Delta X_{2,\text{LAT}}| < Z},$$
$$f_{\text{LON}}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{1}_{|\Delta X_{1,\text{LON}} - \Delta X_{2,\text{LON}}| < Z},$$
$$f(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{1}_{\sqrt{(\Delta X_{1,\text{LAT}} - \Delta X_{2,\text{LAT}})^2 + (\Delta X_{1,\text{LON}} - \Delta X_{2,\text{LON}})^2} < Z}.$$

Thus, the functions are 1 when the sources change their relative position by less than Z (along some axis or both), and 0 otherwise. To compare the results of our analysis with current methods, we consider the joint rate $R_1 + R_2$ where $\mathbf{X}_1$ is communicated at rate R_1 and $\mathbf{X}_2$ is communicated at rate R_2. There are several means of rate reduction summarized in Figure 5-2.

First, the most common (in practice) means of communication is to actually communicate the full index of the value. This means that if $\mathbf{X}_1$ takes M_1 possible values and $\mathbf{X}_2$ takes M_2 possible values, each source will communicate those values using $\log M_1$ bits and $\log M_2$ bits, respectively. Thus, the joint rate is $\log M_1 M_2$. This is clearly inefficient.

Second, we can immediately reduce the rate by compressing each source before communication. Therefore the rate for $\mathbf{X}_1$ would be $H_1 = H(\mathbf{X}_1)$, and the rate for $\mathbf{X}_2$ would be $H_2 = H(\mathbf{X}_2)$. The joint rate would be $H_1 + H_2$. This is strictly better than the first method unless the sources are uniformly distributed.

Third, we can further reduce the rate using correlation, or Slepian-Wolf, encoding.

We could use any of the techniques already developed to achieve near optimal rates, such as DISCUS codes [21] and source-splitting [8]. The joint rate would be $H_{12} = H(\mathbf{X}_1, \mathbf{X}_2)$. This is strictly better than the second method unless the sources are independent.

Fourth, we could use our coloring techniques from Section 3.2. If we considered each source communicating its color to the central receiver, the joint rate will be $H_1^\chi + H_2^\chi = H(c_1(\mathbf{X}_1)) + H(c_2(\mathbf{X}_2))$. This may not be better than the above third method, though it certainly will be for independent sources. It will always be better than the second method.

Finally, we could use Slepian-Wolf coding over the colors to achieve a joint rate of $H_{12}^\chi = H(c_1(\mathbf{X}_2), c_2(\mathbf{X}_2))$, which will be strictly better than the third method unless c_1 and c_2 are both injective and strictly better than the fourth method unless $c_1(\mathbf{X}_1)$ and $c_2(\mathbf{X}_2)$ are independent. Thus, the rate relations are as follows:

$$\log M_1 M_2 \geq H_1 + H_2 \geq \begin{array}{c} H_{12} \\ \\ H_1^\chi + H_2^\chi \end{array} \geq H_{12}^\chi .$$

In our simulations, we test various values of Z to see how the likelihood $p = P[f_{\text{LON}}(\mathbf{X}_1, \mathbf{X}_2) = 1]$, which changes with Z affects the rate reduction.[3] Intuitively, we expect that as p becomes more extreme and approaches either 0 or 1, the rate reduction will become more extreme and approach 100%. (Because if $f_{\text{LON}} = 1$ or $f_{\text{LON}} = 0$ with probability 1, there is nothing to communicate and the rate required is 0. This is shown in Figure 5-3 where we plot the empirical probability $p = p(Z)$ versus the rate H_{12}^χ.

We expect it would be more symmetric about 1/2 if we used optimal encoding schemes. However, we are only considering $G = G^1$ (no power graphs) when coloring, as well as a quite simple coloring algorithm. Had we used power graphs, our rate gains would be higher, though the computational complexity would increase exponentially with n. Our coloring algorithm was a simple greedy algorithm that did not use any of the probability information nor was it an ε-coloring. We expect better gains with

[3] We only show our results for f_{LON} for brevity. The intuition remains for f and f_{LAT}.

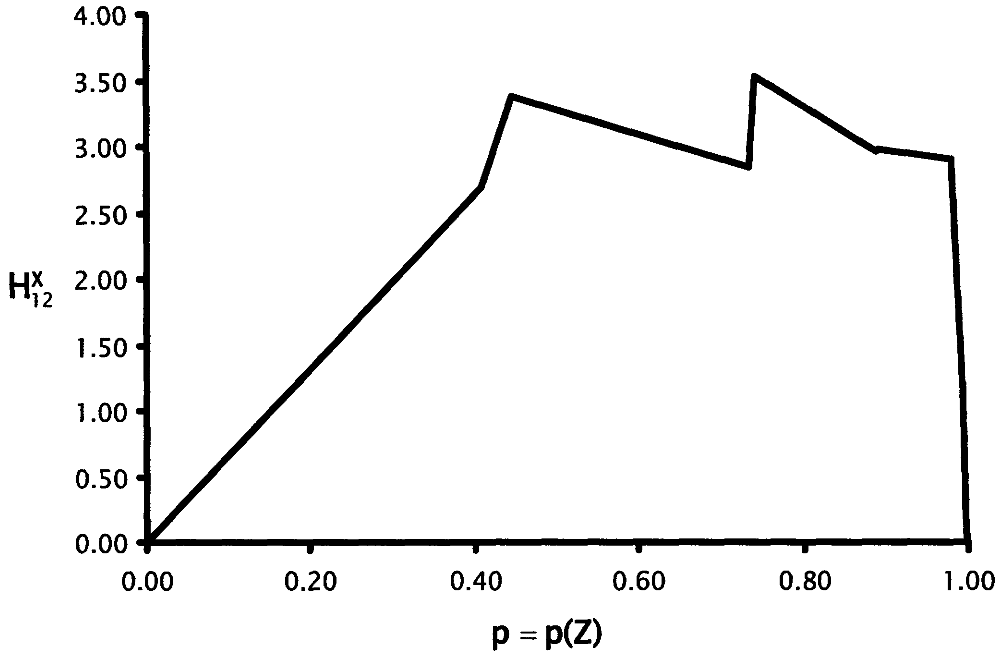

Figure 5-3: Empirical probability versus rate.

more advanced graph coloring schemes.

In Table 5.1, we present the rate results for the various stages of compression in Figure 5-2. All units are in bits. In the table, we use the values of Z that provide the lowest rate reductions; in other words, we use the worst case rates by testing various Z as in Figure 5-3. The percentage next to each number shows the percentage decrease in rate. Thus, for the first column, we see 0%, and in the second we see

$$1 - \frac{7.93}{9.32} \approx 0.149 \approx 15\%.$$

We can see that the sources are close to independent, as H_{12} is only slightly smaller than $H_1 + H_2$. Therefore, there is not much gain when considering the correlation between the sources. Nevertheless, the coloring provides a great deal of coding gain. For the simpler f_{LAT} and f_{LON}, the rate has been reduced almost threefold.

This provides evidence that our techniques can indeed lead to large rate gains. For the more simple functions, the rate has been reduced over 60%. Further, considering

Table 5.1: Empirical results for our encoding scheme.

Rates	$\log M_1M_2$	$H_1 + H_2$	H_{12}	$H_1^X + H_2^X$	H_{12}^X
$f(\mathbf{X}_1, \mathbf{X}_2)$	9.32 (0%)	7.93 (15%)	7.78 (17%)	5.44 (42%)	5.29 (43%)
$f_{\mathrm{LAT}}(\mathbf{X}_1, \mathbf{X}_2)$	9.32 (0%)	7.93 (15%)	7.78 (17%)	3.38 (64%)	3.37 (64%)
$f_{\mathrm{LON}}(\mathbf{X}_1, \mathbf{X}_2)$	9.32 (0%)	7.93 (15%)	7.78 (17%)	3.55 (62%)	3.53 (62%)

that the indices are often sent without compression, it is worth noting that even simple compression is 15% better.

5.2 Privacy

Privacy is a difficult notion to define. Samuel Warren and Louis Brandeis famously wrote privacy as "the right to be let alone" [26]. Nevertheless, the existence of vast electronic troves of data such as census and medical databases requires a definition of privacy that can be interpretted by a computer. Tore Dalenius [10] maintained that "nothing should be learnable from the database that cannot be learned without access to the database" [14].

These semantic definition do not lend themselves to implementable schemes to achieve privacy in databases. In fact, it has been proven that Dalenius's scheme is impracticle in the face of outside information [14]. There have been few systematic attempts to quantify privacy for databases.

Consider databases $D = \{X^i\}_{i=1}^m$, where m is the size of the population and each $X^i \in \mathcal{X}$, some finite set. Thus, one way to treat databases is as a realization of a variable $D \in \mathcal{D}$ where $\mathcal{D}$ is the collection of all possible datasets, and D is drawn according to some population distribution. Some database manager (like the government) is in control of the data. The manager wants the data to provide utility to researchers who would, say, conduct research on the efficacy of public policies. The public, on the other hand, wants to protect their data.

There is a fundamental tension between the privacy and the utility of data in a database. The public can achieve perfect privacy by providing no data. The public can allow perfect utility by providing complete data. The goal is to quantify the tradeoff. We consider two approaches, both of which quantify in different ways the amount of privacy any given function will give.

First, given some desired level of privacy, what statistics should be allowed. In other words, some quantum level of privacy determines the set of all statistics, or functions, that can be made available to the general public by the benevolent central body. This is the approach taken by Cynthia Dwork [14] and described in the next section.

Second, given some statistics that must be computable, encode the data so as to guarantee that only those statistics are computable. Conversely, given some undesired statistics, encode the data so as to make computation of those statistics impossible. Further, this is done at the citizen-level so each citizen only gives up as much information as required, and no more. In that sense, the central body can be less than trustworthy for this scheme to work. Functional compression is applicable here, and this is the approach taken by us in Section 5.2.2.

5.2.1 Dwork's Differential Privacy

Dwork defines the differential privacy of randomized functions.[4]

Definition 5.1 (Dwork's Differential Privacy [14]). *A randomized function $\mathcal{K}$ gives ε-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq \mathcal{K}(\mathcal{D})$,*

$$P[\mathcal{K}(D_1) \in S] \leq \exp(\varepsilon) \times P[\mathcal{K}(D_2) \in S].$$

We modify the definition slightly to shift the emphasis from the function to the quantum of privacy.

[4] Dwork, her co-authors, and others have done extensive work on database privacy beyond the papers of immediate interest. See `http://research.microsoft.com/research/sv/DatabasePrivacy/` for a list of papers on the subject by this group.

Definition 5.2. *The differential privacy of any randomized function $\mathcal{K}$ is:*

$$\varepsilon = \min_{\delta} \{\mathcal{K} \text{ gives } \delta\text{-differential privacy}\}.$$

The main idea of differential privacy is to ensure that the removal of one's own data from the data set does not make any output from the function more or less likely. Consider first what the definition means for deterministic functions K.

Example 5.3. *For a deterministic function, K, $P[K(D_1) \in S] \leq \exp(\varepsilon) \times P[K(D_2) \in S]$ is equivalent to saying $K(D_1) = K(D_2)$ for any $\varepsilon \geq 0$. This is because we can consider $S = \{K(D_1)\}$. This means $P[K(D_1) \in S] = 1$. Because for $\varepsilon \geq 0$, $\exp(\varepsilon) \geq 1$, this means $P[K(D_2) \in S] \geq 1$, which means $K(D_1) = K(D_2)$.*

Thus, ε-differential privacy, or just differential privacy because ε is irrelevant here, implies that for any D_1 and D_2 differing on at most one element, $K(D_1) = K(D_2)$. This is similar to our Zigzag Condition, but is considerably more strict. □

Most statistics of interest are not randomized. Thus, Dwork et al. [15] came up with a "sanitizing" algorithm that would take any deterministic function and ensure any level of privacy by randomizing it in an appropriate way.

Their sanitization is a randomized function $\mathcal{K}_f(D)$ where $D \in \mathcal{D}$ such that $P[\mathcal{K}_f(D) = a] \propto \exp(-|f(D) - a|/\sigma)$. The idea is for any query f on the database D, return a realization of $\mathcal{K}_f(D)$. Dwork et al. prove that $\mathcal{K}_f$ has differential privacy at most $\Delta f/\sigma$ where Δf is the $\mathcal{L}_1$-sensitivity of a function f defined as $\Delta f = \max_{D_1,D_2} |f(D_1) - f(D_2)|$ for all D_1 and D_2 differing on at most one element. Consider a simple example.

Example 5.4. *Consider some $f : \{00, 01, 10, 11\}^M \rightarrow \{0, 1\}$, where f is the sum of its arguments modulo 2. Then, $\mathcal{K}_f(D)$ is drawn according to $P[\mathcal{K}_f(D) = a] \propto \exp(-|f(D) - a|/\sigma)$. This provides $1/\sigma$-differential privacy because $\Delta f = 1$. In order to provide maximum privacy, we let σ go to ∞, making the data useless. In order to provide minimal privacy, we let σ go to 0, and return $f(D)$ exactly.* □

Thus, this explicitly shows the tradeoff between privacy and utility. If a query function compromises too much privacy, then adding the random noise to the value will decrease the utility of that value such that privacy is maintained at a fixed level.

5.2.2 A New Definition of Privacy

We approach the issue of privacy from a different angle. Information theoretically, everybody in the population is holding some bits: their information. The database manager (e.g., the U.S. Census) mandates that they need to be able to compute some statistics for various reasons. We want to quantify how many bits the public must reveal in order for the manager to compute its statistics (and make those statistics available). We begin with an assumption.

Assumption 5.5. *Assume that Theorem 3.5 applies to M sources under an extension of the Zigzag Condition. Further, assume the extended Zigzag Condition applies to the databases we consider.*

We have not proven a completely tight bound for the distributed functional compression problem in Section 3.2, and certainly have not extended it to M sources. Nevertheless, we proceed with a proof of concept. If Assumption 5.5 is false, the results would still provide a one-sided bound. This means that the bits revealed by the population would be necessary, but perhaps not sufficient, to compute the given statistics. Therefore, for the purposes of this section, assume there is some coloring solution for this problem.

Consider the database $D = \{X_i\}_{i=1}^{M}$ as M sources drawn from the same population distribution $p(x)$. These are drawn independently, so it is reasonable to assume any extension of the Zigzag Condition applies. Then, by Assumption 5.5, there is some rate-optimal encoding of this information through a coloring scheme. Further, because every source is drawn according to the same distribution, we need only consider the coloring of a single graph G. Call this rate-optimal coloring c.

We propose the following scheme for computation of a function f:

1. The database manager determines the population distribution. This could be

done by looking at historical data for the population.[5] The manager could also achieve this goal by sampling the population relegating a subset of the population to have zero privacy by revealing all their bits.

2. The manager reveals a public coloring scheme. Because each member of the population is drawn from the same distribution given in the above step, a single public coloring scheme can be made available. This coloring scheme would be rate-optimal at $R = H^{\chi}_{G(f)}(X)$. Further, any atypical member of the population would randomly (proportionate to $p(c(x))$) choose a color to report.

3. Members of the population reveal their colors to the manager. They each do this at rate $H^{\chi}_{G(f)}(X)$.

This scheme has several nice properties. First, any atypical member of the population does not reveal their data. In that sense, no "singleton" could ever be discovered by any function. Such a scheme is also privacy-optimal in the sense that members of the population reveal as few of their bits as possible to the manager, who may be untrustworthy, to compute the statistics.

This leads directly to a different quantitative definition of privacy.

Definition 5.6. *The privacy allowed by a function f is*

$$\frac{H(X) - H^{\chi}_{G(f)}(X)}{H(X)}.$$

We can consider this as a percentage. Thus, zero privacy would mean $R = H(X)$ and total (100%) privacy would be $R = 0$. Consider the following example.

Example 5.7. *Consider again $f : \{00, 01, 10, 11\}^M \to \{0, 1\}$, where f is the sum of its arguments modulo 2. We consider randomness on the database. Suppose the data is drawn according to $p(x) = 1/3 - \varepsilon$ for $x \in \{01, 10, 11\}$ and $p(00) = \varepsilon$. Considering ε-colorings on $G(f)$, the optimal coloring is $c(01) = c(11) = c_1$ and $c(10) = c_2$. Then, clearly $H^{\chi}_{G(f)}(X) = \frac{1}{3}\ln 3 + \frac{2}{3}\ln\frac{3}{2}$. For $\varepsilon \in (0, 0.1)$, the privacy is $H(X) - H^{\chi}_{G(f)}(X) \in (42\%, 49\%)$.*

[5] This fails only if there are large changes in the population over time.

This approach necessarily places utility of the data over the privacy of the individuals and is, in a sense, dual to the Dwork approach. Whereas the Dwork approach determines the set of allowable functions for a given level of privacy, our approach determines the level of privacy for a given set of functions.[6]

[6] We do not require the functions to be cardinal, so more concrete comparisons remain difficult.

Chapter 6

Conclusion

This thesis considered the problem of coding for computing in new contexts. We considered the functional compression problem with side information and gave novel solutions for both the zero and nonzero distortion cases. These algorithms gave an explicit decoupling of the computing from the correlation between the sources as a graph coloring problem. We proved that this decoupling is rate optimal. We extended this encoding scheme to the distributed functional compression with zero distortion. We gave an inner bound to the rate region, and gave the conditions under which the decoupling is optimal in the distributed case.

We never considered the nonzero distortion distributed functional compression problem mainly because even the case of $f(x,y) = (x,y)$ is unsolved. Nevertheless, it is our hope that the methods discussed in this thesis will yield new results for the more general problem.

All of our results concern two sources. An extension of these results to M sources seems plausible. However, the graph constructs used rely heavily on the two source structure and would need to be modified to deal with M sources. We leave that to future work.

Finally, we examined the applicability of our results to two scenarios. For the Blue Force Tracking scenario, we saw that even simple coloring schemes yielded large compression gains (64%). For the database scenario, we presented a new notion of privacy using functional compression. Our privacy can be thought of as dual to existing notions of database privacy.

In summary, this thesis is about modeling the distillation of relevant information from disparate sources. We hope the work presented herein serves as a step towards more research in this area.

Appendix A

Information Theory Fundamentals

This appendix states some important well-known results from information theory without proof. These are provided as a backdrop to the results of this paper. Throughout this section, consider discrete memoryless sources. We recall Shannon's fundamental results on data compression.

Theorem A.1 (Shannon, 1948 [23]). *Consider discrete memoryless sources* $\mathbf{X} \in \mathcal{X}^n$ *and* $\mathbf{Y} \in \mathcal{Y}^n$*. For any* $\varepsilon > 0$*, there exists an* N *and a coding scheme* (e_n, d_n) *for some* $n > N$ *where* $e_n : \mathcal{X}^n \rightarrow \left\{1, \ldots, 2^{n(R+\varepsilon)}\right\}$ *and* $d_n : \left\{1, \ldots, 2^{n(R_x+\varepsilon)}\right\} \times \mathcal{Y}^n \rightarrow \mathcal{X}^n$ *such that* $P\left[\mathbf{X} \neq d_n(e_n(\mathbf{X}))\right] < \varepsilon$*, as long as* $R_x \geq H(X|Y)$*.*

In other words, this theorem states that one can recover source X given source Y at rate $H(X|Y)$. If Y is independent of X, the rate required is $H(X)$. This provides intuition about the concept of entropy. This concept of entropy is not explanatory, however, when the question is changed so that recovery of a function of the source is desired. That requires a new definition of entropy to be introduced later in Definition 2.3.

This source coding theorem above is the fundamental basis for much of the work in this thesis. The arguments used to prove it are based in random coding and typicality. We next define typicality and give some standard results. These can be found in Cover and Thomas, 1991 [9, p. 358].

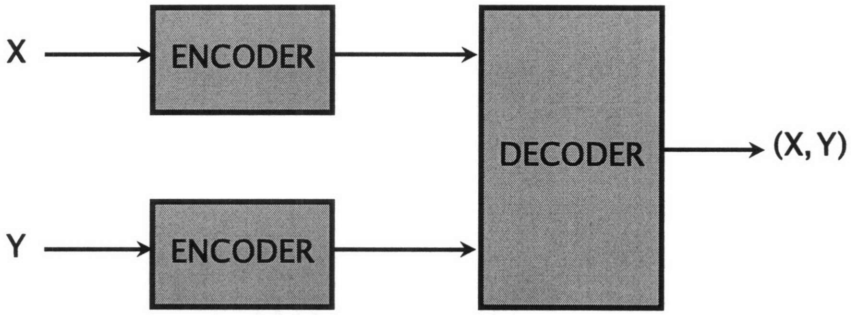

Figure A-1: The Slepian-Wolf problem.

Definition A.2. *For any n, and any sequence $\mathbf{x}$ of length n, define the empirical frequency of the symbol x in $\mathbf{x}$ as*

$$\nu_{\mathbf{x}}(x) = \frac{|\{i : x_i = x\}|}{n}.$$

Then a sequence $\mathbf{x}$ is ε-strongly typical for $\varepsilon > 0$ if for all $x \in \mathcal{X}$ with $p(x) > 0$,

$$|\nu_{\mathbf{x}}(x) - p(x)| \leq \frac{\varepsilon}{|\mathcal{X}|},$$

and for all $x \in \mathcal{X}$ with $p(x) = 0$, $\nu_{\mathbf{x}}(x) = 0$. The set of all such ε-strongly typical sequences, called the ε-typical set will be denoted as $T_\varepsilon^n(X)$, or T_ε^n when the variables are clear from context. A similar definition naturally extends for the case of joint variables.

Next, we discuss the result of Slepian and Wolf which gives the achievable rate region for the distributed compression problem.

The problem is that of separately encoding $\mathbf{X}$ and $\mathbf{Y}$ such that both can be recovered at the receiver. See Figure A-1. The goal in this problem is to determine the set closure of the set of all rate pairs (R_x, R_y) that are achievable in the sense that

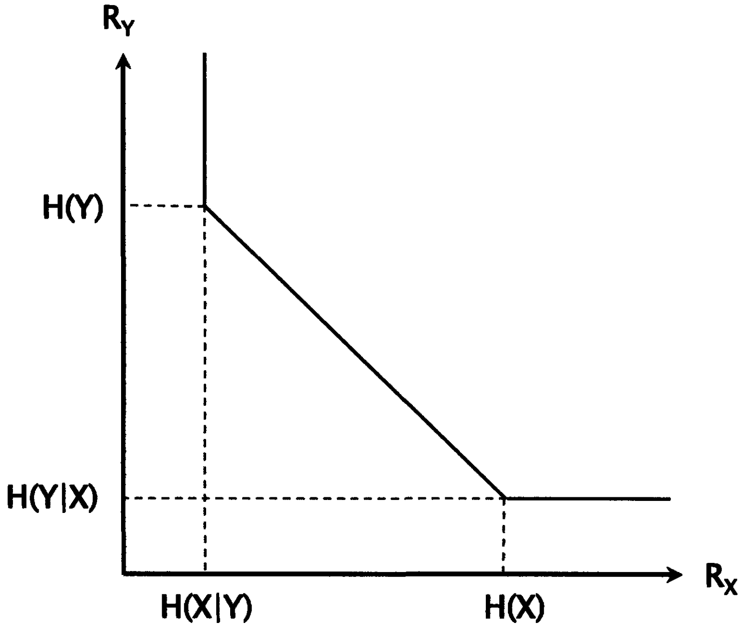

Figure A-2: The Slepian-Wolf rate region.

there are some encodings (and a decoding) at those rates such that the probability of error in decoding is neglibly small.

Theorem A.3 (Slepian and Wolf, 1973 [24]). *The achievable rate region for the distributed compression problem is the set of all rates rates* (R_x, R_y) *where*

$$\begin{aligned} R_x &\geq H(X|Y) \\ R_y &\geq H(Y|X) \\ R_x + R_y &\geq H(X,Y). \end{aligned}$$

A plot of the rate region is provided in Figure A-2. We consider a similar problem where the goal is to discover the achievable rate region when the receiver computes a general $f(X,Y)$, not just $f(X,Y) = (X,Y)$. We provide a rate region for this problem. See Figure 3-3.

Wyner and Ziv considered the problem depicted in Figure A-3. They wished to

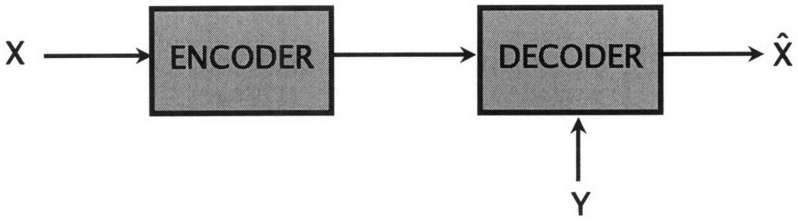

Figure A-3: The Wyner-Ziv problem.

determine the rate distortion function, i.e. the rate R_x required for the decoder to recover $\mathbf{X}$ at the receiver. We will consider a similar problem where the goal is to recover a function of the source and side information within a distortion criterion. We provide both the rate distortion function for both zero and non-zero distortion. The result of Wyner and Ziv is given below [28].

Theorem A.4 (Wyner-Ziv Rate Distortion Function with Side Information [28]). *Given some distortion function $d : \mathcal{X}^2 \to \mathbb{R}^+$, the rate distortion function is:*

$$R(D) = \min_{p(w|x)} \min_{g} I(X; W|Y)$$

where $W - X - Y$ forms a Markov chain and $g : \mathcal{Y} \times \mathcal{W} \to \mathcal{X}$ is such that

$$E[d(X, g(Y, W))] \leq D.$$

Bibliography

[1] Rudolf F. Ahlswede and János Körner. Source coding with side information and a converse for degraded broadcast channels. *IEEE Trans. Inform. Theory*, 21(6):629–637, November 1975.

[2] Noga Alon and Alon Orlitsky. Source coding and graph entropies. *IEEE Trans. Inform. Theory*, 42(5):1329–1339, September 1996.

[3] João Barros and Sergio Servetto. On the rate-distortion region for separate encoding of correlated sources. In *IEEE Symposium on Information Theory (ISIT)*, page 171, Yokohama, Japan, 2003.

[4] Toby Berger and Raymond W. Yeung. Multiterminal source encoding with one distortion criterion. *IEEE Trans. Inform. Theory*, 35(2):228–236, March 1989.

[5] G. Campers, O. Henkes, and J. P. Leclerq. Graph coloring heuristics: A survey, some new propositions and computational experiences on random and "Leighton's" graphs. In G. K. Rand, editor, *Operations Research '87*, pages 917–932. 1988.

[6] Jean Cardinal, Samuel Fiorini, and Gilles Van Assche. On minimum entropy graph colorings. In *ISIT 2004*, page 43, June–July 2004.

[7] Shuchi Chawla, Cynthia Dwork, Frank McSherry, and Adam Smith. Toward privacy in public databases. In *2nd Theory of Cryptography Conference–TCC 2005*, pages 363–385, 2005.

[8] Todd P. Coleman, Anna H. Lee, Muriel Médard, and Michelle Effros. Low-complexity approaches to Slepian-Wolf near-lossless distributed data compression. *IEEE Trans. Inform. Theory*, 52(8):3546–3561, August 2006.

[9] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[10] Tore Dalenius. Towards a methodology for statistical disclosure control. *Statistik Tidskrift 15*, pages 429–444, 1977.

[11] Vishal Doshi, Devavrat Shah, and Muriel Médard. Source coding with distortion through graph coloring. In *2007 International Symposium on Information Theory*, Nice, France, June 2007.

[12] Vishal Doshi, Devavrat Shah, Muriel Médard, and Sidharth Jaggi. Graph coloring and conditional graph entropy. In *2006 Asilomar Conference on Signals, Systems, and Computers*, pages 2137–2141, Asilomar, CA, October-November 2006.

[13] Vishal Doshi, Devavrat Shah, Muriel Médard, and Sidharth Jaggi. Distributed functional compression through graph coloring. In *2007 Data Compression Conference*, pages 93–102, Snowbird, UT, March 2007.

[14] Cynthia Dwork. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming-ICALP 2006, Part II*, pages 1–12, 2006. Invited paper.

[15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *3rd Theory of Cryptography Conference–TCC 2006*, pages 265–284, 2006.

[16] Hanying Feng, Michelle Effros, and Serap Savari. Functional source coding for networks with receiver side information. In *Proceedings of the Allerton Conference on Communication, Control, and Computing*, pages 1419–1427, September 2004.

[17] János Körner. Coding of an information source having ambiguous alphabet and the entropy of graphs. In *6th Prague Conference on Information Theory*, pages 411–425, 1973.

[18] János Körner and Katalin Marton. How to encode the modulo-two sum of binary sources. *IEEE Trans. Inform. Theory*, 25(2):219–221, March 1979.

[19] Colin McDiarmid. Colourings of random graphs. In Roy Nelson and Robin J. Wilson, editors, *Graph Colourings*, Pitman Research Notes in Mathematics Series, pages 79–86. Longman Scientific & Technical, 1990.

[20] Alon Orlitsky and James R. Roche. Coding for computing. *IEEE Trans. Inform. Theory*, 47(3):903–917, March 2001.

[21] S. Sandeep Pradhan and Kannan Ramchandran. Distributed source coding using syndromes (DISCUS): design and construction. *IEEE Trans. Inform. Theory*, 49(3):626–643, March 2003.

[22] Claude E. Shannon. The zero error capacity of a noisy channel. *IEEE Trans. Inform. Theory*, 2(3):8–19, September 1956.

[23] Claude E. Shannon and Warren Weaver. *A Mathematical Theory of Communication.* University of Illinois Press, Champaign, IL, USA, 1963.

[24] David Slepian and Jack K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inform. Theory*, 19(4):471–480, July 1973.

[25] Aaron B. Wagner, Saurabha Tavildar, and Pramod Viswanath. Rate region of the quadratic gaussian two-terminal source-coding problem. *IEEE Trans. Inform. Theory.* Submitted February 2006.

[26] Samuel D. Warren and Louis D. Brandeis. The right to privacy. *Harvard Law Review*, 4(5):193–220, December 1890.

[27] Hans S. Witsenhausen. The zero-error side information problem and chromatic numbers. *IEEE Trans. Inform. Theory*, 22(5):592–593, September 1976.

[28] Aaron Wyner and Jacob Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Trans. Inform. Theory*, 22(1):1–10, January 1976.

[29] Hirosuke Yamamoto. Wyner-Ziv theory for a general function of the correlated sources. *IEEE Trans. Inform. Theory*, 28(5):803–807, September 1982.