8-2013

Common Information and Decentralized Inference with Dependent Observations

Ge Xu

Abstract

Wyner's common information was originally defined for a pair of dependent discrete random variables. This thesis generalizes its definition in two directions: the number of dependent variables can be arbitrary, so are the alphabets of those random variables. New properties are determined for the generalized Wyner's common information of multiple dependent variables. More importantly, a lossy source coding interpretation of Wyner's common information is developed using the Gray-Wyner network. It is established that the common information equals to the smallest common message rate when the total rate is arbitrarily close to the rate distortion function with joint decoding if the distortions are within some distortion region.

The application of Wyner's common information to inference problems is also explored in the thesis. A central question is under what conditions does Wyner's common information capture the entire information about the inference object. Under a simple Bayesian model, it is established that for infinitely exchangeable random variables that the common information is asymptotically equal to the information of the inference object. For finite exchangeable random variables, connection between common information and inference performance metrics are also established.

The problem of decentralized inference is generally intractable with conditional dependent observations. A promising approach for this problem is to utilize a hierarchical conditional independence model. Utilizing the hierarchical conditional independence model, we identify a more general condition under which the distributed detection problem becomes tractable, thereby broadening the classes of distributed detection problems with dependent observations that can be readily solved.

We then develop the sufficiency principle for data reduction for decentralized inference. For parallel networks, the hierarchical conditional independence model is used to obtain conditions such that local sufficiency implies global sufficiency. For tandem networks, the notion of conditional sufficiency is introduced and the related theory and tools are developed. Connections between the sufficiency principle and distributed source coding problems are also explored. Furthermore, we examine the impact of quantization on decentralized data reduction. The conditions under which sufficiency based data reduction with quantization constraints is optimal are identified. They include the case when the data at decentralized nodes are conditionally independent as well as a class of problems with conditionally dependent observations that admit conditional independence structure through the hierarchical conditional independence model.

Common Information and Decentralized Inference with Dependent Observations

by

Ge Xu

M.S. Xidian University, Xi'an, China, 2008

B.S. Xidian University, Xi'an, China, 2005

DISSERTATION

Submitted in partial fulfillment of the requirements for the

degree of Doctor of Philosophy in Electrical and Computer Engineering

in the Graduate School of Syracuse University

August 2013

Acknowledgement

Foremost, I would like to express my gratitude to my advisor, Dr. Biao Chen. I feel very fortunate and privileged to have been a student of his. I truly thank him for his guidance and continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and technical insight. The weekly meeting with him has been an invaluable resource of inspiration. A somewhat vague idea in my mind on the research problem often turns out to be much clearer after discussion with him. I am deeply indebted to his suggestions and insights. His understanding, encouraging and personal guidance have provided a good basis for the present thesis. His profound thinking, generosity and integrity will be an inspiring role model for my future career.

I would also give thanks to all my committee members for carefully reading my thesis and giving me helpful suggestions. They are (in alphabetic order) Dr. M. Cenk Gursoy, Dr. Jing Lei, Dr. Yingbin Liang, Dr. Bin Liu, Dr. Pramod Varshney and Dr. Senem Velipasalar. I would also like to thank Dr. Hao Chen for his comments, suggestions and discussions during my research work. I have benefited a lot from stimulating discussions with him.

Thanks to my officemates and friends: Xiaohu Shang, Jin Xu, Yi Cao, Wei Liu, Minna Chen, Fangfang Zhu, Kapil Borle, Fangrong Peng, Pengfei Yang, Yu Zhao and Shengyu Zhu with whom I have shared a great time in my study. I also like to thank Renbin Peng who helped me a lot when I first arrived at U.S.

Special thanks to Wei Liu for coauthoring Chapter 2 and Shengyu Zhu for coauthoring Chapter 7 in this thesis.

Finally, I am deeply thankful to my family for all their love, support and sacrifices. Without them, I would never have been able to finish this dissertation. I dedicate this dissertation to the memory of my father Leipeng Xu, whose role in my life was, and remains, immense. I also owe a great debt of gratitude to my husband Yuanming Geng, who has been with me and encouraged me all these years. Thank you.

I acknowledge Syracuse University Graduate Fellowship program for providing funding for my Ph.D. study, and additional support from the National Science Foundation under award 0925854, by the Air Force Office of Sponsored Research under award FA9550-10-1-0458 and by the Army Research Office under award W911NF-12-1-0383.

Contents

List of Figures

Chapter 1

Introduction

Correlated observations occur in many engineering applications even if samples may be collected at decentralized nodes. The presence of data dependence may be due to a common phenomenon that produces the data. Often times, the objective is to understand such common phenomenon when it is subject to various distortion or observation noises. This thesis focuses on 1) the quantitative characterization of data dependence and the physical interpretation of such quantities; 2) statistical inference in decentralized systems with dependent data. Thus, the thesis can be loosely separated into two parts. The first part deals with the generalization of Wyner's common information to multi-variate random variables of arbitrary alphabet. Motivated by some interesting property associated with the generalized common information, we then explore the use of common information in decentralized inference. The second part of the thesis addresses several research problems in decentralized inference with an emphasis on problems involving dependent observations across different sensors.

Quantifying the information that is common between two dependent random vari-

ables has been a classical problem both in information theory and in mathematical statistics [1–4]. The most widely used notion is Shannon's mutual information [1], which measures the amount of uncertainty reduction in one variable by observing the other. Other notions of information between a pair of dependent variables include Gács and Körner's common randomness [2] and Wyner's common information [4].

Gács and Körner's common randomness is defined as the maximum number of common bits per symbol that can be independently extracted from the correlated random variables. On the other hand, Wyner's common information can be defined as the number of common random bits per symbol that are needed to generate a sequence of random variable pairs with the specified joint distribution. While Wyner's common information was originally defined for two discrete random variables, the expression (c.f. Equation 1.1) can be evaluated for any pair of random variables with arbitrary alphabets. However, the operational meanings available in existing literature are largely confined to that for discrete alphabets. These include the minimum common rate for the Gray-Wyner lossless source coding problem under a sum rate constraint, the minimum rate of a common input of two independent random channels for distribution approximation [4], and the strong coordination capacity of a two-node network without common randomness and with actions assigned at one node [5].

This thesis generalizes Wyner's common information along two directions. The first is to generalize it to that of multiple dependent random variables. The second is to generalize it to that of continuous random variables. For the first direction, Wyner's common information is defined by introducing a conditional independence structure among the multiple random variables, which is equivalent to the Markov

chain condition for two dependent variables. For this case, it is shown that Wyner's original interpretations in [4] can be directly extended to that involving multiple variables. This generalization to multiple dependent random variables also reveals a surprising monotone property of Wyner's common information in the number of variables involved.

For the second direction, we provide a new lossy source coding interpretation using the Gray-Wyner network. Specifically, we show that, for the Gray-Wyner network, Wyner's common information is precisely the smallest common message rate for a certain range of distortion constraints when the total rate is arbitrarily close to the rate distortion function with joint decoding. This new operational interpretation justifies the generalization of Wyner's common information to that of continuous random variables. Computing Wyner's common information is known to be a challenging problem and it was only resolved for several special cases described in [4, 6]. Along with our generalizations of Wyner's common information, we provide two new examples where we can explicitly evaluate the common information of multiple dependent variables. In particular, we provide closed-form expressions of Wyner's common information of bivariate Gaussian random variables as well as multi-variate Gaussian random variables with a particular correlation structure.

Motivated by the monotonicity property of Wyner's common information in the number of variables, we explore the application of Wyner's common information to inference problems and its connection with some performance metrics. The central question we are going to answer is under what conditions does Wyner's common information capture the entire information about the inference object under a simple Bayesian model.

In the second part of the thesis, we address several research problems in decentralized inference with dependent observations. Decentralized inference refers to the decision making process involving multiple sensors [7]. Each sensor summarizes its observation and sends a message to the fusion center, which makes the final decision based on the messages it receives. Tremendous efforts have been devoted to this problem that leads to many fundamental results. However, most of the results obtained for this model are under the assumption of conditional independence. Therefore, we focus on the problem of decentralized inference with conditionally dependent observations in this thesis.

We first study the problem of distributed detection with conditionally dependent observations. Following the hierarchical conditional independence model proposed in [8], we identify a more general condition under which the structure of optimal local sensor decision rules can be specified. This enables us to tackle a much broader class of distributed detection problems with dependent observations.

We then develop the sufficiency principle that guides local data reduction in decentralized inference. For decentralized inference, data reduction is done locally without access to the global data. Therefore, the contrasting notions of local sufficiency and global sufficiency [9] need to be treated with care. We consider two classical inference networks, parallel networks and tandem networks, in this thesis. For the parallel networks, we obtain conditions such that local sufficiency implies global sufficiency for conditionally dependent observations. For the tandem networks, we introduce the notion of conditional sufficiency and develop related theory and tools associated with the new notion.

Finally, we investigate decentralized data reduction when each sensor is subject

to a quantization constraint under the Bayesian inference framework. We show that sufficiency based data reduction is structurally optimal for decentralized inference with conditionally independent observations. For decentralized inference with conditionally dependent observations, quantizing sufficient statistics, even global ones, need not be optimal. We proceed to identify conditions under which sufficiency based data reduction is structurally optimal as well as establish a unifying condition that encompasses both the independent and the dependent observation cases.

In the following, we provide some of the background knowledge for this thesis and the main contributions of the thesis. We start by introducing the notion of common information in Section 1.1. In Section 1.2, we review the sufficiency principle for centralized inference. The outline of the thesis is given in Section 1.3 and the main contributions are summarized in Section 1.4. Finally, we conclude this chapter with the notations used in this dissertation in Section 1.5.

1.1 Common information

Consider a pair of dependent random variables X and Y with joint distribution $p(x, y)$ which denotes either the probability density function if X and Y are continuous or the probability mass function if X and Y are discrete. Quantifying the information that is common between X and Y has been a classical problem both in information theory and in mathematical statistics [1, 2, 4, 10]. There are three widely used notions in the literature.

1.1.1 Mutual information

The most widely used notion is Shannon's mutual information, defined as

$$I(X;Y) = E\left[\log \frac{p(x,y)}{p(x)p(y)}\right],$$

where $p(x)$ and $p(y)$ are the marginal distribution of X and Y corresponding to the joint distribution $p(x,y)$ and $E[\cdot]$ denotes expectation taken with respect to $p(x,y)$. Shannon's mutual information measures the amount of uncertainty reduction in one variable by observing the other. In the case that X and Y are independent, mutual information $I(X;Y) = 0$, indicating that observing one variable X does not give any information about Y and vice versa. The significance of $I(X;Y)$ lies in its applications to a broad range of problems in which concrete operational meanings of $I(X;Y)$ can be established. These include both source and channel coding problems in information and communication theory [11] and hypothesis testing problems in statistical inference [12].

Generalization of mutual information to $N > 2$ random variables was first reported in [13]. The generalization is obtained from the observation that for a pair of random variables, computing $I(X;Y)$ is consistent with the Venn diagram for set operations [12, 14].

1.1.2 Gács and Körner's common randomness

Gács and Körner's common randomness is defined as the maximum number of common bits per symbol that can be independently extracted from discrete random variables X and Y whose joint distribution is specified by $p(x,y)$. That is

$$K(X,Y) = \sup H(V),$$

where the supremum is taken over all the finite random variables V satisfying

$$V = f(X) = g(Y),$$

for some functions $f : \mathcal{X} \to \mathcal{V}$ and $g : \mathcal{Y} \to \mathcal{V}$. Quite naturally, $K(X, Y)$ has found extensive applications in secure communications, e.g., for key generation [15–17]. More recently, a new interpretation of $K(X, Y)$ using the Gray-Wyner source coding network was given in [18]. It was noted in [2] [19] that the definition of $K(X, Y)$ is rather restrictive in that $K(X, Y)$ equals 0 in most cases except for the special case when $X = (X', V)$ and $Y = (Y', V)$ and X', Y', V are independent variables or those (X, Y) pair that can be converted to such a dependence structure through relabeling the realizations, i.e., whose distribution is a permutation of the original joint distribution matrix. Note that $I(X; Y) = K(X; Y) = H(V)$ for this case.

Gács and Körner's common randomness has been generalized to multiple random variables in [20], which extends the encoding process in the definition of common randomness to that of N terminals.

1.1.3 Wyner's common information

Assume X, Y are two dependent discrete random variables with distribution $p(x, y)$, Wyner defined the common information as follows:

$$C(X, Y) = \inf_{X-W-Y} I(X, Y; W). \tag{1.1}$$

Here, the infimum is taken over all auxiliary random variables W such that X, W, and Y form a Markov chain.

Wyner provided two operational meanings for the above definition. The first approach is based on a simple source coding network first studied by Gray and Wyner

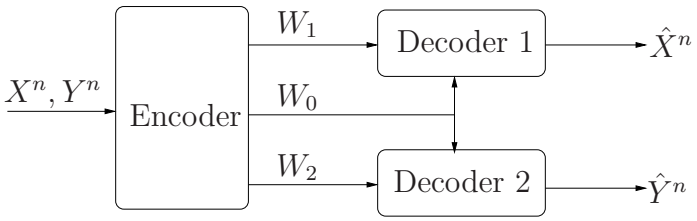


Figure 1.1: Gray-Wyner source coding network.

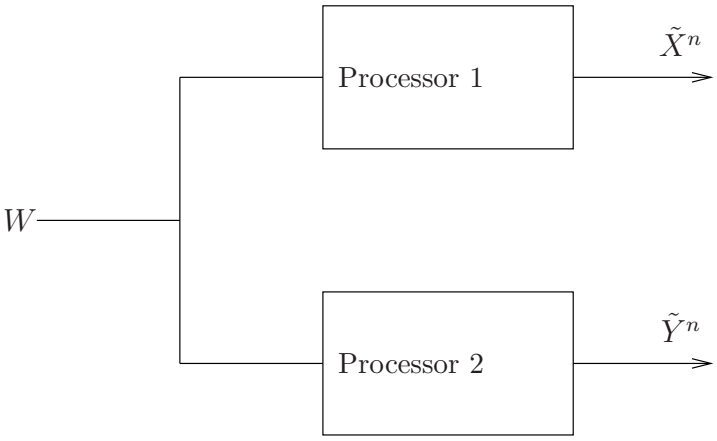


Figure 1.2: Distribution approximation.

[21] and illustrated in Fig. 1.1. In this system, the encoder observes a sequence of independent and identically distributed (i.i.d.) random variable pairs (X^n, Y^n), and maps them to three messages W_0, W_1, W_2, taking values in alphabets of respective sizes $2^{nR_0}, 2^{nR_1}$ and 2^{nR_2}. Decoder 1, upon receiving (W_0, W_1), needs to reproduce X^n with high reliability while decoder 2, upon receiving (W_0, W_2), needs to reproduce Y^n with high reliability. Define

$$\Delta = \frac{1}{2n}\left(E[d_H(X^n, \hat{X}^n)] + E[d_H(Y^n, \hat{Y}^n)]\right), \tag{1.2}$$

where $d_H(\cdot, \cdot)$ is the Hamming distortion. Let C_1 be the the infimum of all achievable R_0 for the system in Fig. 1.1 such that for any $\epsilon > 0$, there exists, for n sufficiently large, a source code with the total rate $R_0 + R_1 + R_2 \leq H(X, Y) + \epsilon$ and $\Delta \leq \epsilon$.

The second approach is shown in Fig. 1.2. In this approach, the joint distribution of the i.i.d sequences (X^n, Y^n)

$$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i), \tag{1.3}$$

is approximated by the output distribution of a pair of random number generators. Specifically, a common input W, uniformly distributed on $\mathcal{W} = \{1, \cdots, 2^{nR_0}\}$ is sent to two separate processors which are independent of each other. These processors (random number generators) generate i.i.d sequence according to two distributions $q_1(x^n|w)$ and $q_2(y^n|w)$ respectively. The output sequences of the two processors are denoted by $\tilde{X}^n$ and $\tilde{Y}^n$ respectively and the joint distribution of the output sequences is given by

$$q(x^n, y^n) = \sum_{w \in \mathcal{W}} \frac{1}{|\mathcal{W}|} q_1(x^n|w) q_2(y^n|w). \tag{1.4}$$

Let

$$D_n(q,p) = \frac{1}{n} \sum_{x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n} q(x^n, y^n) \log \frac{q(x^n, y^n)}{p(x^n, y^n)}. \tag{1.5}$$

Let C_2 be the infimum of rate R_0 for the common input such that for any $\epsilon > 0$, there exists a pair of distribtions $q_1(x^n|w)$, $q_2(y^n|w)$ and n such that $D_n(q,p) \leq \epsilon$.

Wyner proved in [4] that

$$C_1 = C_2 = C(X,Y). \tag{1.6}$$

We emphasize that the above operational meanings of $C(X,Y)$, both their definition and the proofs, are confined to discrete X and Y. However, the expression in (1.1) can be easily evaluated for any pair of random variables with discrete or continuous alphabet. In addition, the two operational interpretations have natural extensions to the multi-variate case. This motivates our work in generalizing Wyner's common information along two directions: that involve multiple variables and variables with arbitrary alphabets. For multi-variate generalization, we show that both of Wyner's operational interpretations hold. For the common information defined for continuous random variables, we provide a new lossy source coding interpretation using the Gray-Wyner network.

Finally, for a pair of discrete random variables, it was shown in [4] that the mutual information, Gács and Körner's common randomness and Wyner's common information satisfy the following relationship

$$K(X,Y) \leq I(X;Y) \leq C(X,Y), \tag{1.7}$$

and the equalities hold *if and only if* it is possible to write $X = (X', V)$ and $Y = (Y', V)$ where X', Y' are conditionally independent given V.

1.2 Sufficiency principle

In this section, we review the basic sufficiency principle for centralized inference. Sufficiency principle is a guiding principle for data reduction. A sufficient statistic is a function of the data, chosen so that it 'should summarize the whole of the relevant information supplied by the sample' [22]. A classical example is in binary hypothesis testing where the likelihood ratio can be shown to be a sufficient statistic of the unknown hypothesis, thus can be used instead of the raw data for subsequent decision making [23]. Another example is the waveform channel with additive white Gaussian channel as often assumed in digital communications [24]. It can be easily established that the outputs of simple correlators (or equivalently, that of matched filters) form a sufficient statistic for the unknown input signals. In both examples, the original data, often of high or infinite dimensions, is reduced to low dimension statistics which greatly facilitate the subsequent inference. Indeed, the sufficiency principle has played a prominent role in designing various data processing methods for statistical inference and it encompasses numerous results that have been developed since Fisher's original work [22, 25, 26].

Suppose θ is the parameter of inference interest and $\mathbf{X} \triangleq \{X_1, \cdots, X_n\}$ is a random vector observation, whose distribution is given by $p(\mathbf{x}|\theta)$. The sufficiency principle states that a function (or statistic) of $\mathbf{X}$, denoted by $T(\mathbf{X})$, is a sufficient statistic for θ if the inference outcome does not change when either $\mathbf{x}$ or $\mathbf{y}$ is observed as long as $T(\mathbf{x}) = T(\mathbf{y})$ [25]. If $T(\mathbf{X})$ is a sufficient statistic for θ, then any inference about θ should depend on $\mathbf{X}$ only through $T(\mathbf{X})$.

A useful tool to identify sufficient statistics is the Neyman-Fisher factorization

theorem [25] which states that a statistic $T(\mathbf{X})$ is sufficient for θ *if and only if* there exist functions $g(t|\theta)$ and $h(\mathbf{x})$ such that

$$p(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}). \tag{1.8}$$

If the parameter θ is itself random, the sufficiency principle can be elegantly reframed using the data processing inequality, assisted with the use of Shannon's mutual information [11]. That is, a function $T(\mathbf{X})$ is a sufficient statistic *if and only if* the following Markov chain holds

$$\theta - T(\mathbf{X}) - \mathbf{X}, \tag{1.9}$$

which is equivalent to the mutual information equation

$$I(\theta; \mathbf{X}) = I(\theta, T(\mathbf{X})). \tag{1.10}$$

The following lemma is a straightforward result from the definition of Markov chain.

Lemma 1. *Let* $\mathbf{X} \sim p(\mathbf{x}|\theta)$ *where* θ *is a random parameter. If* $T(\mathbf{X})$ *is a sufficient statistic for* θ *with respect to* $\mathbf{X}$*, then*

$$p(\theta|\mathbf{x}) = p(\theta|T(\mathbf{x})). \tag{1.11}$$

Proof. As $T(\mathbf{X})$ is a function of $\mathbf{X}$, $\theta - \mathbf{X} - T(\mathbf{X})$ form a Markov chain. Together with (1.9) we thus have

$$p(\theta|\mathbf{x}) = p(\theta|\mathbf{x}, T(\mathbf{x})) = p(\theta|T(\mathbf{x})). \tag{1.12}$$

■

Sufficient statistics are not unique, therefore, it is natural to ask whether one sufficient statistic is better than another. The sufficient statistic that achieves the

maximum data reduction is called the *minimal sufficient statistic*. That is, the minimal sufficient statistic is a sufficient statistic that is a function of all other sufficient statistics.

As with sufficient statistics, a minimal sufficient statistic can also be characterized through the use of Markov chains by the data-processing inequality. Assume θ is random, a statistic $M(\mathbf{X})$ is a minimal sufficient statistic if the following Markov chain is satisfied

$$\theta - M(\mathbf{X}) - T(\mathbf{X}) - \mathbf{X}, \tag{1.13}$$

for every other sufficient statistic $T(\mathbf{X})$.

1.3 Outline of thesis

The first part of the thesis, consisting of Chapters 2, 3 and 4, develops Wyner's common information for multiple dependent random variables with arbitrary alphabets and explores its connection to some statistical inference problems.

In Chapter 2, we generalize Wyner's common information of a pair of discrete random variables to that of N random variables with arbitrary alphabets. We provide coding theorems showing that Wyner's original interpretations in [4] can be directly extended to that involving multiple variables. We establish a monotone property of Wyner's common information in the number of variables which is in contrast to that of mutual information or common randomness.

In Chapter 3, we develop a lossy source coding interpretation of Wyner's common information using the Gray-Wyner network. We show that for the Gray-Wyner network, Wyner's common information is precisely the smallest common message rate

for a certain range of distortion constraints when the total rate is arbitrarily close to the rate distortion function with joint decoding. As the common information is only a function of the joint distribution, this smallest common rate remains constant even if the distortion constraints vary, as long as they are in a specific distortion region. Furthermore, we establish that for successive refinement sources, given a total rate of rate distortion function, it is optimal to use a two-stage encoding scheme in the Gray-Wyner network: first encode the common message with rate of common information, then encode the two private messages with extra rates. In the same chapter, we also provide examples of how to compute Wyner's common information for these extensions. Specifically, we consider the binary sources and Gaussian sources. For the Gaussian case, we derive, through an estimation theoretic approach, the common information for a bivariate Gaussian source and its extension to the multi-variate case with a certain correlation structure. In addition, we characterize the distortion regions where the common information equals to the smallest common message rate in the Gray-Wyner network for both cases.

Chapter 4 explores the application of Wyner's common information to various inference problems. The inference problems considered in this chapter arise from symmetric simple Bayesian models. We study the common information of exchangeable random variable and show that for infinite exchangeable sequences, the common information is asymptotically equal to the information object, i.e., the hidden variable in the Bayesian model. For finite exchangeable sequences, while this result is no longer true in general, we identify two important cases such that the result still holds. For these two cases, one with binary and the other with Gaussian observations, we further establish the relationship between common information and relevant

performance metrics for the underlying inference problems.

Chapter 5 considers the problem of distributed detection with conditionally dependent observations. Under the Bayesian detection framework and utilizing a recently proposed Hierarchical Conditional Independence (HCI) model [8], we identify a more general condition associated with the hidden variable for the continuous HCI model which enables us to tackle a broader class of distributed detection problems with dependent observations.

Chapter 6 develops the sufficiency principle that guides local data reduction in networked inference with dependent observations for two classes of inference networks: the parallel network and the tandem network. For the parallel networks, the HCI model is used to obtain conditions such that local sufficiency implies global sufficiency. For the tandem networks, we introduce the notion of conditional sufficiency and developed related theory and tools.

Chapter 7 investigates the decentralized data reduction problem when each sensor is subject to a quantization constraint. We show that sufficiency based data reduction is structurally optimal under the Bayesian inference framework for decentralized inference with conditionally independent observations. For decentralized inference with conditionally dependent observations, utilizing the HCI model, we provide a suitable way of finding optimal data reduction if it exists. We also establish a unifying condition that encompasses both the independent and the dependent observation cases.

We conclude the thesis in Chapter 8 where we summarize our major contributions and point to future research directions.

1.4 Main contributions

In this section, we briefly summarize the main contributions of this thesis. We classify them into two areas: one related to Wyner's common information and the other on decentralized inference with dependent observations. For Wyner's common information, the main contributions are:

- We have generalized Wyner's common information of a pair of discrete random variables to that of multiple random variables with arbitrary alphabets. It is shown that Wyner's original interpretations directly extend to that involving multiple variables.

- We have provided a new lossy source coding interpretation of Wyner's common information which can be applied to both discrete and continuous random variables. Wyner's common information is precisely the smallest common message rate for a certain range of distortion constraints when the total rate is arbitrarily close to the rate distortion function with joint decoding.

- We have solved the computation of Wyner's common information for bivariate Gaussian random variables.

- We have established the connection between Wyner's common information and the symmetric Bayesian inference model.

In the area of decentralized inference with dependent observations, the main contributions are:

- We have identified a more general condition associated with the hidden variable

for the continuous HCI model which enables us to tackle a broader class of distributed detection problems with dependent observations.

- We have explored the connection of local sufficiency and global sufficiency with dependent observations for both parallel and tandem networks.
- We have proposed a new notion, conditional sufficiency, for tandem networks and developed related theory and tools.
- We have established the conditions under which sufficiency based data reduction with quantization constraints is optimal for both parallel and tandem networks.

1.5 Notations

In the following, we introduce notations which will be used throughout this thesis.

Term	Desciption
X	a random variable
x	a realization of a random variable X
$\mathcal{X}$	sample space of random variable X
X^n	a vector of i.i.d random variables, $\{X_1, \cdots, X_n\}$
$\mathbf{X}$	a vector of dependent variables $\{X_i\}$
$\mathbf{X}^A$	a vector of random variables $\{X_i\}$, $i \in A$ where A is a set of integers.
$p(\cdot)$	pmf or pdf of a random variable
$q(\cdot)$	pmf or pdf of a random variable
$X \sim p(x)$	the pmf or pdf of X is $p(x)$
$X \sim \mathcal{N}(\mu, \sigma^2)$	X is Gaussian distributed with mean μ and variance σ^2
$E[\cdot]$	expectation
$d_H(\cdot, \cdot)$	Hamming distortion
$H(\cdot)$	entropy function
$h(a)$	binary entropy function
$h(X)$	differential entropy of random variable X
$I(\cdot;\cdot)$	mutual information
$K(\cdot, \cdot)$	Gács and Körner's common randomness
$C(\cdot, \cdot)$	Wyner's common information
BSC	binary symmetric channel
DSBS	doubly symmetric binary source
CI	conditional independence
HCI	hierarchical conditional independence
DHCI	discrete HCI
CHCI	continuous HCI
HHCI	hybrid HCI
BCDF	Bayesian cost density function

Chapter 2

Common Information of N Random Variables

In Chapter 1, we introduced the concept of Wyner's common information and its two operational meanings. However, the original notion of common information was limited to two discrete random variables. The first question is how does the notion extend to that of multiple random variables. In addition, definition in (1.1) tells that the common information can be evaluated for any pair of random variables with arbitrary alphabets. Therefore, in this chapter, we will generalize Wyner's common information to *multiple* random variables with *arbitrary alphabets* and provide the corresponding operational interpretations.

In this chapter, we will discuss the generalization to multiple random variables and introduce the corresponding coding theorems. The generalization to that of arbitrary alphabet will be developed in Chapter 3.

The common information for N random variables is defined through a condi-

tional independence structure which is equivalent to the Markov chain condition for two dependent variables. We establish a monotone property of Wyner's common information in the number of variables. In addition, we prove that Wyner's original interpretations in [4] can be directly extended to the multi-variate case.

The rest of this chapter is organized as follows. Definition of Wyner's common information for N dependent random variables with arbitrary alphabets is given in Section 2.1 along with some associated properties. Section 2.2 generalizes the Gray-Wyner network in Fig. 1.1 to include N source sequences and N decoders. We also characterize the lossless and lossy rate regions for the generalized network in this section. In Section 2.3, the operational meanings of Wyner's common information are extended to that of N discrete dependent random variables. Section 2.4 concludes this chapter.

2.1 Common information of N random variables

2.1.1 Definition

Wyner's original definition of the common information in (1.1) assumes a Markov chain $X - W - Y$. This Markov chain is equivalent to stating that X and Y are conditionally independent given W. This conditional independence structure can be naturally generalized to that of N dependent random variables. Let $X_1, \cdots, X_N$ be N dependent random variables that take values in some arbitrary (finite, countable, or continuous) spaces $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$. The joint distribution of $X_1, \cdots, X_N$ is denoted as $p(x_1, \cdots, x_N)$. We now give the definition of the common information for

N dependent random variables.

Definition 1. *Let $\{X_1, \cdots, X_N\}$ be a random vector with joint distribution $p(x_1, \cdots, x_N)$. The common information of N random variables $X_1, \cdots, X_N$ is defined as*

$$C(X_1, \cdots, X_N) \triangleq \inf I(X_1, \cdots, X_N; W), \tag{2.1}$$

where the infimum is taken over all the joint distributions of $(X_1, \cdots, X_N, W)$ such that

1. *the marginal distribution for $X_1, \cdots, X_N$ is $p(x_1, \cdots, x_N)$,*

2. *$X_1, \cdots, X_N$ are conditionally independent given W, i.e.,*

$$p(x_1, \cdots, x_N|w) \;=\; \prod_{i=1}^{N} p(x_i|w). \tag{2.2}$$

2.1.2 Properties

We now discuss several properties associated with the definition given in (2.1).

Wyner's common information of two random variables (X, Y) satisfies the following inequality

$$I(X;Y) \leq C(X,Y) \leq \min\{H(X), H(Y)\}. \tag{2.3}$$

A similar inequality for the common information of N random variables can be derived. Denote by $\mathbf{X} \triangleq \{X_1, \cdots, X_N\}$, we have the following lemma.

Lemma 2. *Let $\mathbf{X} \sim p(x_1, \cdots, x_N)$, $A \subseteq \mathcal{N} = \{1, 2, \cdots, N\}$ and $\bar{A} = \mathcal{N} \backslash A$. We have*

$$\max_{A}\{I(\mathbf{X}^A; \mathbf{X}^{\bar{A}})\} \leq C(\mathbf{X}) \leq \min_{j}\{H(\mathbf{X}^{-j})\}, \tag{2.4}$$

where $\mathbf{X}^{-j} \triangleq \mathbf{X}^{\mathcal{N}\backslash\{j\}} = \{X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_N\}$ for $j \in \mathcal{N}$.

Proof. To verify the upper bound, for any $j \in \mathcal{N}$, let $W_j = \mathbf{X}^{-j}$. Thus, $X_1, \cdots, X_N$ are conditionally independent given W_j, and

$$I(\mathbf{X}; W_j) = I(\mathbf{X}; \mathbf{X}^{-j}) = H(\mathbf{X}^{-j}). \tag{2.5}$$

Thus $C(\mathbf{X}) \leq H(\mathbf{X}^{-j})$ for all $j \in \mathcal{N}$.

For the lower bound, since X_1, $\cdots$, X_N are conditionally independent given W, we have the Markov chain $\mathbf{X}^A - W - \mathbf{X}^{\bar{A}}$ for any subset $A \subseteq \mathcal{N}$. Hence,

$$I(\mathbf{X}; W) \geq I(\mathbf{X}^A; W) \geq I(\mathbf{X}^A; \mathbf{X}^{\bar{A}}), \tag{2.6}$$

where the second inequality is by the data processing inequality.

Therefore,

$$I(\mathbf{X}; W) \geq \max_A \{I(\mathbf{X}^A; \mathbf{X}^{\bar{A}})\}. \tag{2.7}$$

This completes the proof of Lemma 2. ■

In the following, we show that the common information defined in (2.1) also satisfies a monotone property. Let us first consider a ternary example with

$$\begin{aligned} X &= (X', U, V), \\ Y &= (Y', U, W), \\ Z &= (Z', V, W), \end{aligned}$$

where (X', Y', Z', U, V, W) are mutually independent random variables. It is easy to show that for this example

$$C(X, Y, Z) = H(U, V, W),$$

$$C(X,Y) \;=\; H(U).$$

This simple example is surprising in that

$$C(X,Y,Z) > C(X,Y),$$

i.e., the common information of the three variables is greater than that of two variables. In other words, the inclusion of an additional variable increases the common information. Indeed, we have the following lemma:

Lemma 3. *Let* $\mathbf{X} \sim p(\mathbf{x})$. *For any two sets* A, B *that satisfy* $A \subseteq B \subseteq \mathcal{N} = \{1, 2, \cdots, N\}$, *we have*

$$C(\mathbf{X}^A) \leq C(\mathbf{X}^B), \tag{2.8}$$

Proof. Let W' be the auxiliary variable that achieves $C(\mathbf{X}^B)$, i.e., $I(\mathbf{X}^B; W') = \inf_W I(\mathbf{X}^B; W)$. Since $A \subseteq B$, $\mathbf{X}^B$ being conditionally independent given W' implies that $\mathbf{X}^A$ are conditionally independent given W'. Thus

$$\begin{aligned} I(\mathbf{X}^B; W') &\geq I(\mathbf{X}^A; W'), \\ &\geq \inf I(\mathbf{X}^A; W), \end{aligned}$$

where the infimum is taken over all W such that $\mathbf{X}^A$ is independent given W. ■

The above monotone property of the common information is contrary to what the name implies: conceptually, the information in common ought to decrease when new variables are included in the set of random variables. Such is the case for Gács and Körner's common randomness, i.e., $K(\mathbf{X}^A) \geq K(\mathbf{X}^B)$. As a consequence, we have that for any N random variables $C(\mathbf{X}) \geq K(\mathbf{X})$. The fact that the common information $C(\mathbf{X})$ increases as more variables are involved suggests that it may have

potential applications in statistical inference problems. This will be explored in detail in Chapter 5.

2.2 Generalized Gray-Wyner networks

In Chapter 1, we reviewed the fact that Wyner's common information has its operational interpretation in Gray-Wyner network as shown in Fig. 1.1. Therefore, it is quite natural to consider a generalized Gray-Wyner source coding network with N source sequences for common information of N random variables. To facilitate this extension, we first explore the generalized Gray-Wyner network in detail in this section.

Consider the Gray-Wyner source coding network [21] with one encoder and N decoders as shown in Fig. 2.1.

The encoder observes a sequence $\{X_1^n, X_2^n, \cdots, X_N^n\}$, which is the sequence of n independent drawings of N random variables $(X_1, \cdots, X_N)$, $X_1 \in \mathcal{X}_1, \cdots, X_N \in \mathcal{X}_N$, from a distribution $p(x_1, \cdots, x_N)$. The source alphabets $(\mathcal{X}_1, \cdots, \mathcal{X}_N)$ can be either discrete or continuous. The sequence can also be expressed as $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ where each $\mathbf{X}_k = \{X_{1k}, \cdots, X_{Nk}\}$, $k = 1, \cdots, n$, is a length-N vector with joint distribution $p(x_1, \cdots, x_N)$.

There are a total of N receivers, with the ith receiver only interested in recovering the ith component sequence X_i^n. The encoder encodes the source into $N+1$ messages, one is a public message available at all receivers while the other N messages are private messages only available at the corresponding receivers.

For $m = 1, 2, \cdots$, let $I_m = \{0, 1, 2, \cdots, m-1\}$. We define an $(n, M_0, M_1, \cdots, M_N)$

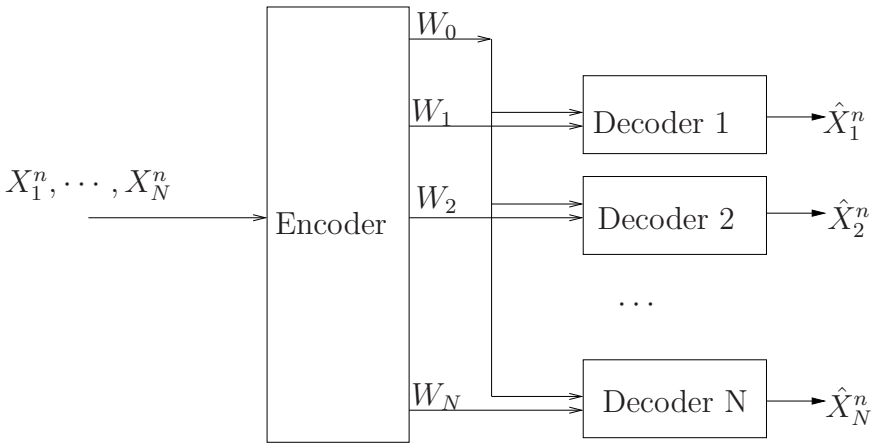


Figure 2.1: Generalized Gray-Wyner source coding network.

code corresponding to the generalized Gray-Wyner model.

Definition 2. *An* $(n, M_0, M_1, \cdots, M_N)$ *code consists of the following:*

- *an encoder mapping* $f : \mathcal{X}_1^n \times \cdots \times \mathcal{X}_N^n \to I_{M_0} \times I_{M_1} \times \cdots I_{M_N}$,
- N *decoder mappings* $g_i : I_{M_i} \times I_{M_0} \to \hat{\mathcal{X}}_i^n, \quad i = 1, 2, \cdots, N.$

Here $\hat{\mathcal{X}}_i$ is the reproducing alphabets for X_i, $i = 1, \cdots, N$. For an $(n, M_0, M_1, \cdots, M_N)$ code defined above, let

$$f(X_1^n, \cdots, X_N^n) = (W_0, W_1, \cdots, W_N),$$

where $X_i^n \in \mathcal{X}_i^n$, $i = 1, \cdots, N$ and $(W_0, W_1, \cdots, W_N)$ is the tuple of indices. Then set

$$\hat{X}_i^n = g_i(W_i, W_0), i = 1, 2, \cdots, N,$$

where $\hat{X}_i^n \in \hat{\mathcal{X}}_i^n$, $i = 1, \cdots, N$.

We now discuss below the lossless and lossy cases for the generalized Gray-Wyner network respectively.

2.2.1 Lossless Gray-Wyner source coding

If the source alphabets $(\mathcal{X}_1, \cdots, \mathcal{X}_N)$ are finite sets, we define the probability of error of an $(n, M_0, M_1, \cdots, M_N)$ code as

$$P_e^{(n)} = \frac{1}{nN} \sum_{i=1}^{N} E[d_H(X_i^n, \hat{X}_i^n)], \tag{2.9}$$

where $\hat{X}_i^n = g_i(W_i, W_0) \in \mathcal{X}_i^n$ for $i = 1, \cdots, N$ $d_H(u^n, \hat{u}^n)$ is the Hamming distance between u^n and $\hat{u}^n$.

A rate tuple $(R_0, R_1, \cdots, R_N)$ is said to be *achievable* if for any $\epsilon > 0$, there exists, for n sufficiently large, an $(n, M_0, M_1, \cdots, M_N)$ code such that

$$M_i \leq 2^{n(R_i+\epsilon)}, \quad i = 0, 1, \cdots, N, \tag{2.10}$$

$$P_e^{(n)} \leq \epsilon. \tag{2.11}$$

Denote by $\mathcal{R}_1$ the region of all achievable rate tuples $(R_0, R_1, \cdots, R_N)$. The rate region $\mathcal{R}_1$ of the lossless source coding problem for generalized Gray-Wyner network is given in the following theorem.

Theorem 1. *$\mathcal{R}_1$ is the union of all rate tuples $(R_0, R_1, \cdots, R_N)$ that satisfy*

$$R_0 \geq I(X_1, \cdots, X_N; W), \tag{2.12}$$

$$R_i \geq H(X_i|W), \quad i = 1, 2, \cdots, N, \tag{2.13}$$

for some $W \sim p(w|x_1, \cdots, x_N)$.

2.2.2 Lossy Gray-Wyner source coding

In this section, we consider a more general case for the source coding problem of Gray-Wyner network where we require that the source sequence to be reproduced to within a certain fidelity criterion.

Specifically, for source sequence with arbitrary alphabets, let $\mathbf{d}(\mathbf{x}, \hat{\mathbf{x}}) \triangleq \{d_1(x_1, \hat{x}_1), \cdots, d_N(x_N, \hat{x}_N)\}$ be a compound distortion measure. For an $(n, M_0, M_1, \cdots, M_N)$ code, define $\Delta_i, i = 1, \cdots, N$ to be the average distortion between the ith component sequence of the encoder input and the ith decoder output,

$$\Delta_i \triangleq E[d_i(X_i^n, \hat{X}_i^n)] = \frac{1}{n}\sum_{k=1}^{n} E[d_i(X_{ik}, \hat{X}_{ik})]. \tag{2.14}$$

The vector of average distortions is defined as

$$\mathbf{\Delta} \triangleq \{\Delta_1, \cdots, \Delta_N\}. \tag{2.15}$$

An $(n, M_0, M_1, \cdots, M_N)$ code with an average distortion vector $\mathbf{\Delta}$ is said to be an $(n, M_0, M_1, \cdots, M_N, \mathbf{\Delta})$ rate distortion code. Let $\mathbf{D} \triangleq \{D_1, D_2, \cdots, D_N\} \in \mathbb{R}_+^N$. A rate tuple $(R_0, R_1, \cdots, R_N)$ is said to be $\mathbf{D}$-*achievable* if for arbitrary $\epsilon > 0$, there exists, for n sufficiently large, an $(n, M_0, M_1, \cdots, M_N, \mathbf{\Delta})$ code such that

$$M_i \leq 2^{n(R_i+\epsilon)}, \quad i = 0, 1, \cdots, N, \tag{2.16}$$

$$\mathbf{\Delta} \leq \mathbf{D} + \epsilon. \tag{2.17}$$

Let $\mathcal{R}_2(\mathbf{D})$ be the region of all $\mathbf{D}$-achievable rate tuples $(R_0, R_1, \cdots, R_N)$.

Theorem 2. *$\mathcal{R}_2(\mathbf{D})$ is the union of all rate tuples $(R_0, R_1, \cdots, R_N)$ that satisfy*

$$R_0 \geq I(X_1, \cdots, X_N; W), \tag{2.18}$$

$$R_i \geq R_{X_i|W}(D_i), \quad i = 1, 2, \cdots, N, \tag{2.19}$$

for some $W \sim p(w|x_1, \cdots, x_N)$.

Here, $R_{X_i|W}(D_i)$ is the conditional rate distortion function defined as in [27]

$$R_{X_i|W}(D_i) = \min_{p_t(\hat{x}_i|x_i,w):Ed_i(X_i,\hat{X}_i)\leq D_i} I(X_i; \hat{X}_i|W). \tag{2.20}$$

Theorems 1 and 2 are direct extensions of Theorems 4 and 8 in [21] for Gray-Wyner network with two receivers. Note that in [21], the authors proved only the discrete case for [21, Theorem 8], the proof for continuous alphabets can be constructed in a similar fashion.

2.3 Operational meaning of Wyner's common information for N variables

Section 1.1 describes two operational interpretations of Wyner's common information for two discrete random variables based on the Gray-Wyner network and distribution approximation. These operational interpretations can also be extended to the common information of N dependent random variables. In this section, we will show that the common information of N random variables defined in Definition 1 has operational significance in the generalized Gray-Wyner network and distribution approximations.

2.3.1 Gray-Wyner network interpretation

For the first approach, we consider the lossless Gray-Wyner network with N terminals discussed in Section 2.2.1. For this Gray-Wyner source coding network, a number R_0 is said to be *achievable* if for any $\epsilon > 0$, there exists, for n sufficiently large, an $(n, M_0, M_1, \cdots, M_N)$ code (c.f. Definition 2) with

$$\begin{aligned} M_0 &\leq 2^{nR_0}, && (2.21)\\ \frac{1}{n}\sum_{i=0}^{N} \log M_i &\leq H(X_1, \cdots, X_N) + \epsilon, && (2.22)\\ P_e^{(n)} &\leq \epsilon, && (2.23) \end{aligned}$$

where $P_e^{(n)}$ is given in (2.9). As with the case for two random variables, we define C_1 as the infimum of all achievable R_0.

Theorem 3. *For N discrete random variables $X_1, \cdots, X_N$ with joint distribution $p(x_1, \cdots, x_N)$,*

$$C_1 = C(X_1, \cdots, X_N). \tag{2.24}$$

The proof of Theorem 3 is a direct extension of the proof for two discrete random variables in [4] and hence is omitted.

The above theorem stated that the common information of N discrete random variables is the minimum common information rate R_0 needed for the generalized Gray-Wyner network to recover the intended sources *losslessly* while keeping the total rate close to the entropy bound. The counterpart to this is the lossy source coding interpretation applicable to the Gray-Wyner network with discrete and continuous alphabet source sequences. This will be explored in Chapter 3. In the following, we explore another operational meaning of Wyner's common information for multiple variables, namely that of distribution approximation.

2.3.2 Distribution approximation interpretation

The second approach of interpreting the common information of N discrete random variable uses distribution approximation. Let $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ be i.i.d. copies of $\mathbf{X}$ with distribution $p(\mathbf{x}) = p(x_1, \cdots, x_N)$, i.e., the joint distribution for $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ is

$$p^{(n)}(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \prod_{k=1}^{n} p(\mathbf{x}_k). \tag{2.25}$$

An (n, M, Δ) *generator* consists of the following:

- a message set $\mathcal{W}$ with cardinality M;
- for all $w \in \mathcal{W}$, probability distributions $q_i^{(n)}(x_i^n|w)$, for $i = 1, 2, \cdots, N$.

Define the probability distribution on $\mathcal{X}_1^n \times \mathcal{X}_2^n \times \cdots \times \mathcal{X}_N^n$

$$q^{(n)}(\mathbf{x}_1, \cdots, \mathbf{x}_n) = \sum_{w \in \mathcal{W}} \frac{1}{M} \prod_{i=1}^{N} q_i^{(n)}(x_i^n|w). \tag{2.26}$$

Let

$$\Delta = D_n\left(q^{(n)}(\mathbf{x}_1, \cdots, \mathbf{x}_n); p^{(n)}(\mathbf{x}_1, \cdots, \mathbf{x}_n)\right) = \sum_{\mathbf{x}^n} \frac{1}{n} q^{(n)}(\mathbf{x}_1, \cdots, \mathbf{x}_n) \log \frac{q^{(n)}(\mathbf{x}_1, \cdots, \mathbf{x}_n)}{p^{(n)}(\mathbf{x}_1, \cdots, \mathbf{x}_n)}, \tag{2.27}$$

where $p^{(n)}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ is defined in (2.25) and $q^{(n)}(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ is defined as in (2.26).

A number R is said to be *achievable* if for all $\epsilon > 0$, if for n sufficiently large there exists an (n, M, Δ) generator with $M \leq 2^{nR}$ and $\Delta \leq \epsilon$. Define C_2 as the infimum of all achievable R.

Theorem 4. *For N discrete random variables $X_1, \cdots, X_N$ with joint distribution $p(x_1, \cdots, x_N)$,*

$$C_2 = C(X_1, \cdots, X_N). \tag{2.28}$$

The proof can be constructed in the same way as that of [4, Theorems 5.2 and 6.2], hence is omitted.

2.4 Summary

This chapter generalized Wyner's common information, defined for a pair of discrete random variables, to that of N random variables with arbitrary alphabets. Common information for N random variables is defined through a conditional independence

structure which is equivalent to the Markov chain condition for two dependent variables. We established a monotone property of Wyner's common information in the number of variables. We also introduced the coding theorems related to the lossless and lossy source coding problems for the generalized Gray-Wyner network involving multiple dependent random sequences.

We then established that the common information of N random variables is the minimum common information rate R_0 needed for the generalized Gray-Wyner network to recover the intended sources losslessly while keeping the total rate close to the entropy bound. It is also equal to the smallest rate of the common input to N independent processors (random number generators) such that the output distribution is arbitrarily close to the specified joint distribution.

Chapter 3

Lossy Source Coding Interpretation of Wyner's Common Information

The common information defined in (2.1) equally applies to that of continuous random variables. Such definitions are only meaningful when they are associated with concrete operational interpretations. However, the interpretations provided in the previous chapter only apply to discrete random variables. In this chapter, we will provide an operational meaning for the common information of continuous random variables.

This new interpretation is motivated by the connection between the common information of discrete random variables and the losslessly source coding of Gray-Wyner network. For continuous random variables, it is quite natural to explore the connection between the common information and the lossy source coding of Gray-Wyner network in a manner analogous to that of the discrete counterpart. This is

the focus of the present chapter.

Specifically, we will show that for the Gray-Wyner network, Wyner's common information is precisely the smallest common message rate for a certain range of distortion constraints when the total rate is arbitrarily close to the rate distortion function with joint decoding. As the common information is only a function of the joint distribution, this smallest common rate remains constant even if the distortion constraints vary, as long as they are in a specific distortion region. While this new interpretation holds for the general case of N dependent random variable, we elect to present coding theorems involving only a pair of dependent variables for ease of notion and presentation.

The rest of this chapter is organized as follows. Section 3.1 reviews the concepts of joint, marginal and conditional rate distortion functions and their relations. Section 3.2 develops a lossy source coding interpretation of the common information utilizing the Gray-Wyner network. Section 3.3 uses two examples, the doubly symmetric binary source and the bivariate Gaussian source, to illustrate the lossy source coding interpretation of Wyner's common information. The common information for bivariate Gaussian source and its extension to the multi-variate case is also derived in Section 3.3. Section 3.4 concludes this chapter.

3.1 Joint, marginal and conditional rate distortion functions

The rate distortion function of a source represents the minimum rate required to describe the source within a fidelity criterion. In this section, we review the joint, marginal and conditional rate distortion functions and their relations. These results will be used to show our main result in subsequent sections. Two-dimensional sources will be considered in this section and the results can be generalized immediately to N-dimensional vector sources.

3.1.1 Definitions

The joint, marginal and conditional rate functions are defined as follows. Given a two-dimensional source (X_1, X_2) with probability distribution $p(x_1, x_2)$ and two distortion measures $d_1(x_1, \hat{x}_1)$ and $d_2(x_2, \hat{x}_2)$ defined on $\mathcal{X}_1 \times \hat{\mathcal{X}}_1$ and $\mathcal{X}_2 \times \hat{\mathcal{X}}_2$, the joint rate distortion function of (X_1, X_2) is given by

$$R_{X_1X_2}(D_1, D_2) = \min_{p_t(\hat{x}_1\hat{x}_2|x_1x_2)} I(X_1, X_2; \hat{X}_1, \hat{X}_2), \tag{3.1}$$

where the random variables $\hat{X}_1, \hat{X}_2$ are defined by the test channels $p_t(\hat{x}_1\hat{x}_2|x_1x_2)$ and the minimum is taken over all test channels $p_t(\hat{x}_1\hat{x}_2|x_1x_2)$ such that $Ed_1(X_1, \hat{X}_1) \leq D_1$, $Ed_2(X_2, \hat{X}_2) \leq D_2$. Similarly, the marginal rate distortions are defined by

$$R_{X_1}(D_1) = \min_{p_t(\hat{x}_1|x_1):Ed_1(X_1,\hat{X}_1)\leq D_1} I(X_1; \hat{X}_1), \tag{3.2}$$

$$R_{X_2}(D_2) = \min_{p_t(\hat{x}_2|x_2):Ed_2(X_2,\hat{X}_2)\leq D_2} I(X_2; \hat{X}_2). \tag{3.3}$$

Given a two-dimensional source (X, Y) with probability distribution $p(x, y)$ and distortion measure $d(x, \hat{x})$ on $\mathcal{X} \times \hat{\mathcal{X}}$, we can define the conditional rate distortion

function as

$$R_{X|Y}(D) = \min_{p_t(\hat{x}|x,y):Ed(X,\hat{X})\leq D} I(X;\hat{X}|Y), \tag{3.4}$$

where the expectations are over both X and Y. The conditional rate distortion function of X given Y is the rate needed to transmit the source X within a fidelity criterion when Y is available both at the encoder and the decoder. More detailed discussions can be found in [27].

3.1.2 Rate distortion function relations

In this section, we discuss the properties and relations among joint, marginal and conditional rate distortion functions.

Lemma 4. *[28, 29] Given a two-dimensional source (X_1, X_2) with joint distribution $p(x_1, x_2)$ and two distortion measures $d_1(x_1, \hat{x}_1)$, $d_2(x_2, \hat{x}_2)$ defined respectively on $\mathcal{X}_1 \times \hat{\mathcal{X}}_1$ and $\mathcal{X}_2 \times \hat{\mathcal{X}}_2$, the rate distortion functions satisfy the following inequalities*

$$R_{X_1X_2}(D_1, D_2) \geq R_{X_1|X_2}(D_1) + R_{X_2}(D_2), \tag{3.5a}$$

$$R_{X_1|X_2}(D_1) \geq R_{X_1}(D_1) - I(X_1; X_2), \tag{3.5b}$$

$$R_{X_1X_2}(D_1, D_2) \geq R_{X_1}(D_1) + R_{X_1}(D_2) - I(X_1; X_2). \tag{3.5c}$$

$$R_{X_1}(D_1) \geq R_{X_1|X_2}(D_1), \tag{3.6a}$$

$$R_{X_1}(D_1) + R_{X_2}(D_2) \geq R_{X_1X_2}(D_1, D_2). \tag{3.6b}$$

Sufficient conditions for equality in (3.5) are that the optimum backward test channels for the functions on the left side of each equation factor appropriately, i.e., for (3.5a) $p_b(x_1x_2|\hat{x}_1\hat{x}_2) = p(x_1|\hat{x}_1x_2)p(x_2|\hat{x}_2)$, for (3.5b) $p_b(x_1|\hat{x}_1x_2) = p(x_1|\hat{x}_1)$ and for (3.5c)

that $p_b(x_1x_2|\hat{x}_1\hat{x}_2) = p(x_1|\hat{x}_1)p(x_2|\hat{x}_2)$. *Equalities hold in (3.6)* if and only if X_1 *and* X_2 *are independent.*

Lemma 4 shows that the relations among rate distortion functions are analogous to that of the corresponding entropies. Specifically, the entropies have the following relations:

$$H(X_1, X_2) = H(X_1|X_2) + H(X_2) \tag{3.7a}$$

$$H(X_1|X_2) = H(X_1) - I(X_1; X_2) \tag{3.7b}$$

$$H(X_1, X_2) = H(X_1) + H(X_2) - I(X_1; X_2) \tag{3.7c}$$

$$H(X_1) \geq H(X_1|X_2) \tag{3.8a}$$

$$H(X_1) + H(X_2) \geq H(X_1X_2), \tag{3.8b}$$

with equality in (3.8) *if and only if* X_1 and X_2 are independent. Therefore, (3.6) is exactly the same as (3.8) and (3.5) is resemblance to (3.7) except that the rate distortion function relations involve inequalities in stead of equalities [28].

Furthermore, Gray has shown that under quite general conditions, equalities hold in (3.5) for small values of distortion. This is because the marginal, joint and conditional rate distortion functions equal to their Extended Shannon Lower Bounds (ESLB) [27, 28] under suitable conditions. These ESLB, denoted by $R_X^{(L)}(D)$ for a rate distortion function $R_X(D)$, satisfy the property stated in Lemma 5.

We use the following notations in Lemma 5. Denote by $\mathcal{D}$ a surface in the m-dimensional space and Δ an m-dimensional vector. The inequality $\Delta \leq \mathcal{D}$ means that there exists a vector $\beta \in \mathcal{D}$ such that $\Delta \leq \beta$. If there is no such a vector, $\Delta > \mathcal{D}$. Likewise, $\mathcal{D}_1 \leq \mathcal{D}_2$ means that $\beta \leq \mathcal{D}_2$ for any $\beta \in \mathcal{D}_1$ [28].

Lemma 5. *[28] Given a two-dimensional source (X_1, X_2) with joint distribution $p(x_1, x_2)$ such that for $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$, $p(x_2|x_1) > 0$, reproduction alphabets $\hat{\mathcal{X}}_1 = \mathcal{X}_1$, $\hat{\mathcal{X}}_2 = \mathcal{X}_2$ and two per-letter distortion measures $d_1(x_1, \hat{x}_1)$ and $d_2(x_2, \hat{x}_2)$ that satisfy*

$$d_i(x_i, \hat{x}_i) \;>\; d_i(x_i, x_i) = 0, x_i \neq \hat{x}_i, i = 1, 2, \tag{3.9}$$

there exist strictly positive surfaces $\mathcal{D}(X_1X_2)$, $\mathcal{D}(X_1|X_2)$, $\mathcal{D}(X_1)$ and $\mathcal{D}(X_2)$ such that

$$\begin{aligned}
&R_{X_1X_2}(D_1, D_2) = R^{(L)}_{X_1X_2}(D_1, D_2), && \text{if } (D_1, D_2) \leq \mathcal{D}(X_1X_2),\\
&R_{X_1|X_2}(D_1) = R^{(L)}_{X_1|X_2}(D_1), && \text{if } D_1 \leq \mathcal{D}(X_1|X_2),\\
&R_{X_1}(D_1) = R^{(L)}_{X_1}(D_1), && \text{if } D_1 \leq \mathcal{D}(X_1),\\
&R_{X_2}(D_2) = R^{(L)}_{X_2}(D_2), && \text{if } D_2 \leq \mathcal{D}(X_2),
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{D}(X_1|X_2) &\leq \mathcal{D}(X_1),\\
\mathcal{D}(X_1X_2) &\leq (\mathcal{D}(X_1|X_2), \mathcal{D}(X_2)) \leq (\mathcal{D}(X_1), \mathcal{D}(X_2)).
\end{aligned}$$

Finally,

$$\begin{aligned}
R^{(L)}_{X_1X_2}(D_1, D_2) &= R^{(L)}_{X_1|X_2}(D_1) + R^{(L)}_{X_2}(D_2), && (3.10)\\
&= R^{(L)}_{X_1}(D_1) + R^{(L)}_{X_2}(D_2) - I(X_1; X_2). && (3.11)
\end{aligned}$$

It is apparent that when the rate distortion functions equal to their corresponding ESLB, equations (3.10) and (3.11) imply equalities in (3.5a)-(3.5c). Therefore, given a two-dimensional source satisfying the conditions in Lemma 5, there exists a strictly positive surface $\mathcal{D}(X_1X_2)$ such that (3.5a)-(3.5c) of Lemma 4 hold with equality if $(D_1, D_2) \leq \mathcal{D}(X_1X_2)$.

3.2 A lossy source coding interpretation of common information

In this section, we develop a lossy source coding interpretation of Wyner's common information using the Gray-Wyner network.

In the previous chapter, a quantity C_1 with respect to the lossless source coding of Gray-Wyner network was defined and was shown to be equivalent to the common information of discrete random variables. A natural approach for continuous random variables is thus to examine the Gray-Wyner network with lossy source coding to see if an analogous interpretation exists. In this section, we first define a quantity $C_3(D_1, D_2)$ with respect to the lossy source coding of Gray-Wyner network in a way similar to that of C_1 for a two-dimensional source. We then establish the connection between $C_3(D_1, D_2)$ and the common information which provides a lossy source coding interpretation for the common information of arbitrary alphabets.

While the result in this section is based on a pair of random variables for ease of notation, the conclusion extends to multiple random variables directly.

3.2.1 Common message rate of lossy Gray-Wyner source coding

First, we give the definition of $C_3(D_1, D_2)$. Given a two-dimensional source $(X_1, X_2) \sim p(x_1, x_2)$, for any $(D_1, D_2) \geq 0$, a number R_0 is said to be (D_1, D_2)-*achievable* if for any $\epsilon > 0$, there exists, for n sufficiently large, an $(n, M_0, M_1, M_2, \Delta_1, \Delta_2)$ code (as

defined in Section 2.2 for $N = 2$) with

$$M_0 \leq 2^{nR_0}, \tag{3.12}$$

$$\textstyle\sum_{i=0}^{2} \frac{1}{n} \log M_i \leq R_{X_1X_2}(D_1, D_2) + \epsilon, \tag{3.13}$$

$$\Delta_1 \leq D_1 + \epsilon, \Delta_2 \leq D_2 + \epsilon. \tag{3.14}$$

$C_3(D_1, D_2)$ is defined as the infimum of all R_0's that are (D_1, D_2)-achievable. Thus, $C_3(D_1, D_2)$ is the minimum common message rate for the Gray-Wyner network with sum rate $R_{X_1X_2}(D_1, D_2)$ while satisfying the distortion constraint. Since $R_{X_1X_2}(D_1, D_2)$ is always (D_1, D_2)-achievable, it is obvious that

$$C_3(D_1, D_2) \leq R_{X_1X_2}(D_1, D_2). \tag{3.15}$$

The following theorem gives a precise characterization of $C_3(D_1, D_2)$.

Theorem 5.

$$C_3(D_1, D_2) = \tilde{C}(D_1, D_2), \tag{3.16}$$

where $\tilde{C}(D_1, D_2)$ is the solution to the following optimization problem:

$$\begin{aligned} \inf \quad & I(X_1, X_2; W) \\ \textit{subject to} \quad & R_{X_1|W}(D_1) + R_{X_2|W}(D_2) + I(X_1, X_2; W) = R_{X_1X_2}(D_1, D_2). \end{aligned} \tag{3.17}$$

Proof. See Appendix A. ■

The authors in [30] gave an alternative characterization of $C_3(D_1, D_2)$. Define

$$C^*(D_1, D_2) = \inf I(X_1, X_2; W),$$

where the infimum is taken over all joint distributions for $X_1, X_2, X_1^*, X_2^*, W$ such that

$$X_1^* - W - X_2^*, \tag{3.18}$$

$$(X_1, X_2) - (X_1^*, X_2^*) - W, \tag{3.19}$$

where (X_1^*, X_2^*) achieves $R_{X_1X_2}(D_1, D_2)$. It was shown in [30] that $C_3(D_1, D_2) = C^*(D_1, D_2)$. This, combined with Theorem 5, establishes that

$$\tilde{C}(D_1, D_2) = C^*(D_1, D_2). \tag{3.20}$$

$\tilde{C}(D_1, D_2)$ is derived from the rate distortion region $\mathcal{R}_2(D_1, D_2)$ given in Theorem 2 while the authors in [30] chose to derive $C^*(D_1, D_2)$ from an alternative characterization of $\mathcal{R}_2(D_1, D_2)$ given in [31]. In Appendix B, we provide a direct proof of (3.20) for completeness. Also, as given in Appendix B, a necessary condition for the equality condition in the optimization problem (3.17) is

$$R_{X_1X_2|W}(D_1, D_2) = R_{X_1|W}(D_1) + R_{X_2|W}(D_2). \tag{3.21}$$

3.2.2 Lossy source coding interpretation

Given our characterization of $C_3(D_1, D_2)$ in Theorem 5, we now establish its connection with $C(X_1, X_2)$ which leads to a new interpretation of Wyner's common information. We begin with the following two lemmas.

Lemma 6. *Let W be the random variable that achieves the common information of X_1 and X_2. If*

$$R_{X_1X_2|W}(D_1, D_2) + C(X_1, X_2) = R_{X_1X_2}(D_1, D_2),$$

then

$$C_3(D_1, D_2) \leq C(X_1, X_2). \tag{3.22}$$

Lemma 6 is a direct consequence of Theorem 5 as the Markov chain $X_1 - W - X_2$ implies $R_{X_1X_2|W}(D_1, D_2) = R_{X_1|W}(D_1) + R_{X_2|W}(D_2)$. Thus, the equality constraint in (3.17) is satisfied. Inequality (3.22) follows as

$$C_3(D_1, D_2) = \tilde{C}(D_1, D_2) \leq I(X_1, X_2; W) = C(X_1, X_2).$$

The next lemma gives a sufficient condition under which $C_3(D_1, D_2) \geq C(X_1, X_2)$.

Lemma 7. *For any distortion pair (D_1, D_2), if the rate distortion function satisfies*

$$R_{X_1X_2}(D_1, D_2) = R_{X_1}(D_1) + R_{X_2}(D_1) - I(X_1; X_2), \tag{3.23}$$

then we have

$$C_3(D_1, D_2) \geq C(X_1, X_2).$$

Proof. See Appendix C. ■

Lemmas 6 and 7, together with the relations of marginal, joint and conditional rate distortion functions described in Lemmas 4 and 5, allow us to determine a region where $C_3(D_1, D_2) = C(X_1, X_2)$ as stated in Theorem 6.

Theorem 6. *Let random variables X_1, X_2 be distributed as $p(x_1, x_2)$ on $\mathcal{X}_1 \times \mathcal{X}_2$ such that for $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$, $p(x_2|x_1) > 0$. Let the reproduction alphabets $\hat{\mathcal{X}}_1 = \mathcal{X}_1$, $\hat{\mathcal{X}}_2 = \mathcal{X}_2$. The two per-letter distortion measures $d_1(x_1, \hat{x}_1)$, $d_2(x_2, \hat{x}_2)$ satisfy*

$$d_i(x_i, \hat{x}_i) \;>\; d_i(x_i, x_i) = 0, \quad x_i \neq \hat{x}_i, i = 1, 2. \tag{3.24}$$

Then there exists a strictly positive surface $\gamma \triangleq (\gamma_1, \gamma_2)$ *such that, for* $(D_1, D_2) \leq \gamma$,

$$C_3(D_1, D_2) = C(X_1, X_2). \tag{3.25}$$

Proof. See Appendix D. ■

Theorem 6 shows that Wyner's common information is precisely the smallest common message rate $C_3(D_1, D_2)$ of Gray-Wyner network for a certain range of distortion constraints when the total rate is arbitrarily close to the rate distortion function with joint decoding. As the common information is only a function of the joint distribution, hence is fixed for a given $p(x_1, x_2)$, it is surprising that the smallest common rate $C_3(D_1, D_2)$ remains constant even if the distortion constraints vary, as long as they are in a specific distortion region.

3.2.3 Discussions

While Theorem 6 establishes that $C_3(D_1, D_2) = C(X_1, X_2)$ for $(D_1, D_2) \leq \gamma$, it does not specify the value of the positive distortion vector γ. Let $\mathcal{D}^c \triangleq (D_1^c, D_2^c)$ be the two-dimensional distortion surface such that $R_{X_1X_2}(D_1^c, D_2^c) = C(X_1, X_2)$, then we must have that $\gamma \leq \mathcal{D}^c$. This is because if $\gamma > \mathcal{D}^c$, then there exists (D_1, D_2) such that $\gamma \geq (D_1, D_2) > \mathcal{D}^c$ and $C_3(D_1, D_2) \leq R_{X_1X_2}(D_1, D_2) < R_{X_1X_2}(D_1^c, D_2^c) = C(X_1, X_2)$, which contradicts Theorem 6. Now let us consider a particular point on the surface $\mathcal{D}^c$.

Lemma 8. *Let* W *be an auxiliary random variable that achieves* $C(X_1, X_2)$. *Suppose there exists a distortion pair* (D_1^0, D_2^0) *satisfying, for* $i = 1, 2$,

$$\begin{aligned} R_{X_i}(D_i^0) &= I(X_i; W), \\ D_i^0 &= \inf_{\hat{x}_i(w)} Ed_i(X_i, \hat{X}_i^0(W)), \end{aligned} \tag{3.26}$$

where $\hat{x}_1^0(w), \hat{x}_2^0(w)$ are deterministic functions. Then, we have

$$C_3(D_1^0, D_2^0) = R_{X_1X_2}(D_1^0, D_2^0) = C(X_1, X_2) = I(X_1, X_2; W). \tag{3.27}$$

It is apparent that (D_1^0, D_2^0) is on the distortion surface $\mathcal{D}^c$.

Proof. To prove (3.27), we first show that $R_{X_1X_2}(D_1^0, D_2^0) = I(X_1, X_2; W)$. From the definition of (D_1^0, D_2^0) in (3.26), we have

$$R_{X_1X_2}(D_1^0, D_2^0) \geq R_{X_1}(D_1^0) + R_{X_2}(D_2^0) - I(X_1; X_2) = I(X_1, X_2; W), \tag{3.28}$$

where the first inequality is from (3.5c). On the other hand,

$$R_{X_1X_2}(D_1^0, D_2^0) \leq I(X_1, X_2; \hat{X}_1^0, \hat{X}_2^0) \leq I(X_1, X_2; W). \tag{3.29}$$

Therefore, $R_{X_1X_2}(D_1^0, D_2^0) = I(X_1, X_2; \hat{X}_1^0, \hat{X}_2^0) = I(X_1X_2; W)$.

Furthermore, we can show

$$C_3(D_1^0, D_2^0) = C(X_1, X_2), \tag{3.30}$$

using Lemma 7 and the fact that $C_3(D_1^0, D_2^0) \leq R_{X_1X_2}(D_1^0, D_2^0)$. ■

This means that given total rate of $C(X_1, X_2)$, the optimal scheme for the Gray-Wyner network to transmit the pair of sources (X_1^n, X_2^n) within distortion constraints (D_1^0, D_2^0) is to communicate W to the two receivers using the common channel.

Let us now decrease the distortion constraints from (D_1^0, D_2^0) to $(D_1, D_2) \leq (D_1^0, D_2^0)$. The question is whether the rate $C(X_1, X_2)$ is still $(D_1, D_2)-$achieveble, i.e., if it is possible to transmit the sources (X_1^n, X_2^n) with smaller distortions (D_1, D_2) with the sum rate at $R_{X_1X_2}(D_1, D_2)$ while keeping the common rate at $C(X_1, X_2)$.

In the following, we identify a sufficient condition for $C_3(D_1, D_2) = C(X_1, X_2)$ for such (D_1, D_2) pair for successively refinable sources. A source X with distortion measure $d(x, \hat{x})$ is said to be successively refinable from a coarser distortion δ_1 to a finer distortion δ_2 ($\delta_1 \geq \delta_2$) if it can be encoded in two stages in which the optimal descriptions at the second stage is a refinement of the optimal descriptions at the first stage [32]. Similar definition can be applied to vector sources with individual distortion constraints and the details can be found in [33].

Theorem 7. *Let W be the auxiliary variable that achieves $C(X_1, X_2)$ and (D_1^0, D_2^0) be a distortion pair satisfying (3.26). If the source (X_1, X_2) is successively refinable from (D_1^0, D_2^0) to (D_1, D_2) for $(D_1, D_2) \leq (D_1^0, D_2^0)$, and X_i is successively refinable from D_i^0 to D_i for $D_i \leq D_i^0$, $i = 1, 2$, then,*

$$C_3(D_1, D_2) = C(X_1, X_2),$$

for any $(D_1, D_2) \leq (D_1^0, D_2^0)$.

Proof. See Appendix E. ■

Theorem 7 gives a sufficient condition under which $C_3(D_1, D_2) = C(X_1, X_2)$ for any $(D_1, D_2) \leq (D_1^0, D_2^0)$. This sufficient condition ensures the optimality of a two-stage encoding scheme in the Gray-Wyner network: first encode the common message with rate $C(X_1, X_2)$ to obtain a coarse distortion (D_1^0, D_2^0), then encode the two private messages with rates $R_{X_1|W}(D_1)$ and $R_{X_2|W}(D_2)$. The successive refinement assumption guarantees that the two-step approach can achieve the distortion (D_1, D_2) and the sum rate does not exceed the total rate $R_{X_1X_2}(D_1, D_2)$.

In the following section, we will consider two examples involving successively refinable sources: the binary random variables and bivariate Gaussian variables. For

these two cases, we compute explicitly the function $C_3(D_1, D_2)$ and establish its connection with $C(X_1, X_2)$. The distortion pair (D_1^0, D_2^0) satisfying (3.26) are identified for both cases, providing practical significance to Theorem 7.

3.3 Common information for two examples

The computation of common information for a given source is known to be a challenging problem even for the original definition of a pair of dependent discrete random variables. To date, only several special cases have been resolved [4, 6].

We have generalized the definition of Wyner's common information to that involving multiple dependent variables of arbitrary alphabets, it is natural to consider the computation of the common information for the generalized setting. In this section, we discuss the computation of the common information for two cases: the binary and the Gaussian sources. For both cases, we evaluate the common information involving multiple variables. In particular, we derive, through an estimation theoretic approach, the common information for a bivariate Gaussian source and its extension to the multi-variate case with a certain correlation structure. In addition, we characterize the distortion region where the common information equals to the smallest common message rate in the Gray-Wyner network for both cases.

3.3.1 Binary random variables

In this section, we consider the common information of multiple exchangeable binary random variables.

Let S be a Bernoulli random variable with successive probability β for $0 \leq \beta \leq 1$,

denoted from now on as $S \sim Bernoulli(\beta)$, i.e., $S \in \{0,1\}$ and $P(S=1)=\beta$. Let X_i, $i=1,\cdots,N$, be the output of a Binary Symmetric Channel (BSC) with crossover probability a_1 ($0 \le a_1 \le \frac{1}{2}$) and with S as input. The BSCs are independent of each other. Thus, we have the conditional distribution of the output

$$p(x_1,\cdots,x_N|s) = \prod_{i=1}^{N} p(x_i|s), \tag{3.31}$$

where each term $p(x_i|s)$ is a BSC channel with

$$p(x_i|s) = \begin{cases} 1-a_1, & \text{if } x_i = s, \\ a_1, & \text{otherwise,} \end{cases} \tag{3.32}$$

for $x_i \in \{0,1\}$. Therefore, the joint distribution of $(X_1, X_2, \cdots, X_N)$ is given by

$$\begin{aligned} p(x_1,x_2,\cdots,x_N) &= \sum_{s\in\{0,1\}} p(s) \prod_{i=1}^{N} p(x_i|s), & (3.33)\\ &= \beta a_1^{t_N}(1-a_1)^{N-t_N} + (1-\beta)(1-a_1)^{t_N} a_1^{N-t_N}, & (3.34) \end{aligned}$$

where $t_N = \sum_{i=1}^{N} x_i$.

For $N=2$, the joint distribution of (X_1, X_2) is given by the following probability matrix,

$$\begin{bmatrix} \beta(1-a_1)^2 + (1-\beta)a_1^2 & a_1(1-a_1) \\ a_1(1-a_1) & \beta a_1^2 + (1-\beta)(1-a_1)^2 \end{bmatrix}. \tag{3.35}$$

It has been shown by Witsenhausen [6] that for the binary source (X_1, X_2) with joint distribution of the form (3.35), the common information is achieved with W being S with the corresponding common information

$$C(X_1,X_2) = I(X_1X_2;S) = H(X_1,X_2) - 2h(a_1), \tag{3.36}$$

where $h(\cdot)$ is the binary entropy function, defined as $h(a_1) = -a_1 \log(a_1) - (1 - a_1)\log(1-a_1)$. Note that when $\beta = \frac{1}{2}$, (X_1, X_2) is a Doubly Symmetric Binary Source (DSBS) whose common information was also derived by Wyner using a different approach in [4].

We now compute the common information for N variables.

Proposition 1. *Let $S \sim$ Bernoulli(β) and let X_i, $i = 1, \cdots, N$, be the output of independent BSCs with common input S and crossover probability $0 \leq a_1 \leq 1/2$. Then for any $N \geq 2$, the common information for $X_1, \cdots, X_N$ is given as*

$$C(X_1, \cdots, X_N) = I(X_1, \cdots, X_N; S). \tag{3.37}$$

Proof. That $C(X_1, \cdots, X_N) \leq I(X_1, \cdots, X_N; S)$ follows from the definition of the common information (2.1). The inequality $C(X_1, \cdots, X_N) \geq I(X_1, \cdots, X_N; S)$ can be proved by contradiction. Suppose there exists a W such that

$$C(X_1, \cdots, X_N) = I(X_1, \cdots, X_N; W) < I(X_1, \cdots, X_N; S), \tag{3.38}$$

i.e., $C(X_1, \cdots, X_N)$ is achieved by W and it is strictly less than $I(X_1, \cdots, X_N; S)$. Since W induces conditional independence of $X_1, \cdots, X_N$, we have, from (3.38),

$$\sum_{i=1}^{N} H(X_i|W) > \sum_{i=1}^{N} H(X_i|S). \tag{3.39}$$

Thus, there must exist two random variables X_k, X_j, $k, j \in \{1, \cdots, N\}$ such that

$$H(X_k|W) + H(X_j|W) > H(X_k|S) + H(X_j|S). \tag{3.40}$$

Given that the sequence $\{X_1, \cdots, X_N\}$ is exchangeable (See Definition 3 in Chapter 4), $p(x_k, x_j)$ has the same joint distribution as $p(x_1, x_2)$. Thus,

$$C(X_1, X_2) = C(X_k, X_j) = I(X_k, X_j; W) < I(X_k, X_j; S) = I(X_1, X_2; S). \tag{3.41}$$

This, however, contradicts the fact that S achieves $C(X_1, X_2)$. Thus the proposition is proved. ■

We now study the asymptotic value of the common information for binary sources with distribution defined in (3.34). First, we note that if $a_1 = 1/2$, $X_1, X_2, \cdots, X_N$ are mutually independent for any N, so $C(X_1, X_2, \cdots, X_N) = 0$.

Corollary 1. *For $a_1 < 1/2$,*

$$\lim_{N\to\infty} C(X_1, X_2, \cdots, X_N) = H(S) \tag{3.42}$$

Proof. From Proposition 1, $C(X_1, \cdots, X_N) = I(X_1, \cdots, X_N; S)$. First we have

$$I(X_1, \cdots, X_N; S) = H(S) - H(S|X_1, \cdots, X_N) \leq H(S), \tag{3.43}$$

for any N. On the other hand, for any ϵ, it can be established that

$$H(S|X_1, \cdots, X_N) < \epsilon, \tag{3.44}$$

for N sufficiently large. This is because by the law of large number, one can construct an estimate of S using $X_1, \cdots, X_N$ that converges to S in probability. Hence inequality (3.44) is a consequence of applying Fano's inequality. Thus, we have

$$\lim_{N\to\infty} C(X_1, X_2, \cdots, X_N) = H(S). \tag{3.45}$$

■

We now characterize the minimum common rate $C_3(D_1, D_2)$ for a DSBS.

Proposition 2. *Consider a DSBS (X_1, X_2) with distribution*

$$p(x_1, x_2) = \begin{cases} \frac{1}{2}(1 - a_0), & \text{if } x_1 = x_2, \\ \frac{1}{2}a_0, & \text{otherwise,} \end{cases} \tag{3.46}$$

where, without loss of generality, $0 \leq a_0 \leq 1/2$. *Let* a_1 *be such that* $a_0 = 2a_1(1 - a_1), 0 \leq a_1 \leq 1/2$. *With Hamming distortion* $d_1 = d_2 = d_H$, *we have*

$$C_3(D_1, D_2) = \begin{cases} C(X_1, X_2), & (D_1, D_2) \in \mathcal{E}_{10}, \\ R_{X_1X_2}(D_1, D_2), & (D_1, D_2) \in \mathcal{E}_2 \cup \mathcal{E}_3, \\ 0, & (D_1, D_2) \geq (\frac{1}{2}, \frac{1}{2}), \end{cases} \tag{3.47}$$

$$C(X_1, X_2) \leq C_3(D_1, D_2) \leq R_{X_1X_2}(D_1, D_2), \quad (D_1, D_2) \in \mathcal{E}_{11}, \tag{3.48}$$

where

$$\begin{aligned} \mathcal{E}_{10} &= \{(D_1, D_2) : 0 \leq D_i \leq a_1, i = 1, 2\}, \\ \mathcal{E}_{11} &= \mathcal{E}_{10}^c \cap \{(D_1, D_2) : D_1 + D_2 - 2D_1D_2 \leq a_0\}, \\ \mathcal{E}_2 &= \mathcal{E}_{10}^c \cap \mathcal{E}_{11}^c \cap \left\{(D_1, D_2) : \max\left\{\tfrac{D_1 - D_2}{1 - 2D_2}, \tfrac{D_2 - D_1}{1 - 2D_1}\right\} \leq a_0\right\}, \\ \mathcal{E}_3 &= \mathcal{E}_{10}^c \cap \mathcal{E}_{11}^c \cap \mathcal{E}_2^c \cap \left\{(D_1, D_2) : D_i \leq \tfrac{1}{2}, i = 1, 2\right\}. \end{aligned} \tag{3.49}$$

Proof. For $X_i \sim Bernoulli(1/2), i = 1, 2$ with Hamming distortion, the rate distortion function is

$$R_{X_i}(D_i) = \begin{cases} 1 - h(D_i), & 0 \leq D_i \leq \frac{1}{2}, \\ 0, & D_i \geq \frac{1}{2}. \end{cases} \tag{3.50}$$

The joint rate distortion function of the DSBS (X_1, X_2) is given by [33]

$$R_{X_1X_2}(D_1, D_2) = \begin{cases} 1 + h(a_0) - h(D_1) - h(D_2), & (D_1, D_2) \in \mathcal{E}_1, \\ 1 - (1 - a_0)h\left(\frac{D_1 + D_2 - a_0}{2(1 - a_0)}\right) - a_0 h\left(\frac{D_1 - D_2 + a_0}{2a_0}\right), & (D_1, D_2) \in \mathcal{E}_2, \\ 1 - h\left(\min\{D_1, D_2\}\right), & (D_1, D_2) \in \mathcal{E}_3. \end{cases} \tag{3.51}$$

where $\mathcal{E}_1 = \mathcal{E}_{10} \cup \mathcal{E}_{11}$ with $\mathcal{E}_{10}, \mathcal{E}_{11}, \mathcal{E}_2$ and $\mathcal{E}_3$ defined in (3.49). Therefore, for this DSBS, $R_{X_1}(D_1) + R_{X_2}(D_2) - I(X_1; X_2) = R_{X_1X_2}(D_1, D_2)$, for $(D_1, D_2) \in \mathcal{E}_1$. From

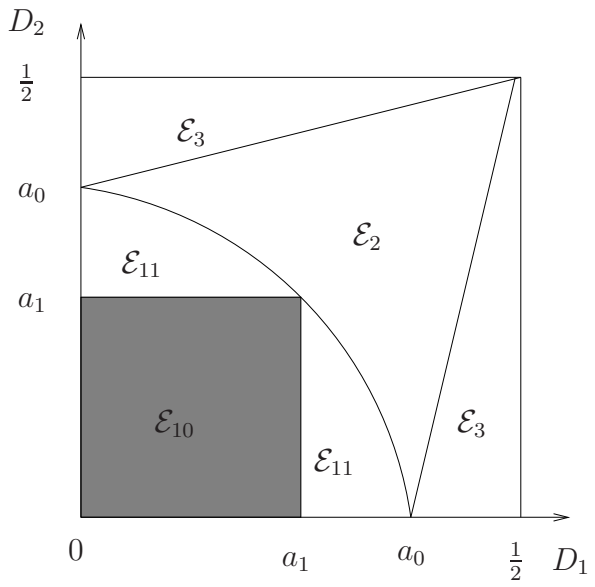


Figure 3.1: The distortion regions $\mathcal{E}_{10}, \mathcal{E}_{11}, \mathcal{E}_2$ and $\mathcal{E}_3$ for the DSBS. $C_3(D_1, D_2) = C(X_1, X_2)$ in the shaded region.

Lemma 7, we have for $(D_1, D_2) \in \mathcal{E}_1$,

$$C_3(D_1, D_2) \geq C(X_1, X_2). \tag{3.52}$$

On the other hand, the conditional rate distortion function $R_{X_i|S}(D_i)$, $i = 1, 2$, is given by [28]

$$R_{X_i|S}(D_i) = \begin{cases} h(a_1) - h(D_i), & 0 \leq D_i \leq a_1, \\ 0, & D_i \geq a_1. \end{cases} \tag{3.53}$$

Therefore, $R_{X_1|S}(D_1) + R_{X_2|S}(D_2) + I(X_1, X_2; S) = R_{X_1X_2}(D_1, D_2)$ is satisfied for $(D_1, D_2) \in \mathcal{E}_{10}$. By Theorem 5, $C_3(D_1, D_2) \leq C(X_1, X_2)$ for $(D_1, D_2) \in \mathcal{E}_{10}$. Together with (3.52) and given that $\mathcal{E}_{10} \subset \mathcal{E}_1$, we have proved that for $(D_1, D_2) \in \mathcal{E}_{10}$,

$$C_3(D_1, D_2) = C(X_1, X_2). \tag{3.54}$$

For $(D_1, D_2) \in \mathcal{E}_2$, we only need to show that $C_3(D_1, D_2) \geq R_{X_1X_2}(D_1, D_2)$. It was

shown in [33] that the backward test channel that achieves $R_{X_1X_2}(D_1, D_2)$ is given by

$$\begin{aligned} X_1 &= \hat{X}_1 + Z_1, \\ X_2 &= \hat{X}_2 + Z_2, \end{aligned} \tag{3.55}$$

where both $\hat{X}_1, \hat{X}_2$ and Z_1, Z_2 are binary vectors independent of each other with the probability mass functions given respectively as

$$P_{\hat{X}_1\hat{X}_2} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \qquad P_{Z_1Z_2} = \frac{1}{2}\begin{bmatrix} 2 - a_0 - D_1 - D_2 & D_2 - D_1 + a_0 \\ D_1 - D_2 + a_0 & D_1 + D_2 - a_0 \end{bmatrix}. \tag{3.56}$$

Therefore, $(\hat{X}_1, \hat{X}_2)$ that achieves $R_{X_1X_2}(D_1, D_2)$ satisfies

$$\hat{X}_2 = \hat{X}_1. \tag{3.57}$$

For the characterization $C^*(D_1, D_2)$ of $C_3(D_1, D_2)$, any W satisfying the Markov chain $\hat{X}_1 - W - \hat{X}_1$ must satisfy $H(\hat{X}_1|W) = 0$. Thus, $\hat{X}_1$ is a function of W and we have

$$I(X_1, X_2; W) = I(X_1, X_2; W, \hat{X}_1) \geq I(X_1, X_2; \hat{X}_1) = R_{X_1X_2}(D_1, D_2). \tag{3.58}$$

Therefore, $C_3(D_1, D_2) = R_{X_1X_2}(D_1, D_2)$.

The region $\mathcal{E}_3$ is a degenerated one. For example, $R_{X_1X_2}(D_1, D_2) = R_{X_1}(D_1)$ if $a_0 < \frac{D_2 - D_1}{1 - 2D_1}$ and $D_i \leq \frac{1}{2}, i = 1, 2$. This implies that the optimal coding scheme is to ignore X_2 and optimally compress X_1. Then $\hat{X}_2$ can be estimated from $\hat{X}_1$ with distortion less than D_2. The case of $a_0 < \frac{D_1 - D_2}{1 - 2D_2}$ is dealt with similarly. Hence, similar to the region $\mathcal{E}_2$, $C_3(D_1, D_2) = R_{X_1X_2}(D_1, D_2)$. ■

The characterization of $C_3(D_1, D_2)$ is plotted in Fig. 3.1 as a function of the distortion constraints. $C_3(D_1, D_2) = C(X_1, X_2)$ in the shaded region. For the symmetric

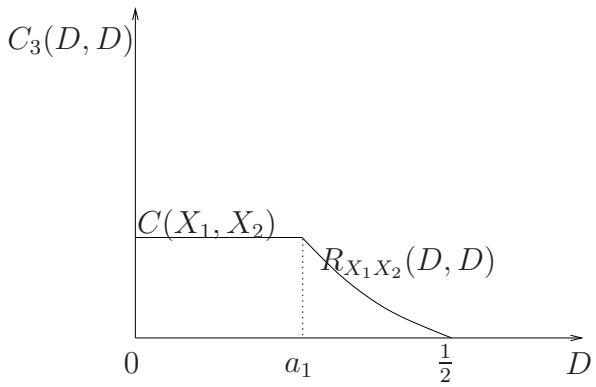


Figure 3.2: The relationship between $C_3(D,D)$ and D for the DSBS with $D_1 = D_2 = D$.

distortion constraint, $D_1 = D_2 = D$, the relation of $C_3(D,D)$ and D for the DSBS is given in Fig. 3.2. For the DSBS, $C_3(D,D)$ is a constant and equals to the common information as long as $D \leq a_1$.

The claim $C_3(D_1,D_2) = C(X_1,X_2)$ for $(D_1,D_2) \in \mathcal{E}_{10}$ can also be proved using Theorem 7. $R_{X_1X_2}(a_1,a_1)$ is achieved by the backward test channel $p_b(x_1,x_2|s) = p(x_1|s)p(x_2|s)$. The vector source (X_1,X_2) is successively refinable for any $(D_1,D_2) \leq (a_1,a_1)$ [33] and the scalar source X_i is successively refinable for any $D_i \leq a_1$, $i = 1,2$ [32]. Thus by Theorem 7, $C_3(D_1,D_2) = C(X_1,X_2)$ for $(D_1,D_2) \leq (a_1,a_1)$.

For the DSBS, not only the common information $C(X_1,X_2)$ is (D_1,D_2)-achievable for any $(D_1,D_2) \leq (a_1,a_1)$, but also the rate $R_{X_1X_2}(D_1',D_2')$ is also (D_1,D_2)-achievable for any (D_1',D_2') satisfying $(D_1,D_2) \leq (D_1',D_2') \leq (a_1,a_1)$.

This can be shown as follows. The backward test channel that achieves $R_{X_1X_2}(D_1',D_2')$ can be decomposed as $p_b(x_1,x_2|\hat{x}_1\hat{x}_2) = p_b(x_1|\hat{x}_1)p_b(x_2|\hat{x}_2)$ where

$$p_b(x_i|\hat{x}_i) = \begin{cases} 1 - D_i', & \text{if } x_i = \hat{x}_i, \\ D_i', & \text{Otherwise.} \end{cases} \tag{3.59}$$

for $i = 1, 2$. Then for $(D_1, D_2) \leq (D_1', D_2') \leq (a_1, a_1)$, let the rates R_0, R_1, R_2 in Theorem 2 be

$$\begin{aligned} R_0 &= R_{X_1X_2}(D_1', D_2') && (3.60)\\ &= 1 + h(a_0) - h(D_1') - h(D_2'), && (3.61)\\ R_i &= R_{X_i|\hat{X}_1\hat{X}_2}(D_i) && (3.62)\\ &= R_{X_i|\hat{X}_i}(D_i) && (3.63)\\ &= h(D_i') - h(D_i), i = 1, 2, && (3.64) \end{aligned}$$

where (3.63) is because of the Markov chain $X_i - \hat{X}_i - \hat{X}_1\hat{X}_2$. Since R_0, R_1 and R_2 in (3.61) and (3.64) sum up to $R_{X_1X_2}(D_1, D_2)$, $R_{X_1X_2}(D_1', D_2')$ is (D_1, D_2)-achievable.

3.3.2 Gaussian random variables

In this section we consider bivariate Gaussian random variables X_1, X_2 with zero mean and covariance matrix

$$K_2 = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{3.65}$$

The common information between this pair of Gaussian random variables is given in the following theorem.

Theorem 8. *For two joint Gaussian random variables X_1, X_2 with covariance matrix K_2, the common information is*

$$C(X_1, X_2) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}. \tag{3.66}$$

Proof. See Appendix F. ■

As the common information of (X_1, X_2) is only a function of the correlation coefficient ρ, we consider, without loss of generality, the covariance matrix

$$K_2' = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{3.67}$$

The above result generalizes to multi-variate Gaussian random variables satisfying a certain covariance matrix structure, the proof of which can be constructed in a similar fashion.

Corollary 2. *For N joint Gaussian random variables $X_1, X_2, \cdots, X_N$ with covariance matrix K_N,*

$$K_N = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \cdot & \cdot & \cdots & \cdot \\ \rho & \rho & \cdots & 1 \end{bmatrix}, \tag{3.68}$$

the common information is

$$C(X_1, \cdots, X_N) = \frac{1}{2} \log \left(1 + \frac{N\rho}{1-\rho} \right). \tag{3.69}$$

We now characterize the minimum common rate $C_3(D_1, D_2)$ in the Gray-Wyner lossy source coding network for bivariate Gaussian random variables with covariance matrix K_2' in equation (3.67). It was shown in [30] that for symmetric distortion, i.e., $D_1 = D_2 = D$,

$$C_3(D, D) = \begin{cases} C(X_1, X_2), & 0 \le D \le 1-\rho, \\ R_{X_1X_2}(D, D), & 1-\rho \le D \le 1, \\ 0, & D \ge 1. \end{cases} \tag{3.70}$$

We characterize $C_3(D_1, D_2)$ for general distortion (D_1, D_2) in the following proposition.

Proposition 3. *For bivariate Gaussian random variables X_1, X_2 with zero mean, covariance matrix K_2' and squared error distortion, we have that*

$$C_3(D_1, D_2) = \begin{cases} C(X_1, X_2), & (D_1, D_2) \in \mathcal{D}_{10}, \\ R_{X_1X_2}(D_1, D_2), & (D_1, D_2) \in \mathcal{D}_2 \cup \mathcal{D}_3, \\ 0, & (D_1, D_2) \geq (1, 1), \end{cases} \tag{3.71}$$

$$C(X_1, X_2) \leq C_3(D_1, D_2) \leq R_{X_1X_2}(D_1, D_2), \quad (D_1, D_2) \in \mathcal{D}_{11}, \tag{3.72}$$

where

$$\begin{aligned} \mathcal{D}_{10} &= \{(D_1, D_2) : 0 \leq D_i \leq 1 - \rho, i = 1, 2\}, \\ \mathcal{D}_{11} &= \mathcal{D}_{10}^c \cap \{(D_1, D_2) : D_1 + D_2 - D_1D_2 \leq 1 - \rho^2\}, \\ \mathcal{D}_2 &= \mathcal{D}_{10}^c \cap \mathcal{D}_{11}^c \cap \left\{(D_1, D_2) : \min\left\{\tfrac{1-D_1}{1-D_2}, \tfrac{1-D_2}{1-D_1}\right\} \geq \rho^2\right\}, \\ \mathcal{D}_3 &= \mathcal{D}_{10}^c \cap \mathcal{D}_{11}^c \cap \mathcal{D}_2^c \cap \{(D_1, D_2) : D_i \leq 1, i = 1, 2\}. \end{aligned} \tag{3.73}$$

Proof. The joint rate distortion function for Gaussian random variables with squared error distortion [33–35] is given by

$$R_{X_1X_2}(D_1, D_2) = \begin{cases} \frac{1}{2}\log\frac{1-\rho^2}{D_1D_2}, & (D_1, D_2) \in \mathcal{D}_1, \\ \frac{1}{2}\log\frac{1-\rho^2}{D_1D_2 - \left(\rho - \sqrt{(1-D_1)(1-D_2)}\right)^2}, & (D_1, D_2) \in \mathcal{D}_2, \\ \frac{1}{2}\log\frac{1}{\min\{D_1, D_2\}}, & (D_1, D_2) \in \mathcal{D}_3, \end{cases} \tag{3.74}$$

where $\mathcal{D}_1 = \mathcal{D}_{10} \cup \mathcal{D}_{11}$. The marginal rate distortion function for $X_i \sim \mathcal{N}(0, 1), i = 1, 2$, is

$$R_{X_i}(D_i) = \begin{cases} \frac{1}{2}\log\frac{1}{D_i}, & 0 \leq D_i \leq 1, \\ 0, & D_i \geq 1. \end{cases} \tag{3.75}$$

Therefore, $R_{X_1}(D_1) + R_{X_2}(D_2) - I(X_1; X_2) = R_{X_1X_2}(D_1, D_2)$, for $(D_1, D_2) \in \mathcal{D}_1$. From Lemma 7, for $(D_1, D_2) \in \mathcal{D}_1$,

$$C_3(D_1, D_2) \geq C(X_1, X_2). \tag{3.76}$$

On the other hand, the random variable W in the following decomposition of X_1 and X_2 achieves the common information

$$X_i = \sqrt{\rho}W + \sqrt{1-\rho}N_i, \ i = 1, 2. \tag{3.77}$$

where W, N_1, N_2 are mutually independent standard Gaussian random variables. The conditional distribution of X given W is Gaussian distribution with variance $1 - \rho$. Hence, for $i = 1, 2$, the conditional rate distortion function is

$$R_{X_i|W}(D_i) = \begin{cases} \frac{1}{2}\log\frac{1-\rho}{D_i}, & 0 \leq D_i \leq 1-\rho, \\ 0, & D_i \geq 1-\rho. \end{cases} \tag{3.78}$$

The condition $R_{X_1|W}(D_1) + R_{X_2|W}(D_2) + I(X_1, X_2; W) = R_{X_1X_2}(D_1, D_2)$ is satisfied for $(D_1, D_2) \in \mathcal{D}_{10}$. From Theorem 5, $C_3(D_1, D_2) \leq C(X_1, X_2)$ for $(D_1, D_2) \in \mathcal{D}_{10}$. Since, $\mathcal{D}_{10} \in \mathcal{D}_1$, we proved that for $(D_1, D_2) \in \mathcal{D}_{10}$,

$$C_3(D_1, D_2) = C(X_1, X_2). \tag{3.79}$$

For $(D_1, D_2) \in \mathcal{D}_2$, it was shown in [33] that the pair $(\hat{X}_1, \hat{X}_2)$ achieving $R_{X_1X_2}(D_1, D_2)$ satisfies

$$\hat{X}_2 = \sqrt{\frac{1-D_2}{1-D_1}}\hat{X}_1. \tag{3.80}$$

Hence, using the characterization $C^*(D_1, D_2)$, it is easy to show that the W satisfying the Markov chains (3.18) and (3.19) must satisfy two Markov chains

$$X_1X_2 - \hat{X}_1 - W - \hat{X}_2, \tag{3.81}$$

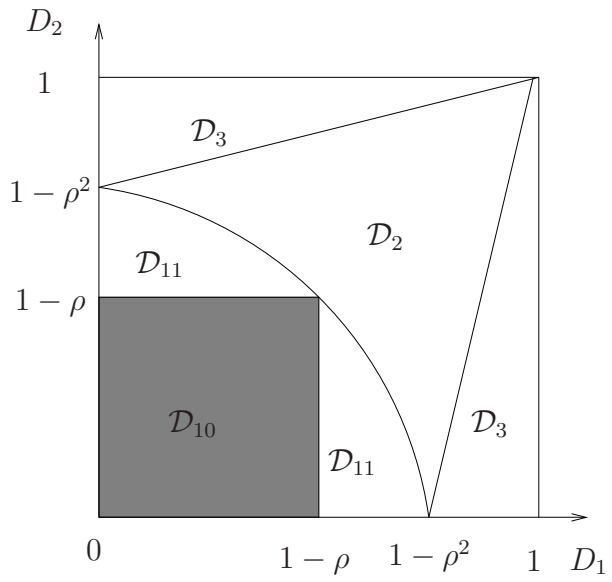


Figure 3.3: The distortion regions $\mathcal{D}_{10}, \mathcal{D}_{11}, \mathcal{D}_2$ and $\mathcal{D}_3$ for bivariate Gaussian random variables. $C_3(D_1, D_2) = C(X_1, X_2)$ in the shaded region.

$$X_1 X_2 - \hat{X}_2 - W - \hat{X}_1. \tag{3.82}$$

Therefore, we have

$$I(X_1, X_2; W) = I(X_1, X_2; \hat{X}_1) = I(X_1, X_2; \hat{X}_1, \hat{X}_2), \tag{3.83}$$

which proved $C_3(D_1, D_2) = R_{X_1 X_2}(D_1, D_2)$.

The region $\mathcal{D}_3$ is a degenerated one. For example, $R_{X_1 X_2}(D_1, D_2) = R_{X_1}(D_1)$ if $\frac{1-D_2}{1-D_1} < \rho^2$, this means that the correlation between X_1 and X_2 is so strong that the optimal coding scheme is to encode X_1 to within distortion D_1 and ignore X_2. Then $\hat{X}_2$ can be estimated from $\hat{X}_1$. We have

$$\hat{X}_2 = \rho \hat{X}_1. \tag{3.84}$$

The case of $\frac{1-D_1}{1-D_2} < \rho^2$ is dealt with similarly. Hence, we have $C_3(D_1, D_2) = R_{X_1 X_2}(D_1, D_2)$. ■

The characterization of $C_3(D_1, D_2)$ is plotted in Fig. 3.3 as a function of the distortion constraints. $C_3(D_1, D_2) = C(X_1, X_2)$ in the shaded region.

Similar to the binary case, the claim $C_3(D_1, D_2) = C(X_1, X_2)$ for $(D_1, D_2) \in \mathcal{D}_{10}$ can also be proved using Theorem 7. This is because for the bivariate Gaussian random variables with covariance matrix K_2', $R_{X_1X_2}(1-\rho, 1-\rho)$ is achieved by the backward test channel $p_b(x_1, x_2|w) = p(x_1|w)p(x_2|w)$, (X_1, X_2) is successively refinable for any $(D_1, D_2) \leq (1-\rho, 1-\rho)$ [33] and X_i is successively refinable for $D_i \leq 1-\rho$, $i = 1, 2$ [32].

Let $(D_1, D_2) \leq (D_1', D_2') \leq (1-\rho, 1-\rho)$, then the rate $R_{X_1X_2}(D_1', D_2')$ is (D_1, D_2)-achievable in the Gray-Wyner network. This is because for $(D_1', D_2') \in \mathcal{E}_{10}$, the joint rate distortion function $R_{X_1X_2}(D_1', D_2')$ is achieved by Gaussian distributed $(\hat{X}_1, \hat{X}_2)$ satisfying $X_1 - \hat{X}_1 - \hat{X}_2 - X_2$ where the covariance matrix of $(\hat{X}_1, \hat{X}_2)$ is [33]

$$K_{\hat{X}_1\hat{X}_2} = \begin{bmatrix} 1-D_1' & \rho \\ \rho & 1-D_2' \end{bmatrix}. \tag{3.85}$$

Then for $(D_1, D_2) \leq (D_1', D_2') \leq (1-\rho, 1-\rho)$, let the rates R_0, R_1, R_2 in Theorem 2 be as follows:

$$\begin{aligned} R_0 &= R_{X_1X_2}(D_1', D_2') = \tfrac{1}{2}\log\tfrac{1-\rho^2}{D_1'D_2'}, \\ R_i &= R_{X_i|\hat{X}_1\hat{X}_2}(D_i) = R_{X_i|\hat{X}_i}(D_i) = \tfrac{1}{2}\log\tfrac{D_i'}{D_i}, i = 1, 2. \end{aligned} \tag{3.86}$$

R_0, R_1 and R_2 in (3.86) sum up to $R_{X_1X_2}(D_1, D_2)$, so $R_{X_1X_2}(D_1', D_2')$ is (D_1, D_2)-achievable.

Therefore, in the Gray-Wyner network, we can use the rate allocation in (3.86) to achieve the distortion $(D_1, D_2) \leq (1-\rho, 1-\rho)$ for any $(D_1, D_2) \leq (D_1', D_2') \leq (1-\rho, 1-\rho)$. The minimal R_0 satisfying (3.86) is exactly $C(X_1, X_2)$, which is achieved by letting $(D_1', D_2') = (1-\rho, 1-\rho)$.

3.4 Summary

In this chapter, we showed that, for lossy source coding using the Gray-Wyner network, Wyner's common information is precisely the smallest common message rate for a certain range of distortion constraints when the total rate is arbitrarily close to the rate distortion function with joint decoding. As the common information is only a function of the joint distribution, this smallest common rate remains constant even if the distortion constraints vary, as long as they are in a specific distortion region.

Furthermore, we have shown that for successive refinement sources, given a total rate of rate distortion function, it is optimal to use a two-stage encoding scheme in the Gray-Wyner network: first encode the common message with rate of common information to obtain a coarse distortion, then encode the two private messages with extra rates to achieve the desired distortion.

We also discussed the common information for two examples: the binary sources and Gaussian sources. For both cases, we evaluated the common information of multiple variables. In particular, we derived, through an estimation theoretic approach, the common information for a bivariate Gaussian source and its extension to the multi-variate case with a certain correlation structure. In addition, we characterized the distortion region where the common information equals to the smallest common message rate in the Gray-Wyner network for both cases.

Chapter 4

Common Information and Statistical Inference

4.1 Introduction

In Chapter 2, we have seen that the inclusion of an additional variable increases the common information. This seems to contradict the intuition that common information should decrease as the number of random variables increase as is the case for the generalization of Gács and Körner's common randomness [2, 36].

This monotonicity property of Wyner's common information motivates us to explore the application of it to inference problems: it is expected that any notion of information, if it is relevant to any inference problems, ought to be non-decreasing as more observations come in.

To further motivate our study, consider the following generalization of the DSBS. Let W be a Bernoulli random variable with successive probability β. Let the sequence

of observations $X_1, X_2, \cdots$, be generated from independent BSC with crossover probability not equal to 1/2. We have shown in Corollary 1 that

$$\lim_{n\to\infty} C(X_1, \cdots, X_n) = H(W). \tag{4.1}$$

That is , the common information asymptotically captures the entire information of the hidden variable W.

In addition, Wyner's common information satisfies both the data processing inequality and the additivity property:

- If $X - Y - Z$ forms a Markov chain, then $C(X, Z) \leq \min\{C(X, Y), C(Y, Z)\}$ [6].
- If $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ is an independent sequence of random variables pairs, then

$$C(X_1, \cdots, X_n, Y_1, \cdots, Y_n) = \sum_{i=1}^{n} C(X_i, Y_i). \tag{4.2}$$

These properties are necessary for any notion of information that is relevant to statistical inference.

A natural model that arises is the simple hierarchical model as in Fig. 4.1, where θ is a random variable with distribution $p(\theta)$ while the observations X_i, $i = 1, 2, \cdots, n$, are independent noisy realizations governed by transition probability $p_i(x_i|\theta)$. The binary example is a special case of this hierarchical model. Thus, the joint distribution of θ and $X_1, \cdots, X_n$ satisfies

$$p(\theta, x_1, \cdots, x_n) = p(\theta) \prod_{i=1}^{n} p_i(x_i|\theta). \tag{4.3}$$

This simple Bayesian model often arises in various inference problems where one wants to infer about θ with $p(\theta)$ serving as its prior while the observations $X_1, \cdots, X_n$

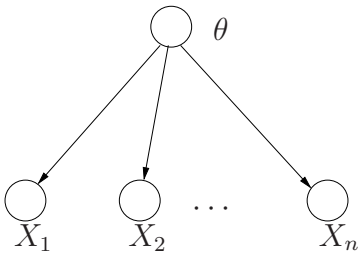


Figure 4.1: A simple Bayesian graphical model.

supply the samples for inference. The conditional independence assumption is directly motivated by the conditional independence structure in defining the common information.

Clearly, given the conditional independence assumption of the Bayesian model in (4.3), the dependence among the observations come entirely from the common variable θ. A natural question is that, for such a simple Bayesian inference problem, does the common information capture the entire information about θ that is contained in $X_1, \cdots, X_n$? Or, mathematically, does the equation

$$C(X_1, \cdots, X_n) = I(\theta; X_1, \cdots, X_n) \tag{4.4}$$

hold? Notice that since $I(\theta; X_1, \cdots, X_n)$ carries the meaning of uncertainty reduction in θ by observing $X_1, \cdots, X_n$, it can be interpreted as the information about θ contained in the data $X_1, \cdots, X_n$.

This is, however, not true in general. Consider a simple example where θ is $Bernoulli(1/2)$ and X_1 and X_2 are the output of two independent BSCs with crossover probabilities $\epsilon_1 = 1/2$ and $\epsilon_2 \neq 1/2$. Clearly, $C(X_1, X_2) = 0$ (since X_1 and (θ, X_2) are independent) but $I(X_1X_2; \theta) = I(X_2; \theta) = 1 - h(\epsilon_2) > 0$.

However, if we are to impose some symmetric condition in the model, i.e., if all X_i's are not only conditionally independent but are also identically distributed, stronger

and concrete connections between the common information and inference problems can be established.

With a symmetric model, the observations generated by the simple Bayseian model constitute an exchangeable sequence. We will show that for infinite exchangeable random variables, the common information is asymptotically equal to the information of the inference object θ. Such statement, however, is not true for finite n, i.e., $C(X_1, \cdots, X_n)$ is not always equal to $I(X_1, \cdots, X_n; \theta)$ even if the random variables are infinitely extendable, by which we mean a finite sequence that can be extended to an infinite exchangeable sequence. However, there exist some special cases, including both the binary and Gaussian cases, such that equality still holds for finite n. For these two cases, we will also establish concrete connections between common information and inference performance metrics.

The rest of this chapter is organized as follows. In Section 4.2, the results of common information for both infinite and finite exchangeable random variables are given. In Section 4.3, the connection between the common information and inference performance metrics are established for the binary and Gaussian cases. Section 4.4 concludes this chapter.

4.2 Common information for exchangeable random variables

We start by introducing the concept of exchangeable random variables.

4.2.1 Exchangeable random variables

Definition 3. *An exchangeable sequence of random variables is a finite or infinite sequence* $X_1, X_2, X_3, \cdots$ *of random variables such that for every permutation* σ *of the indices* $1, 2, 3, \cdots$*, the joint probability distribution of the permuted sequence*

$$X_{\sigma(1)}, X_{\sigma(2)}, X_{\sigma(3)}, \cdots$$

is the same as the joint probability distribution of the original sequence.

For finite n, $(X_1, \cdots, X_n)$ is called $n-$*exchangeable*. A $k-$exchangeable sequence $(X_1, \cdots, X_k)$ is $n-$*extendable* if it has the same distribution as the k first of an $n-$exchangeable variables $(X_1, \cdots, X_n)$. In particular, a $k-$exchangeable sequence is *infinite-extendable* if it has the same distribution as the k first variables of an infinite exchangeable sequence.

It was first shown by de Finetti, later generalized by Hewitt and Savage in [37] that any infinite exchangeable probability measure is a unique mixture of i.i.d product measures. Specifically, a sequence of random variables $X_1, X_2, X_3, \cdots$ is infinite exchangeable *if and only if* there exists $(p(\theta), p(x|\theta))$ such that, for any k,

$$p(x_1, \cdots, x_k) = \int_\theta \prod_{i=1}^{k} p(x_i|\theta) p(\theta) d\theta. \tag{4.5}$$

Furthermore, the $p(x|\theta)$ and $p(\theta)$ pair is unique for an infinite exchangeable sequence.

Clearly, from the definition of exchangeable sequences, if a finite exchangeable sequence $(X_1, \cdots, X_n)$ is infinitely extendable, then there must exist $(p(\theta), p(x|\theta))$ such that

$$p(x_1, \cdots, x_n) = \int_\theta \prod_{i=1}^{n} p(x_i|\theta) p(\theta) d\theta. \tag{4.6}$$

Any sequence that is generated according to a symmetric Baysian model is an infinite exchangeable sequence if it is infinite or an infinite extendable exchangeable sequence if it is finite. The converse is not true, however, for finite sequences. That is, if a sequence is finite exchangeable, there may not exist a $p(\theta)$ and $p(x|\theta)$ pair that gives rise to the sequence via a simple Baysian model.

As our intent is to study the relevance of the common information in inference problems, we consider in the following exchangeable sequences that are actually generated by a symmetric Bayesian model as described in Fig. 4.1. The inference problem is therefore to infer about θ given the sample $(X_1, \cdots, X_n)$. As the observations are independent conditional on θ, it is intuitive that dependence among X_i's comes solely from θ. Given that $I(\theta; X_1, \cdots, X_n)$ carries the interpretation of uncertainty reduction in θ by observing $(X_1, \cdots, X_n)$, one natural question is if $C(X_1, \cdots, X_n) = I(\theta; X_1, \cdots, X_n)$? As we shall see, this is not always the case and our endeavor is to identify conditions such that the equality holds.

4.2.2 Common information for infinite exchangeable sequences

For infinite exchangeable sequences, we have the following result.

Theorem 9. *Let $X_1, \cdots, X_n$ be generated by the Bayesian model described above, Asymptotically, we have*

$$\lim_{n\to\infty} C(X_1, \cdots, X_n) = \lim_{n\to\infty} I(X_1, \cdots, X_n; \theta). \tag{4.7}$$

In addition, if Θ is finite where Θ is the alphabet of θ, $\lim_{n\to\infty} C(X_1, \cdots, X_n) > 0$ and there exists a consistent estimator $\hat{\theta}(X_1, \cdots, X_n)$ of θ, then

$$\lim_{n\to\infty} C(X_1, \cdots, X_n) = H(\theta). \tag{4.8}$$

Proof. First, by the symmetry of the problem, the hidden variable W that achieves $\lim_{n\to\infty} C(X_1, \cdots, X_n)$ induces identical $p(x_i|w)$. By the de Finetti Theorem, the hidden variable W that induces conditional i.i.d. sequences $X_1, X_2, \cdots$ is unique and given the observations arise from the Bayesian model, we must have $W = \theta$, i.e., θ achieves the common information as $n \to \infty$.

Given that Θ is finite, $\lim_{n\to\infty} C(X_1, \cdots, X_n) > 0$ and there exists a consistent estimate of θ, $\hat{\theta}(X_1, \cdots, X_n)$, we have

$$\lim_{n\to\infty} P\{\hat{\theta}(X_1, \cdots, X_n) = \theta\} = 1, \forall\theta \in \Theta.$$

Hence, $H(\theta|\hat{\theta}(X_1, \cdots, X_n)) \to 0$. From the Markov chain

$$\theta - (X_1, \cdots, X_n) - \hat{\theta}(X_1, \cdots, X_n),$$

we have

$$\lim_{n\to\infty} H(\theta|X_1, \cdots, X_n) = 0. \tag{4.9}$$

■

Therefore, for infinite exchangeable sequences generated from a simple Bayesian model, the data informs about the unknown parameter perfectly. In such a case, the common information is precisely the amount of information in the Bayesian prior as defined using the Shannon entropy.

4.2.3 Common information for finite exchangeable sequences

While it is tempting to speculate that the same conclusion holds for the finite exchangeable sequence generated from a simple symmetric Bayesian model, the following example shows that this is not the case.

Example 1. *Let θ be uniformly distributed over $[0,1]$ and $p(x|\theta) = \theta^x(1-\theta)^{1-x}$ for $x \in \{0,1\}$, i.e., θ is the success probability for the Bernoulli random variable X. For $n = 2$, the joint distribution of X_1, X_2 is*

$$p(x_1, x_2) = \int_0^1 \theta^t(1-\theta)^{2-t} d\theta = B(t+1, 3-t), \tag{4.10}$$

where $t = x_1 + x_2$ and $B(a,b)$ is the beta function defined as

$$B(a,b) = \int_0^1 y^{a-1}(1-y)^{b-1} dy.$$

It is straightforward to simplify the joint distribution and obtain

$$p(x_1, x_2) = \frac{1}{3}\delta_{x_1,x_2} + \frac{1}{6}(1 - \delta_{x_1,x_2}), \tag{4.11}$$

where

$$\delta_{x_1,x_2} = \begin{cases} 1 & \text{if } x_1 = x_2, \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

This is a DSBS whose common information was given in Section 3.3.1. The hidden variable W that achieves the common information is a Bernoulli(1/2) random variable and X_1 and X_2 are connected to W through two independent BSCs with crossover probability $a_1 = 0.2113$, i.e., a_1 satisfies

$$a_1 * a_1 = 2a_1(1-a_1) = 1/3.$$

The common information can be computed to be $C(X_1, X_2) = I(X_1, X_2; W) = 0.430$ whereas $I(X_1, X_2; \theta) = 0.476$.

From the above example, it is clear that the same marginal distribution $p(x_1, x_2)$ may arise from two different Bayesian models. This, of course, does not contradict

de Finetti Theorem because of the finite number of samples. Indeed, the marginal distributions will diverge for these two models for $n > 3$. Another observation is that, for the model that the common information is indeed the same as mutual information between the prior and the observations, the observations are connected to the prior through an additive channel (modulo 2 sum). Indeed, this result can be generalized to any Bernoulli variables. Before proceeding, we first introduce the following lemma.

Lemma 9. *Let $(X_1, \cdots, X_m)$, $m > 2$, be an exchangeable sequence generated from a simple Bayesian model. Let W be a variable the induces conditional independence of $(X_1, \cdots, X_m)$. If W achieves the common information of $(X_1, \cdots, X_n)$ for any n with $m \geq n \geq 2$, it also achieves the common information of $(X_1, \cdots, X_m)$.*

Proof. We prove it by contradiction. Since W achieves the common information of $(X_1, \cdots, X_n)$,

$$C(X_1, \cdots, X_n) = I(X_1, \cdots, X_n; W).$$

Suppose that there exists another W' such that

$$C(X_1, \cdots, X_m) = I(X_1, \cdots, X_m; W') < I(X_1, \cdots, X_m; W), \tag{4.13}$$

i.e., $C(X_1, \cdots, X_m)$ is achieved by W' and is strictly less than $I(X_1, \cdots, X_m; W)$. Since W' induces conditional independence of $(X_1, \cdots, X_m)$, and by our assumption that W also induces conditional independence of $(X_1, \cdots, X_m)$, we have, from (4.13),

$$\sum_{i=1}^{m} H(X_i|W) < \sum_{i=1}^{m} H(X_i|W'). \tag{4.14}$$

Thus, there must exist a subset of $\{1, \cdots, m\}$ with size n, denoted by $\{k_1, \cdots, k_n\}$, such that

$$\sum_{i=1}^{n} H(X_{k_i}|W) < \sum_{i=1}^{n} H(X_{k_i}|W'). \tag{4.15}$$

Given that the sequence $(X_1, \cdots, X_m)$ is exchangeable, $p(x_{k_1}, \cdots, x_{k_n})$ has the same joint distribution as $p(x_1, \cdots, x_n)$. Thus,

$$C(X_1, \cdots, X_n) = C(X_{k_1}, \cdots, X_{k_n}). \tag{4.16}$$

This, however, contradicts the fact that W achieves $C(X_1, \cdots, X_n)$ since from (4.15),

$$I(X_{k_1}, \cdots, X_{k_n}; W') < I(X_{k_1}, \cdots, X_{k_n}; W). \tag{4.17}$$

Thus the lemma is proved. ■

Given the above lemma, we now examine two examples. We show that for the two special cases, the binary and the Gaussian variables, common information captures the entire information about the variables that represents the prior that generates the additive symmetric Bayesian model.

First, consider the exchangeable binary sequences obtained via an additive symmetric Bayesian model.

Proposition 4. *Let $\theta \sim Bernoulli(\beta)$ and let X_i, $i = 1, \cdots, n$ be the output of independent BSCs with common input θ and crossover probability not equal to $1/2$, then we have*

$$C(X_1, \cdots, X_n) = I(X_1, \cdots, X_n; \theta). \tag{4.18}$$

Proposition 4 is proved in Proposition 1 and it can also be proved by Lemma 9 using the fact that the common information of the source for $n = 2$ is achieved with W being θ.

A similar result can be obtained for exchangeable Gaussian random variables.

Proposition 5. *Let $Y \sim \mathcal{N}(0, P)$ and let Z_i, $i = 1, \cdots, n$ be i.i.d $\sim \mathcal{N}(0, \sigma^2)$. Let $X_i = Y + Z_i$. Then,*

$$C(X_1, \cdots, X_n) = I(X_1, \cdots, X_n; Y). \tag{4.19}$$

Proof. Notice that X_i's form an exchangeable sequence with pairwise correlation coefficient

$$\rho = \frac{P}{P + \sigma^2}. \tag{4.20}$$

Then by applying Corollary 1, the common information can be obtained to be

$$C(X_1, \cdots, X_n) = \frac{1}{2} \log \left(1 + \frac{n\rho}{1 - \rho} \right) = \frac{1}{2} \log \left(1 + \frac{nP}{\sigma^2} \right). \tag{4.21}$$

Straightforward calculation shows that

$$C(X_1, \cdots, X_n) = I(X_1, \cdots, X_n; Y). \tag{4.22}$$

■

4.3 Common information and inference

In this section, we will consider the Bayesian estimation problem. The inference goal is to estimate the variable θ in Fig. 4.1 that gives rise to the observation model. For the two special cases, namely the binary and the Gaussian cases, the common information captures the entire information about the hidden variable that generates the additive symmetric Bayesian model. As such, the common information is expected to be intimately related the inference performance of such models. In this section, we explore connections between the common information and the respective inference performance metrics.

For the binary case, it is natural to use error probability as a performance metric. Let θ be a Bernoulli(β) variable and X_is are the output of independent and identical BSC(a_1) with common input θ. Let $P_e^{(n)}$ be the minimum probability of error of estimating θ from observations $X_1, \cdots, X_n$. We have

Proposition 6.

$$H(\theta) - H(P_e^{(n)}) \leq C(X_1, \cdots, X_n) \leq H(\theta) - \frac{P_e^{(n)}}{2}. \tag{4.23}$$

Proof. Estimating θ is equivalent to testing a simple hypothesis

$$H_0 \ : \ \theta = 0$$

$$H_1 \ : \ \theta = 1$$

with prior probability β and $1-\beta$. For $\theta = 0$, $X_1, \cdots, X_n$ are i.i.d Bernoulli distributed with success probability a_1, while for $\theta = 1$, $X_1, \cdots, X_n$ are i.i.d Bernoulli distributed with success probability $1 - a_1$. Without loss of generality, assume $0 < a_1 < 1/2$. The likelihood ratio is

$$\frac{p(x_1, \cdots, x_n|\theta = 0)}{p(x_1, \cdots, x_n|\theta = 1)} = \frac{(1-a_1)^{t_n} a_1^{n-t_n}}{a_1^{t_n}(1-a_1)^{n-t_n}}.$$

where $t_n = sum_{i=1}^n x_i$. The *maximum a posteriori probability* detector that minimizes the probability of error is

$$\hat{\theta} = \begin{cases} 1 & \text{if } t_n > \frac{n}{2} + \frac{1}{2}\frac{\log \frac{\beta}{1-\beta}}{\log \frac{1-a_1}{a_1}}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.24}$$

It was shown by Rényi in [38] that for this test,

$$\frac{P_e^{(n)}}{2} \leq H(\theta|X_1, \cdots, X_n) \leq H(P_e^{(n)}). \tag{4.25}$$

where the second inequality is Fano's inequality. Together with (4.18), this proved (4.23). ■

It was shown in [38] that $\lim_{n\to\infty} P_e^{(n)} = 0$ *if and only if* $\lim_{n\to\infty} H(\theta|X_1, \cdots, X_n) = 0$. Therefore, $\lim_{n\to\infty} P_e^{(n)} = 0$ if and only if $\lim_{n\to\infty} C(X_1, \cdots, X_n) = H(\theta)$, which is consistent with Theorem 9.

Let us now consider the additive Gaussian model. For $i = 1, \cdots, n$,let $X_i = Y + Z_i$, where $Y \sim \mathcal{N}(0, P)$ and Z_i be i.i.d $\mathcal{N}(0, \sigma^2)$. The inference problem is to estimate Y using $X_1, \cdots, X_n$ and the performance metric is the usual mean squared error (MSE). We have

Proposition 7.

$$C(X_1, \cdots, X_n) = \frac{1}{2} \log \frac{P}{\mathcal{E}} \tag{4.26}$$

where

$$\mathcal{E} = \frac{P\sigma^2}{\sigma^2 + nP} \tag{4.27}$$

is the MMSE of estimating Y *using* $X_1, \cdots, X_n$.

Proof. Through direct computation. ■

Proposition 7 can also be proved using the following alternative approaches.

- The MMSE of estimating Y using $X_1, \cdots, X_n$ is bounded by [11]

$$\mathcal{E} \geq \frac{1}{2\pi e} e^{2h((Y|X_1, \cdots, X_n)} \tag{4.28}$$

  where equality is hold *if and only if* $\hat{Y} = E[Y|X_1, \cdots, X_n]$ and $Y, X_1, \cdots, X_n$ are jointly Gaussian. For the additive Gaussian model, the inequality is tight

and using (4.27) and the fact that $h(Y) = \frac{1}{2}\log(2\pi eP)$, we can directly obtain (4.26).

- From [39], the MMSE and the mutual information between input and the output of a Gaussian channel satisfies

$$\frac{dI(snr)}{snr} = \frac{1}{2}mmse(snr) \tag{4.29}$$

Since

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{4.30}$$

is a sufficient statistic for Y given $X_1, \cdots, X_n$, $I(Y;\bar{X}) = I(Y;X_1,\cdots,X_n)$. The channel from Y to $\bar{X}$ is an additive Gaussian channel with noise distributed according to $\mathcal{N}(0,\sigma^2/n)$. Using (4.29) and that Y achieves the common information of $X_1, \cdots, X_n$, we can obtain (4.26).

4.4 Summary

Motivated by the monotonicity property of common information with respect to the number of random variables, we explored the application of Wyner's common information to various inference problems. The inference problems considered in this chapter arise from symmetric simple Bayesian models. As such models give rise to exchangeable random variables, we studied the common information of exchangeable random variable. It was shown that for infinite exchangeable sequences, the common information is asymptotically equal to the information object, i.e., the hidden variable in the Bayesian model. For finite exchangeable sequences, while this result is

no longer true in general, we identify two important cases such that the result still holds. For these two cases, one binary and the other Gaussian, we further established relationship between the common information and that of the inference performance metrics.

Chapter 5

Distributed Detection with Dependent Observations

The second part of the thesis deals with decentralized inference with a particular emphasis on problems involving conditionally dependent observations. In the present chapter, we focus on the canonical distributed detection model shown in Fig. 5.1. A fusion center makes a decision regarding the hypothesis H using outputs $U_1, \cdots, U_K$ from the K sensors. Different from a centralized system, the observation at each sensor needs to be quantized separately prior to being sent to the fusion center to make a final decision. The reason for quantization is that the observations are typically collected remotely and the communication between the sensors and the fusion center may be severely bandlimited. Tremendous effort has been devoted to this problem that leads to many fundamental results (see [40–43] and references therein).

While the optimum fusion rule is known to be the likelihood-ratio test (LRT) at the fusion center [44–46], finding the optimal local senor decision rules is much more

challenging because of the distributed nature. Most of the results obtained are under the assumption that local sensor observations are conditionally independent given the underlying hypothesis, i.e., the joint distribution of the observations obeys

$$p(x_1, \cdots, x_K|H_l) = \prod_{k=1}^{K} p(x_k|H_l), l = 0, 1, \cdots, L-1. \tag{5.1}$$

Under this assumption the problem simplifies significantly: it was shown that the optimal local sensor decision are threshold quantizers that operate on the likelihood ratio of the observations under different scenarios [42,47,48]. Therefore, the problems reduces to finding the quantizer thresholds for which a person-by-person optimization (PBPO) methodology can be adopted [40]. In the simple case of a binary hypotheses testing (L=2) where each sensor sends a single bit to the fusion center, the optimal decision rule at each sensor is simply an LRT.

Without the conditional independence assumption, the problem of finding the optimal local decision rule becomes intractable in general. Such situation arises if one detects a random signal in independent noises or a deterministic signal in correlated noises. It was shown in [49] that the problem becomes NP complete when the observations are conditionally dependent. In such a case, the form of the optimal local decision rule is often unknown and is coupled with other sensor rules and the fusion rule. LRTs at local sensors often are not optimal even for the binary hypotheses and binary sensor output. For example, for the simple problem of two sensors observing a shift-in-mean correlated Gaussian random variables [50] [51], the optimality of LRT for this problem can only be established for certain parameter regions.

A potentially promising framework for distributed detection with dependent and independent observations is the HCI model [8]. The main idea is to inject a hidden

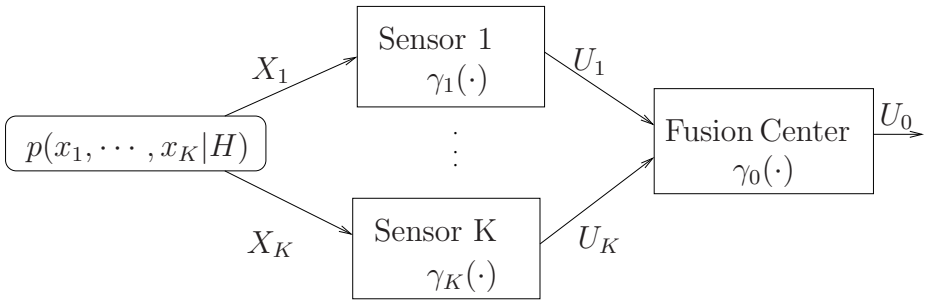


Figure 5.1: A canonical distributed detection system.

random variable **W** such that the sensor observations are conditionally independent with respect to this new variable regardless of the dependence structure of the original model. This new model unifies existing results for distributed detection with dependent observations and is also useful in solving new problems that otherwise seem quite formidable.

Two classes of distributed detection problem with dependent observations were identified in [8] whose optimal local decision rules are reminiscent in structure to the conditional independence case. When **W** has a finite alphabet, it was shown that the form of the optimal local decision rule can be determined independently of any other sensor decision rule and the fusion rule, and is essentially of the same form as that of the conditionally independent case. However, when **W** is a continuous random variable, it is not clear when the optimal decision structure resembles that of the conditional independence case. A sufficient condition was proposed in [8] under which single threshold quantizers at local sensors are optimal for binary hypothesis testing with binary sensor outputs. The utility of this condition can be demonstrated by its treatment of a previously known problem, namely the two sensor Gaussian problem [51].

However, this condition is limited in that it applies to the case where the final form of the test at local sensors amounts to directly quantizing the observations. In this chapter, we propose a more general sufficient condition such that single threshold quantizers of functions (i.e., statistics) of the observations at local sensors are optimal for binary hypotheses testing and binary sensor outputs. This includes the previous result as a special case. Furthermore, we illustrate the usefulness of this new result using the problem of detecting a random signal in independent noises. While the problem can be readily solved using the result of the present chapter, the original approach provided in [8] appears to be inadequate because of the restrictive conditions therein.

The rest of this chapter is organized as follows. Section 5.1 describes the problem of Bayesian distributed detection. Section 5.2 gives the hierarchical conditional independence model for Bayesian distributed inference problem. Section 5.3 generalized the sufficient condition under which single threshold quantizers are optimal for binary hypotheses testing and binary sensor outputs for a class of problems with dependent observations. An example of detection of random signals in independent noises is given in Section 5.4 to illustrate the usefulness of the obtained result. Section 5.5 concludes the chapter.

5.1 Bayesian distributed detection

Consider the parallel distributed M-ary hypothesis testing system with K sensors as in Fig.5.1. The observations at local sensors are denoted as X_k, $k = 1, \cdots, K$ and

their joint conditional density $p(x_1, \cdots, x_K|H)$[1], $H \in \{0, 1, \cdots, M-1\}$, is assumed known. Based on its own observation X_k, sensor k makes a local decision $U_k = \gamma_k(X_k) \in \{0, 1, \cdots, L-1\}$. Local decisions from all sensors are transmitted to the fusion center where a global decision is made $U_0 = \gamma_0(U_1, \cdots, U_K) \in \{0, 1, \cdots, M-1\}$. Let the prior probability of the hypothesis H be π_H. Denote by $\mathbf{X} = \{X_1, \cdots, X_K\}$, $\mathbf{x} = \{x_1, \cdots, x_K\}$.

In general, the variables involved in this model satisfy the following Markov chain

$$H - \mathbf{X} - \mathbf{U} - U_0, \tag{5.2}$$

and

$$p(\mathbf{u}|\mathbf{x}) = \prod_{k=1}^{K} p(u_k|x_k). \tag{5.3}$$

The objective of a Bayesian hypothesis testing problem for a parallel network is to obtain the set of decision rules $\{\gamma_0, \gamma_1, \cdots, \gamma_K\}$ that minimizes the expected Bayesian cost. Let $c_{u_0,h}$ be the Bayesian cost of deciding $U_0 = u_0$ when $H = h$ is true. Denote by $\mathbf{X}^k = \mathbf{X} \backslash X_k = \{X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_K\}$.

The average Bayesian cost C that needs to be minimized for the hypothesis testing problem is

$$C = \sum_{u_0=0}^{M-1} \sum_{h=0}^{M-1} c_{u_0,h} \pi_h p(u_0|h) \tag{5.4}$$

$$= \int_{\mathbf{X}} \sum_{\mathbf{u}} \sum_{u_0=0}^{M-1} \sum_{h=0}^{M-1} c_{u_0,h} \pi_h p(u_0|\mathbf{u}) p(\mathbf{u}|\mathbf{x}) p(\mathbf{x}|h) d\mathbf{x}, \tag{5.5}$$

[1]We use $p(\cdot)$ to denote both probability density function and probability mass function. Its meaning shall become clear in the context of where it appears.

where (5.5) is from the Markov chain $H - \mathbf{X} - \mathbf{U} - U_0$. Expanding C with respect to sensor k, we have

$$C = \int_{X_k} \sum_{u_k} p(u_k|x_k) f_k(u_k, x_k) dx_k, \tag{5.6}$$

where $f_k(u_k, x_k)$ is the Bayesian Cost Density Function (BCDF) for the kth sensor making decision u_k while observing x_k and is defined as

$$f_k(u_k, x_k) \triangleq \sum_{u_0=0}^{M-1} \sum_{h=0}^{M-1} \sum_{\mathbf{u}^k} c_{u_0,h} \pi_h p(u_0|\mathbf{u}^k, u_k) \int_{\mathbf{X}^k} p(\mathbf{u}^k|\mathbf{x}^k) p(\mathbf{x}^k, x_k|h) d\mathbf{x}^k \tag{5.7}$$

$$= \sum_{u_0=0}^{M-1} \sum_{h=0}^{M-1} c_{u_0,h} \pi_h p(x_k|h) p(u_0|u_k, x_k, h), \tag{5.8}$$

where (5.8) is because

$$p(u_0|u_k, x_k, h) = \sum_{\mathbf{u}^k} p(u_0|\mathbf{u}^k, u_k) \int_{\mathbf{X}^k} p(\mathbf{u}^k|\mathbf{x}^k) p(\mathbf{x}^k|x_k, h) d\mathbf{x}^k. \tag{5.9}$$

From equation (5.18), to minimize the expected Bayesian cost C, the optimal kth sensor decision rule given fixed decision rules at all other sensors and the fusion center is to make a decision u_k such that $f_k(u_k, x_k)$ is minimized. Since the BCDF $f_k(U_k, X_k)$ is coupled with the fusion rule $\gamma_0(\cdot)$ and other sensor decision rules $\gamma_i(\cdot)$, $i \neq k$, the problem of finding the optimal decision rule is difficult in general.

When the observations follow a Conditional Independence (CI) model where the observations at local sensors follow a joint distribution that satisfies

$$p(x_1, \cdots, x_K|H) = \prod_{k=1}^{K} p(x_k|H), H \in \{0, 1, \cdots, M-1\}, \tag{5.10}$$

then $p(u_0|u_k, x_k, h)$ in (5.8) reduces to $p(u_0|u_k, h)$ and the BCDF in (5.8) becomes to

$$f_k(u_k, x_k) = \sum_{h=0}^{M-1} \alpha_k(u_k, h) p(x_k|h), \tag{5.11}$$

where

$$\alpha_k(u_k, h) \triangleq \sum_{u_0=0}^{M-1} c_{u_0,h}\pi_h p(u_0|u_k, h), \tag{5.12}$$

is a scalar function of the sensor output $U_k = u_k$ and the underlying hypothesis $H = h$. The optimal decision rule at the kth sensor is

$$U_k = \gamma_k(X_k) = \arg\min_{u_k} \sum_{h=0}^{M-1} \alpha_k(u_k, h)p(x_k|h). \tag{5.13}$$

Thus, the optimal kth sensor decision rule $\gamma_k(X_k)$ reduces to an optimal M-ary Bayesian hypotheses test with Bayesian cost coefficients $\alpha_k(u_k, h)$ for M hypotheses and L possible decisions.

5.2 Hierarchical Conditional Independence (HCI) model

A HCI model is defined by introducing a new random variable $\mathbf{W}$ (can be a scalar variable or a vector variable) such that [8]

1. the following Markov chain holds

$$H - \mathbf{W} - \mathbf{X} - \mathbf{U} - U_0. \tag{5.14}$$

2. $X_1, \cdots, X_K$ are conditionally independent given $\mathbf{W}$, i.e.,

$$p(x_1, \cdots, x_K|\mathbf{w}) = \prod_{k=1}^{K} p(x_k|\mathbf{w}). \tag{5.15}$$

The inclusion of the hidden variable $\mathbf{W}$ induces conditional independence of the sensor observations with respect to this new variable regardless of the dependence

structure of the original model. Although it seems that the HCI model is less general, it has been shown in [8] that any distributed detection model in Fig.5.1 satisfying (5.2) can be represented as a HCI model and vice versa. Therefore, this HCI model provides a unified framework for analyzing distributed detection problems under various dependence assumptions. Besides, although this HCI model is proposed for distributed detection problem, it can be used for general decentralized inference problems.

This HCI model unifies existing results for distributed detection with dependent observations and is also useful in solving new problems that otherwise seem quite formidable. The HCI model can be classified into three categories according to the support set of $\mathbf{W}$: "Discrete" HCI (DHCI) model, "Continuous" HCI(CHCI) model and "Hybrid" HCI (HHCI) model.

Under the HCI model, the conditional distribution of $X_1, \cdots, X_k$ given H can be expanded as (assume $\mathbf{W}$ is of finite alphabet)

$$p(\mathbf{x}|H) = \sum_{\mathbf{w}} p(\mathbf{x}, \mathbf{w}|H) \tag{5.16}$$

$$= \sum_{\mathbf{w}} p(\mathbf{w}|H) \prod_{k=1}^{K} p(x_k|\mathbf{w}). \tag{5.17}$$

If $\mathbf{W}$ is of continuous alphabet, the summation in the above equations is replaced by integration.

5.2.1 DHCI model

Now let us consider the Bayesian hypothesis testing problem for the HCI model. By the discussion in Section 5.1, the average Bayesian cost C that needs to be minimized

for the hypothesis testing problem is

$$C = \int_{X_k} \sum_{u_k} p(u_k|x_k) f_k(u_k, x_k) dx_k, \tag{5.18}$$

Here $f_k(u_k, x_k)$ is the BCDF for the kth sensor making decision u_k while observing x_k and is equal to

$$f_k(u_k, x_k) = \sum_{u_0=0}^{M-1} \sum_{h=0}^{M-1} c_{u_0,h} \pi_h p(x_k|h) p(u_0|u_k, x_k, h). \tag{5.19}$$

For the DHCI model where W is a finite alphabet scalar variable, e.g., $W \in \{0, 1, \cdots, N-1\}$, by substituting (5.17) into (5.19), the BCDF can be simplified as:

$$f_k(u_k, x_k) = \sum_{w=0}^{N-1} \beta_k(u_k, w) p(x_k|w), \tag{5.20}$$

where

$$\beta_k(u_k, w) \triangleq \sum_{u_0=0}^{M-1} \sum_{h=0}^{M-1} c_{u_0,h} \pi_h p(u_0|u_k, w) p(w|h), \tag{5.21}$$

is a scalar function of the sensor output U_k and $W = w$. The optimal decision rule at the kth sensor thus is

$$U_k = \gamma(X_k) = \arg\min_{u_k} \sum_{y=0}^{N-1} \beta_k(u_k, w) p(x_k|w). \tag{5.22}$$

Similar to the CI case, $\gamma_k(X_k)$ at sensor k under the DHCI model is also an optimal multiple Bayesian hypotheses test of N hypotheses and L decisions with Bayesian cost coefficients $\beta_k(u_k, w)$ for $u_k = 0, \cdots, L-1, y = 0, \cdots, N-1$.

5.2.2 CHCI model

For the CHCI model when W is a continuous scalar random variable, similar to the DHCI model, we can have the BCDF as follows:

$$f_k(u_k, x_k) = \int_W \beta_k(u_k, w) p(x_k|w) dw \tag{5.23}$$

where $\beta_k(u_k, w)$ is similarly defined as in (5.21) except that $p(\cdot)$ now denotes pdf instead of pmf.

However, unlike the DHCI model, the BCDF for this model can not be described completely by a set of finite parameters. Hence, unlike the optimal design problem under the DHCI model, it is not clear when the optimal decision structure resembles that of the conditional independence case.

In [8], by imposing additional constraints on W for the CHCI model, or more specifically, on $p(x_k|w)$ and $p(w|h)$, a class of CHCI model was determined where the optimal local decision rules are the threshold quantizers of local observations. The result is given in the following.

Proposition 8. *[8, Propostion 1] Consider a distributed binary hypothesis testing system with scalar sensor observations and binary sensor outputs. Suppose that the distributed hypothesis testing problem is equivalent to a CHCI model where W is a scalar random variable, and the following three conditions are satisfied:*

1. *The fusion center implements a monotone fusion rule that satisfies*

$$P(U_0 = 1|U_k = 1, w) \geq P(U_0 = 1|U_k = 0, w);$$

2. *The ratio $\frac{p(w|H=1)}{p(w|H=0)}$ is a nondecreasing function of w;*

3. *The ratio $\frac{p(x_k|w)}{p(x'_k|w)}$ is also a nondecreasing function of w for any $x_k > x'_k$.*

Then there exists a single threshold quantizer at sensor k, i.e.,

$$U_k = \begin{cases} 1, & \text{if } X_k \geq \tau_k, \\ 0, & \text{if } X_k < \tau_k, \end{cases}$$

for some suitable τ_k, that minimizes the error probability P_e.

Proposition 8 provides a new tool in addressing distributed detection with dependent observations. For example, it provides a new approach to solve the problem of shift-in-mean dependent Gaussian random variables [8]. However, for many detection problems that require a test statistic that is not necessarily monotone in the observations, directly using Proposition 8 is futile. One can pivot such problems using transformed data but except for the case where the transformation is 1-1, further justification is needed to preserve the optimality.

5.3 More general condition for CHCI model

We generalize Proposition 8 in this section to include cases where quantization at local sensors may operate on a general statistic instead of directly on the data. Sufficient conditions are derived for such quantizers to be the optimal form of local sensor decision rules.

A real-parameter family of densities $p_\theta(w)$, is said to have Monotone Likelihood Ratio (MLR) in $T(w)$ if there exists a real-valued function $T(w)$ such that for any $\theta < \theta'$ the distributions p_θ and $p_{\theta'}$ are distinct and $\frac{p_{\theta'}(T(w))}{p_\theta(T(w))}$ is a nondecreasing function of $T(w)$ [52]. In Proposition 8, if we treat the hypothesis H as a parameter, the MLR is used in condition 2 for a special case: $T(w) = w$. We will extend the result where a general statistic $T(w)$ is used in the MLR. Correspondingly, the optimal quantizers will operate on some statistics instead of the original observations.

If the ratio $\frac{p(w|H=1)}{p(w|H=0)}$ is a nondecreasing function of $T(w)$, $T(W)$ is a sufficient statistic for H. If $T(w)$ is a one-to-one mapping of w, then the situation reduces to that considered in Proposition 8. We now consider the general case when $T(w)$ is

monotone in w in disjoint intervals.

Let the sample spaces of X_k and W be subsets of the real line, i.e., $\mathcal{X} \in \mathcal{R}$, $\mathcal{W} \in \mathcal{R}$, and let $A_1, \cdots, A_m$ are disjoint intervals of $\mathcal{R}$. Suppose the function $T(\cdot)$ on $\mathcal{R}$ is monotone in each interval, i.e.,

$$T(r) = T_i(r), \text{ for } r \in A_i$$

and $T_i(r)$ is monotone in r for $r \in A_i$.

To ease notation and presentation, we focus on binary hypothesis testing, scalar sensor observations and binary sensor outputs. The general M-ary hypotheses testing with $L > 2$ sensor outputs can be treated similarly. We have the following theorem.

Theorem 10. *Consider a distributed binary hypothesis testing system with scalar sensor observations and binary sensor outputs. Suppose that the distributed hypothesis testing problem is equivalent to a CHCI model where the hidden random variable Y is a scalar random variable. Furthermore,*

1. *The fusion center implements a monotone fusion rule that satisfies*

$$P(U_0 = 1|U_k = 1, w) \geq P(U_0 = 1|U_k = 0, w).$$

2. *The ratio $\frac{p(w|H=1)}{p(w|H=0)}$ is a nondecreasing function of $T(w)$, $Z_i(t) \triangleq T_i^{-1}(t)$ has a continuous derivative on t and the set $\mathcal{T}_w \triangleq \{t : t = T_i(w)\}$ for some $w \in A_i$ is the same for each $i = 1, \cdots, m$.*

3. *For any $S(x_k) > S(x'_k)$, $x_k, x'_k \in A_i$, $i = 1, \cdots, m$, $\frac{\lambda(x_k,t)}{\lambda(x'_k,t)}$ is a nondecreasing function of t for $t \in \mathcal{T}_w$, where*

$$\lambda(x_k, t) \triangleq \sum_{i=1}^{m} \left(p_{X_k|W}(x_k|Z_i(t)) \left| \tfrac{d}{dt} Z_i(t) \right| h(Z_i(t)) \right) \tag{5.24}$$

$$h(w) = (P(U_0 = 1|U_k = 1, w) - P(U_0 = 1|U_k = 0, w))\pi_1 p(w|H = 0) \quad (5.25)$$

and $S(\cdot)$ is a function on $\mathcal{R}$ such that $S(r)$ has the monotone property for $r \in A_i$.

Then there exists a single threshold quantizer at sensor k such that

$$U_k = \begin{cases} 1 & \text{if } S(x_k) \geq \tau_k \\ 0 & \text{otherwise} \end{cases} \quad (5.26)$$

for some suitable τ_k that minimizes the error probability P_e.

Proof. See Appendix G. ■

$S(X_k)$ may not be unique, i.e., there may be multiple functions that satisfy the conditions in the Theorem. Often times, setting $S(\cdot) = T(\cdot)$ may satisfy the specified conditions. Proposition 8 is a special case of Theorem 10 with $T(w) = w$, $S(x_k) = x_k$. For this case, $\lambda(x_k, t)$ reduces to $p(x_k|w)h(w)$ where $h(y)$ is positive. So $\frac{p(x_k|w)}{p(x'_k|w)}$ being nondecreasing of w for $x_k \geq x'_k$ is equivalent to that $\frac{\lambda(x_k,t)}{\lambda(x_k,t)}$ being nondecreasing of t for $S(x_k) > S(x'_k)$.

5.4 Detection of a random signal in Gaussian noise

Consider the detection of a common random signal S in Gaussian noise using K sensors. The observations at the kth sensor is

$$X_k = a_k S + N_k,$$

where a_k is a deterministic attenuation factor and N_k is the observation noise at the kth sensor with Gaussian distribution $N_k \sim N(0, \sigma^2)$. The hypotheses test is

$$H = 0 : S \sim N(0, \sigma_0^2)$$

$$H = 1 : S \sim N(0, \sigma_1^2)$$

where $0 < \sigma_0^2 < \sigma_1^2$.

For this problem, each sensor makes a binary decision and sends it to a fusion center which makes a final decision. The problem of finding optimal local decision rules was studied in [8] where a separate proof was given as Proposition 8 does not apply. We show in the following that one can directly apply Theorem 10 to solve this problem.

Let $W = S$, so we have $H - W - \mathbf{X}$ and $\mathbf{X}$ are conditionally independent given W. Thus, this is a CHCI model with hidden variable being the random signal S. Assume that a monotone fusion rule, e.g. AND rule, is used at the fusion center.

First we notice that $\frac{p(w|H=1)}{p(w|H=0)}$ is a monotone function of $|w|$, this is because $\frac{p(w|H=1)}{p(w|H=0)} = \frac{\sigma_0}{\sigma_1} \exp\left[\frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right) w^2\right]$, $\sigma_0 < \sigma_1$.

Thus, let $T(w) = |w|$. Both $\mathcal{X}$ and $\mathcal{W}$ are the real lines. We divide the real line into two intervals: $A_1 = [0, +\infty)$, $A_2 = (-\infty, 0)$ where $T(w)$ is increasing on A_1 and decreasing on A_2.

By the definition of $\lambda(x_k, t)$ in (5.24), we have

$$\lambda(x_k, t) = p(x_k|t)h(t) + p(x_k|-t)h(-t), \tag{5.27}$$

Also, we can obtain that $h(t) = h(-t) \geq 0$ by the symmetry of the fusion rule and the symmetry of $p(w|H = 0)$. Therefore,

$$\lambda(x_k, t) = (p(x_k|t) + p(x_k|-t))h(t). \tag{5.28}$$

Choose $S(x_k) = |x_k|$, and one can verify that condition 3 in Theorem 10 is satisfied. That is $\frac{\lambda(x_k,t)}{\lambda(x_k',t)}$ is a nondecreasing function of t for $|x_k| > |x_k'|$ where $x_k, x_k' \in A_i$, $i = 1, 2$.

For $x_k, x'_k \in A_1$, $|x_k| > |x'_k|$ implies $x_k > x'_k \geq 0$. We now verify that $\frac{\lambda(x_k,t)}{\lambda(x'_k,t)}$ is a nondecreasing function of t.

$$\frac{\lambda(x_k,t)}{\lambda(x'_k,t)} = \frac{p(x_k|t)+p(x_k|-t)}{p(x'_k|t)+p(x'_k|-t)} \tag{5.29}$$

$$= \frac{e^{\frac{x_k a_k t}{\sigma^2}} + e^{\frac{-x_k a_k t}{\sigma^2}}}{e^{\frac{x'_k a_k t}{\sigma^2}} + e^{\frac{-x'_k a_k t}{\sigma^2}}} e^{-\frac{x_k^2 - x_k'^2}{2\sigma^2}}. \tag{5.30}$$

Differentiate $\ln \frac{\lambda(x_k,t)}{\lambda(x'_k,t)}$ with respect to t, we have

$$\frac{d}{dt} \ln \frac{\lambda(x_k,t)}{\lambda(x'_k,t)} \geq 0 \tag{5.31}$$

for $x_k > x'_k \geq 0$ and $t > 0$. So $\frac{\lambda(x_k,t)}{\lambda(x'_k,t)}$ is a nondecreasing function of t for $x_k, x'_k \in A_1$.

Similarly, for $x_k, x'_k \in A_2$, $|x_k| > |x'_k|$ implies $x_k < x'_k < 0$. One can verify that $\frac{\lambda(x_k,t)}{\lambda(x'_k,t)}$ is also a nondecreasing function of t.

Hence, by Theorem 10, the optimal local decision rule is

$$U_k = \begin{cases} 1 & \text{if } |x_k| \geq \tau_k \\ 0 & \text{otherwise} \end{cases} \tag{5.32}$$

Clearly, the choice of $S(\cdot)$ is not unique here. For example, any monotone functions of $|x_k|$, e.g., x_k^2, can be used instead.

5.5 Summary

The problem of distributed detection with conditionally dependent observations was considered in this chapter. Utilizing the so-called HCI model, we identify more general conditions under which the distributed detection problem becomes tractable. This proposed generalization enables us to tackle a much broader class of distributed detection problems with dependent observations. The problem of detecting a random signal in independent noises is used to illustrate the advantage of such an approach.

Chapter 6

Sufficiency Principle for Decentralized Data Reduction

6.1 Introduction

The sufficiency principle has played a prominent role in designing data processing methods for statistical inference. The primary goal of sufficiency-based data reduction is dimensionality reduction to facilitate subsequent inferences based on the reduced data [22, 25, 26].

This chapter studies data reduction in decentralized inference and extends the sufficiency principle to systems where data reduction needs to be done locally. Decentralized inference refers to the decision making process involving multiple sensors [7]. Parallel networks and tandem networks, illustrated in Figs. 6.1 and 6.2, are two canonical models for decentralized inference.

For decentralized inference, data reduction is done locally without access to the

global data. Therefore, the contrasting notions of local sufficiency and global sufficiency need to be treated with care [9]. A sufficient statistic defined with respect to local data is referred to as a local sufficient statistic; if a collection of local statistics form a global sufficient statistic, they are said to be globally sufficient.

For the special case when data are conditionally independent given the inference parameter, local sufficient statistics are known to be globally sufficient. This result was established for parallel networks in [9, 53, 54] and it is straightforward to show that the same result holds for tandem networks.

However, for the general case when data are conditionally dependent, a set of local sufficient statistics need not be globally sufficient and vice versa. In this chapter, we develop theories and tools for decentralized data reduction with conditionally dependent observations for both parallel and tandem networks. We show that global sufficiency of local statistics is not determined solely by the statistical characterization of local data but also depends on the statistical property of the global data as well as the structure of the network.

For parallel networks, we investigate the sufficiency principle under the HCI model, which is a framework proposed to deal with distributed detection with conditionally dependent observations [8]. Suitable conditions are identified under this HCI model such that local sufficiency implies global sufficiency.

For tandem networks such as that described in Fig. 6.2, $\mathbf{X}_2$ is fully available at the decision node. We define a novel notion of conditional sufficiency to capture the difference in network structure with that of the parallel network.

Data reduction through sufficiency statistics has application beyond that of statistical inference problems. It was shown, for example, that sufficient statistic based

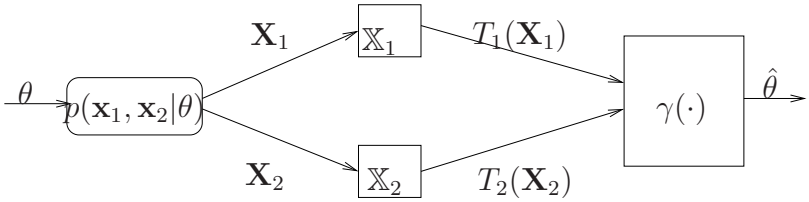


Figure 6.1: Parallel network.

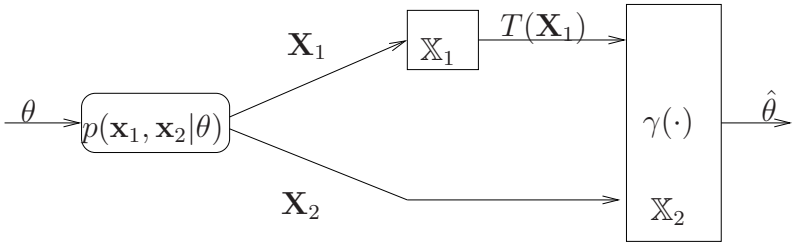


Figure 6.2: Tandem network.

data reduction achieves the same rate distortion function as the original data for the point to point remote rate distortion problem [55]. In this chapter, we apply the sufficient statistics to two classical distributed source coding problems. There, sufficiency-based data reduction prior to a source encoder is shown to incur no penalty on the corresponding rate region or the rate distortion function.

The rest of the paper is organized as follows. Section 6.2 develops the sufficiency principle in parallel networks with emphasis on conditionally dependent observations. Section 6.3 deals with tandem networks where the notion of conditional sufficiency is introduced and associated theories are developed. In Section 6.4, the connection between the developed sufficiency principle and two distributed source coding problems is explored. Section 6.5 concludes the chapter.

6.2 Sufficient principle for parallel network

This section considers decentralized data reduction in a parallel network as illustrated in Fig. 6.1. Let $\theta \sim p(\theta)$ be the parameter of interest and $\mathbf{X}_i$ the local observation at sensor i for $i = 1, 2$. For a decentralized system, there is a need to distinguish the notions of local versus global sufficient statistics [9]. When θ is random, for $i = 1, 2$, $T_i(\mathbf{X}_i)$ is a local sufficient statistic if

$$\theta - T_i(\mathbf{X}_i) - \mathbf{X}_i, \tag{6.1}$$

form a Markov chain, i.e., sufficiency is defined with respect to the local observation $\mathbf{X}_i$. On the other hand, we call $(T_1(\mathbf{X}_1), T_2(\mathbf{X}_2))$ a global sufficient statistic if the Markov chain

$$\theta - (T_1(\mathbf{X}_1), T_2(\mathbf{X}_2)) - (\mathbf{X}_1, \mathbf{X}_2), \tag{6.2}$$

holds. It is apparent that for the general case, the two individual Markov chains (6.1) and (6.2) do not imply each other.

6.2.1 Conditionally independent observations

For the conditional independence case, it can be easily established that local sufficiency implies global sufficiency. The converse also holds for the conditional independence case, which is given in the following proposition.

Proposition 9. *Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be conditionally independent observations given the random parameter θ. If $(T_1(\mathbf{X}_1), T_2(\mathbf{X}_2))$ form a global sufficient statistic for θ, then both $T_1(\mathbf{X}_1)$ and $T_2(\mathbf{X}_2)$ are respectively local sufficient statistics with respect to the observations $\mathbf{X}_1$ and $\mathbf{X}_2$.*

We first state some useful properties of Markov chains [56] that will be used for subsequent proofs:

- Symmetry: $X - Z - Y \Rightarrow Y - Z - X$;
- Decomposition: $X - Z - YW \Rightarrow X - Z - Y$;
- Weak Union:$X - Z - YW \Rightarrow X - ZW - Y$;
- Contraction: $X - Z - Y$ and $X - ZY - W \Rightarrow X - Z - YW$;
- Intersection: $X - ZW - Y$ and $X - ZY - W \Rightarrow X - Z - YW$.

Proof. Since $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent given θ, $\mathbf{X}_1 - \theta - \mathbf{X}_2$ form a Markov chain and so does $(\mathbf{X}_1, T_1(\mathbf{X}_1)) - \theta - \mathbf{X}_2$ as $T_1(\mathbf{X}_1)$ is a function of $\mathbf{X}_1$. Using the weak union property, we have that $\mathbf{X}_1 - (\theta, T_1(\mathbf{X}_1)) - \mathbf{X}_2$ form a Marokov chain. That $(T_1(\mathbf{X}_1), T_2(\mathbf{X}_2))$ is globally sufficient implies that (6.2) holds and thus $\mathbf{X}_1 - (T_1(\mathbf{X}_1), T_2(\mathbf{X}_2)) - \theta$ form a Markov chain according to the decomposition and symmetry properties. Combining $\mathbf{X}_1 - (\theta, T_1(\mathbf{X}_1)) - \mathbf{X}_2$ and $\mathbf{X}_1 - (T_1(\mathbf{X}_1), T_2(\mathbf{X}_2)) - \theta$, and using the intersection property we get the Markov chain $\mathbf{X}_1 - T_1(\mathbf{X}_1) - (\theta, T_2(\mathbf{X}_2))$ whenever $p(\mathbf{x}_1, T_1(\mathbf{x}_1), T_2(\mathbf{x}_2), \theta)$ is positive. Thus $T_1(\mathbf{X}_1)$ is a local sufficient statistic for θ. That $T_2(\mathbf{X}_2)$ is locally sufficient for θ can be established similarly. ■

6.2.2 Conditionally dependent observations

While the above establishes that global and local sufficient statistics imply each other for conditionally independent observations, the same is not true for the dependent case. Consider the following trivial example.

Example 2. *Let* $\mathbf{X}_1 = \mathbf{X}_2$ *in Fig. 6.1. It is clear that* $(T_1(\mathbf{X}_1) = \mathbf{X}_1, T_2(\mathbf{X}_2) = \varnothing)$ *is globally sufficient for* θ *while* $T_2(\mathbf{X}_2) = \varnothing$ *is not locally sufficient.*

The rest of this section is devoted to the question of how to identify global sufficient statistics at distributed nodes with conditionally dependent observations. Our approach leverages the HCI model, which is a framework developed for distributed detection with conditionally dependent observations, as discussed in Section 5.2. An HCI model is constructed by introducing a hidden variable $\mathbf{W}$ such that the following Markov chains hold:

$$\begin{aligned} &\mathbf{X}_1 - \mathbf{W} - \mathbf{X}_2, \\ &\theta - \mathbf{W} - (\mathbf{X}_1, \mathbf{X}_2). \end{aligned} \tag{6.3}$$

That is, $\mathbf{W}$ induces conditional independence between $\mathbf{X}_1$ and $\mathbf{X}_2$ as well as conditional independence between the inference parameter θ and the sensor observations $(\mathbf{X}_1, \mathbf{X}_2)$. Any general distributed inference model is equivalent to an HCI model and vice versa. We notice here that while we only illustrate the HCI model using the two sensor system, the framework is applicable to that involving any arbitrary number of sensors where we replace the Markov chain $\mathbf{X}_1 - \mathbf{W} - \mathbf{X}_2$ with the equivalent conditional independence assumption.

Notice that the second Markov chain in defining the HCI model implies that the information about the inference parameter θ in the data $(\mathbf{X}_1, \mathbf{X}_2)$ is preserved entirely in $\mathbf{W}$. This is formalized in the following lemma.

Lemma 10. *Let* $\mathbf{X}_1, \mathbf{X}_2 \sim p(\mathbf{x}_1, \mathbf{x}_2|\theta)$ *and suppose that there exists a random variable* $\mathbf{W}$ *such that*

$$\theta - \mathbf{W} - (\mathbf{X}_1, \mathbf{X}_2). \tag{6.4}$$

A statistic $T(\mathbf{X}_1, \mathbf{X}_2)$ that is sufficient for $\mathbf{W}$ is also sufficient for θ.

Proof. The Markov chain (6.4) implies that $\theta - \mathbf{W} - (\mathbf{X}_1, \mathbf{X}_2, T(\mathbf{X}_1, \mathbf{X}_2))$ forms a Markov chain for any statistics $T(\mathbf{X}_1, \mathbf{X}_2)$. That $T(\mathbf{X}_1, \mathbf{X}_2)$ is sufficient for $\mathbf{W}$ implies the Markov chain $\mathbf{W} - T(\mathbf{X}_1, \mathbf{X}_2) - (\mathbf{X}_1, \mathbf{X}_2)$. It is straightforward to show that these two Markov chains give rise to a long Markov chain

$$\theta - \mathbf{W} - T(\mathbf{X}_1, \mathbf{X}_2) - (\mathbf{X}_1, \mathbf{X}_2). \tag{6.5}$$

Therefore, $T(\mathbf{X}_1, \mathbf{X}_2)$ is sufficient for θ. ■

Lemma 10 is not useful in itself as $T(\mathbf{X}_1, \mathbf{X}_2)$ is a function of the global data which is not available in either of the nodes. Its use is mainly for establishing the following result.

Theorem 11. *Let $\mathbf{X}_1, \mathbf{X}_2 \sim p(\mathbf{x}_1, \mathbf{x}_2|\theta)$ and suppose there exists a random variable $\mathbf{W}$ such that $\theta - \mathbf{W} - (\mathbf{X}_1, \mathbf{X}_2)$. Let $T(\mathbf{W})$ be a sufficient statistic for θ, i.e., $\theta - T(\mathbf{W}) - \mathbf{W}$.*

1. *If a pair of statistics $(T_1(\mathbf{X}_1), T_2(\mathbf{X}_2))$ are globally sufficient for $T(\mathbf{W})$, they are globally sufficient for θ.*

2. *If $T(\mathbf{W})$ induces conditional independence between $\mathbf{X}_1$ and $\mathbf{X}_2$ and $(T_1(\mathbf{X}), T_2(\mathbf{X}_2))$ are locally sufficient for $T(\mathbf{W})$, then $(T_1(\mathbf{X}_1), T_2(\mathbf{X}_2))$ are globally sufficient for θ.*

Proof. To prove the first result, from Lemma 10, we only need to show that the Markov chain $\theta - T(\mathbf{W}) - (\mathbf{X}_1, \mathbf{X}_2)$ holds. Note first that the Markov chain $T(\mathbf{W}) - (\theta, \mathbf{W}) - (\mathbf{X}_1, \mathbf{X}_2)$ forms a Markov chain as $T(\mathbf{W})$ is a function of $\mathbf{W}$. Together

with $\theta - \mathbf{W} - (\mathbf{X}_1, \mathbf{X}_2)$ we obtain the Markov chain $(\theta, T(\mathbf{W})) - \mathbf{W} - (\mathbf{X}_1, \mathbf{X}_2)$ using the contraction property. Combined with the Markov chain $\theta - T(\mathbf{W}) - \mathbf{W}$, we get $\theta - T(\mathbf{W}) - \mathbf{W} - (\mathbf{X}_1, \mathbf{X}_2)$ which implies $\theta - T(\mathbf{W}) - (\mathbf{X}_1, \mathbf{X}_2)$.

To prove the second result, since conditional independence ensures that local sufficient statistics are globally sufficient, $(T_1(\mathbf{X}_1), T_2(\mathbf{X}_2))$ are thus sufficient for $T(\mathbf{W})$. The first result then establishes that they are also sufficient for θ. ■

Applying Theorem 11 to the HCI model, we have the following corollary.

Corollary 3. *For an HCI model, local sufficiency with respect to the hidden variable implies global sufficiency.*

Corollary 3 suggests that a way to obtain global sufficient statistics at individual nodes is to ensure local sufficiency of the statistics with respect to the hidden variable $\mathbf{W}$ in the HCI model. However, as we shall illustrate, the approach is meaningful only if the hidden variable $\mathbf{W}$ is chosen appropriately. For example, choosing $\mathbf{W} = (\mathbf{X}_1, \mathbf{X}_2)$ ensures that the Markov chains used to define the HCI model in (6.3) are always satisfied, yet it does not lead to any meaningful data reduction. We now use a simple example to show how Corollary 3 can be used for data reduction through an appropriately chosen $\mathbf{W}$.

Example 3. *For $i = 1, \cdots, n$, let*

$$\begin{aligned} X_{1i} &= \theta + Z + U_i, \\ X_{2i} &= \theta + Z + V_i, \end{aligned}$$

where $Z, U_1, \cdots, U_n, V_1, \cdots, V_n$ are mutually independent Gaussian random variables such that $Z \sim \mathcal{N}(0, \rho)$, $U_j \sim \mathcal{N}(0, 1-\rho)$, $V_j \sim \mathcal{N}(0, 1-\rho)$. Thus, we need to estimate

a parameter θ in the presence of a constant interference Z and independent noises U_i and V_i. Since $X_{1i}, X_{2i} \sim N(\theta, \theta, 1, 1, \rho)$, given θ, $\mathbf{X}_1$ and $\mathbf{X}_2$ are not independent conditioned on θ.

Choose the hidden variable $W = \theta + Z$. One can verify easily that W satisfies the Markov chains $\theta - W - (\mathbf{X}_1, \mathbf{X}_2)$ and $\mathbf{X}_1 - W - \mathbf{X}_2$ required by the HCI model. For Gaussian observations, it is also clear that $\sum_i X_{1i}$ and $\sum_i X_{2i}$ are locally sufficient for W. Therefore, from Corollary 3, $(\sum_i X_{1i}, \sum_i X_{2i})$ is globally sufficient for θ.

The next example is motivated by the cooperative spectrum sensing problem [57].

Example 4. *Consider the hypothesis testing problem involving K sensors where the two hypotheses under test with observations are*

$$H_0 : X_k = N_k, \tag{6.6}$$

$$H_1 : X_k = h_k S + N_k, \tag{6.7}$$

where X_k, $k = 1, \cdots, K$, is the observation at sensor k, h_k's are circularly symmetric complex Gaussian and independent of each other and of other variables, S is a signal taking values in the set $\mathcal{S} = \{s_m = r_m e^{j\theta_m}, m = 1, \cdots, M\}$ with probability $p(S = s_m) = \pi_m$, and N_k is the observation noise at the kth sensor which is circularly complex Gaussian distributed and is independent of each other. This hypothesis testing problem can be used to describe the baseband model of detecting the presence of a QAM signal in independent Rayleigh fading channels using K sensors. Each sensor makes a local decision $U_k = \gamma(X_k)$ and sends it to a fusion center which makes a final decision regarding the hypothesis under test.

The observations are not conditionally independent under H_1 given that the observations contain a common random signal S. Again, taking a Baysian viewpoint

where we assume that the true hypothesis H is a binary random variable, then $H-S-(X_1,\cdots,X_K)$ form a Markov chain since the observations depend on the hypothesis only through the signal. It is easy to verify that the statistic $|S|$ is sufficient for H given S. Thus, the Markov chain $H-|S|-S-(X_1,\cdots,X_K)$ holds. On the other hand, given $|S|$, the observations are conditionally independent of each other under the independent Rayleigh fading assumption. Therefore, $|S|$ serves as the hidden variable $\mathbf{W}$ *for the HCI model corresponding to this decentralized hypothesis testing problem.*

For any k, $|X_k|$ is a minimal sufficient statistic for $|S|$. This can be easily verified by writing out the ratio $\frac{p(x_k||s|)}{p(x_k'||s|)}$ for two sample points x_k and x_k'. Therefore, from Corollary 3, $\{|X_k|\}, k=1,\cdots,K$, are globally sufficient for H.

6.3 Sufficiency principle for tandem network

A tandem network, as illustrated in Fig. 6.2, is one such that compressed data are transmitted to a node which also has its own observation. The second node will then make a final decision using its own data and the input from the first node. Knowing that $\mathbf{X}_2$ is available at the fusion center even without directly observing $\mathbf{X}_2$ should have an impact on how node $\mathbb{X}_1$ summarizes its own data $\mathbf{X}_1$. A natural way of extending the sufficiency principle to this network is as follows: the inference performance should remain the same whether the inference is based on $(\mathbf{X}_1,\mathbf{X}_2)$ or $(T(\mathbf{X}_1),\mathbf{X}_2)$. From the data processing inequality, the sufficiency of $T(\mathbf{X}_1)$ can thus be characterized using the Markov chain $\theta-(T(\mathbf{X}_1),\mathbf{X}_2)-(\mathbf{X}_1,\mathbf{X}_2)$. Given that $T(\mathbf{X}_1)$ is a function $\mathbf{X}_1$, it is straightforward to show that that the Markov chain $\theta-(T(\mathbf{X}_1),\mathbf{X}_2)-(\mathbf{X}_1,\mathbf{X}_2)$ is equivalent to $\theta-(T(\mathbf{X}_1),\mathbf{X}_2)-\mathbf{X}_1$. This motivates the

following definition of conditional sufficiency.

Definition 4. *A statistic $T(\mathbf{X}_1)$ is a conditional sufficient statistic for θ, conditioned on $\mathbf{X}_2$, if the conditional distribution of the sample $\mathbf{X}_1$ given the value of $T(\mathbf{X}_1)$ and $\mathbf{X}_2$ does not depend on θ.*

The definition allows us to generalize a number of classical results related to sufficient statistics.

Theorem 12. *Let $\mathbf{X}_1, \mathbf{X}_2$ be distributed according to $p(\mathbf{x}_1, \mathbf{x}_2|\theta)$. Let $q(T(\mathbf{x}_1), \mathbf{x}_2|\theta)$ be the joint distribution of $T(\mathbf{X}_1)$ and $\mathbf{X}_2$, then $T(\mathbf{X}_1)$ is a conditional sufficient statistic for θ, conditioned on $\mathbf{X}_2$, if for every $(\mathbf{X}_1, \mathbf{X}_2)$ pair, the ratio $\frac{p(\mathbf{x}_1,\mathbf{x}_2|\theta)}{q(T(\mathbf{x}_1),\mathbf{x}_2|\theta)}$ is constant as a function of θ.*

Similarly, the Neyman-Fisher factorization theorem can also be generalized to the conditional case.

Theorem 13. *Let $\mathbf{X}_1, \mathbf{X}_2$ be distributed according to $p(\mathbf{x}_1, \mathbf{x}_2|\theta)$. A statistic $T(\mathbf{X}_1)$ is conditionally sufficient for θ, conditioned on $\mathbf{X}_2$,* if and only if *there exist functions $g(t, \mathbf{x}_2|\theta)$ and $h(\mathbf{x}_1, \mathbf{x}_2)$ such that,*

$$p(\mathbf{x}_1, \mathbf{x}_2|\theta) = g(T(\mathbf{x}_1), \mathbf{x}_2|\theta)h(\mathbf{x}_1, \mathbf{x}_2), \tag{6.8}$$

for all sample points $(\mathbf{x}_1, \mathbf{x}_2)$ and all parameter values θ.

The proof can be constructed similarly to that of the factorization theorem in [25, Theorem 6.2.6].

Minimal sufficient statistic plays a prominent role in statistical inference as it attains maximum data reduction without compromising inference performance. Sim-

ilar to the definition of minimal sufficient statistic [25], we can define the notion of minimal conditional sufficient statistic as follows.

Definition 5. *A conditional sufficient statistic $T(\mathbf{X}_1)$ is a minimal conditional sufficient statistic if it is a function of any other conditional sufficient statistic $U(\mathbf{X}_1)$.*

The following theorem provides a meaningful way to find minimal conditional sufficient statistics.

Theorem 14. *Let $\mathbf{X}_1, \mathbf{X}_2$ be distributed according to $p(\mathbf{x}_1, \mathbf{x}_2|\theta)$. Suppose there exists a function $T(\mathbf{X}_1)$ such that for every two sample points $\mathbf{x}_1$, $\hat{\mathbf{x}}_1$, and $\mathbf{x}_2$, the ratio $\frac{f(\mathbf{x}_1,\mathbf{x}_2|\theta)}{f(\hat{\mathbf{x}}_1,\mathbf{x}_2|\theta)}$ is constant as a function of θ* if and only if *$T(\mathbf{x}_1) = T(\hat{\mathbf{x}}_1)$. Then $T(\mathbf{X}_1)$ is a minimal conditional sufficient statistic for θ given $\mathbf{X}_2$.*

The proof follows the same line of proof for Theorem 6.2.13 in [25].

The definition of conditional sufficiency is more general than global sufficiency. This is because if there exist a pair of statistics $(T_1(\mathbf{X}_1), T_2(\mathbf{X}_2))$ that are globally sufficient for θ, then $T_1(\mathbf{X}_1)$ must be conditionally sufficient for θ, conditioned on $\mathbf{X}_2$.

Example 5. *Let $\{X_{1i}, X_{2i}\}$, $i = 1, \cdots, n$ be i.i.d according to $p(x_1, x_2|\theta)$, where*

$$p(x_1, x_2|\theta) = \begin{cases} 2 & \theta < x_1 < \theta + 1, \theta < x_2 < x_1, \\ 0 & \textit{otherwise.} \end{cases}$$

The marginal distribution of X_1 and X_2 are therefore,

$$\begin{aligned} p(x_1|\theta) &= 2(x_1 - \theta), \quad \theta < x_1 < \theta + 1, \\ p(x_2|\theta) &= 2(\theta + 1 - x_2), \quad \theta < x_2 < \theta + 1. \end{aligned}$$

It can be easily shown that no data reduction is possible using the marginal distribution, i.e., no meaningful locally sufficient statistics can be found other than the

data themselves. Note that X_1 is uniformly distributed on the interval $(x_2, \theta+1)$, therefore, we have

$$p(\mathbf{x}_1|\mathbf{x}_2,\theta) = \frac{1}{\prod_{i=1}^{n}(\theta+1-x_{2i})}, x_{2i} < x_{1i}, (\max_i\{x_{1i}\}-1) < \theta.$$

Thus, $\max_i\{X_{1i}\}$ is a conditional sufficient statistic for θ, conditioned on $\mathbf{X}_2$. Similarly, we can obtain that $\min_i\{X_{2i}\}$ is a conditional sufficient statistic of $\mathbf{X}_2$, conditioned on $\mathbf{X}_1$. This is consistent with the fact that $(\max_i\{X_i\}, \min_i\{Y_i\})$ is globally sufficient given both $\mathbf{X}_1$ and $\mathbf{X}_2$.

6.4 Sufficient statistics and distributed source coding

In this section, we study the connection between the sufficiency principle and two distributed source coding problems: the lossless source coding with side information problem and the remote source coding with side information available both at encoder and decoder. We show that for these two problems, sufficient statistic based data reduction achieves the same rate distortion function as the original data.

6.4.1 Source coding with side information

Consider the lossless source coding problem in Fig. 6.3. An i.i.d. sequence of source pairs (X^n, Y^n) are encoded separately with rates (R_1, R_2) and the descriptions are sent to a decoder where only X^n is to be recovered with asymptotically vanishing probability of error. A rate pair (R_1, R_2) is achievable if there exists a lossless source code with rates (R_1, R_2). The rate region $\mathcal{R}$ is defined as the closure of the set of all

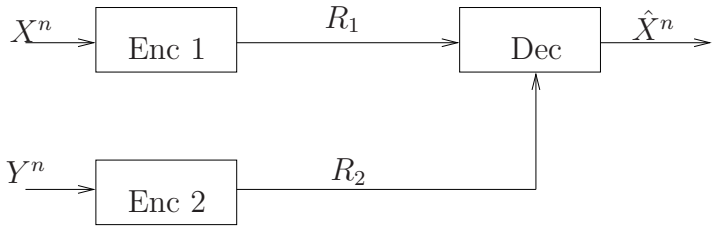


Figure 6.3: Source coding with side information.

achievable rate pairs and was shown to be [4, 58],

$$\mathcal{R} = \{(R_1, R_2) : R_1 \geq H(X|U), R_2 \geq I(Y;U), X - Y - U\}. \tag{6.9}$$

Assume $T(Y)$ is a sufficient statistic for X, i.e., $X - T(Y) - Y$. Define

$$\mathcal{R}' = \{(R_1, R_2) \quad : \quad R_1 \geq H(X|U), R_2 \geq I(T(Y);U), X - T(Y) - U\}, \tag{6.10}$$

which is the rate region for encoding $(X^n, T^n(Y^n))$ where $T^n(Y^n)$ is the i.i.d sequence $T(Y_i), i = 1, \cdots, n$. The following theorem shows that encoding reduced data $T^n(Y^n)$ achieves the same rate region as encoding the original data.

Theorem 15.

$$\mathcal{R} = \mathcal{R}'. \tag{6.11}$$

Proof. It is straightforward to show $\mathcal{R} \supseteq \mathcal{R}'$. To show $\mathcal{R} \subseteq \mathcal{R}'$, let $(R_1, R_2) \in \mathcal{R}$, then there exists a U such that $X - Y - U$, $R_1 \geq H(X|U), R_2 \geq I(Y;U)$. Since $(X, T(Y)) - Y - U$ and $X - T(Y) - Y$, the Markov chain $X - T(Y) - Y - U$ holds. Therefore, $R_1 \geq H(X|U)$, $R_2 \geq I(Y;U) \geq I(T(Y);U)$ by the data processing inequality. Thus, $(R_1, R_2) \in \mathcal{R}'$. ■

A direct consequence of Theorem 15 is that the corner point of the rate region $(R_1 = H(X|Y), R_2 = H(T(Y))$ may be strictly smaller than $(R_1 = H(X|Y), R_2 =$

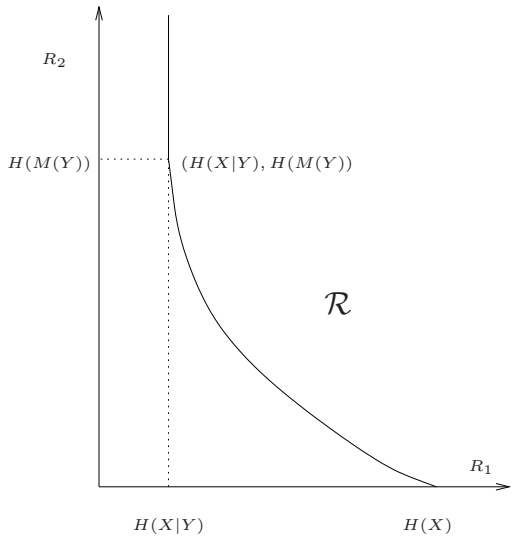


Figure 6.4: The corner points of the rate region for the source coding with side information problem.

$H(Y)$. This observation was first reported in [18]. Specifically, the corner point can be obtained by finding the smallest admissible R_2 when $R_1 = H(X|Y)$ and it was shown that [18]

$$\begin{aligned} \inf\{R_2 : (H(X|Y), R_2) \in \mathcal{R}\} &= \inf_{X-Y-U, X-U-Y} I(Y;U), \\ &= H(\Phi_Y^X). \end{aligned}$$

As it turns out, the quantity Φ_Y^X is precisely the minimal sufficient statistic of X given Y, $M(Y)$. The corner point of the rate region for the source coding with side information problem is shown in Fig. 6.4.

6.4.2 Remote source coding with side information

In this section, we examine the application of the conditional sufficient statistics in a remote rate distortion problem with side information.

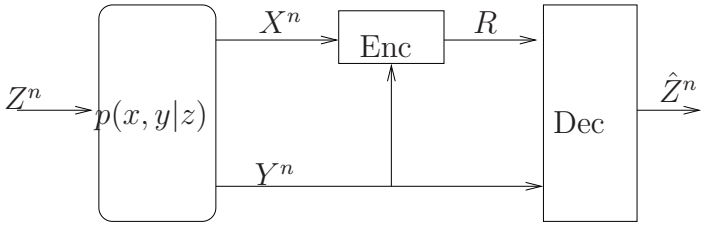


Figure 6.5: Remote source coding with side information.

Consider a model in Fig 6.5, which is the remote source coding with side information available at both the encoder and decoder. We will show that in this problem, the rate distortion function will not change by encoding a conditional sufficient statistic $T(X)$.

Let $(X, Y, Z) \sim p(x, y, z)$ and $d(z, \hat{z})$ be a given distortion function. Let (X^n, Y^n, Z^n) be i.i.d sequences drawn from (X, Y, Z). Upon receiving the sequences (X^n, Y^n), the encoder generates a description of the sources with rate R and sends it to the decoder who has the side information Y^n and wishes to reproduce Z^n with distortion D. The rate distortion function $R(D)$ is the infimum of rate R such that there exist maps $f_n : \mathcal{X}^n \times \mathcal{Y}^n \to \{1, \cdots, 2^{nR}\}$, $g_n : \mathcal{Y}^n \times \{1, \cdots, 2^{nR}\} \to \hat{Z}^n$ such that

$$\limsup_{n\to\infty} Ed(Z^n, g_n(Y^n, f_n(X^n, Y^n))) \leq D. \tag{6.12}$$

It is easy to show that the rate distortion function $R(D)$ is:

$$R(D) = \min_{p(u|x,y)} \min_{f} I(X; U|Y), \tag{6.13}$$

where the minimum is taken over all $p(u|x, y)$ and functions $\hat{z} = f(u, y)$ such that

$$E_1[d(Z, \hat{Z})] \triangleq \sum_{x,y,z,u} p(x, y, z)p(u|x, y)d(z, f(u, y)) \leq D. \tag{6.14}$$

Let $T(X)$ be a conditional sufficient statistic for the remote source Z, conditioned

on Y (i.e., $Z-(T(X),Y)-(X,Y)$). Define

$$R'(D)=\min_{p(u|t,y)}\min_{f} I(T(X);U|Y), \tag{6.15}$$

where the minimum is taken over all $p(u|t,y)$ and functions $\hat{z}=f(u,y)$ such that

$$E_2[d(Z,\hat{Z})]\triangleq\sum_{t,y,z,u} p(t,y,z)p(u|t,y)d(z,f(u,y))\le D. \tag{6.16}$$

$R'(D)$ is the rate distortion function when we have $(T^n(X^n),Y^n)$ instead of (X^n,Y^n) at the encoder, where $T^n(X^n)$ is the i.i.d sequence $T(X_i), i=1,\cdots,n$.

Theorem 16.

$$R(D)=R'(D). \tag{6.17}$$

Proof. It is obvious that $R(D)\le R'(D)$.

We now show $R(D)\ge R'(D)$. For any U that achieves $R(D)$, since $T(X)$ is a function of X, we have the Markov chain $(T(X),Y)-(X,Y)-U$, hence

$$\begin{aligned} I(X;U|Y) &= H(U|Y)-H(U|X,Y) && (6.18)\\ &\ge H(U|Y)-H(U|T(X),Y) && (6.19)\\ &= I(T(X);U|Y). && (6.20)\end{aligned}$$

Given that $T(X)$ is a conditional sufficient statistic for Z, we have the following

$$\begin{aligned} D &\ge E_1[d(Z,\hat{Z})]\\ &= \sum_{y,z,u} d(z,f(u,y))\left(\sum_x p(x,y,z)p(u|x,y)\right) && (6.21)\\ &= \sum_{y,z,u} d(z,f(u,y))\left(\sum_x p(z|x,y)p(x,y,u)\right) && (6.22)\\ &= \sum_{y,z,u} d(z,f(u,y))\left(\sum_t p(z|t,y)\sum_{x:T(x)=t} p(x,y,u)\right) && (6.23)\end{aligned}$$

$$= \sum_{y,z,u} d(z, f(u,y)) \left(\sum_t p(z|t,y)p(t,y,u) \right) \tag{6.24}$$

$$= \sum_{y,z,u} d(z, f(u,y)) \left(\sum_t p(u|t,y)p(t,y,z) \right) \tag{6.25}$$

$$= \mathrm{E}_2[d(Z, \hat{Z})] \tag{6.26}$$

where (6.23) comes from the definition of conditional sufficiency and (6.24) is true by defining $p(t,y,u) = \sum_{x:T(x)=t} p(x,y,u)$. This shows that for any $p(u|x,y)$ and $f(u,y)$ satisfying (6.14) there exist $p(u|t,y)$ and $f(u,y)$ such that (6.16) is satisfied. Thus, $R(D) \geq R'(D)$. ■

6.5 Summary

This chapter develops the sufficiency principle that guides local data reduction in networked inference with dependent observations for two classes of inference networks: parallel network and tandem network.

For the parallel network, the HCI model is used for conditional dependent observations to obtain conditions such that local sufficiency implies global sufficiency. For the tandem network, the notion of conditional sufficiency is proposed and related theories and tools associated with this new sufficiency concept are developed.

Finally, we established that data reduction using suitable notions of sufficiency incurs no penalty on the rate region for two distributed source coding problems.

Chapter 7

Decentralized Data Reduction with Quantization Constraints

7.1 Introduction

Sufficiency based data reduction ensures no loss of inference performance using the reduced data. While the sufficiency principle often results in maximum dimensionality reduction, communicating a one-dimensional real data may still be infeasible when communication is subject to a finite capacity constraint.

In this chapter, we consider the simple case where each sensor node communicates only a finite number of bits to the fusion center. Directly quantizing the raw data, especially if the data is of high dimension and quantizers operate in a decentralized fashion, is often a formidable task [59]. As such, it is often desirable to achieve maximum data reduction at each node prior to quantization.

We are then led to the question: *is it optimal to implement data reduction by*

forming a collection of global sufficient statistics followed by the design of optimal quantizers using the reduced data? Alternatively, one can consider the sufficiency principle to be the ubiquitous principle for data reduction in a 'lossless' sense, that is, complete information in the original data needs to be retained in the statistics. When practical constraints such as finite-bit quantization are imposed which result in inevitable loss of information, is sufficient principle still the guiding principle for data reduction?

Unfortunately, as seen from Example 6, the answer to this question is negative in general. However, there exist results where quantizing sufficient statistics is known to be optimal. The classical example is distributed detection with conditionally independent observations where the local likelihood ratios form a set of global sufficient statistics. Indeed, Tsitsiklis established in [60] that likelihood ratio quantizers (LRQ's) are optimal for a broad class of performance criteria, even with non-ideal, possibly coupling channels between the sensors and the fusion center [61, 62]. There also exist instances where quantizing local sufficient statistics is globally optimal for certain parameter regimes in the dependent observation case [63].

The objective of this chapter is thus to identify, for decentralized inference involving dependent data, conditions under which data reduction using sufficient statistics is still optimal when quantization is required at each node. While the result includes that of [60] as its special case, the approach differs from that of [60] as we do not start with an explicit form of quantizers thus can not explore the structural information of the statistics as that of [60]. Instead, our approach utilizes the Markovian structure implied in sufficient statistics. On the other hand, our optimality is strictly in the sense of minimizing a Bayesian cost as opposed to that of [60] which includes

a broader class of performance criteria. We discuss the problem in details for parallel networks and generalize the results to tandem networks.

Related to the work in this chapter is the quantizer design for distributed estimation in [59] and [64] where necessary conditions for optimal quantizers are derived. We do not explicitly address the quantizer design problem. Instead, we derive sufficient conditions such that sufficient statistics based data reduction followed by quantization is structurally optimal. Various optimal quantizer design approaches can then be applied to the reduced data which is often much more tractable than dealing with the raw data.

The rest of this chapter is organized as follows. Section 7.2 establishes the structural optimality of sufficiency based data reduction for centralized inference when quantization is required. In Section 7.3, the sufficiency principle is re-examined in decentralized inference when quantization is necessary at each node. Both conditionally independent and conditionally dependent observations are considered. We establish the structural optimality of sufficiency based data reduction followed by quantizers for the independent case. For the dependent case, we identify a class of problems where we prove that sufficiency based data reduction is still optimal in the presence of quantizers. Also we obtain a unifying condition under which the sufficiency based data reduction is optimal, which includes the independence and dependence conditions as its special cases in this section. Section 7.4 discuss the centralized inference with quantization constraints for tandem networks. Section 7.5 concludes the chapter.

7.2 Centralized inference with quantization

In this section, we consider a simple centralized inference system where the entire data is available at a single node. We establish the optimality of sufficiency based data reduction when quantization is required.

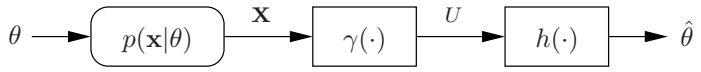
Consider a centralized inference system in which quantization is required, as shown in Fig. 7.1(a). Here, θ is the parameter of inference interest with distribution $p(\theta)$, $\mathbf{X}$ is the random vector observation, $\gamma(\cdot)$ is the quantizer directly operating on the data $\mathbf{X}$ and the output of the quantizer is $U = \gamma(\mathbf{X}) \in \{0, \ldots, L-1\}$ where L is the number of possible outputs. The estimator at the fusion center is denoted by the function $h(\cdot)$ whose input is the quantizer output.

Let $T(\mathbf{X})$ be any sufficient statistic for θ. To establish the optimality of sufficiency based data reduction with a quantization constraint, we will show in the following that the two systems in Fig. 7.1 achieve the same optimal performance where the second system applies data reduction to obtain $T(\mathbf{X})$ prior to a quantization operation. The quantizer and estimator in Fig. 7.1(b) are similarly defined by $U' = \gamma'(T(\mathbf{X}))$ and $h'(U')$. Note that for a centralized system there is no distinction between local and global sufficient statistics.

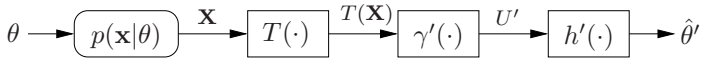
Let $d(\theta, \hat{\theta})$ be a given cost function between the parameter θ and the estimator output $\hat{\theta}$. For the model in Fig. 7.1(a), $\hat{\theta} = h(U) = h(\gamma(\mathbf{X}))$. The Bayesian cost is the expected cost function given by

$$C = E[d(\theta, h(\gamma(\mathbf{X})))], \tag{7.1}$$

where the expectation is taken with respect to both the random parameter θ and the



(a)



(b)

Figure 7.1: Centralized inference systems with quantizers operating on (a) the raw data $\mathbf{X}$, (b) a statistic $T(\mathbf{X})$.

observation $\mathbf{X}$. Let

$$C_{\min} = \min_{\gamma,h} C. \tag{7.2}$$

For the model in Fig. 7.1(b), $\hat{\theta}' = h'(\gamma'(T(\mathbf{X}))$ and the Bayesian cost is given by

$$C' = E[d(\theta, h'(\gamma'(T(\mathbf{X}))))], \tag{7.3}$$

where again the expectation is taken with respect to θ and $\mathbf{X}$. Let

$$C'_{\min} = \min_{\gamma',h'} C'. \tag{7.4}$$

We now establish that the system described in Fig. 7.1(b) is structurally optimal, i.e., it can achieve the same inference performance as that of Fig. 7.1(a), hence quantizing the sufficient statistic achieves the same minimum Bayesian cost as quantizing the observation in centralized inference.

Theorem 17. *For the Bayesian cost in (7.1) and (7.3),*

$$C_{\min} = C'_{\min}. \tag{7.5}$$

Proof. Apparently, $C'_{\min} \geq C_{\min}$ as one can always define a new quantizer $\gamma(\mathbf{X}) = \gamma'(T(\mathbf{X}))$ for any given $\gamma'(\cdot)$, thus converting any system described by Fig. 7.1(b) to that of Fig. 7.1(a) whose performance is no better than $C_{\min}$.

Next we establish $C'_{\min} \leq C_{\min}$ by showing that for any pair of $(\gamma(\cdot), h(\cdot))$ that achieves $C_{\min}$, there exists corresponding $(\gamma'(\cdot), h'(\cdot))$ pair that can achieve the same cost.

Expanding C in (7.1) with respect to the observation $\mathbf{X}$, we have

$$\begin{aligned} C &= \int_\theta \int_{\mathbf{X}} d(\theta, h(\gamma(\mathbf{x})))p(\mathbf{x}, \theta) d\mathbf{x} d\theta && (7.6) \\ &= \int_{\mathbf{X}} \int_\theta d(\theta, h(\gamma(\mathbf{x})))p(\theta|\mathbf{x})p(\mathbf{x}) d\theta d\mathbf{x} && (7.7) \\ &= \int_{\mathbf{X}} f(u, \mathbf{x})p(\mathbf{x}) d\mathbf{x}, && (7.8) \end{aligned}$$

where (7.8) is by the following definitions

$$\begin{aligned} u &\triangleq \gamma(\mathbf{x}) && (7.9) \\ f(u, \mathbf{x}) &\triangleq \int_\theta d(\theta, h(u))p(\theta|\mathbf{x}) d\theta. && (7.10) \end{aligned}$$

From (7.8), to minimize the Bayesian cost function C, the optimal quantizer given fixed estimator $h(\cdot)$ is to make a decision u such that $f(u, \mathbf{x})$ is minimized, that is

$$U = \gamma(\mathbf{X}) = \arg\min_u f(u, \mathbf{X}). \tag{7.11}$$

On the other hand, since $T(X)$ is the sufficient statistic of θ, by Lemma 1, we have

$$f(u, \mathbf{x}) = \int_\theta d(\theta, h(u))p(\theta|T(\mathbf{x})) d\theta. \tag{7.12}$$

Therefore, given $h(\cdot)$ being the optimal estimator, the optimal quantizer decides $U = i \in \{0, \ldots, L-1\}$ if

$$\begin{aligned} 0 &\geq f(i, \mathbf{x}) - f(j, \mathbf{x}) \\ &= \int_\theta \left(d(\theta, h(i)) - d(\theta, h(j)) \right) p(\theta|T(\mathbf{x})) d\theta, && (7.13) \end{aligned}$$

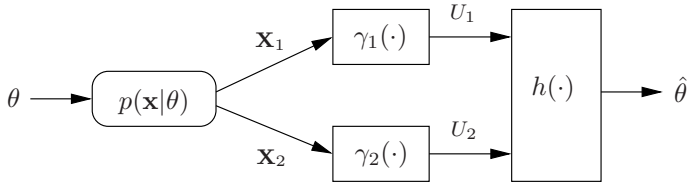
for any $j \in \{0, \ldots, L-1\}$. Note that (7.13) depends on $\mathbf{X}$ only through $T(\mathbf{X})$, hence can be realized by a $\gamma'(T(\mathbf{X}))$. Such an $\gamma'(\cdot)$, together with $h'(\cdot) = h(\cdot)$, can also achieve $C_{\min}$. Thus, the proof is complete. ■

The above result is not surprising in view of the fact that a sufficient statistic captures all the information about θ contained in the data. Indeed, the above theorem can be viewed as a simple instantiation of the sufficiency principle for the Bayesian cost. Other inference objective functions can also be used. Consider for example the "indirect rate distortion problem" [65] where a noisy version of a source sequence is observed at the encoder while the decoder tries to minimize the end-to-end distortion subject to a rate constraint between the encoder and the decoder. It was shown in [55] that data reduction using a sufficient statistic at the encoder does not affect the rate distortion function.

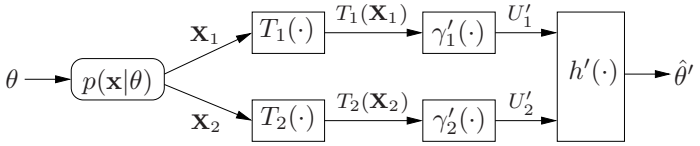
In decentralized inference, however, the same statement is not necessarily true, i.e., sufficient statistics based data reduction may not be optimal when quantization is required at individual nodes.

7.3 Decentralized data reduction with quantization constraints in parallel networks

We now consider decentralized inference where quantization is required at each node. For simplicity and ease of presentation, we assume a simple two-node system, as illustrated in Fig. 7.2. The result extends to systems with more than two nodes in a straightforward manner.



(a)



(b)

Figure 7.2: Decentralized inference systems with quantizers operating on (a) the raw data $\mathbf{X}_i, i = 1, 2$, (b) statistics $T_i(\mathbf{X}_i), i = 1, 2$.

Let $\theta \sim p(\theta)$ be the parameter of interest and $\mathbf{X}_i$ the local observation at sensor i with a likelihood function $p(\mathbf{x}_i|\theta)$, for $i = 1, 2$. Statistics and quantizers at local nodes, as well as the estimator at the fusion center are defined in a similar fashion as that in Section 7.2. Let $d(\theta, \hat{\theta})$ be the cost function where θ is the true parameter and $\hat{\theta}$ its estimate. The Bayesian costs for Fig. 7.2(a) and Fig. 7.2(b) are given respectively by

$$C = E[d(\theta, h(U_1, U_2))], \tag{7.14}$$

$$C' = E[d(\theta, h'(U_1', U_2'))], \tag{7.15}$$

where $U_i = \gamma_i(\mathbf{X}_i) \in \{0, \ldots, L-1\}$ and $U_i' = \gamma_i'(T_i(\mathbf{X}_i)) \in \{0, \ldots, L-1\}$.

The additional constraint that a quantizer is used at each sensor node may lead to inevitable information loss. As such, it is not clear whether global sufficient statistics based data reduction is still optimal. That is, even if $(T_1(\mathbf{X}_1), T_2(\mathbf{X}_2))$ form a global sufficient statistic, can the system in Fig. 7.2(b) achieve the same performance as that

of Fig. 7.2(a)?

The answer, unfortunately, is *no*, as can be seen from the following simple example.

Example 6. *Consider the degenerate case where* $\mathbf{X}_1 = \mathbf{X}_2$ *and* U_i *is constrained to be of one bit. Clearly* $(T_1(\mathbf{X}_1) = \mathbf{X}_1, T_2(\mathbf{X}_2) = \phi)$ *is a global sufficient statistic. However it is trivial to see that quantizing such constructed* $T_1(\mathbf{X}_1)$ *and* $T_2(\mathbf{X}_2)$ *using* 1*-bit each can be strictly suboptimal compared with quantizing the data directly, with the former equivalent to a* 1*-bit quantizer of the data whereas the latter a* 2*-bit quantizer.*

The above example involves data that are conditionally dependent given the parameter of interest. It turns out when data are conditionally independent given θ, the answer is indeed the affirmative, i.e., quantizing sufficient statistics is structurally optimal.

7.3.1 Conditionally independent observations

Theorem 18. *For the Bayesian costs in (7.14) and (7.15) when* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are conditionally independent given* θ*,*

$$\min_{\gamma_1,\gamma_2,h} C = \min_{\gamma_1',\gamma_2',h'} C'. \tag{7.16}$$

Note that for conditionally independent observations, there is no need to distinguish between local and global sufficient statistics. We now establish Theorem 18 using the Bayesian cost for a two-sensor system.

Proof. Let

$$C_{\min} = \min_{\gamma_1,\gamma_2,h} C, \tag{7.17}$$

where the minimum Bayesian cost is achieved by the optimal quantizers $\gamma_i^*(\cdot)$ and estimator $h^*(\cdot)$. It is easy to see that Fig. 7.2(b) can not achieve a better performance than $C_{\min}$. Thus we only need to show that $C_{\min}$ can be achieved by Fig. 7.2(b), i.e., one can find $(\gamma_1'(\cdot), \gamma_2'(\cdot), h'(\cdot))$ that achieve $C_{\min}$ for the given sufficient statistics $T_1(\mathbf{X}_1)$ and $T_2(\mathbf{X}_2)$. Similar to the proof for the centralized case, it suffices to show that the optimal quantizers $\gamma_i^*(\mathbf{X}_i)$ achieving $C_{\min}$ depends on $\mathbf{X}_i$ only through $T_i(\mathbf{X}_i)$.

As $\mathbf{X}_1$ and $\mathbf{X}_2$ are conditionally independent,

$$p(\mathbf{x}_1, \mathbf{x}_2, \theta) = p(\theta)p(\mathbf{x}_1|\theta)p(\mathbf{x}_2|\theta) \tag{7.18}$$

$$= p(\mathbf{x}_1)p(\theta|\mathbf{x}_1)p(\mathbf{x}_2|\theta) \tag{7.19}$$

$$= p(\mathbf{x}_1)p(\theta|T_1(\mathbf{x}_1))p(\mathbf{x}_2|\theta). \tag{7.20}$$

The last step comes from the fact that $T_1(\mathbf{X}_1)$ is sufficent for the data $\mathbf{X}_1$ and Lemma 1. Expanding C with respect to $\mathbf{X}_1$, we get

$$C = \int_\theta \int_{\mathbf{X}_1} \int_{\mathbf{X}_2} d(\theta, h(\gamma_1(\mathbf{x}_1), \gamma_2(\mathbf{x}_2)))p(\mathbf{x}_1, \mathbf{x}_2, \theta)d\mathbf{x}_2 d\mathbf{x}_1 d\theta \tag{7.21}$$

$$= \int_\theta \int_{\mathbf{X}_1} \int_{\mathbf{X}_2} d(\theta, h(\gamma_1(\mathbf{x}_1), \gamma_2(\mathbf{x}_2)))p(\mathbf{x}_1)p(\theta|T_1(\mathbf{x}_1))p(\mathbf{x}_2|\theta)d\mathbf{x}_2 d\mathbf{x}_1 d\theta \tag{7.22}$$

$$= \int_{\mathbf{X}_1} f_1(u_1, \mathbf{x}_1)p(\mathbf{x}_1)d\mathbf{x}_1, \tag{7.23}$$

where

$$f_1(u_1, \mathbf{x}_1) \triangleq \int_{\mathbf{x}_2} \int_\theta d(\theta, h(u_1, \gamma_2(\mathbf{x}_2)))p(\theta|T_1(\mathbf{x}_1))p(\mathbf{x}_2|\theta)d\theta d\mathbf{x}_2. \tag{7.24}$$

Let $\gamma_2(\cdot)$ and $h(\cdot)$ take the form of the optimal $\gamma_2^*(\cdot)$ and $h^*(\cdot)$, $\gamma_1^*(\cdot)$ must be chosen such that the corresponding $f_1(u_1, \mathbf{x}_1)$ is minimized. The condition for making $U_1 = \gamma_1^*(\mathbf{x}_1) = i \in \{0, \ldots, L-1\}$ given $\mathbf{X}_1 = \mathbf{x}_1$ is

$$0 \geq f_1(i, \mathbf{x}_1) - f_1(j, \mathbf{x}_1),$$

$$= \int_{\mathbf{x}_2} \int_{\theta} \left[d(\theta, h^*(i, \gamma_2^*(\mathbf{x}_2))) - d(\theta, h^*(j, \gamma_2^*(\mathbf{x}_2))) \right] p(\theta|T_1(\mathbf{x}_1)) p(\mathbf{x}_2|\theta) d\theta d\mathbf{x}_2, \quad (7.25)$$

for any $j \in \{0, \ldots, L-1\}$. Therefore, (7.25) depends on $\mathbf{X}_1$ only through $T_1(\mathbf{X}_1)$.

The optimal quantizer $\gamma_2^*(\cdot)$ at the second node, given that $\gamma_1(\cdot)$ and $h(\cdot)$ take the form of $\gamma_1^*(\cdot)$ and $h^*(\cdot)$, can be similarly shown to be a function of the sufficient statistic $T_2(\mathbf{X}_2)$. Thus we have established that both $\gamma_1^*(\cdot)$ and $\gamma_2^*(\cdot)$ can be equivalently expressed as functions of $T_1(\mathbf{X}_1)$ and $T_2(\mathbf{X}_2)$ respectively, i.e., there exist $\gamma_1'(\cdot)$ and $\gamma_2'(\cdot)$ such that

$$\gamma_1'(T_1(\mathbf{X}_1)) = \gamma_1^*(\mathbf{X}_1), \quad (7.26)$$

$$\gamma_2'(T_2(\mathbf{X}_2)) = \gamma_2^*(\mathbf{X}_2). \quad (7.27)$$

Thus, the above $\gamma_1'(\cdot)$ and $\gamma_2'(\cdot)$, together with $h'(\cdot) = h^*(\cdot)$, achieves $C_{\min}$ for Fig. 7.2(b).

■

The fact that likelihood ratio quantizer is optimal for decentralized detection with conditionally independent observations can be naturally derived from the above general result.

Example 7. *Let $\theta \in \{0, 1\}$ and its estimate $\hat{\theta} \in \{0, 1\}$. The observations $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent given θ. Let $d(\cdot)$ take the form of $0-1$ cost, i.e., $d(\theta, \hat{\theta}) = 0$ when $\theta = \hat{\theta}$ and 1 otherwise. It is a trivial exercise to show $T_i(\mathbf{x}_i) = \frac{p(\mathbf{x}_i|\theta=1)}{p(\mathbf{x}_i|\theta=0)}$ is a sufficient statistic for θ with respect to $\mathbf{X}_i$. Thus quantizing $T_i(\mathbf{X}_i)$ is structurally optimal, which is consistent with [60] as the inference problem is precisely a hypothesis testing problem.*

7.3.2 Conditionally dependent observations

While the previous section establishes the optimality of sufficiency based data reduction for conditionally independent observations even with quantization constraints, Example 6 indicates that such is not the case with conditionally dependent observations. Nevertheless, in this section, we establish that within the problems involving dependent observations, there exist a class of problems such that quantizing sufficient statistics is still structurally optimal. Here we again utilize the HCI model [8].

Theorem 19. *Let* $\mathbf{W}$ *be a hidden variable such that the conditions for HCI model (6.3) are true. If* $T_1(\mathbf{X}_1)$ *and* $T_2(\mathbf{X}_2)$ *are local statistics that are sufficient with respect to* $\mathbf{W}$*, then quantizing* $T_1(\mathbf{X}_1)$ *and* $T_2(\mathbf{X}_2)$ *at the respective sensor is structurally optimal for the decentralized inference problem.*

Note that the first Markov chain in (6.3) indicates that $\mathbf{X}_1$ and $\mathbf{X}_2$ are conditionally independent given $\mathbf{W}$. If $T_1(\mathbf{X}_1)$ and $T_2(\mathbf{X}_2)$ are locally sufficient for $\mathbf{W}$, $(T_1(\mathbf{X}_1), T_2(\mathbf{X}_2))$ is globally sufficient for $\mathbf{W}$ and hence for θ by Corollary 3.

Proof. Let $C_{\min}$ be the minimum Bayesian cost achieved by Fig. 7.2(a) with the corresponding optimal quantizers $\gamma_i^*(\cdot)$, $i = 1, 2$, and estimator $h^*(\cdot)$. We show that $\gamma_i^*(\mathbf{X}_i)$ is necessarily a function of the sufficient statistic $T_i(\mathbf{X}_i)$.

Without loss of generality, we assume that $\mathbf{W}$ is continuous. From (6.3), we have

$$p(\mathbf{x}_1, \mathbf{x}_2|\theta) = \int_{\mathbf{W}} p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{w}|\theta) d\mathbf{w} \tag{7.28}$$

$$= \int_{\mathbf{W}} p(\mathbf{x}_1|\mathbf{w}) p(\mathbf{x}_2|\mathbf{w}) p(\mathbf{w}|\theta) d\mathbf{w} \tag{7.29}$$

$$= \int_{\mathbf{W}} \frac{p(\mathbf{w}|\mathbf{x}_1) p(\mathbf{x}_1)}{p(\mathbf{w})} p(\mathbf{x}_2|\mathbf{w}) p(\mathbf{w}|\theta) d\mathbf{w}. \tag{7.30}$$

$$= \int_{\mathbf{W}} \frac{p(\mathbf{w}|T_1(\mathbf{x}_1))p(\mathbf{x}_1)}{p(\mathbf{w})} p(\mathbf{x}_2|\mathbf{w})p(\mathbf{w}|\theta)d\mathbf{w}. \tag{7.31}$$

Expanding C with respect to $\mathbf{X}_1$, we obtain

$$C = \int_\theta \int_{\mathbf{X}_1} \int_{\mathbf{X}_2} d(\theta, h(\gamma_1(\mathbf{x}_1), \gamma_2(\mathbf{x}_2)))p(\mathbf{x}_1, \mathbf{x}_2, \theta)d\mathbf{x}_2 d\mathbf{x}_1 d\theta \tag{7.32}$$

$$= \int_\theta \int_{\mathbf{X}_1} \int_{\mathbf{X}_2} \int_{\mathbf{W}} d(\theta, h(\gamma_1(\mathbf{x}_1), \gamma_2(\mathbf{x}_2))) \frac{p(\mathbf{w}|T_1(\mathbf{x}_1))}{p(\mathbf{w})} p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{w})p(\mathbf{w}|\theta)d\mathbf{w}d\mathbf{x}_2 d\mathbf{x}_1 d\theta \tag{7.33}$$

$$\triangleq \int_{\mathbf{X}_1} f_1'(u_1, \mathbf{x}_1)p(\mathbf{x}_1)d\mathbf{x}_1, \tag{7.34}$$

where

$$f_1'(u_1, \mathbf{x}_1) \triangleq \int_\theta \int_{\mathbf{X}_2} \int_{\mathbf{W}} d(\theta, h(u_1, \gamma_2(\mathbf{x}_2))) \frac{p(\mathbf{w}|T(\mathbf{x}_1))}{p(\mathbf{w})} p(\mathbf{x}_2|\mathbf{y})p(\mathbf{w}|\theta)d\mathbf{w}d\mathbf{x}_2 d\theta. \tag{7.35}$$

Therefore, given $\gamma_2^*(\cdot)$ and $h^*(\cdot)$, for $\gamma_1^*(\cdot)$ to achieve $C_{\min}$, $\gamma_1^*(\mathbf{x}_1)$ must be such that $f_1'(u_1, \mathbf{x}_1)$ is minimized, i.e., $U_1 = i \in \{0, \ldots, L-1\}$ if

$$i = \arg\min_{u_1} f_1'(u_1, \mathbf{x}_1). \tag{7.36}$$

From (7.35), $\gamma_1^*(\mathbf{X}_1)$ depends on $\mathbf{X}_1$ only through $T_1(\mathbf{X}_1)$. Similar argument shows that $\gamma_2^*(\cdot)$ is also a function of the sufficient statistic $T_2(\mathbf{X}_2)$. ■

The key to applying the above result depends largely on a well chosen $\mathbf{W}$ for the HCI model. For example, the naïve choice of $\mathbf{W} = (\mathbf{X}_1, \mathbf{X}_2)$, while satisfying the defining Markov chains, does not result in any data reduction as the sufficient statistics for the data are nothing but the original data. The next two examples illustrate that carefully chosen $\mathbf{W}$ can indeed lead to meaningful data reduction without performance loss.

Example 8. *Consider Example 3 in Chapter 6 under the constraint of quantization, i.e., we need to estimate θ based on the quantized version of $\mathbf{X}_1$ and $\mathbf{X}_2$. Since $\sum_j X_{1j}$ and $\sum_j X_{2j}$ are locally sufficient for the hidden variable W, quantizing $\sum_j X_{1j}$ and $\sum_j X_{2j}$ is structurally optimal by Theorem 19.*

Example 9. *Consider Example 4 in Chapter 6 when quantization is needed at each node. In Example 4 we have shown that $\{|X_k|\}, k = 1, \cdots, K$ are globally sufficient for H. Therefore, from Theorem 19, quantizing $|X_k|$ at the kth sensor is structurally optimal. This result is consistent with that in [57] which shows that the optimal detector at each local sensor is an energy detector for the corresponding cooperative spectrum sensing problem, i.e., in the form of a threshold test using $|X_k|^2$.*

7.3.3 A general condition

Theorems 18 and 19 establish the structural optimality of sufficiency based data reduction with independent data and with dependent data under a given HCI dependence structure, respectively. In this section, we provide a unifying framework for these two cases. To proceed, we note that in Theorems 18 and 19 the joint distribution $p(\mathbf{x}_1, \mathbf{x}_2, \theta)$ can be expressed in both cases as the product of $p(\mathbf{x}_1)$ and a nonnegative function of $T_1(\mathbf{x}_1)$, $\mathbf{x}_2$ and θ. We show that this factorization is indeed what is needed to establish that quantizing $T_1(\mathbf{X}_1)$ achieves the same optimal inference performance as quantizing $\mathbf{X}_1$ given that the optimal quantizer $\gamma_2^*(\cdot)$ and the optimal estimator $h^*(\cdot)$ are used at the second sensor and at the fusion center respectively.

Theorem 20. *If there exist two nonnegative functions $g(\cdot)$ and $f(\cdot)$ and a statistic*

$T_1(\mathbf{X}_1)$ *such that*

$$p(\mathbf{x}_1, \mathbf{x}_2, \theta) = g(\mathbf{x}_1) f(T_1(\mathbf{x}_1), \mathbf{x}_2, \theta), \tag{7.37}$$

then quantizing $T_1(\mathbf{X}_1)$ *achieves the same optimal inference performance as quantizing* $\mathbf{X}_1$.

From (7.37), if we marginalize $\mathbf{X}_2$ on both sides, we have

$$p(\mathbf{x}_1, \theta) = g(\mathbf{x}_1) \int_{\mathbf{x}_2} f(T_1(\mathbf{x}_1), \mathbf{x}_2, \theta) d\mathbf{x}_2. \tag{7.38}$$

Thus, by the factorization theorem [25], (7.37) implies that $T_1(\mathbf{X}_1)$ is a local sufficient statistic for θ.

Proof. Let $C_{\min}$ be the minimum Bayesian cost achieved by Fig. 7.2(a) with quantizer $\gamma_i^*(\cdot)$ and estimator $h^*(\cdot)$. We show that, if (7.37) holds, then $\gamma_1^*(\mathbf{X}_1)$ depends on $\mathbf{X}_1$ only through the sufficient statistic $T_1(\mathbf{X}_1)$.

Again, expanding C with respect to $\mathbf{X}_1$, we get

$$\begin{aligned} C &= \int_\theta \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} d(\theta, h(\gamma_1(\mathbf{x}_1), \gamma_2(\mathbf{x}_2))) p(\mathbf{x}_1, \mathbf{x}_2, \theta) d\mathbf{x}_2 d\mathbf{x}_1 d\theta & (7.39) \\ &= \int_\theta \int_{\mathbf{x}_1} \int_{\mathbf{x}_2} d(\theta, h(\gamma_1(\mathbf{x}_1), \gamma_2(\mathbf{x}_2))) g(\mathbf{x}_1) f(T_1(\mathbf{x}_1), \mathbf{x}_2, \theta) d\mathbf{x}_2 d\mathbf{x}_1 d\theta & (7.40) \\ &\triangleq \int_{\mathbf{x}_1} \alpha_1(u_1, \mathbf{x}_1) g(\mathbf{x}_1) d\mathbf{x}_1, & (7.41) \end{aligned}$$

where

$$\alpha_1(u_1, \mathbf{x}_1) \triangleq \int_\theta \int_{\mathbf{x}_2} d(\theta, h(u_1, \gamma_2(\mathbf{x}_2))) f(T_1(\mathbf{x}_1), \mathbf{x}_2, \theta) d\mathbf{x}_2 d\theta. \tag{7.42}$$

Given the optimal second quantizer $\gamma_2^*(\cdot)$ and estimator $h^*(\cdot)$, $\gamma_1^*(\cdot)$ must be such that it minimizes $\alpha_1(u_1, \mathbf{x}_2)$, i.e., $U_1 = \gamma_1^*(\mathbf{x}_1) = i \in \{0, \ldots, L-1\}$ if

$$0 \geq \alpha_1(i, \mathbf{x}_1) - \alpha_1(j, \mathbf{x}_1), \tag{7.43}$$

for any $t \in \{0, \ldots, L-1\}$. The proof is thus complete by recognizing that $\alpha_1(u_1, \mathbf{x}_1)$ depends on $\mathbf{x}_1$ only through $T(\mathbf{x}_1)$. ■

The fact that (7.37) implies that $T_1(\mathbf{X}_1)$ is a sufficient statistic for $\mathbf{X}_1$ does not mean $T_1(\mathbf{X}_1)$ being a sufficient statistic is a necessary condition for optimality. This is because (7.37) itself is only a sufficient condition for optimality. Given below is a trivial example illustrating that a local statistic which achieves optimality is not necessarily a sufficient statistic.

Example 10. *For $i = 1, \cdots, n$, let*

$$\begin{aligned} X_{1i} &= \theta + W_i, \\ X_{2i} &= \theta + V_i, \end{aligned}$$

where $\theta, W_1, \cdots, W_n, V_1, \cdots, V_n$ are mutually independent Gaussian random variables such that $\theta \sim \mathcal{N}(0,1)$, $W_j \sim \mathcal{N}(0,1)$, $V_j \sim \mathcal{N}(0,1)$. Then $\mathbf{X}_1$ and $\mathbf{X}_2$ are conditionally independent given θ. It is also clear that $\sum_i X_{1i}$ and $\sum_i X_{2i}$ are locally sufficient for θ, thus quantizing $\sum_i X_{1i}$ and $\sum_i X_{2i}$ can achieve the optimal inference with corresponding quantizers $\gamma_1^*(\cdot)$ and $\gamma_2^*(\cdot)$ and the optimal estimator $h^*(\cdot)$.

Now consider another local statistic $U(\mathbf{X}_1) = \gamma_1^*(\sum_i X_{1i}) \in \{0,1\}$. If we quantize this statistic instead of $\sum_i X_{1i}$ at the first node while using $\gamma_2^*(\cdot)$ at the second node and $h^*(\cdot)$ at the fusion center, the optimal inference is also guaranteed. But it is straightforward to see that $U(\mathbf{X}_1)$ is not a sufficient statistic for θ.

The next example shows how to find local statistics for data reduction using Theorem 20.

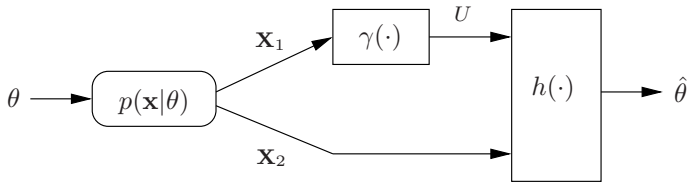
Example 11. *Let us reconsider Example 6 where* $\mathbf{X}_1 = \mathbf{X}_2$*. Now that quantizing* $(\mathbf{X}_1, \varnothing)$*, which is globally sufficient, does not achieve the optimal inference, one might ask that what local statistics can be used to achieve the same optimal inference performance as the raw data. Since this problem is equivalent to a centralized inference problem with a 2-bit quantizer, Theorem 17 implies that quantizing the minimal sufficient statistic* $M(\mathbf{X}_1)$ *at each sensor can achieve the optimal inference. But a minimal sufficient statistic is a function of any other sufficient statistic [25], thus any local sufficient statistics* $(T_1(\mathbf{X}_1), T_2(\mathbf{X}_2))$ *at each node attains the structural optimality.*

The same results can also be obtained using Theorem 20. As $\mathbf{X}_1 = \mathbf{X}_2$*, we have*

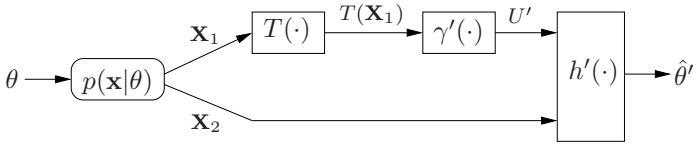
$$p(\mathbf{x}_1, \mathbf{x}_2, \theta) = p(\mathbf{x}_1, \theta) \overset{(a)}{=} p(\mathbf{x}_1) p(\theta | T_1(\mathbf{x}_1)), \tag{7.44}$$

where (a) *is from Lemma 1. We see that (7.44) is exactly in the same form as in Theorem 20. It follows that quantizing* $T_1(\mathbf{X}_1)$ *is sufficient to achieve the optimal inference given the second optimal quantizer* $\gamma_2^*(\cdot)$ *and the optimal estimator* $h^*(\cdot)$*. Similarly, if we let* $T_2(\mathbf{X}_2)$ *be any sufficient statistic with respect to* $\mathbf{X}_2$ *and rewrite* $p(\mathbf{x}_1, \mathbf{x}_2, \theta)$ *as* $p(\mathbf{x}_2) p(\theta | T_2(\mathbf{x}_2))$*, it is straightforward to see that* $T_2(\mathbf{X}_2)$ *is also sufficient for the optimal inference. Therefore, we may use* $T_1(\mathbf{X}_1)$ *and* $T_2(\mathbf{X}_2)$ *at each sensor to achieve data reduction prior to quantization and still attains the optimal inference.*

Note that while any local sufficient statistics $(T_1(\mathbf{X}_1), T_2(\mathbf{X}_2))$ *preserve the optimal inference performance for this degraded observation model, they may not achieve the same degree of data reduction as that of the minimal sufficient statistic* $M(\mathbf{X}_1)$*.*



(a)



(b)

Figure 7.3: Decentralized inference systems for tandem networks with quantizer operating on (a) the raw data $\mathbf{X}_1$, (b) a statistic $T_1(\mathbf{X}_1)$.

7.4 Decentralized data reduction with quantization constraints in tandem networks

In this section, we consider decentralized data reduction with quantization constraints in tandem networks as illustrated in Fig. 7.3. In a tandem network, local decisions propagate sequentially until they reach the last sensor which also serves as the fusion center. We consider a simple two-sensor tandem in this section. As one can see, the tandem network in Fig. 7.3(a) has the whole observation $\mathbf{X}_2$ available at the fusion center while the parallel network in Fig. 7.2(a) only has access to a function of $\mathbf{X}_2$ at the fusion center. Since $\mathbf{X}_2$ is a function of itself, one may guess that quantizing a sufficient statistic $T(\mathbf{X}_1)$ at the first node is structurally optimal. Indeed, this is true under the general condition as in Theorem 20, which includes both conditionally independent and dependent observations.

Parallel to the parallel networks, we have the following Theorem for tandem networks.

Theorem 21. *If there exist two nonnegative functions $g(\cdot)$ and $f(\cdot)$ and a statistic $T_1(\mathbf{X}_1)$ such that*

$$p(\mathbf{x}_1, \mathbf{x}_2, \theta) = g(\mathbf{x}_1) f(T_1(\mathbf{x}_1), \mathbf{x}_2, \theta), \tag{7.45}$$

then for the tandem network as in Fig. 7.3, quantizing $T_1(\mathbf{X}_1)$ achieves the same optimal inference performance as quantizing $\mathbf{X}_1$.

Proof. Similar to the parallel networks case, we can define the Bayasian costs for Fig. 7.3(a) and Fig. 7.3(b) as follows

$$C = E[d(\theta, h(\gamma(\mathbf{X}_1), \mathbf{X}_2))], \tag{7.46}$$

$$C' = E[d(\theta, h'(\gamma'(T(\mathbf{X}_1)), \mathbf{X}_2))]. \tag{7.47}$$

Let $C_{\min} = \min_{\gamma,h} C$, then the proof follows exactly the same steps as the proof of Theorem 20 except that $\gamma_2(\mathbf{X}_2)$ is replaced with $\mathbf{X}_2$. ■

Since the condition (7.45) in Theorem 21 includes the conditionally independent data and dependent data under a given HCI model as special cases, similar to the results in Theorem 18 and 19 for parallel networks, we have the following conclusions for tandem networks.

- If $\mathbf{X}_1$ and $\mathbf{X}_2$ are conditionally independent given θ, then quantizing a locally sufficient statistic $T(\mathbf{X}_1)$ is optimal.
- If $\mathbf{X}_1, \mathbf{X}_2$ and θ satisfy a HCI model with hidden variable $\mathbf{W}$ and $T(\mathbf{X}_1)$ is a locally sufficient statistic with respect to $\mathbf{W}$, then quantizing $T(\mathbf{X}_1)$ at the first sensor is structurally optimal for the decentralized inference problem.

Example 12. *Consider Example 7 under the tandem setting. Since the local likelihood ratio* $T_1(\mathbf{x}_1) = \frac{p(\mathbf{x}_1|\theta=1)}{p(\mathbf{x}_1|\theta=0)}$ *is a sufficient statistic for* θ *with respect to* $\mathbf{X}_1$*, quantizing* $T(\mathbf{X}_1)$ *is structurally optimal by Theorem 21, which is consistent with the result in [66].*

One may ask if the general condition (7.25) in Theorem 21 is the same as the definition of conditional sufficiency discussed in Chapter 6. Actually, we can show that the condition (7.25) is a special case of conditional sufficiency. This is because (7.25) implies (6.8) in Theorem 13. Also, (7.25) implies that $T_1(\mathbf{X}_1)$ is a locally sufficient statistic for θ while for the definition of conditional sufficiency, $T(\mathbf{X}_1)$ is not necessarily a locally sufficient statistic.

7.5 Summary

In this chapter we have investigated the decentralized data reduction problem when each sensor is subject to a quantization constraint. We do not address explicit quantizer design in this work; instead, we find sufficient conditions such that a separation approach, namely data reduction followed by a quantizer, is structurally optimal under the Bayesian inference framework for both centralized inference and decentralized inference with conditionally independent observations. We consider the problem under both the parallel and tandem network frameworks. For decentralized inference with conditionally dependent observations, quantizing sufficient statistics, even global ones, need not be optimal. Nevertheless, utilizing the HCI model, we have provided a suitable way of finding optimal data reduction if it exists. We have also established a unifying condition that encompasses both the independent and the dependent ob-

servation cases.

Chapter 8

Conclusion and Future Research

8.1 Conclusion

Correlated observations are often present in many engineering applications. In this thesis, we have focused on the characterization of the dependence of correlated observations and on decentralized inference problems with dependent observations. This thesis consists of two part. In the first part, we attempt to make progress toward a better understanding of Wyner's common information among random variables, both in its generalization to much more general settings and in its operational interpretation that has not been discovered before. In the second part, we address decentralized inference involving dependent observations with an emphasis on the development of the sufficiency principle for distributed data reduction.

For the first part, we have generalized Wyner's common information, defined originally for a pair of discrete random variables, to that of multiple random variables with arbitrary alphabets. We show that Wyner's original interpretations of common

information can be directly extended to that involving multiple variables. We then developed a new interpretation of Wyner's common information using the Gray-Wyner network that applies to continuous random variables. That is, for the Gray-Wyner network, Wyner's common information is precisely the smallest common message rate for a certain range of distortion constraints when the total rate is arbitrarily close to the rate distortion function with joint decoding. As the common information is only a function of the joint distribution, this smallest common rate remains constant even if the distortion constraints vary, as long as they are in a specific distortion region.

Evaluating the generalized common information has been studied for the two special but very important examples: the binary sources and Gaussian sources. In particular, we derived, through an estimation theoretic approach, the common information for a bivariate Gaussian source and its extension to the multi-variate case with a certain correlation structure.

We established a monotone property of Wyner's common information in the number of variables which is in contrast to other notions of common information. The application of Wyner's common information to simple Bayesian inference models was explored where the observations are assumed to be exchangeable random variable. It is shown that for infinite exchangeable sequences, the common information is asymptotically equal to the information of the inference object, i.e., the hidden variable in the Bayesian model. For finite exchangeable sequences, while this result is no longer true in general, we identify two important cases such that the result still holds. For these two cases, one binary and the other Gaussian, we further established the relationship between the common information and various inference performance metrics.

For the second part, we first considered the problem of distributed detection with

conditionally dependent observations utilizing the hierarchical conditional independence model. Under the Bayesian detection framework, we identified a more general condition associated with the hidden variable for the CHCI model which enables us to tackle a much broader class of distributed detection problems with dependent observations.

We developed the sufficiency principle that guides local data reduction in networked inference with dependent observations for both the parallel network and tandem network. For the parallel network, the HCI model is used to obtain conditions such that local sufficiency implies global sufficiency. For the tandem network, a new notion of conditional sufficiency was proposed to capture the structure of tandem network.

Finally, we have studied decentralized data reduction in distributed inference when each sensor is subject to a quantization constraint in both the parallel and tandem networks. The sufficiency based data reduction was shown to be structurally optimal under the Bayesian inference framework for decentralized inference with conditionally independent observations. For decentralized inference with conditionally dependent observations, utilizing the HCI model, we provided a suitable way of finding optimal data reduction if it exists. Finally, a unifying condition that encompasses both the independent and the dependent observation cases was established.

8.2 Future work

Wyner provided two approaches to interpret Wyner's common information: one is based on lossless source coding for the Gray-Wyner network and the other on a

distribution approximation problem. We have shown that the common information admits a lossy source coding interpretation using the Gray-Wyner network, thus provide justification for the generalization to continuous random variables. It is natural to ask if Wyner's second approach, that of distribution approximation, also applies to continuous random variables.

In Chapter 4, we concluded that the common information for finite exchangeable sequences does not equal to the information of the inference object in a simple Bayesian inference model. Nevertheless, for the binary and the Gaussian cases, the equality holds. A possible direction of future work is to find more general condition under which the common information does capture the entire information of the inference object. A promising approach to examine the class of exchangeable random variables that correspond to additive Bayesian model as is the case for the binary and Gaussian cases.

In Chapter 6, we have defined the notion of conditional sufficiency intended for application in tandem networks. On the other hand, Theorem 21 in Chapter 7 provides a sufficient condition under which the sufficiency based data reduction is structurally optimal for tandem networks. As discussed in Section 7.4, the condition in Theorem 21 is a special case of conditional sufficiency. Our future work is to explore if the conditional sufficiency based data reduction can be proved to be optimal for tandem networks.

We have developed the sufficiency principle for inference networks with and without quantization constraints. Besides sufficient statistics, the notions of ancillary and complete statistics are also of great importance in statistical inference. Ancillary statistics are functions of observations that are independent of the parameter of infer-

ence interest and play an important role for inference problems involving the class of complete distributions. One direction of research is to develop the theory of ancillary and complete statistics for networked inference problems.

Appendix A

Proof of Theorem 5

We first show that $C_3(D_1, D_2) \geq \tilde{C}(D_1, D_2)$. Let R_0 be (D_1, D_2)-achievable, then there exists an (n, M_0, M_1, M_2) code such that (3.12)-(3.14) are satisfied. Define $R_i = \frac{1}{n} \log M_i$ for $i = 1, 2$. Since (R_0, R_1, R_2) is (D_1, D_2)-achievable, from Theorem 2, there exists a W such that

$$\begin{aligned} R_0 &\geq I(X_1, X_2; W), \\ R_i &\geq R_{X_i|W}(D_i), \quad i = 1, 2 \end{aligned}$$

and for any $\epsilon > 0$,

$$\sum_{i=0}^{2} R_i \leq R_{X_1X_2}(D_1, D_2) + \epsilon. \tag{A.1}$$

Therefore,

$$\begin{aligned} R_{X_1X_2}(D_1, D_2) + \epsilon &\geq \sum_{i=0}^{2} R_i && \text{(A.2)} \\ &\geq I(X_1, X_2; W) + \sum_{i=1}^{2} R_{X_i|W}(D_i) && \text{(A.3)} \\ &\geq I(X_1, X_2; W) + R_{X_1X_2|W}(D_1, D_2) && \text{(A.4)} \end{aligned}$$

$$\geq R_{X_1X_2}(D_1, D_2) \tag{A.5}$$

where (A.4) is from (3.6b) and (A.5) comes from (3.5b). Thus, we have

$$I(X_1, X_2; W) + R_{X_1|W}(D_1) + R_{X_2|W}(D_2) = R_{X_1X_2}(D_1, D_2). \tag{A.6}$$

Hence, if R_0 is (D_1, D_2)-achievable, there exists a W such that $R_0 \geq I(X_1, X_2; W)$ and (A.6) is true. It shows that $C_3(D_1, D_2) \geq \tilde{C}(D_1, D_2)$.

Next we show $C_3(D_1, D_2) \leq \tilde{C}(D_1, D_2)$. Let W' be the random variable that achieves $\tilde{C}(D_1, D_2)$. For any $R_0 > \tilde{C}(D_1, D_2)$ and $\epsilon > 0$, let

$$\epsilon_1 = \min\left\{\frac{\epsilon}{3}, R_0 - \tilde{C}(D_1, D_2)\right\}, \tag{A.7}$$

and hence $\epsilon_1 > 0$. From Theorem 2, there exists an (n, M_0, M_1, M_2) code with $Ed_1(X_1, \hat{X}_1) \leq D_1$, $Ed_2(X_2, \hat{X}_2) \leq D_2$, and

$$\begin{aligned} \frac{1}{n}\log M_0 &\leq I(X_1, X_2; W') + \epsilon_1 = \tilde{C}(D_1, D_2) + \epsilon_1 \leq R_0, & \text{(A.8)} \\ \frac{1}{n}\log M_i &\leq R_{X_i|W'}(D_i) + \epsilon_1, & \text{(A.9)} \end{aligned}$$

for $i = 1, 2$. Sum over (A.8) and (A.9), we get

$$\begin{aligned} \sum_{i=0}^{2}\frac{1}{n}\log M_i &\leq I(X_1, X_2; W') + \sum_{i=1}^{2} R_{X_i|W'}(D_i) + 3\epsilon_1 \\ &\leq R_{X_1X_2}(D_1, D_2) + \epsilon, & \text{(A.10)} \end{aligned}$$

where inequality (A.10) comes from (A.7) and definition of $\tilde{C}(D_1, D_2)$.

This proves that R_0 is (D_1, D_2)-achievable, thus completes the proof of $C_3(D_1, D_2) \leq \tilde{C}(D_1, D_2)$.

Appendix B

Direct proof of

$\tilde{C}(D_1, D_2) = C^*(D_1, D_2)$

First we show that $\tilde{C}(D_1, D_2) \geq C^*(D_1, D_2)$. Let W be the variable that achieves $\tilde{C}(D_1, D_2)$ and let $\hat{X}_1, \hat{X}_2$ be random variables that achieve $R_{X_1|W}(D_1)$ and $R_{X_2|W}(D_2)$, i.e.,

$$
\begin{aligned}
I(X_1, X_2; W) + R_{X_1|W}(D_1) + R_{X_2|W}(D_2) &= R_{X_1X_2}(D_1, D_2), && \text{(B.1)} \\
R_{X_1|W}(D_1) &= I(X_1; \hat{X}_1|W), && \text{(B.2)} \\
R_{X_2|W}(D_2) &= I(X_2; \hat{X}_2|W), && \text{(B.3)} \\
E[d_1(X_1, \hat{X}_1)] &\leq D_1, && \text{(B.4)} \\
E[d_2(X_2, \hat{X}_2)] &\leq D_2. && \text{(B.5)}
\end{aligned}
$$

Without loss of generality, we can assume that the joint distribution of $(X_1, X_2, \hat{X}_1, \hat{X}_2, W)$ factors as $p(x_1, x_2, \hat{x}_1, \hat{x}_2, w) = p(x_1, x_2, w)p(\hat{x}|x, w)p(\hat{y}|y, w)$ because the distortion

D_1 is independent of X_2 and D_2 is independent of X_1. We now establish

$$R_{X_1X_2|W}(D_1, D_2) = R_{X_1|W}(D_1) + R_{X_2|W}(D_2). \tag{B.6}$$

This is from (B.1) and the inequalities

$$\begin{aligned} R_{X_1X_2|W}(D_1, D_2) + I(X_1, X_2; W) &\geq R_{X_1X_2}(D_1, D_2), & \text{(B.7)} \\ R_{X_1|W}(D_1) + R_{X_2|W}(D_2) &\geq R_{X_1X_2|W}(D_1, D_2), & \text{(B.8)} \end{aligned}$$

from Lemma 4. Therefore, together with (B.1)-(B.5), we have

$$\begin{aligned} R_{X_1X_2|W}(D_1, D_2) &= I(X_1; \hat{X}_1|W) + I(X_2; \hat{X}_2|W) & \text{(B.9)} \\ &= H(\hat{X}_1|W) + H(\hat{X}_2|W) - H(\hat{X}_1|X_1, W) - H(\hat{X}_2|X_2, W) & \text{(B.10)} \\ &\geq H(\hat{X}_1, \hat{X}_2|W) - H(\hat{X}_1|X_1, W) - H(\hat{X}_2|X_2, W) & \text{(B.11)} \\ &= H(\hat{X}_1, \hat{X}_2|W) - H(\hat{X}_1|W, X_1, X_2) - H(\hat{X}_2|W, X_1, X_2) & \text{(B.12)} \\ &= I(X_1, X_2; \hat{X}_1, \hat{X}_2|W) & \text{(B.13)} \\ &\geq R_{X_1X_2|W}(D_1, D_2). & \text{(B.14)} \end{aligned}$$

As the left-hand side (LHS) and right-hand side (RHS) of the above inequalities are the same, all the inequalities must be equalities so we have

$$I(\hat{X}_1; \hat{X}_2|W) = 0. \tag{B.15}$$

Then we have

$$\begin{aligned} & R_{X_1X_2}(D_1, D_2) \\ = \; & I(X_1, X_2; W) + I(X_1; \hat{X}_1|W) + I(X_2; \hat{X}_2|W) & \text{(B.16)} \\ = \; & I(X_1, X_2; W, \hat{X}_1, \hat{X}_2) - I(X_1, X_2; \hat{X}_1, \hat{X}_2|W) + I(X_1; \hat{X}_1|W) + I(X_2; \hat{X}_2|W) & \text{(B.17)} \end{aligned}$$

$$
\begin{aligned}
&= I(X_1, X_2; \hat{X}_1, \hat{X}_2) + I(X_1, X_2; W|\hat{X}_1, \hat{X}_2) && \text{(B.18)}\\
&\geq I(X_1, X_2; \hat{X}_1, \hat{X}_2) && \text{(B.19)}\\
&\geq R_{X_1X_2}(D_1, D_2).
\end{aligned}
$$

As the LHS and RHS of the above inequalities are the same, all the inequalities must be equalities so we have

$$
\begin{aligned}
I(X_1, X_2; W|\hat{X}_1, \hat{X}_2) &= 0, && \text{(B.20)}\\
I(X_1, X_2; \hat{X}_1, \hat{X}_2) &= R_{X_1X_2}(D_1, D_2). && \text{(B.21)}
\end{aligned}
$$

Therefore, $X_1, X_2, \hat{X}_1, \hat{X}_2, W$ satisfy the Markov chains in (3.18) and (3.19) and $\hat{X}_1, \hat{X}_2$ achieve $R_{X_1X_2}(D_1, D_2)$. Thus, $\tilde{C}(D_1, D_2) \geq C^*(D_1, D_2)$.

Next we show that $\tilde{C}(D_1, D_2) \leq C^*(D_1, D_2)$. Let $X_1, X_2, X_1^*, X_2^*, W$ achieve $C^*(D_1, D_2)$. Therefore, they satisfy the Markov chains in (3.18) and (3.19) and $I(X_1, X_2; X_1^*, X_2^*) = R_{X_1X_2}(D_1, D_2)$ and $E[d_1(X_1, X_1^*)] \leq D_1, E[d_2(X_2, X_2^*)] \leq D_2$.

$$
\begin{aligned}
& R_{X_1X_2}(D_1, D_2)\\
&= I(X_1, X_2; X_1^*, X_2^*) && \text{(B.22)}\\
&= I(X_1, X_2; W, X_1^*, X_2^*) && \text{(B.23)}\\
&= I(X_1, X_2; W) + I(X_1, X_2; X_1^*, X_2^*|W) && \text{(B.24)}\\
&= I(X_1, X_2; W) + H(X_1^*|W) + H(X_2^*|W) - H(X_1^*, X_2^*|X_1, X_2, W) && \text{(B.25)}\\
&= I(X_1, X_2; W) + I(X_1; X_1^*|W) + I(X_2; X_2^*|W) + H(X_1^*|X_1, W) && \text{(B.26)}\\
&\quad + H(X_2^*|X_2, W) - H(X_1^*, X_2^*|X_1, X_2, W) && \text{(B.27)}\\
&\geq I(X_1, X_2; W) + I(X_1; X_1^*|W) + I(X_2; X_2^*|W) + H(X_1^*|X_1, X_2, W) && \text{(B.28)}\\
&\quad + H(X_2^*|X_1, X_2, W) - H(X_1^*, X_2^*|X_1, X_2, W) && \text{(B.29)}
\end{aligned}
$$

$$= I(X_1, X_2; W) + I(X_1; X_1^*|W) + I(X_2; X_2^*|W) + I(X_1^*; X_2^*|X_1, X_2, W) \quad \text{(B.30)}$$

$$\geq I(X_1, X_2; W) + I(X_1; X_1^*|W) + I(X_2; X_2^*|W) \quad \text{(B.31)}$$

$$\geq I(X_1, X_2; W) + R_{X_1|W}(D_1) + R_{X_2|W}(D_2) \quad \text{(B.32)}$$

$$\geq I(X_1, X_2; W) + R_{X_1X_2|W}(D_1, D_2) \quad \text{(B.33)}$$

$$\geq R_{X_1X_2}(D_1, D_2), \quad \text{(B.34)}$$

where (B.23) is from the Markov chain $(X_1, X_2) - (X_1^*, X_2^*) - W$, (B.25) is from the Markov chain $X_1^* - W - X_2^*$, (B.29) is because conditioning reduces entropy, (B.33) and (B.34) are by the properties of rate distortion functions. As the LHS and RHS of the above inequalities are the same, all the inequalities must be equalities so we have

$$I(X_1, X_2; W) + R_{X_1|W}(D_1) + R_{X_2|W}(D_2) = R_{X_1X_2}(D_1, D_2). \quad \text{(B.35)}$$

Therefore, $C^*(D_1, D_2) = I(X_1, X_2; W) \geq \tilde{C}(D_1, D_2)$.

Appendix C

Proof of Lemma 7

Let W be the random variable that achieves $C_3(D_1, D_2)$. Thus,

$$C_3(D_1, D_2) = I(X_1, X_2; W), \tag{C.1}$$

with

$$R_{X_1|W}(D_1) + R_{X_2|W}(D_2) + I(X_1, X_2; W) = R_{X_1X_2}(D_1, D_2). \tag{C.2}$$

Combined with (3.23), we have that

$$\begin{aligned}
& R_{X_1}(D_1) + R_{X_2}(D_2) - I(X_1; X_2) \\
= \;& R_{X_1|W}(D_1) + R_{X_2|W}(D_2) + I(X_1, X_2; W) && \text{(C.3)} \\
\geq \;& R_{X_1}(D_1) - I(X_1; W) + R_{X_2}(D_2) - I(X_2; W) && \text{(C.4)} \\
& + I(X_1, X_2; W) && \text{(C.5)} \\
= \;& R_{X_1}(D_1) + R_{X_2}(D_2) - I(X_1; X_2) + I(X_1; X_2|W) && \text{(C.6)} \\
\geq \;& R_{X_1}(D_1) + R_{X_2}(D_2) - I(X_1; X_2), && \text{(C.7)}
\end{aligned}$$

where equation (C.3) is from equations (C.2) and (3.23), inequality (C.5) comes from Lemma 4, (C.6) is by the chain rule and inequality (C.7) is by the fact that $I(X_1; X_2|W) \geq 0$.

Because the LHS of (C.3) is the same as the RHS of (C.7), we can conclude that all the inequalities above should be equalities. This implies $I(X_1; X_2|W) = 0$, i.e., X_1 and X_2 are conditional independent given W. Therefore,

$$C(X_1, X_2) \leq I(X_1, X_2; W) = C_3(D_1, D_2), \tag{C.8}$$

the proof is complete.

Appendix D

Proof of Theorem 6

Let W be the random variable that achieves the common information of X_1, X_2. By Lemma 5, there exists a strictly positive surface $\mathcal{D}(X_1X_2|W)$ such that for any $0 \leq (D_1, D_2) \leq \mathcal{D}(X_1X_2|W)$,

$$I(X_1, X_2; W) + R_{X_1X_2|W}(D_1, D_2) = R_{X_1X_2}(D_1, D_2). \tag{D.1}$$

Also by Lemma 5, there exists a strictly positive surface $\mathcal{D}(X_1X_2) \geq \mathcal{D}(X_1X_2|W)$ such that for any $0 \leq (D_1, D_2) \leq \mathcal{D}(X_1X_2)$,

$$R_{X_1}(D_1) + R_{X_2}(D_2) - I(X_1; X_2) = R_{X_1X_2}(D_1, D_2). \tag{D.2}$$

Since $\mathcal{D}(X_1X_2|W) \leq \mathcal{D}(X_1X_2)$, let $\gamma = \mathcal{D}(X_1X_2|W)$, both equalities (D.1) and (D.2) hold for $0 \leq (D_1, D_2) \leq \gamma$. Therefore, from Lemmas 6 and 7, $C_3(D_1, D_2) = C(X_1, X_2)$ for $0 \leq (D_1, D_2) \leq \gamma$.

Appendix E

Proof of Theorem 7

First we show that for any $(D_1, D_2) \leq (D_1^0, D_2^0)$,

$$R_{X_1X_2|W}(D_1, D_2) + I(X_1, X_2; W) = R_{X_1X_2}(D_1, D_2). \tag{E.1}$$

In Proposition 8, we have shown that $R_{X_1X_2}(D_1^0, D_2^0) = I(X_1, X_2; \hat{X}_1^0, \hat{X}_2^0) = I(X_1X_2; W)$.

Now, let $(\hat{X}_1, \hat{X}_2)$ achieve $R_{X_1X_2}(D_1, D_2)$. As the vector source (X_1, X_2) is successively refinable under individual distortion constraints [33], we have the Markov chain $X_1X_2 - \hat{X}_1\hat{X}_2 - \hat{X}_1^0\hat{X}_2^0$. Therefore,

$$\begin{aligned} R_{X_1X_2}(D_1, D_2) - I(X_1, X_2; W) &= I(X_1, X_2; \hat{X}_1, \hat{X}_2) - I(X_1, X_2; \hat{X}_1^0, \hat{X}_2^0) && \text{(E.2)} \\ &= I(X_1X_2; \hat{X}_1\hat{X}_2|\hat{X}_1^0, \hat{X}_2^0) && \text{(E.3)} \\ &\geq R_{X_1X_2|\hat{X}_1^0,\hat{X}_2^0}(D_1, D_2) && \text{(E.4)} \\ &\geq R_{X_1X_2|W}(D_1, D_2), && \text{(E.5)} \end{aligned}$$

where the last inequality is from the Markov chain $X_1X_2 - W - \hat{X}_1^0, \hat{X}_2^0$. On the other

hand, by Lemma 4, we have

$$R_{X_1X_2|W}(D_1, D_2) + I(X_1X_2; W) \geq R_{X_1X_2}(D_1, D_2). \tag{E.6}$$

This establishes (E.1). Thus, from Lemma 6, $C_3(D_1, D_2) \leq C(X_1; X_2)$.

To complete the proof, we only need to show

$$R_{X_1}(D_1) + R_{X_2}(D_2) - I(X_1; X_2) = R_{X_1X_2}(D_1, D_2). \tag{E.7}$$

From Lemma 4,

$$R_{X_1}(D_1) + R_{X_2}(D_2) - I(X_1; X_2) \leq R_{X_1X_2}(D_1, D_2). \tag{E.8}$$

Therefore, we only need to establish the other direction. For $i = 1, 2$, let $\hat{X}_i$ achieve $R_{X_i}(D_i)$, then by the definition of a successively refinable scalar source [32], we have the Markov chain $X_i - \hat{X}_i - \hat{X}_i^0$ for $D_i \leq D_i^0$. Therefore,

$$\begin{aligned} R_{X_i}(D_i) - I(X_i; W) &= I(X_i; \tilde{X}_i) - I(X_i; \hat{X}_i^0) && \text{(E.9)} \\ &= I(X_i; \hat{X}_i | \hat{X}_i^0) && \text{(E.10)} \\ &\geq R_{X_i|\hat{X}_i^0}(D_i) && \text{(E.11)} \\ &\geq R_{X_i|W}(D_i), && \text{(E.12)} \end{aligned}$$

where (E.12) is from the Markov chain $X_i - W - \hat{X}_i^0$. Using (E.12), we have

$$\begin{aligned} & R_{X_1}(D_1) + R_{X_2}(D_2) - I(X_1; X_2) && \text{(E.13)} \\ \geq\ & R_{X_1|W}(D_1) + I(X_1; W) + R_{X_2|W}(D_1) + I(X_2; W) - I(X_1; X_2) && \text{(E.14)} \\ =\ & R_{X_1|W}(D_1) + R_{X_2|W}(D_2) + I(X_1X_2; W) && \text{(E.15)} \\ =\ & R_{X_1X_2|W}(D_1, D_2) + I(X_1X_2; W) && \text{(E.16)} \\ =\ & R_{X_1X_2}(D_1, D_2), && \text{(E.17)} \end{aligned}$$

which completes the proof.

Appendix F

Derivation of Wyner's Common Information for Bivariate Gaussian Sources

First, we will show that the common information of (X_1, X_2) is only a function of the correlation coefficient ρ. To show this, let $\tilde{X}_i = \frac{1}{\sigma_i} X_i$, $i = 1, 2$, thus $\tilde{X}_1, \tilde{X}_2$ are joint Gaussian distributed with zero mean and covariance matrix

$$\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

We have the Markov chain that $\tilde{X}_1 - X_1 - X_2 - \tilde{X}_2$ and by the data processing inequality for Wyner's common information [6], $C(\tilde{X}_1, \tilde{X}_2) \leq C(X_1, X_2)$. On the other hand, we have the Markov chain that $X_1 - \tilde{X}_1 - \tilde{X}_2 - X_2$ and $C(\tilde{X}_1, \tilde{X}_2) \leq C(X_1, X_2)$. Thus, $C(\tilde{X}_1, \tilde{X}_2) = C(X_1, X_2)$. Without loss generality, we will consider $\sigma_1^2 = \sigma_2^2 = 1$ in the following.

Let

$$X_i = \sqrt{\rho}W + \sqrt{1-\rho}N_i,\ i = 1, 2, \tag{F.1}$$

where W, N_1, N_2 are mutually independent standard Gaussian random variables. It is clear that X_1, X_2 are bivariate Gaussian with correlation coefficient ρ,

$$C(X_1, X_2) \leq I(X_1, X_2; W) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}. \tag{F.2}$$

Next we will show that

$$C(X_1, X_2) \geq \frac{1}{2}\log\frac{1+\rho}{1-\rho}. \tag{F.3}$$

For any U that satisfies the Markov chain $X_1 - U - X_2$, let D_1 be the minimum mean square error (MMSE) of estimating X_1 using U, thus, $D_1 = E(X_1 - E(X_1|U))^2$. Similarly, let $D_2 = E(X_2 - E(X_2|U))^2$. We now show that $I(X_1X_2; U) \geq \frac{1}{2}\log\frac{1+\rho}{1-\rho}$.

$$\begin{aligned}
I(X_1X_2; U) &= H(X_1X_2) - H(X_1|U) - H(X_2|U) && \text{(F.4)}\\
&= I(X_1; U) + I(X_2; U) - I(X_1; X_2) && \text{(F.5)}\\
&\geq I(X_1; E(X_1|U)) + I(X_2; E(X_2|U)) - I(X_1; X_2) && \text{(F.6)}\\
&\geq R_{X_1}(D_1) + R_{X_2}(D_2) - I(X_1; X_2) && \text{(F.7)}\\
&= \frac{1}{2}\log\frac{1-\rho^2}{D_1D_2}, && \text{(F.8)}
\end{aligned}$$

for $D_1 \leq 1, D_2 \leq 1$, where (F.5) is from the chain rule, (F.6) is from the Markov chains $X_1 - U - E(X_1|U)$, $X_2 - U - E(X_2|U)$ and (F.7) is by the definition of rate distortion function.

Next we show that $D_1 + D_2 \leq 2(1-\rho)$, $D_1 \leq 1$, $D_2 \leq 1$.

$2(1-\rho)$

$$= E(X_1 - X_2)^2 \tag{F.9}$$

$$= E[X_1 - E(X_1|U) + E(X_1|U) - X_2]^2 \tag{F.10}$$

$$= E[X_1 - E(X_1|U)]^2 + E[E(X_1|U) - X_2]^2 + 2E[(X_1 - E(X_1|U))(E(X_1|U) - X_2)] \tag{F.11}$$

$$= E[X_1 - E(X_1|U)]^2 + E[E(X_1|U) - X_2]^2 \tag{F.12}$$

$$= E[X_1 - E(X_1|U)]^2 + E[E(X_1|U) - E(X_2|U) + E(X_2|U) - X_2]^2 \tag{F.13}$$

$$= E[X_1 - E(X_1|U)]^2 + E[X_2 - E(X_2|U)]^2 + E[E(X_2|U) - E(X_1|U)]^2 \tag{F.14}$$

$$+ E[(X_2 - E(X_2|U))(E(X_2|U) - E(X_1|U))] \tag{F.15}$$

$$= E[X_1 - E(X_1|U)]^2 + E[X_2 - E(X_2|U)]^2 + E[E(X_2|U) - E(X_1|U)]^2 \tag{F.16}$$

$$\geq D_1 + D_2 \tag{F.17}$$

where (F.12) is from

$$E[(X_1 - E(X_1|U))(E(X_1|U) - X_2)]$$

$$= E[(X_1 - E(X_1|U))E(X_1|U)] - E[(X_1 - E(X_1|U))X_2] \tag{F.18}$$

$$= -E[(X_1 - E(X_1|U))X_2] \tag{F.19}$$

$$= -E_{UX_2}[X_2 E_{X_1|U}[X_1 - E(X_1|U)]] \tag{F.20}$$

$$= -E_{UX_2}[X_2(E(X_1|U) - E(X_1|U))] \tag{F.21}$$

$$= 0, \tag{F.22}$$

and (F.16) is from

$$E[(X_2 - E(X_2|U))(E(X_2|U) - E(X_1|U))]$$

$$= E[(X_2 - E(X_2|U))E(X_2|U)] - E[(X_2 - E(X_2|U))E(X_1|U)] \tag{F.23}$$

$$= 0 \tag{F.24}$$

In addition, we have $D_1 = E[X_1 - E(X_1|U)]^2 = EX_1^2 - E[E(X_1|U)^2] \leq EX_1^2 = 1$.

Thus, we have

$$
\begin{aligned}
I(X_1X_2;U) &\geq \frac{1}{2}\log\frac{1-\rho^2}{D_1D_2} && \text{(F.25)}\\
&\geq \frac{1}{2}\log\frac{1-\rho^2}{\left(\frac{D_1+D_2}{2}\right)^2} && \text{(F.26)}\\
&\geq \frac{1}{2}\log\frac{1-\rho^2}{(1-\rho)^2} && \text{(F.27)}\\
&= \frac{1}{2}\log\frac{1+\rho}{1-\rho}, && \text{(F.28)}
\end{aligned}
$$

which complete the proof.

Appendix G

Proof of Theorem 10

Given that all the other rules are fixed, the optimal kth local sensor rule that minimizes the Bayesian cost is

$$U_k = \gamma_k(X_k) = \arg\min_{u_k} f_k(U_k, X_k), \tag{G.1}$$

where the BCDF $f_k(U_k, X_k)$ is given in (5.19). As the expected Bayesian cost C is the error probability P_e, we have $c_{11} = c_{00} = 0$ and $c_{01} = c_{10} = 1$. The coefficient $\beta(u_k, w)$ can be expanded as

$$\begin{aligned}
\beta(u_k, w) &= \pi_0 P(U_0 = 1|u_k, w)p(w|H = 0) + \pi_1 P(U_0 = 0|u_k, w)p(w|H = 1), &\text{(G.2)}\\
&= (\pi_0 p(w|H = 0) - \pi_1 p(w|H = 1))P(U_0 = 1|u_k, w) + \pi_1 p(w|H = 1). &\text{(G.3)}
\end{aligned}$$

Then

$$\begin{aligned}
f_k(1, x_k) - f_k(0, x_k) &= \int_Y p_{X_k|W}(x_k|w)(\beta(1, w) - \beta(0, w))dy &\text{(G.4)}\\
&= \int_Y p_{X_k|W}(x_k|w)\left(\pi_0 p(w|H = 0) - \pi_1 p(w|H = 1)\right)\\
&\quad \left(P(U_0 = 1|U_k = 1, w) - P(U_0 = 1|U_k = 0, w)\right) dw &\text{(G.5)}
\end{aligned}$$

$$= -\int_W p_{X_k|W}(x_k|w)\phi(w)dw \tag{G.6}$$

where

$$\begin{aligned}
\phi(w) &\triangleq (P(U_0=1|U_k=1,w) - P(U_0=1|U_k=0,w))\left(\pi_1 p(w|H=1) - \pi_0 p(w|H=0)\right), \\
&= (P(U_0=1|U_k=1,w) - P(U_0=1|U_k=0,w))\,\pi_1 p(w|H=0)\left(\frac{p(w|H=1)}{p(w|H=0)} - \frac{\pi_0}{\pi_1}\right), \\
&= h(w)g(w), \qquad \text{(G.7)}
\end{aligned}$$

and $h(w)$ is defined in (5.25), $g(w)$ is defined by

$$g(w) \triangleq \frac{p(w|H=1)}{p(w|H=0)} - \frac{\pi_0}{\pi_1}. \tag{G.8}$$

Therefore, the optimal kth sensor rule as

$$U_k = \begin{cases} 1 & \text{if } \int_W p_{X_k|W}(x_k|w)h(w)g(w)dw > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{G.9}$$

By the second condition in Theorem 10, we have that $g(w)$ is a nondecreasing function of $T(w)$. Thus, there exists a nondecreasing function $\varphi(t)$ such that $g(w) = \varphi(T(w))$.

Furthermore, the optimal kth sensor rule can also be expressed as

$$\begin{aligned}
&\int_W p_{X_k|W}(x_k|w)h(w)g(w)dw \\
&= \sum_{i=1}^m \int_{W\in A_i} p_{X_k|W}(x_k|w)h(w)g(w)dw \qquad \text{(G.10)} \\
&= \int_T \sum_{i=1}^m \left(p_{X_k|W}(x_k|Z_i(t)) \left|\frac{d}{dt}Z_i(t)\right| h(Z_i(t))g(Z_i(t))\right) dt \qquad \text{(G.11)} \\
&= \int_T \sum_{i=1}^m \left(p_{X_k|W}(x_k|Z_i(t)) \left|\frac{d}{dt}Z_i(t)\right| h(Z_i(t))\right) \varphi(t)dt \qquad \text{(G.12)} \\
&= \int_T \lambda(x_k,t)\varphi(t)dt \qquad \text{(G.13)}
\end{aligned}$$

where

- Equality (G.11) is by transforming the random variable w to $T(w)$. $Z_i(t) \triangleq T_i^{-1}(t)$ and $T_i(w)$ is monotone of w for $w \in A_i$.

- Equality (G.12) is from the fact that $g(Z_i(t)) = \varphi(T(Z_i(t))) = \varphi(t)$.

- Equality (G.13) is by the definition of $\lambda(x_k, t)$ in (5.24).

To establish the sufficiency of a single threshold quantizer defined in (5.26), it is suffices to show that for any $S(x_k) > S(x'_k), x_k, x'_k \in A_i, i = 1, \cdots, m$,

$$\int_T \lambda(x'_k, t)\varphi(t)dt > 0, \tag{G.14}$$

implies

$$\int_T \lambda(x_k, t)\varphi(t)dt > 0. \tag{G.15}$$

Since $\varphi(t)$ is a nondecreasing function of t, there exists a values $\delta \in [-\infty, +\infty]$ satisfying the equation $\varphi(\delta) = 0$ and have the property that

$$\varphi(t) = \begin{cases} \geq 0 & \text{if } t \geq \delta, \\ \leq 0 & \text{if } t < \delta. \end{cases} \tag{G.16}$$

Also, by the first condition 1 in Theorem 10, $h(y)$ is nonnegative. Therefore, $\lambda(x_k, t)$ is nonnegative.

For $i = 1, \cdots, m$, let $x_k, x'_k \in A_i$, $S(x_k) > S(x'_k)$. By the third condition in Theorem 10, the ratio $\frac{\lambda(x_k,t)}{\lambda(x'_k,t)}$ is a nondecreasing function of t. Let $c_i = \frac{\lambda(x_k,\delta)}{\lambda(x'_k,\delta)}$, we have

$$\begin{cases} \frac{\lambda(x_k,t)}{\lambda(x'_k,t)} \geq c_i & \text{if } t \geq \delta, \\ \frac{\lambda(x_k,t)}{\lambda(x'_k,t)} \leq c_i & \text{if } t \leq \delta. \end{cases} \tag{G.17}$$

We obtain that

$$\int_T \lambda(x_k, t)\varphi(t)dt$$

$$= \int_T \lambda(x'_k,t)\frac{\lambda(x_k,t)}{\lambda(x'_k,t)}\varphi(t)dt \tag{G.18}$$

$$= \int_{t\geq\delta} \lambda(x'_k,t)\frac{\lambda(x_k,t)}{\lambda(x'_k,t)}\varphi(t)dt + \int_{t\leq\delta} \lambda(x'_k,t)\frac{\lambda(x_k,t)}{\lambda(x'_k,t)}\varphi(t)dt \tag{G.19}$$

$$\geq \int_{t\geq\delta} c_i\lambda(x'_k,t)\varphi(t)dt + \int_{t\leq\delta} c_i\lambda(x'_k,t)\varphi(t)dt \tag{G.20}$$

$$= c_i\int_T \lambda(x'_k,t)\varphi(t)dt > 0, \tag{G.21}$$

where the first term of inequality (G.20) is because for $t \geq \delta$, $\varphi(t) \geq 0$, $\frac{\lambda(x_k,t)}{\lambda(x'_k,t)} \geq c_i$ and $\lambda(x'_k,t) \geq 0$. The second term of (G.20) is because for $t \leq \delta$, $\varphi(t) \leq 0$, $\frac{\lambda(x_k,t)}{\lambda(x'_k,t)} \leq c_i$ and $\lambda(x'_k,t) \geq 0$.

This establishes that for $x_k \in A_i$, the optimal kth sensor rule is $U_k = \gamma_k(x_k) = 1$ if $S(x_k) \geq \tau_k$. Since the above proof is true for any $i = 1, \cdots, m$, this completes the proof.

Bibliography

[1] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[2] P. Gács and J. Körner, "Common information is much less than mutual information," *Problems Contr. Inform. Theory*, vol. 2, pp. 149–162, 1973.

[3] R. Ahlswede and J. Körner, "On common information and related characteristics of correlated information sources," in *Proc. of the 7th Prague Conference of Information Theory*, 1974.

[4] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 21, pp. 294–300, May 1975.

[5] P. Cuff, H.Permuter, and T. M. Cover, "Coordination capacity," *IEEE Trans. Inf. Theory*, vol. 56, pp. 4181–4206, Sep. 2010.

[6] H. S. Witsenhausen, "Values and bounds for the common information of two discrete random variables," *SIAM J. Appl. Math.*, vol. 31, no. 2, pp. 313–333, 1976.

[7] R. Radner, "Team decision problems," *Annals of Mathematical Statistics*, vol. 33, pp. 857–881, 1962.

[8] H. Chen, B. Chen, and P.K. Varshney, "A new framework for distributed detection with conditionally dependent observations," *IEEE Trans. Signal Processing*, vol. 60, no. 3, pp. 1409–1419, Mar. 2012.

[9] R. Viswanathan, "A note on distributed estimation and sufficiency," *IEEE Trans.Inf. Theory*, vol. 39, no. 5, pp. 1765–1767, Sep. 1993.

[10] R. Ahlswede and J. Körner, "On common information and related characteristics of correlated information sources," in *Proc. of the 7th Prague Conference of Information Theory*, 1974.

[11] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, New York, 2nd edition, 2006.

[12] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic, New York, 1981.

[13] G. Hu, "On the amount of information," *Teor. Veroyatnost. i Primenen.*, vol. 4, pp. 447–455, 1962.

[14] R. W. Yeung, *Information Theory and Network Coding*, Springer, 2008.

[15] R. Ahlswede and I. Csiszár, "Common randomness in information thoery and cryptography, Part I: Secret sharing," *IEEE Trans. Inf. Theory*, vol. 39, pp. 1121–1132, July 1993.

[16] R. Ahlswede and I. Csiszár, "Common randomness in information thoery and cryptography, Part II: CR capacity," *IEEE Trans. Inf. Theory*, vol. 44, pp. 225–240, Jan. 1998.

[17] U. M. Maurer, "Secret key agreement by public discussion from common information," *IEEE Trans. Inf. Theory*, vol. 39, pp. 733–742, May 1993.

[18] S. Kamath and V. Anantharam, "A new dual to the Gács-Körner common information defined via the Gray-Wyner system," in *Proc. Annual Allerton Conference on Communications, Control and Computing*, Monticello, IL, Sep. 2010.

[19] H. S. Witsenhausen, "On sequences of pairs of dependent random varibles," *SIAM J. Appl. Math.*, vol. 28, pp. 100–113, Jan. 1975.

[20] H. Tyagi, P. Narayan, and P. Gupta, "When is a Function Securely Computable?," *IEEE Trans. Inf. Theory*, vol. 57, pp. 6337–6350, Oct. 2011.

[21] R. M. Gray and A. D. Wyner, "Source coding for a simple network," *Bell Syst. Tech. J.*, vol. 58, pp. 1681–1721, Nov. 1974.

[22] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A*, vol. 222, pp. 309–368, 1922.

[23] S. Kay, *Fundamentals of Statistical Signal Processing II: Detection Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

[24] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*, Wiley, New York, 1965.

[25] G. Casella and R.L. Berger, *Statistical Inference*, Duxbury, Belmont, CA, 1990.

[26] E.L. Lehmann and G. Casella, *Thoery of Point Estimation*, springer, New York, 2nd edition, 1998.

[27] R. M. Gray, "Conditional Rate-Distortion Theory," Tech. Rep. 6502-2, Stanford Electronic labs., Oct. 1972.

[28] R. M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 19, pp. 480–489, Jul. 1973.

[29] B. M. Leiner, "An alternate proof of the composite lower bounds," *Information and Control*, vol. 33, pp. 72–86, 1977.

[30] K. Visvanatha, E. Akyol, and K. Rose, "Lossy common information of two dependent random variables," in *Proc. IEEE Int. Symp. Inform. Thoery*, Combridge, MA, Jul. 2012.

[31] R. Venkataramani, G. Kramer, and V.K. Goyal, "Multiple description coding with many channels," *IEEE Trans. Inform. Thoery*, , no. 9, pp. 2106–2114, Sept. 2003.

[32] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inf. Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.

[33] J. Nayak, E. Tuncel, D. Gunduz, and E. Erkip, "Successive refinement of vector sources under individual distortion criteria," *IEEE Trans. Inf. Theory*, vol. 56, pp. 1769–1781, Apr. 2010.

[34] T. Berger, *Rate Distortion Theory:A mathematical basis for data compression*, Prentice-Hall, Englewood Cliffs, N.J., 1971.

[35] J.-J. Xiao and Z.-Q. Luo, "Compression of correlated Gaussian sources under individual distortion criteria," in *Proc. 43th Annual Allerton Conference on Communication, Control, and Computing*, Sep. 2005, pp. 438–447.

[36] R. Tandon, L.Sankar, and H. V. Poor, "Multi-user privacy: The Gray-Wyner system and generalized common information," in *Proc. IEEE Symp. Inform. Thoery*, St. Petersburg,Russia, Aug. 2011.

[37] E. Hewitt and L. J. Savage, "Symmetric measures on Cartesian products," *Transations of the American Mathematical Society*, vol. 80, no. 1, pp. 470–501, 1955.

[38] A. Rényi, "On some basic problems of statistics from the point of view of information theory," *Proc. 5th Berkely Symp. on Mathematical Statistics and Probability*, vol. 1, pp. 531–543, 1967.

[39] D. Guo, S. Shamai (Shitz), and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.

[40] P.K. Varshney, *Distributed Detection and Data Fusion*, Springer, New York, 1997.

[41] R. Viswanathan and P.K. Varshney, "Distributed detection with multiple sensors: Part I — Fundamentals," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 54–63, Jan. 1997.

[42] J. N. Tsitsiklis, "Decentralized detection," in *Advances in Statistical Signal Processing*, H. V. Poor and Eds. JAI Press J. B. Thomas, Eds., Greenwich, CT, 1993.

[43] R.S. Blum, S.A. Kassam, and H.V. Poor, "Distributed detection with multiple sensor: Part II- Advanced topics," *Proceedings of IEEE*, , no. 1, pp. 64–79, Jan. 1997.

[44] Z. Chair and P.K. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Trans. Aerospace Elect. Sys.*, vol. 22, pp. 98–101, Jan. 1986.

[45] E. Drakopoulos and C.C. Lee, "Optimum multisensor fusion of correlated local decisions," *IEEE Trans. Aerospace Elect. Sys.*, vol. 27, no. 4, pp. 593–605, July 1991.

[46] Q. Zhu M. Kam and W.S. Gray, "Optimal data fusion of correlated local decisions in multiple sensor detection systems," *IEEE Trans. Aerospace Elect. Sys.*, vol. 28, pp. 916–920, July 1992.

[47] I.Y. Hoballah and P.K. Varshney, "Distributed Bayesian signal detection," *IEEE Trans. Inf. Theory*, vol. 35, pp. 995–1000, Sept. 1989.

[48] D.J. Warren and P.K. Willett, "Optimum Quantization for Detector Fusion: Some Proofs, Examples, and Pathology," *Journal of the Franklin Institute*, vol. 336, pp. 323–359, Mar. 1999.

[49] J.N. Tsitsiklis and M. Athans, "On the complexity of decentralized decision making and detection problems," *IEEE Trans. on Automatic Control*, vol. 30, pp. 440–446, May 1985.

[50] P. N. Chen and A. Papamarcou, "Likelihood ratio partitions for distributed signal detection in correlated Gaussian noise," in *Proc. IEEE Int. Symp. Inform. Thoery*, Oct. 1995, p. 118.

[51] P.K. Willett, P.F. Swaszek, and R.S. Blum, "The good, bad, and ugly: distributed detection of a known signal in dependent Gaussian noise," *IEEE Trans. Signal Processing*, vol. 48, pp. 3266–3279, Dec. 2000.

[52] E.L. Lehmann, *Testing Statistical Hypotheses*, Wiley, New York, second edition, 1986.

[53] E. B. Hall, A. E. Wessel, and G. L. Wise, "Some aspects of fusion in estimation theory," *IEEE Trans. Inf. Theory*, vol. 37, pp. 420–422, 1991.

[54] P. Ishwar, R. Puri, K. Ramchandran, and S. S. Pradhan, "On rate-constrained distributed estimation in unreliable sensor networks," *IEEE Journal on Seleted Areas in Communications*, pp. 765–775, April 2005.

[55] K. Eswaran and M. Gastpar, "Rate loss in the CEO problem," in *Proc. of the 39th Conference on Information Sciences and Systems*, Baltimore, MD, Mar. 2005.

[56] J. Pearl, *Causality: Models, Reasoning, and Inference, 1st ed*, Cambridge Univ. Press, Cambridge, U.K., 2000.

[57] F. Peng, H. Chen, and B. Chen, "On energy detection for cooperative spectrum sensing," in *Proc. of the 46th Conference on Information Sciences and Systems*, Princeton, NJ, Mar. 2012.

[58] R. F. Ahlswede and J. Korner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inf. Theory*, vol. 21, no.6, pp. 629–637, Nov. 1975.

[59] W.M. Lam and A. Reibman, "Design of quantizers for decentralized estimation systems," *IEEE Trans. Communications*, vol. 41, pp. 1602–1605, Nov. 1993.

[60] J.N. Tsitsiklis, "Extremal properties of likelihood-ratio quantizers," *IEEE Trans. Commun.*, vol. 41, pp. 550–558, 1993.

[61] B. Chen and P.K. Willett, "On the optimality of likelihood ratio test for local sensor decisions in the presence of non-ideal channels," *IEEE Trans. Inf. Theory*, vol. 51, pp. 693–699, Feb. 2005.

[62] H. Chen, B. Chen, and P.K. Varshney, "Further results on the optimality of likelihood ratio quantizer for distributed detection in non-ideal channels," *IEEE Trans.Inf. Theory*, vol. 55, pp. 828–832, Feb. 2009.

[63] P.K. Willett, P.F. Swaszek, and R.S. Blum, "The good, bad, and ugly: distributed detection of a known signal in dependent Gaussian noise," *IEEE Trans.Signal Processing*, vol. 48, pp. 3266–3279, Dec. 2000.

[64] J. A. Gubner, "Distributed estimation and quantization," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1456–1459, Jul. 1993.

[65] H. S. Witsenhausen, "Indirect rate-distortion problems," *IEEE Trans.Inf. Theory*, vol. IT-26, pp. 518–521, Sept. 1980.

[66] P.K. Varshney, *Distributed Detection and Data Fusion*, Springer, New York, 1997.

VITA

NAME OF AUTHOR: Ge Xu

MAJOR: Electrical and Computer Engineering

EDUCATION:

Ph.D.	Aug. 2013	Syracuse University, NY, USA (expected)
M.S.	Dec. 2012	Syracuse University, NY, USA
M.S.	Mar. 2008	Xidian University,Xi'an, China
B.S.	July 2005	Xidian University, Xi'an, China

PUBLICATIONS:

Journal

1. G. Xu, W. Liu, and B. Chen, "Generalization of common information", submitted to *IEEE Trans. Inform. Theorey*, 2013.

2. G. Xu, S. Zhu, and B. Chen, "Decentralized Data Reduction with Quantization Constraints", submitted to *IEEE Trans. Signal Processing*, 2013.

Conference

1. S. Zhu, G. Xu, B. Chen, "Are global sufficient statistics always sufficient: the impact of quantization on decentralized data reduction", *Proc. Asilomar Conference on Signals, Systems, and Computers*, 2013.

2. G. Xu, H. Chen, B. Chen, "New results on distributed detection with dependent

observations", *Proc. IEEE Global Telecommunications Conference (Globecom)*, Anaheim, USA, Dec 2012.

3. G. Xu, B. Chen, "The sufficiency principle for decentralized data reduction", *Proc. IEEE International Symposium on Information Theory (ISIT)*, Boston, MA, July 2012.

4. G. Xu, B. Chen, "On the sufficiency principle for networked inference problems", in *Information Theory and Application Workshop*, San Diego, CA, Feb. 2012.

5. G. Xu, B. Chen, "Information for inference", *Proc. Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, Sep. 2011.

6. G. Xu, W. Liu and B. Chen,, "Wyner's common information for continuous random variables - A lossy source coding interpretation", *Proc. Annual Conference on Information Sciences and Systems(CISS)*, Baltimore, MD, Mar. 2011.

7. W. Liu, G. Xu, B. Chen, "The Common information of N dependent random variables", *Proc. Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, Sep. 2010.

8. G. Xu, M. Sheng, C. Wu, "Advanced routing in Interplanetary Backbone Network", *the second International Conference on space Information Technology (ICSIT)*, Wuhan, China, Nov. 2007.

9. M. Sheng, G. Xu, X. Fang, "The routing of Interplanetary Internet", *China Communications*, vol. 3, no.6 Dec. 2006.

10. G. Xu, Z. Xu and M. Sheng, "The study of ad hoc networks routing Protocol based on ACO", *Academic annual Conference of Xidian University*, 2005 (in Chinese).

AWARDS AND HONORS

- Syracuse University Graduate Fellowship (2008-2010)
- Xidian University Top Scholarship (2004-2007)
- Xidian University Outstanding (Graduate) Students (2003-2007)
- National scholarship of China (2003)
- 3rd Award in Xidian "Spark Cup" competition (2003)