Lehrstuhl für Nachrichtentechnik

Reachback Communication in Wireless Sensor Networks

João Barros

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor–Ingenieurs

genehmigten Dissertation.

Vorsitzender:	Univ.–Prof. Dr.–Ing. G. Färber
Prüfer der Dissertation:	
	1. Univ.–Prof. Dr.–Ing. J. Hagenauer
	2. Univ.–Prof. Dr.–Ing. Dr. rer. nat. H. Boche, Technische Universität Berlin

Die Dissertation wurde am 21.09.2004 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 04.11.2004 angenommen.

Abstract

This dissertation is concerned with a communications scenario in which a large number of sensor nodes, deployed on a field, take local measurements of some physical process and cooperate to send back the collected data to a far receiver. The first part considers the problem of communicating multiple correlated sources over independent channels and with partial cooperation among encoders, and provides a set of coding theorems that give a complete characterization of the conditions on the sources and the channels under which reliable communication is possible. A key insight from these results is that for a large class of networks separate source and channel coding provides an optimal system architecture. The second part presents new contributions for the long-standing multiterminal source coding problem. Finally, the third part assumes that the sensor nodes use very simple encoders and focuses on the design of practical decoding algorithms under given complexity constraints.

Zusammenfassung

Diese Arbeit betrachtet einige Aspekte der Kommunikation zwischen verteilten Sensorknoten, die gemessene Werte versenden, und einem entfernten Empfänger, der die Daten weiterverarbeitet. Zuerst werden die theoretischen Grenzen für die kooperative Übertragung von korrelierten Quellen über statistisch unabhängige Kanäle bestimmt. Dies erfolgt durch eine Reihe von Codierungstheoremen, die hinreichende und notwendige Bedingungen für nahezu fehlerfreie Kommunikation angeben. Eine wichtige Erkenntnis dieser Arbeit ist, dass für viele nicht triviale Netzwerke getrennte Quellen- und Kanalcodierung eine optimale Systemarchitektur anbietet. Als zweiter Schritt wird eine Verzerrung der Daten mitberücksichtigt, was zu einem lange offenen Problem der Rate-Distortion Theorie führt. Auch hierfür werden neue Ergebnisse präsentiert. Der letzte Teil der Arbeit beschäftigt sich mit dem Entwurf von praktischen Decodierungsalgorithmen unter Einschränkung der algorithmischen Komplexität.

Acknowledgements

Throughout the years of research that lead to this dissertation, I was very fortunate to be able to count on the advice, the encouragement and the friendship of a great number of people.

I would like to begin by thanking Prof. Joachim Hagenauer for his trust in granting me an assistentship at the Lehrstuhl für Nachrichtentechnik (LNT) and the freedom to follow any research path that I set for myself. His professional conduct as a dedicated faculty member and well-respected engineer made me appreciate the importance of, among other aptitudes, thinking in international terms, cultivating a broad scope of interests and being able to explain complex matters with simple examples. It was also thanks to Prof. Hagenauer's trust and support that I could present my research at international conferences, learn how to manage an international graduate program, and supervise students in their diploma and master theses. Not many professors encourage their doctoral students to spend part of their PhD in another lab — I am deeply grateful for being given this opportunity.

The core of the present thesis would not have been possible without the thoughtful guidance, the constant encouragement and the contagious scientific enthusiasm of Prof. Sergio Servetto. In the course of two research stays at Cornell University – kindly supported by Prof. Hagenauer and co-sponsored by the Fulbright Commission in June-September 2002, and again February-March 2003 – and through countless online and offline discussions in the past three years, Sergio has taught me how to find good scientific problems, how to gather the expertise needed to accomplish vital research tasks, and, most importantly, how to lose my natural fear of the complex mathematics that are necessary to solve challenging theoretical problems. By doing so, Sergio has deeply influenced the way I think about scientific research and helped shape and solidify the knowledge base upon which I can build my career as a scientist. Words cannot express my gratitude towards Sergio – his true and honest friendship remains one of the most important gifts from this endeavor.

Prof. Holger Boche has kindly accepted to lend his technical and mathematical expertise as a member of the doctoral committee that must evaluate this thesis. Knowing his many solicitations, I am very grateful for his interest, time and dedication.

The work on multiterminal source coding presented in the fourth chapter was much improved following the insightful comments of Prof. Toby Berger, with whom I spent a few very educative afternoons at Cornell, and Prof. Raymond Yeung, who visited our lab early this year.

Prof. Pedro Guedes de Oliveira at the University of Porto was the first person to encourage me to pursue graduate studies in Germany, and has constantly supported my efforts with helpful advice and recommendation letters, as have Prof. Kristian Kroschel (Karlsruhe), Prof. Dan

Costello (Notre-Dame) and Prof. John B. Anderson (Lund).

I have also profited very much from personal relationships with some of my colleagues at LNT. Michael Tüchler deserves my fondest gratitude for his close friendship and enthusiastic collaboration. It was a great pleasure for me to work with him and our two diploma students, Christoph Hausl and Seong Per Lee, on factor graphs and scalable decoding, yielding the results presented in the fifth chapter of this thesis. Special thanks are also due to my dedicated friend Zaher Dawy for many discussions and for proof-reading the first complete draft of this thesis. I also deeply enjoyed the close collaborations with Andrew Schaefer, Ioannis Oikonomidis and Nicolas Dütsch. Our *sysadmins* Johannes Zangl, Günther Liebl, Stephan Bäro and Markus Kaindl guaranteed an outstanding computer network support at all times. Thanks are also due to Günther Soeder, Klaus Eichin, Dieter Heidner and Doris Dorn for friendly administrative support. Finally, I am much indebted to Silke Jauck, Nicole Rossmann and Pavol Hanus, whose care and dedication ensured the success of our international graduate program and allowed me to find time for my research during critical phases of this work.

Beyond the technical skills that are necessary to accomplish any worthy goal, our efforts are never complete without reaching towards that inner source of inspiration and personal strength from which all creative work flows. In the past four years, I have had the great privilige of learning how to seek this path under the guidance of a true master and a wonderful friend: Ulli Dissmann, the endlessly inspired and inspiring director of the Münchner Sommertheater. I am deeply grateful to Ulli for her enlightened and invaluable teachings on — among many others themes — free play, self awareness, powerful speaking, thoughtful leadership, human nature, and life itself. I hope that my future path is true to these teachings, and that one day I might give my students at least part of what Ulli has taught me.

Since my son, Daniel, was born, I have come to realize what a overwhelming mission it is to raise a child and to take responsibility for a family environment that is nurturing in every possible way. My mother and my father suceeded in finding the necessary balance between setting important rules and encouraging my sister Luisa, my brother Ze and I to search for our own path. One of their best decisions was to send us to the German school in Porto, which broadened our horizons right from early childhood. Words cannot express my gratitude to my parents for all that they have given me.

This work is dedicated, with love and affection, to Ana, who helps me spread my wings anew and always holds me when I fall.

München, August 2004
João Barros

To my wife,

Ana

Contents

List of Figures

1

Introduction

> *First, good Peter Quince, say what the play treats on, then read the names of the actors, and so grow to a point.*
>
> WILLIAM SHAKESPEARE, *A Midsummer Night's Dream*, Act I Scene 2

Why Sensor Networks?

Whether by telephone, on television or over the internet, a substantial part of our daily exchange of information occurs in a virtual world we call cyberspace. In the past decades, by introducing data into these networks and acting upon the collected data, we, the human users of these networks, have been the sole bridges between the physical world we live in and this virtual world we use to communicate. With the advent of tiny, low-cost devices capable of sensing the physical world and communicating over a wireless network, it becomes more and more evident that the status quo is about to change – sensor networks will soon close the gap between cyberspace and the real world.

There are two important reasons why this vision is gaining momentum: (1) its inherent potential to improve our lives — with sensor networks we can expand our environmental monitoring, increase the safety of our buildings, improve the precision of military operations, provide better health care, and give well-targeted rescue aid, among many other applications —(2) the endless possibilities it offers for multi-disciplinary research combining typical areas of electrical and computer engineering (sensor technology, integrated systems, signal processing, wireless communications), with classical computer science subjects (routing, data processing, database management), and all potential application fields (medicine, biology, environmental sciences, agriculture, etc.).

In many ways, sensor networks are significantly different from classical wireless networks like cellular communications systems and wireless LANs:

(a) the design of a sensor network is strongly driven by its particular application,

(b) sensor nodes are highly constrained in terms of power consumption, computational complexity and production cost,

(c) since the network is dense and the nodes share a common objective — to gather and convey information — cooperation can be used to enhace the network's efficiency.

These key features lead to very challenging research problems that are best illustrated with a practical example.

Example 1.1 (Sensor Webs for Precision Farming) *Precision agriculture is about bringing the right amount of water, fertilizers and pesticides to the right plantation site at the right time [42]. Breaking with traditional methods that spread excessive quantities of chemicals uniformly over a field, this new paradigm guarantees a more efficient use of the farm's resources, a strong reduction of undesirable substances on the soil and in the ground water, and ultimately better crops at lower costs. Since fundamental parameters such as soil moisture and concentration of nutrients depend on the measuring spot and can vary fast in time, precision farming requires constant sampling of the site characteristics on a fine spatial grid and a short-time scale — an ideal application for wireless sensor networks. With the aid of a sensor web, the control center can obtain constant updates of the soil conditions site by site and adapt the flows of water, fertilizer and pesticide on the go according to the actual needs of the growing plants, as illustrated in Figure 1.1.*

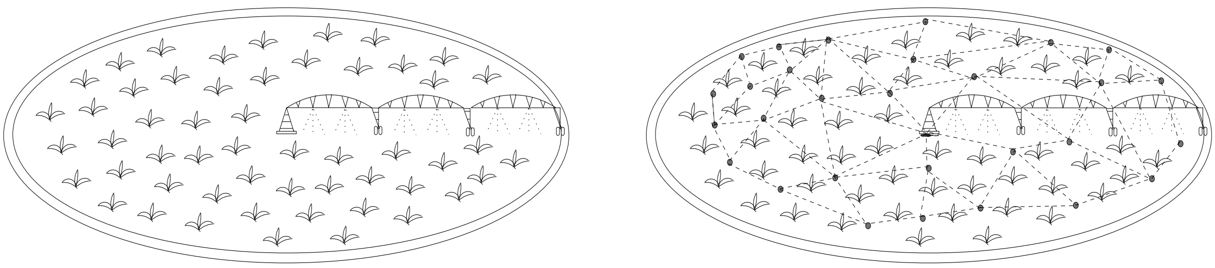


Figure 1.1: A sketch of a center-pivot irrigation system. In the classical application (left drawing) the irrigation arm moves in circles and spreads water, fertilizer and pesticide uniformly over the crop area. To implement precision farming and increase the agricultural efficiency of the plantation, we can use a self-organizing sensor web that monitors the crop (right drawing) and sends the relevant data (e.g. soil humidity and nutrient status) to a fusion center over a wireless network. Based on the acquired information the fusion center can determine the right flow of water and chemicals for each sector of the crop.

While the decision-making process falls within the scope of the agricultural engineer, the design of the supporting sensor web opens a myriad of challenging research problems for the electrical and computer engineer.

The main interest of our present work lies in the communications aspects of wireless sensor networks. For this purpose, we view the sensor network as a collection of transmitters that observe multiple sources of information, encode the picked up data (possibly with an intermediate cooperative step) and *reach back* to a remote fusion center using a wireless channel to transmit the required information — we refer to this setup as *reachback communication*. Based on appropriate models for the sources and the channels, the communications problem becomes finding a suitable system architecture and optimizing the transmission/reception scheme subject to the technological constraints of the sensor nodes.

Prior State of the Art

Although sensor systems and technologies have been the focus of intense research efforts for several decades already, wireless sensor networks have only recently begun to catch the attention of the communications and signal processing communities. At the starting point of the research work that lead to this thesis, little was known about the fundamental limits of communication in wireless sensor networks, and practical implementations were still at an infant stage. The following paragraphs give a brief overview of the knowledge base from which we set out to investigate the sensor reachback problem.

From the theoretical point of view, sensor networks offer the essential practical motivation for a number of challenging problems involving two fundamental ingredients: (1) *correlated* sources of information — an arguably reasonable model for the physical measurements taken by a large number of sensors in a confined area — and (2) collaborative data transmission over noisy channels. The information-theoretic foundations for this class of basic research problems — firmly based on Shannon's *A Mathematical Theory of Communication* [82] — were first laid by Slepian and Wolf [85] in a seminal paper on separate compression of correlated sources, and secondly by Cover, El Gamal and Salehi [26], who provided a partial solution to the problem of communicating correlated sources over a multiple acccess channel. A third important contribution stems from Berger and Tung's work on yet another open issue, namely the rate-distortion extension of the Slepian-Wolf data compression problem ([18], [90]). Thus far, both the capacity region of the multiple access channel with correlated sources and the rate-distortion region of separate encoding of correlated sources remain unknown. The relevant aspect of collaborative transmission was addressed by Willems in a short correspondance [93], which contains the capacity region of the multiple access channel in the case of two independent sources and partial cooperation between encoders.

The aforementioned contributions, like many other results in Shannon theory, are not constructive in the sense that they do not provide a clear description of a *feasible* system design. Recently, the goal of devising practical coding schemes that approach the theoretical limits of Slepian and Wolf has been achieved by several groups (e.g. [38], [1], [69], [87] and [63]), capitalizing on the discovery of highly effective channel coding techniques based on concatenated codes and iterative decoding ([21], [44]). Although said coding schemes perform very well for two or three correlated sources, the complexity of the associated encoding and decoding algorithms renders them unsuitable for reachback networks with hundreds of sensor nodes.

Thesis Outline

This thesis attacks the sensor reachback communications problem in three different ways. In the first part, we study the ultimate performance limits for this class of communications systems using the mathematical tools of network information theory. By modelling the sensor network as a set of multiple correlated sources that are observed by partially cooperating encoders and transmitted over an array of orthogonal channels, we are able to characterize the reachback capacity, i.e., the exact conditions on the sources and the channels under which reliable communication with the far receiver is possible.

The second part of the present work is dedicated to the rate-distortion version of the sensor reachback problem. At this point, we present some progress made towards the solution of the long-standing multiterminal source coding problem, and extend its formulation to the case of partial cooperation between encoders, providing a partial characterization of the corresponding rate-distortion region.

Finally, in the third part, we consider the more practical problem of jointly decoding the correlated data transmitted by hundreds of sensor nodes. After showing that the optimal decoder based on minimum mean square estimation (MMSE) is unfeasible — its complexity grows exponentially with the number of nodes — we present a two-step ”scalable” alternative: (1) approximate the correlation structure of the data with a suitable factor-graph, and (2) perform belief propagation decoding on this graph to produce the desired estimates. Based on this general approach, which can be applied to sensor networks with arbitrary topologies, we provide an exact characterization of the decoding complexity, as well as optimization algorithms for finding optimal factor trees under the Kullback-Leibler criterion.

The rest of the thesis is organized as follows. In **Chapter 2** we review the main information-theoretic results for discrete memoryless sources and channels, and discuss several proof techniques that are required for a thorough understanding of our main contributions. Our goal is not only to give precise mathematical statements, but also to provide the reader with useful intuition, particularly where separate encoding of correlated sources and multiple user communications are concerned.

Chapter 3 is devoted to the sensor reachback problem with perfect reconstruction at the receiver. First, we give a detailed justification of the main modeling assumptions followed by a brief overview of our main results. The latter are then presented step by step, from the simplest instance of two non-cooperating sensor nodes to the most general case of a sensor network with an arbitrary number of cooperating encoders. As a general rule, we include in the main body only those parts of a proof that provide intuition and leave the more technical (and sometimes more tedious) details for the appendix. The insights gained from the solution of the sensor reachback problem, inspire us to conclude this chapter with some reflections on the usefulness of the separation principle in communications networks.

In **Chapter 4**, we assume that perfect reconstruction at the receiver is either not possible or not necessary, and look at separate encoding of correlated sources with distortion constraints. After giving a precise statement of the classical multiterminal source coding problem, we discuss several different bounds for the sought rate-distortion region — for general discrete-memoryless sources and, specifically for the binary case with Hamming distortion. We then extend the problem to account for partial cooperation between encoders, providing a full characterization of the rate region in the lossless case and a new inner and outer bound for the rate-distortion version of the problem.

Chapter 5 looks at the sensor reachback problem from a more practical perspective and addresses the issue of decoding complexity. We begin by proposing a very simple system model for the reachback network and showing that optimal decoding is unfeasible. Then, we propose a scalable solution which tackles the complexity by approximating the correlation structure of the sensor data with a factor tree. The key contribution here is a set of tools for constrained factorization of Gaussian distributions, which greatly simplify the design of the decoder. The last section of this chapter includes some numerical examples that underline the effectiveness of the proposed approach.

Chapter 6 concludes this dissertation with a survey of possible directions for future work. The interested reader will also find a comprehensive appendix with the detailed proofs of the results presented in the main chapters.

Parts of this work have been presented at [11], [14], [13], [9], [12], [15], [16] and [10].

2

Fundamentals of Information Theory

A journey of a thousand miles begins with a single step.
CHINESE PROVERB

We begin this dissertation with a brief overview of some of the fundamental concepts and mathematical tools of information theory. This will allow us to establish some notation and set the stage for the main results presented in Chapters 3 and 4. For a comprehensive introduction to the fundamental concepts and methods of information theory we refer the interested reader to the excellent treatises of Gallager [36], Cover and Thomas [28], and Yeung [97]. In [30] Csiszár and Körner offer a panoply of very useful mathematical tools for discrete memoryless sources and channels.

The rest of the chapter is organized as follows. Section 2.1 begins with Shannon's communications model and gives a precise formulation of the point-to-point problem. Section 2.2 then proceeds with a detailed overview of the proof techniques that are relevant for the present work, followed by a discussion of some of Shannon's fundamental theorems in Section 2.3. This chapter concludes with Section 2.4, which is entirely devoted to some of the most relevant results in network information theory, with special emphasis on separate encoding of correlated sources and multiple access communications.

2.1 The Point-to-Point Communications Problem

2.1.1 Communications Model

The foundations of information theory were laid by Claude E. Shannon in his brilliant 1948 paper entitled "A Mathematical Theory of Communication" [82]. In his own words: *the fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point*. If the message — for example a letter from

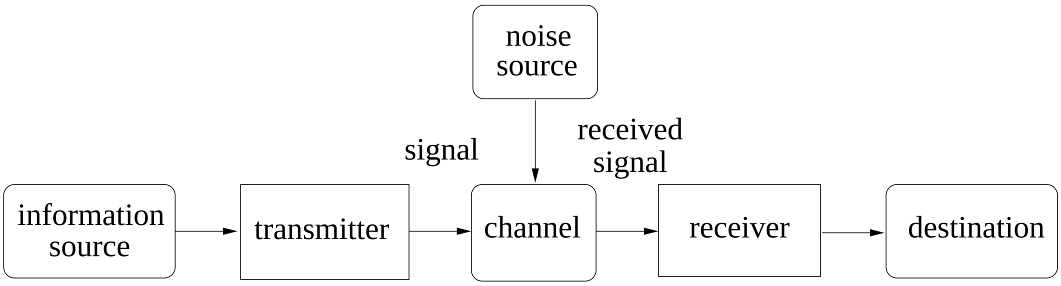


Figure 2.1: Shannon's communications model (from [82]).

the alphabet, the gray level of a pixel or some physical quantity measured by a sensor — is to be reproduced at a remote location with a certain fidelity, some amount of information must be transmitted over a physical channel. This observation is the crux of Shannon's general model for *point-to-point* communication reproduced in *Figure 2.1*. It consists of the following parts:

- The *information source* generates messages at a given rate according to some random process.
- The *transmitter* observes this messages and forms a signal to be sent over the channel.
- The *channel* is governed by a *noise source* which corrupts the original input signal. This models the physical constraints of a communications system, e.g. thermal noise in electronic circuits or multipath fading in a wireless medium.
- The *receiver* takes the received signal, forms a reconstructed version of the original message, and delivers the result to the *destination*.

Given the statistical properties of the information source and the noisy channel, the goal of the communications engineer is to design the transmitter and the receiver in a way that allows the sent information to reach its destination in a *reliable* way. Information theory can help us achieve this goal by characterizing the fundamental mechanisms behind communications systems and providing us with precise mathematical conditions under which reliable communication is possible.

2.1.2 Problem Statement

To give a precise formulation of the point-to-point communications problem, we require rigorous definitions[1] for each of its constituent parts. We assume that the *source* and the *channel* are described by discrete-time random processes, and we determine that the receiver and the transmitter agree on a common *code*, specified by an *encoder* and *decoder* pair. The basic relationship between these entities is illustrated in *Figure 2.2* and described rigorously in the following lines.

[1]We point out that although in this thesis we are mostly concerned with discrete memoryless sources and channels, many of the results presented here and in the following chapters can be extended to account for continuous-valued alphabets, as well as sources and channels with memory.

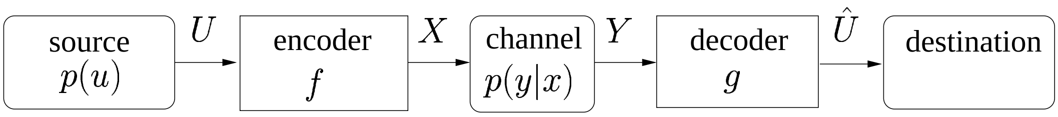


Figure 2.2: Mathematical model of a communications system.

Definition 2.1 (Source) A *discrete memoryless source* denoted U generates a sequence of independent and identically distributed (i.i.d.) *messages*, also referred to as *letters* or *symbols*, from the alphabet $\mathcal{U}$. The messages correspond to independent drawings from the probability distribution[2] $p_U(u)$.

Definition 2.2 (Channel) A *discrete memoryless channel* $(\mathcal{X},\ p(y|x), \mathcal{Y})$ is described by an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$ and a conditional probability distribution $p(y|x)$, such that X and Y denote the *channel input* and the *channel output*, respectively.

Definition 2.3 (Code) A *code* consists of

1. an encoding function $f : \mathcal{U} \rightarrow \mathcal{X}^N$, which maps a message u to a *codeword* x^N with N symbols,
2. a decoding function $g : \mathcal{Y}^N \rightarrow \hat{\mathcal{U}}$, which maps a block of N channel outputs y^N to a message $\hat{u}$ from the reconstruction alphabet $\hat{\mathcal{U}}$. For simplicity, we assume that $\hat{\mathcal{U}} = \mathcal{U}$, i.e. source and reconstruction alphabets are identical.

The *rate* of the code is given by $R = (1/N)\log_2 |\mathcal{U}|$ in *bits per channel use*, where $|\mathcal{U}|$ denotes the size of the alphabet $\mathcal{U}$.

To give a precise statement of the problem, we require one more definition:

Definition 2.4 (Reliable Communication) Given the rate R, reliable communication of the source $U \sim p(u)$ over the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ is possible if there exists a code $x^N(u)$ with rate R and with decoding function $g(y^N)$ such that, as $N \rightarrow \infty$,

$$P_N = p\{g(Y^N) \neq U\} \rightarrow 0,$$

i.e. the source messages are reconstructed with *arbitrarily small probability of error*. If reliable communication is possible at rate R then R is an *achievable rate*.

The main goal of the problem is to give precise conditions for reliable communication based on single-letter information-theoretic quantities that depend only on the given probability distributions and not on the block lengths N. The mathematical tools required for this characterization are the topic of the next section.

Remark 1 Notice that the classical information-theoretic formulation of the point-to-point communications problem does not put any constraints neither on the computational complexity nor

[2]In the sequel we follow the convention that subscripts of a probability distribution are dropped if the subscript is the capitalized version of the argument, i.e. , we simply write $p(u)$ for the probability distribution $p_U(u)$.

on the delay of the encoding and decoding procedures. In other words, the goal is to describe the fundamental limits of communications systems irrespective of their technological limitations.

2.2 Mathematical Tools of Information Theory

The key to the solution of the point-to-point communications problem stated in the previous section — and of many other problems in information theory — is to take into consideration the statistical properties of very long sequences of symbols and divide these into useful classes. The following paragraphs describe a powerful set of tools for this task, which will be of great use in the proofs of Chapters 3 and 4.

2.2.1 Types and Typical Sequences

Let x^N be a sequence of symbols drawn i.i.d. from the alphabet $\mathcal{X} = \{a_1, a_2, \ldots, a_{|\mathcal{X}|}\}$ according to the probability distribution $p(x)$ with $x \in \mathcal{X}$ and $|\mathcal{X}| < \infty$. The sequence x^N can be classified according to its *type* by applying the following definition.

Definition 2.5 (Type) The *type* P_{x^N} of the sequence x^N can be obtained by counting the number of occurrences $N(a|x^N)$ of each symbol $a \in \mathcal{X}$ divided by the total sequence length N, i.e. P_{x^N} is the empirical probability distribution of X obtained from the observed sequence x^N.

It is often useful to group those sequences x^N, whose type or empirical distribution is *close* to the probability distribution $p(x)$. We call the resulting set the *strongly typical set*.

Definition 2.6 (Strongly Typical Set) If the type P_{x^N} of the sequence x^N does not differ from the true probabilities $p(x)$ by more than $\delta/|\mathcal{X}|$ then x^n belongs to the *strongly typical set*. Specifically,

$$\mathcal{T}_\delta^N(X) = \left\{ x^N \in \mathcal{X}^N : \left| \frac{1}{N} N(a|x^N) - p(a) \right| < \frac{\delta}{|\mathcal{X}|} \right\}$$

for every $a \in \mathcal{X}$. We refer to the sequences x^n in this set as *strongly typical sequences*.

Assume now that we have a pair of sequences x^N and y^N drawn i.i.d. from the alphabets $\mathcal{X} = \{a_1, a_2, \ldots, a_{|\mathcal{X}|}\}$ and $\mathcal{Y} = \{c_1, c_2, \ldots, c_{|\mathcal{Y}|}\}$ according to the joint probability distribution $p(xy)$. The next definition extends the concept of *strong typicality* to describe the relationship between two sequences x^N and y^N.

Definition 2.7 (Strong Joint Typicality) If the joint type $P_{x^N y^N} = N(ac|x^N y^N)/N, ac \in \mathcal{X} \times \mathcal{Y}$, of the sequence pair $(x^N y^N)$ does not differ from the true probabilities $p(xy)$ by more than $\delta/(|\mathcal{X}||\mathcal{Y}|)$, then we say x^N and y^N are *strongly jointly typical*. In this case, the strongly typical set is given by

$$\mathcal{T}_\delta^N(XY) = \left\{ x^N \in \mathcal{X}^N, y^N \in \mathcal{Y}^N : \left| \frac{1}{N} N(ac|x^N y^N) - p(ac) \right| < \frac{\delta}{|\mathcal{X}||\mathcal{Y}|} \right\}.$$

What makes the notion of strong typicality so useful, is the so called *asymptotic equipartition property* (AEP). The latter follows naturally from the law of large numbers [97, Chapter 5] and is described by the following theorem.

Theorem 2.1 (Strong AEP [97, p. 74]) Let η be a small positive quantity, such that $\eta \to 0$ as $\delta \to 0$.

1. For N sufficiently large,
$$p\{X^N \in \mathcal{T}_\delta^N(X)\} > 1 - \delta.$$

2. If $x^N \in \mathcal{T}_\delta^N(X)$, then
$$2^{-N(H(X)+\eta)} \leq p(x^N) \leq 2^{-N(H(X)-\eta)}.$$

3. For N sufficiently large,
$$(1-\delta)2^{N(H(X)-\eta)} \leq |\mathcal{T}_\delta^N(X)| \leq 2^{N(H(X)+\eta)}.$$

Here, $H(X)$ denotes the *Shannon entropy* of the random variable X given by

$$H(X) = -\sum_x p(x) \log p(x),$$

where the logarithm is taken to base two[3] and the summation is carried out over the support of $p(x)$.

Proof: See [97, pp. 74-77]. ■

In simple terms, this theorem states that for sufficiently large sequence length N, the probability that a sequence X^N drawn i.i.d. $\sim p(x)$ belongs to the typical set is very close to one. Moreover, for practical purposes we may assume that the probability of any strongly typical sequence is about $2^{-NH(X)}$, and the number of strongly typical sequences is approximately $2^{NH(X)}$. The generalization of the AEP for two random variables X and Y can be obtained in a straightforward manner using the joint entropy

$$H(XY) = -\sum_x \sum_y p(xy) \log p(xy),$$

where once again the summation is carried out over the support of $p(xy)$. For large N, it follows that there exist around $2^{NH(XY)}$ jointly typical sequences x^N and y^N.

Clearly, the notion of *entropy* is far too important to be mentioned in passing. Conceptually, $H(X)$ can be viewed as a measure of the average amount of information contained in X or, equivalently, the amount of uncertainty that subsists until the outcome of X is revealed. Other useful information measures include the conditional entropy of X given Y defined as

$$H(X|Y) = H(XY) - H(Y),$$

[3]Unless otherwise specified, all logarithms in this thesis are taken to base two.

describing the amount of uncertainty that remains about X when Y is revealed, and the mutual information

$$I(X;Y) = H(X) - H(X|Y),$$

which can be interpreted as the reduction in uncertainty about X when Y is given. The relationship between the aforementioned information-theoretic quantities is well explained in [97, Section 2.2]. As a consequence of the AEP, we may assume for N very large that (a) there are about $2^{NH(X|Y)}$ sequences x^N which are jointly typical with a given sequence y^N, and (b) two arbitrarily chosen sequences x^N and y^N are jointly typical with probability $\approx 2^{NH(X|Y)}/2^{NH(X)} = 2^{-NI(X;Y)}$.

By requiring that the relative frequency of each possible symbol be close to the corresponding probability, strong typicality is particularly suitable for information-theoretic problems which involve minimizing a distortion measure between source sequences and reconstruction sequences, as we will see in Chapter 4. In the case of asymptotically *perfect* reconstruction at the receiver (i.e. arbitrarily small probability of error) the only parameter of interest is the coding rate and many problems can be solved using a *weaker* notion of typicality, which only requires that the empirical entropy of a sequence be close to the true entropy of the corresponding random variable.

Definition 2.8 (Weakly Typical Set) The weakly typical set $\mathcal{A}_\epsilon^N(X)$ is the set of sequences $x^N \in \mathcal{X}^N$ such that

$$\left| -\frac{1}{N}\log p(x^N) - H(X) \right| \leq \epsilon.$$

The *weak* version of the AEP then follows naturally from the weak law of large numbers [28, Chapter 3].

Theorem 2.2 (Weak AEP [97, p. 61]) For any $\epsilon > 0$ we have

1. For N sufficiently large,
$$p\{X^N \in \mathcal{A}_\epsilon^N(X)\} > 1 - \epsilon.$$
2. If $x^N \in \mathcal{A}_\epsilon^N(X)$, then
$$2^{-N(H(X)+\epsilon)} \leq p(x^N) \leq 2^{-N(H(X)-\epsilon)}.$$
3. For N sufficiently large,
$$(1-\epsilon)2^{N(H(X)-\epsilon)} \leq |\mathcal{A}_\epsilon^N(X)| \leq 2^{N(H(X)+\epsilon)}.$$

Proof: See [97, pp. 62-63]. ■

Since, for N sufficiently large, any sequence X^N drawn i.i.d. $\sim p(x)$ is very likely to be weakly typical, any property that we prove for weakly typical sequences is true with high probability. In the following, we will generally refer to weakly typical sequences simply as *typical sequences*.

2.2.2 Random Coding and Random Binning

Consider once again the formal statement of the point-to-point communications problem in Section 2.1.2. In Shannon's mathematical model a block of messages is mapped to a sequence of channel input symbols, also called *codeword*. The set of codewords builds the core of the code used by the transmitter and the receiver to communicate reliably over the channel.

Since information theory is primarily concerned with the fundamental limits of reliable communication, it is often useful to prove the existence of codes with certain properties without having to search for explicit code constructions. A simple way to accomplish this task is to perform a *random selection* of codewords. Random selection is often used in mathematics to prove the existence of mathematical objects without actually constructing them. For example, if we want to prove that a real-valued function $h(n)$ takes a value less than c for some n in a given set $\mathcal{S}$, then it suffices to introduce a uniform probability distribution on $\mathcal{S}$ and show that the mean value of $h(n)$ is less than c. When this technique is applied to prove the existence of codes with certain properties, we speak of *random coding*. Based on this simple idea, we can construct a random code for the system model shown in *Figure 2.2* by drawing codewords X^N at random according to the probability distribution $\prod_{i=1}^{N} p(x_i)$. Then, if we want to prove that there exists a code such that the error probability goes to zero for N sufficiently large, it suffices to show that the average of the probability of error taken over all possible random codebooks goes to zero for N sufficiently large — in that case there exists at least one code whose probability of error is below the average.

A different coding technique, which is particularly useful in information-theoretic problems with multiple correlated sources, consists of *throwing* sequences $u^N \in \mathcal{U}^N$ into a finite set of bins, such that the sequences that land in the same bin share a common bin index. If each sequence is assigned a bin at random according to a uniform distribution, then we refer to this procedure as *random binning*. By partitioning the set of sequences into equiprobable bins, we can rest assure that, as long as the number of bins is much larger than the number of typical sequences, the probability that there is more than one typical sequence in the same bin is very, very small [28, pp. 410-411]. This in turn means that each typical sequence is uniquely determined by its corresponding bin index. If *side information* is available and we can distinguish between different typical sequences in the same bin — e.g. when we are given a sequence w^N that is jointly typical with u^N — then we can decrease the number of bins, or equivalently the number of bin indices, and increase the efficiency of our coding scheme (see e.g. [23] and [95]). The following examples illustrate the main idea behind the aforementioned binning mechanism.

Example 2.1 (From [101]) *Assume that Alice must communicate an even number x to Bob and that Bob receives as side information one of the neighbours of x. If Alice has access to Bob's side information then she only needs to send one bit, e.g. indicating whether x is above or below the known neighbour. What if Alice does not know the side information available to Bob? A possible solution for this problem is to place all possible even numbers into two bins: one bin for all multiples of four and one bin for all the others. By sending the bin index (0 or 1) to Bob, Alice is able to achieve the exact same efficiency (1 bit) even without access to the side information.*

Example 2.2 *Let U_1 and U_2 be two correlated sources, which output binary triplets differing at most in one bit, e.g. 000 and 010 or 101 and 001. Assuming that U_1 must be communicated by the transmitter and that U_2 is available as side information at the decoder, how should we encode the former, such that perfect reconstruction is possible with a minimum amount of transmitted bits? Notice that there are only two bits of uncertainty, i.e. enough to indicate in which bit U_1 and U_2 differ, provided of course that U_2 is known at the encoder. Interestingly enough, the same coding efficiency can still be achieved even if U_2 is not known at the encoder. Here is the key idea: it is not necessary for the code to differentiate between U_1 triplets that differ in three positions, because the decoder will count the number of bits that are different from U_2 and only one of the two possible U_1 triplets will be one bit away. Thus, by putting the eight possible realizations of U_1 in four bins and guaranteeing that the elements in one bin differ in three bit positions, we can rest assure that U_1 will be perfectly reconstructed at the decoder.*

2.2.3 The Markov Lemma

Another useful tool for coding problems with side information is provided by the following lemma, which plays a major role in the class of multiterminal source coding problems considered in Chapter 4.

Lemma 2.3 (Markov Lemma, [90, 18]) Let U, X and Y be three random variables that form a Markov chain $U \rightarrow X \rightarrow Y$. If for a given $(u^N, x^N) \in T_\delta^N(UX)$, Y^N is drawn $\sim \prod_{i=1}^N p(y_i|x_i)$, then $p\{(U^N, X^N, Y^N) \in T_\delta^N(UXY)\} \rightarrow 1$ as $N \rightarrow \infty$.

Proof: See [90] and [18]. ■

In more intuitive terms, the Markov property of U, X and Y implies that if we take two sequences u^N and x^N which are strongly jointly typical and generate a third sequence y^N according to the conditional probability $p(y|x)$ — for example, by transmitting the sequence x^N over a memoryless channel— then with high probability u^N, x^N and y^N are strongly jointly typical. This, in turn, implies that with high probability u^N and y^N are strongly jointly typical.

2.2.4 Useful Inequalities

We conclude this section on mathematical tools of information theory with three inequalities which will also prove very useful in Chapters 3 and 4. The corresponding proofs can be found in [28] and [97].

Lemma 2.4 (Conditioning does not increase entropy) Let X and Y be two random variables $\sim p(xy)$. Then,

$$H(X|Y) \leq H(X).$$

In other words this lemma asserts that the knowledge of Y cannot increase our uncertainty[4] about X. Assume now that U is a random variable of interest to us and X is an observation of U, such that our average uncertainty about U given X is $H(U|X)$. If we process X and obtain

[4] In their celebrated textbook Cover and Thomas often refer to this statement as *conditioning reduces entropy* [28]. The interpretation *conditioning does not increase entropy* seems to us somewhat more appropriate.

the random variable Y (e.g. by transmitting X over a memoryless channel), then $U \to X \to Y$ forms a Markov chain and the following statement holds true.

Lemma 2.5 (Data Processing Inequality) Let U, X and Y be three random variables that form a Markov chain $U \to X \to Y$. Then,

$$I(U;X) \leq I(U;Y).$$

It is not difficult to see that

$$\begin{aligned} H(U|Y) &= H(U) - I(U;Y) \\ &\geq H(U) - I(U;X) \\ &\geq H(U|X), \end{aligned}$$

which means that, on average, by processing (or transmitting) X we can only increase our uncertainty about U.

Finally, we assume that X is a random variable and $\hat{X}$ is an estimate of X taking values in the same alphabet $\mathcal{X}$. The following lemma gives a precise description of the relationship between the conditional entropy $H(X|\hat{X})$ and the probability of error $P_e = p\{X \neq \hat{X}\}$.

Lemma 2.6 (Fano's Inequality) Let X and $\hat{X}$ be two random variables with the same alphabet $\mathcal{X}$. Then

$$H(X|\hat{X}) \leq H_b(P_e) + P_e \log(|\mathcal{X}| - 1),$$

where $H_b(P_e)$ is the binary entropy function computed according to

$$H_b(P_e) = -P_e \log P_e - (1 - P_e)\log(1 - P_e).$$

Fano's inequality is the key ingredient of all the converse proofs in this thesis.

2.3 Shannon's Coding Theorems

The typical results in information theory are concerned with the existence of codes with certain asymptotic properties. A theorem that confirms the existence of codes for a class of achievable rates is often referred to as a *direct result* and the arguments that lead to this result constitute the *achievability proof*. On the other hand, when a theorem asserts that codes with certain properties do not exist, we speak of a *converse result* and a *converse proof*. A fundamental result that includes both the achievability and the converse parts is called a *coding theorem* [30].

Having discussed some of the basic proof techniques in information theory, we will turn to four of Shannon's fundamental coding theorems. These results form the basis of classical information theory and are of great use in several of our own proofs.

2.3.1 The Channel Coding Theorem

The *channel coding theorem* gives a complete solution (achievability and converse) for the point-to-point communications problem stated in Section 2.1.2. According to the problem statement, we use a code of rate R to transmit the messages produced by source U over a discrete memoryless channel $(\mathcal{X}, p(y|x), \mathcal{Y})$. If reliable communication is possible at rate R, i.e. the average error probability P_N goes to zero as the block length N goes to infinity, then we say the rate R is achievable. As it turns out, one simple condition is sufficient to fully characterize the set of achievable rates:

Theorem 2.7 (Channel Coding Theorem [97, Section 8.2]) A rate R is achievable for a discrete memoryless channel $p(y|x)$ if and only if $R \leq C$, where C is the channel capacity given by

$$C = \max_{p(x)} I(X;Y).$$

This remarkable result guarantees the existence of a code with arbitrarily small probability of error for all rates below the capacity of the channel. The latter equals the maximum mutual information between channel input X and channel output Y, where the maximization is carried out over all possible input probability distributions $p(x)$.
There are several ways to prove the channel coding theorem and details can be found e.g. [28, Chapter 8] and [97, Chapter 8]. In the following we will give a sketch or proof that highlights those aspects that are relevant for the information-theoretic problems considered in this thesis. The presented proof structure is common to most our proofs.
Sketch of Proof:
We begin with the achievability part, which can be summarized in the following steps:

1. *Codebook Construction:* Construct a random codebook by drawing $|\mathcal{U}|$ codewords $X^N \sim \prod_{i=1}^N p(x_i)$, as explained in Section 2.2.2, such that each message in $\mathcal{U}$ is mapped to a codeword indexed by $X^N(1), X^N(2), \ldots, X^N(|\mathcal{U}|)$. The codebook is known both to the encoder and the decoder.

2. *Encoder:* For message u generated by the source transmit the apropriate codeword $X^N(u)$ through the channel.

3. *Decoder*[5]*:* Upon observing the channel output sequence Y^N generated according to the probability distribution $\prod_{i=1}^N p(y_i|x_i(u))$, the decoder outputs $\hat{u} = u$ if $(X^N(u), Y^N) \in \mathcal{A}_\epsilon^N(XY)$, i.e. $X^N(u)$ and Y^N are jointly typical, and there does not exist another $\tilde{u} \neq u$ such that $(X^N(\tilde{u}), Y^N) \in \mathcal{A}_\epsilon^N(XY)$. Otherwise, the decoder sets $\hat{u}$ equal to some predefined constant.

4. *Error Events:* An error occurs if $X^N(u)$ and Y^N are not jointly typical, or there exists an $\tilde{u} \neq u$ such that $X^N(\tilde{u})$ and Y^N are jointly typical.

[5]The described decoding procedure is called *typical set decoding*.

5. *Analysis of the Probability of Error:* We average the probability of error over all possible random codes generated according to $\prod_{i=1}^{N} p(x_i)$. Since the code construction is symmetric, the average probability of error is the same for all messages $U \in \mathcal{U}$ and we can assume without loss of generality that the sent message is $u = 1$. Based on the AEP we can assume that the probability that $X^N(1)$ and Y^N are not jointly typical goes to zero for N sufficiently large. Moreover since the probability that any $X^N(u)$ is jointly typical with Y^N is approximately $2^{-NI(X;Y)}$, if we use about $2^{NI(X;Y)}$ codewords the probability that there exists an $\tilde{u} \neq 1$ such that $X^N(\tilde{u})$ and Y^N are jointly typical is also negligible.

6. *Capacity:* Choosing the probability distribution $p(x)$ that maximizes $I(X;Y)$, we prove that the channel capacity C can be achieved.

The converse part follows directly from Fano's Inequality and standard information-theoretic identities and inequalities [97, Section 8.3]. ■

2.3.2 The Source Coding Theorem

When the channel is noiseless, i.e. $Y = X$, it may still be useful to encode the messages produced by the source. In this case, the purpose of the code is not to compensate for the impairments caused by the channel, but to achieve a more efficient representation of the source information in terms of bits per message or equivalently bits per source symbol — this proccedure is called *source coding* or *data compression*. The main idea is to consider only a subset $\mathcal{B}$ of all possible source sequences $\mathcal{U}^N$, and assign a different index $i \in \{1, 2, \ldots, |\mathcal{B}|\}$ to each of the sequences u^N in $\mathcal{B}$. If the sequence produces a source sequence $u^N \in \mathcal{B}$, then the encoder outputs the corresponding index i, otherwise i is set to some predefined constant. The decoder receives the index i and outputs the corresponding sequence in $\mathcal{B}$. The rate of the resulting *source code* can be computed according to $R = (1/N) \log |\mathcal{B}|$. The following result gives the minimum rate R at which we can encode the data and still guarantee that the messages can be perfectly reconstructed.

Theorem 2.8 (Source Coding Theorem [97, Section 4.2]) Let U be an information source drawn i.i.d. $\sim p(u)$. For N sufficiently large, there exists a code with arbitrarily small probability of error, whose coding rate R is arbitrarily close to the entropy $H(U)$. Conversely, if $R < H(U)$ the error probability goes to one, as N goes to infinity.

Proof: See e.g. [97, Section 4.2]. ■

The main idea behind this theorem can be stated in very simple terms: since for large N the AEP guarantees that any sequence produced by the source U belongs with high probability to the typical set $\mathcal{A}_\epsilon^N(U)$, we only need to index the approximately $2^{NH(U)}$ typical sequences to achieve arbitrarily small probability of error. Thus, setting $\mathcal{B} = \mathcal{A}_\epsilon^N(U)$, we get $R \approx H(U)$.

Alternatively, the theorem can be proved using a simple random binning argument: if we randomly assign each source sequence to one of a finite number of bins, then as long as the number of bins is larger than $2^{NH(U)}$ we know that the probability of finding more than one typical sequence in the same bin is very small [28, pp. 410-411]. Since each typical sequence

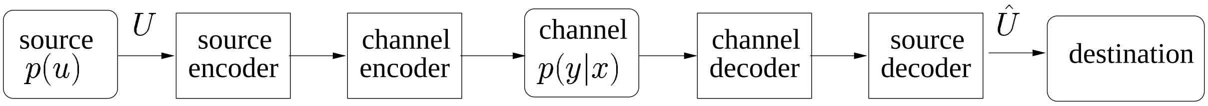


Figure 2.3: A communications system based on the separation principle.

is mapped to a different bin index, arbitrarily small probability of error can be easily achieved by letting the decoder output the typical sequence that corresponds to the received index.

2.3.3 The Separation Theorem

Going back to the point-to-point communications problem, consider the following coding strategy illustrated in *Figure 2.3*:

1. *Source Encoder:* Upon observing a block of source messages u^N, map u^N to an index $i \in \{1, 2, \ldots 2^{NR_s}\}$ thus compressing the source U at rate R_s.

2. *Channel Encoder:* Map the index i to one of $2^{N_c R_c}$ distinct channel codewords X^{N_c}, thus encoding i at rate $R_c > R_s$.

3. *Channel Decoder:* Upon observing the channel output sequence Y^{N_c}, use typical set decoding to produce the codeword estimate $\hat{X}^{N_c}$.

4. *Source Decoder:* Based on $\hat{X}^{N_c}$ look up the corresponding index î and output the sequence estimate $\hat{u}^N(\hat{\imath})$.

This set of procedures describes a *modular* system architecture, in which source coding and channel coding are carried out separately. The conditions for reliable communication based on this approach, can be easily obtained from the previous coding theorems as follows. First, the source coding theorem states that U^N can be reconstructed from the index i with arbitrarily small probability of error if $R_s > H(U)$, and secondly, the channel coding theorem guarantees that i can be transmitted reliably as long as $R_c < C$. Thus, we conclude that reliable communication is possible if

$$H(U) < R_s < R_c < C.$$

Shannon's source-channel coding theorem, often referred to as the *separation theorem*, states that the condition $H(U) < C$, *entropy of the source less than channel capacity*, is not only sufficient, but also *necessary* for reliable communication to be possible.

Theorem 2.9 (Separation Theorem, [28, Section 8.13]) If U is a finite alphabet stochastic process that satisfies the AEP, then for sufficiently large N there exists a code with $P_N \to 0$ if and only if $H(U) < C$.

Proof: The achievability part follows directly from our previous reasoning. The converse proof is based on Fano's Inequality and the Data Processing Lemma. For details see [28, Section 8.13]. ■

Intuitively, this theorem implies that there is nothing to lose by splitting the source coding and channel coding tasks — first, we compress the data to its most efficient representation, and then add redundancy in a controlled fashion to combat the errors caused by the channel. The obvious advantage of this modular architecture is that we can design the source code ignoring the channel, and then design the channel code ignoring the source. If we want to transmit different sources over the same channel we only need to substitute the source code, and similarly to transmit the same source over different channels we just need to adapt the channel code accordingly.

It is worth noting that although separate source and channel coding is *asymptotically* optimal, for finite block lengths a joint source-channel code can lead to better performance [29].

2.3.4 The Rate-Distortion Theorem

Sometimes perfect reconstruction of the source information is not required and we can accept an erroneous representation that satisfies a certain fidelity criterion. Typical scenarios include a television broadcast or a telephone conversation, where we only need to comply with the constrained demands of human perception. The characterization of the trade-off between encoding rate and allowable distortion is the goal of *rate-distortion theory* [17].

For the information source U drawn i.i.d. $\sim p(u)$ and its reconstruction $\hat{U}$, we introduce a single-letter distortion measure

$$d : \mathcal{U} \times \hat{\mathcal{U}} \to \mathbb{R}^+,$$

which maps a $(U, \hat{U})$ pair to a nonnegative real number corresponding to the distortion incurred by describing the message U with the reconstruction $\hat{U}$. In Chapters 4 and 5, we will use the *Hamming distortion* for discrete alphabets according to

$$d(u, \hat{u}) = \begin{cases} 0 & \text{if } u = \hat{u}, \\ 1 & \text{otherwise,} \end{cases}$$

and also the *square-error distortion* for continuous-valued alphabets given by

$$d(u, \hat{u}) = (u - \hat{u})^2.$$

In general, our goal is to minimize the *average* distortion between a source sequence u^N and a reconstruction sequence $\hat{u}^N$ defined as[6]

$$d(u^N, \hat{u}^N) = \frac{1}{N} \sum_{i=1}^{N} d(u_i, \hat{u}_i).$$

We say a rate-distortion pair (R, D) is *achievable* if there exists a code with rate R that satisfies $p\{d(u^N, \hat{u}^N) > D + \epsilon\} \leq \epsilon$, i.e. the probability that the average distortion exceeds the prescribed distortion D is arbitrarily small. The following result gives a complete characterization of the set of achievable (R, D) pairs, also called the *rate-distortion region*.

Theorem 2.10 (Rate-Distortion Theorem, [97, pp. 197-198]) Let U be an information source drawn i.i.d. $\sim p(u)$ and let $d(u, \hat{u})$ be a bounded distortion measure. The rate-distortion pair

[6] For convenience, we use once again the letter *d*, but this abuse of notation will not cause any confusion.

(R, D) is achievable if and only if

$$R \geq \min_{p(\hat{u}|u):\mathrm{E}[d(u,\hat{u})]\leq D} I(U;\hat{U}).$$

The minimum achievable rate R for every distortion $D \geq 0$ is given by the so called *rate-distortion function*, denoted $R(D)$.

Proof: See [97, Chapter 9]. ■

In order to obtain the rate-distortion function, we must minimize the mutual information between the source U and the reconstruction $\hat{U}$ over all possible joint probability distributions $p(u, \hat{u}) = p(u)p(\hat{u}|u)$ whose expected distortion value $\mathrm{E}\left[d(u, \hat{u})\right]$ computed according to

$$\mathrm{E}\left[d(u, \hat{u})\right] = \sum_{(u,\hat{u})} p(u)p(\hat{u}|u)d(u, \hat{u})$$

does not exceed the prescribed distortion D.

If U is to be transmitted over a channel of capacity C then the definition of the rate-distortion function and the separation theorem imply that the rate-distortion pair (R, D) is achievable if and only if $R(D) < C$. Once again, the coding task can be split into two parts: first we find a rate-distortion code that satisfies the distortion constraint and then we add a channel code that operates close to capacity.

2.4 Network Information Theory

The previous results help us characterize the fundamental limits of communication between two users, i.e. one sender and one receiver. However, in many communications scenarios — for example, satellite broadcasting, cellular telephony, the internet or wireless sensor networks — the information is sent by one or more transmitting nodes to one or more receiving nodes over more or less intricate communication networks. The interactions between the users of said networks introduce whole new range of fundamental communications aspects that are not present in the classical point-to-point problem, such as *interference*, *user cooperation* and *feedback*. The central goal of *network information theory* is to provide a thorough understanding of these basic mechanisms, by characterizing the fundamental limits of communications systems with multiple users. In this section, we review some of its most relevant contributions, in particular those which are required for a detailed understanding of our own results, discussed in Chapters 3 and 4.

2.4.1 The Slepian-Wolf Theorem

Assume that two sources U_1 and U_2 drawn i.i.d. $\sim p(u_1 u_2)$ are to be processed by a joint encoder and transmitted to a common destination over two noiseless channels, as shown in *Figure 2.4*. In general, $p(u_1 u_2) \neq p(u_1)p(u_2)$, such that the messages produced by U_1 and U_2 at any given point in time are statistically dependent — we refer to U_1 and U_2 as *correlated sources*. Since the channels do not introduce any errors, we may ask the following question: at what rates R_1 and R_2 can we transmit information generated by U_1 and U_2 with arbitrarily

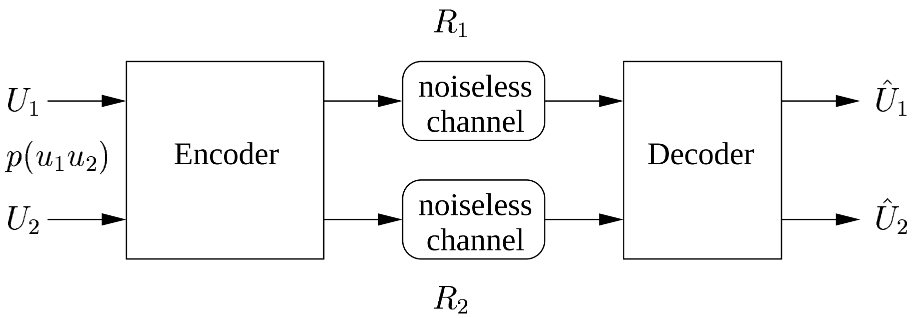


Figure 2.4: Joint encoding of correlated sources.

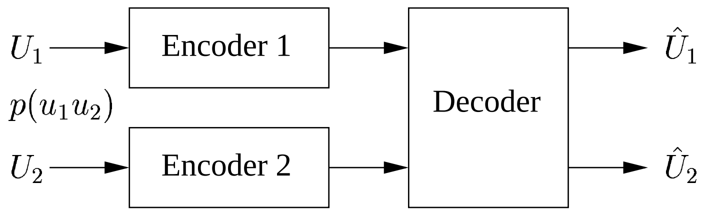


Figure 2.5: Separate encoding of correlated sources (*Slepian-Wolf problem*). The noiseless channels between the encoders and the decoder are omitted.

small probability of error? Not surprisingly, since we have a common encoder and a common decoder, this problem reduces to the classical point-to-point problem and the solution follows naturally from Shannon's source coding theorem: the messages can be perfectly reconstructed at the receiver if and only if

$$R_1 + R_2 > H(U_1U_2),$$

i.e. the sum rate must be greater than the joint entropy of U_1 and U_2.

The problem becomes considerably more challenging if instead of a joint encoder we have two *separate* encoders, as shown in *Figure 2.5*. Here, each encoder observes only the realizations of the one source it is assigned to and does not know the output symbols of the other source. In this case, it is not immediately clear which encoding rates guarantee perfect reconstruction at the receiver. If we encode U_1 at rate $R_1 > H(U_1)$ and U_2 at rate $R_2 > H(U_2)$, then the source coding theorem guarantees once again that arbitrarily small probability of error is possible. But, in this case, the sum rate amounts to $R_1 + R_2 > H(U_1) + H(U_2)$, which in general is greater than the joint entropy $H(U_1U_2)$.

In their landmark paper [85], Slepian and Wolf come to a surprising conclusion: the sum rate required by two separate encoders is the same as that required by a joint encoder, i.e. $R_1+R_2 > H(U_1U_2)$ is sufficient for perfect reconstruction to be possible. In other words, there is no loss in overall compression efficiency due to the fact that the encoders can only observe the realizations of the one source they have been assigned to. However, it is important to point out that the decoder does require a minimum amount of rate from each encoder, specifically the average remaining uncertainty about the messages of one source given the messages of the other source, i.e. $H(U_1|U_2)$ and $H(U_2|U_1)$. The set of achievable compression rates is thus fully characterized by the following theorem.

Theorem 2.11 (Slepian-Wolf Theorem, [85]) Let (U_1U_2) be two correlated sources drawn i.i.d. $\sim p(u_1u_2)$. The compression rates (R_1, R_2) are achievable if and only if

$$\begin{aligned} R_1 &\geq H(U_1|U_2) \\ R_2 &\geq H(U_2|U_1) \\ R_1 + R_2 &\geq H(U_1U_2). \end{aligned}$$

Sketch of Proof: The crux of the proof is joint typicality and random binning [28, Section 14.4]. Since the source sequences u_1^N and u_2^N are drawn i.i.d. $\sim p(u_1u_2)$, we may assume for large N that the two sequences are jointly typical. Thus, to construct the codebooks we randomly map each of the possible U_1^N sequences to one of 2^{NR_1} bins, and each of the possible U_2^N sequences to one of 2^{NR_2} bins, such that upon observing the actual source sequences u_1^N and u_2^N, encoder 1 and encoder 2 send the corresponding bin indices to the joint decoder. In order to obtain the original source sequences, the decoder looks inside the two indicated bins for a pair of sequences u_1^N and u_2^N that are jointly typical. The decoder makes an error if

1. u_1^N and u_2^N are not jointly typical,
2. there exists a $\tilde{u}_1^N \neq u_1^N$ inside the first bin, such that $\tilde{u}_1^N$ and u_2^N are jointly typical,
3. there exists a $\tilde{u}_2^N \neq u_2^N$ inside the second bin, such that u_1^N and $\tilde{u}_2^N$ are jointly typical,
4. there exist a $\tilde{u}_1^N \neq u_1^N$ and a $\tilde{u}_2^N \neq u_2^N$ inside the indicated bins, such that $\tilde{u}_1^N$ and $\tilde{u}_2^N$ are jointly typical.

Clearly, the AEP guarantees that the first type of error has asymptotically small probability. For the other three sources of error, it can be shown [28, Section 14.4] that by ensuring a sufficiently large number of bins, specifically

$$2^{NR_1} > 2^{NH(U_1|U_2)}, 2^{NR_2} > 2^{NH(U_2|U_1)} \text{ and } 2^{N(R_1+R_2)} > 2^{NH(U_1U_2)},$$

it is possible to drive the probability of finding other jointly typical pairs inside the bins arbitrarily close to zero.
The converse is based on Fano's inequality and standard information-theoretic manipulations. For details, see e.g. [85] or [28, Section 14.4]. ■

The three inequalities in the statement of the theorem define the so called *Slepian Wolf* region, containing all the achievable rate pairs (R_1, R_2), as shown in *Figure 2.6*. The implications of this fundamental result are best illustrated with a simple example.

Example 2.3 *Assume that two sensor nodes measure the humidity at two different sites of a plantation and then communicate their observations to a central computer. Depending on whether the measured humidity value exceeds a certain threshold, each sensor i selects a message U_i that can be either* HUMID *or* NOT HUMID *with equal probability. Additionally, we know from long-term observation that the joint probability distribution of the messages U_1 and U_2 is given by*

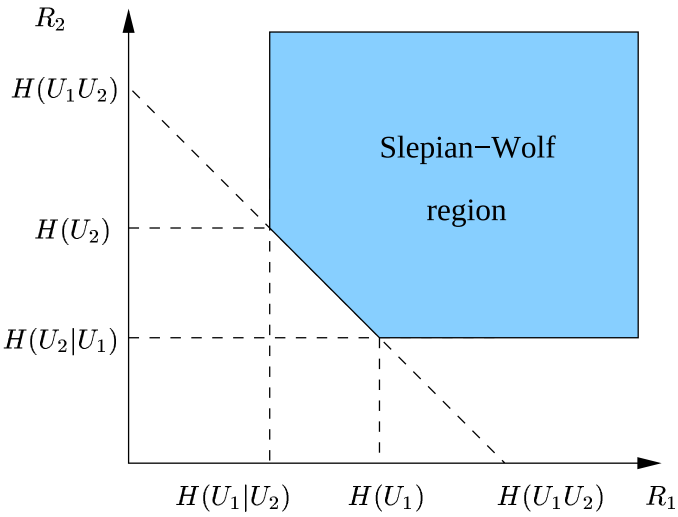


Figure 2.6: The Slepian-Wolf region.

	$p(u_1u_2)$	U_1 HUMID	NOT HUMID
U_2	HUMID	*0.48*	*0.02*
	NOT HUMID	*0.02*	*0.48*

Knowing that the individual entropies are given by $H(U_1) = H(U_2) = H_b(1/2) = 1$ *bits per message, we conclude that if the sensors do not take into consideration the statistical dependencies between their measurements, they must transmit 2000 bits to convey the information of* $N = 1000$ *pairs of measurements. On the other hand, using Slepian-Wolf codes to exploit these dependencies they can reduce the sum rate from* $H(U_1) + H(U_2) = 2$ *to* $H(U_1U_2) \approx 1.242$ *bits per message, thus requiring only 1242 bits on average to convey the exact same information. This means that, if one of the sensor nodes encodes its data close to the corresponding entropy,i.e.* $R_1 > H(U_1) = 1$ *bit per message, the second sensor only needs to transmit* $H(U_1|U_2) \approx 0.242$ *bits per message to guarantee perfect reconstruction at the receiver.*

The Slepian-Wolf theorem can be easily generalized to more than two sources yielding the following result.

Theorem 2.12 (Slepian-Wolf with many sources [28, p. 409]) Let $U_1U_2 \dots U_M$ denote a set of correlated sources drawn i.i.d. $\sim p(u_1u_2 \dots u_M)$. The set of achievable rates is given by

$$R(S) > H(U(S)|U(S^c))$$

for all $S \subseteq \{1, 2, \dots, M\}$, where $R(S) = \sum_{i \in S} R_i$, S^c denotes the complement of S, and $U(S) = \{U_j : j \in S\}$.

Proof: The proof goes along the lines of the case with two sources. Details can be found in [28, Section 14.4]. ∎

2.4.2 The Multiple Access Channel with Independent Sources

In the previous problem, we assumed that the information generated by multiple sources is transmitted over noiseless channels. If this data is to be communicated over a common noisy channel to a single destination, we call this type of channel a *multiple access channel*. The resulting information-theoretic problem, illustrated in *Figure 2.7*, takes into account not only the noise at the receiver, but also the interference caused by different users communicating over a common channel — the mathematical subtlety lies in allowing the channel output Y to depend on the channel inputs X_1 and X_2 according to the conditional probability distribution $p(y|x_1x_2)$. The set of achievable rates at which the different encoders can transmit their data reliably is called the *capacity region* of the multiple access channel.

Assuming independent messages, i.e. $p(u_1u_2) = p(u_1)p(u_2)$, and independent encoders, Ahlswede [3] and Liao [64] were independently able to prove the following result which fully characterizes the set of achievable rates.

Theorem 2.13 (Multiple Access Channel [3, 64]) The capacity region of the discrete multiple access channel is given by the convex hull of the set of points (R_1, R_2) satisfying

$$R_1 \leq I(X_1;Y|X_2) \quad (2.1)$$

$$R_2 \leq I(X_2;Y|X_1) \quad (2.2)$$

$$R_1 + R_2 \leq I(X_1X_2;Y), \quad (2.3)$$

for some joint distribution $p(x_1)p(x_2)$.

Sketch of Proof: The proof goes along the lines of the proof of the channel coding theorem. Each encoder uses an independent random code, and the decoder looks for codewords x_1^N and x_2^N that are jointly typical with the observed channel output sequence y^N. It can be shown that if the encoding rates obey the stated conditions, the average probability of error goes to zero as N goes to infinity (cf. [28, Chapter 14.3]). ■

The boundaries of the capacity region, shown in *Figure 2.8*, can be explained in a very intuitive way. When encoder 1 views the signals sent by encoder 2 as noise, its maximum achievable rate is given by $R_1 \approx I(X_1;Y)$ — a direct consequence of the channel coding theorem. Then, the decoder can estimate the sent codeword x_1^N and subtract it from the channel output sequence y_1^N, thus allowing encoder 2 to achieve a maximum rate of $R_2 \approx I(X_2;Y|X_1)$. This procedure, sometimes referred to as *successive cancellation* [66], leads to the upper corner

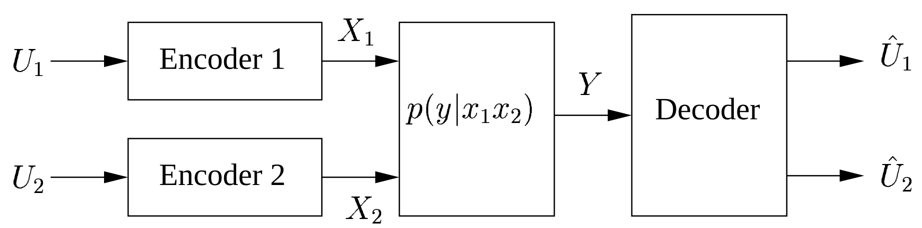


Figure 2.7: The multiple access channel.

point of the capacity region. The lower corner point corresponds to the symmetric case and a time-sharing argument yields the remaining points in the segment between them.

It is also worth noting that conditions (2.1)-(2.3) can be easily generalized for more than two sources. In this case, the capacity region is given by

$$R(S) \leq I(X(S); Y|X(S^c))$$

for all $S \subseteq \{1, 2, \ldots, M\}$, where $R(S) = \sum_{i \in S} R_i$, S^c denotes the complement of S, and $X(S) = \{X_j : j \in S\}$ [28, Chapter 14.3].

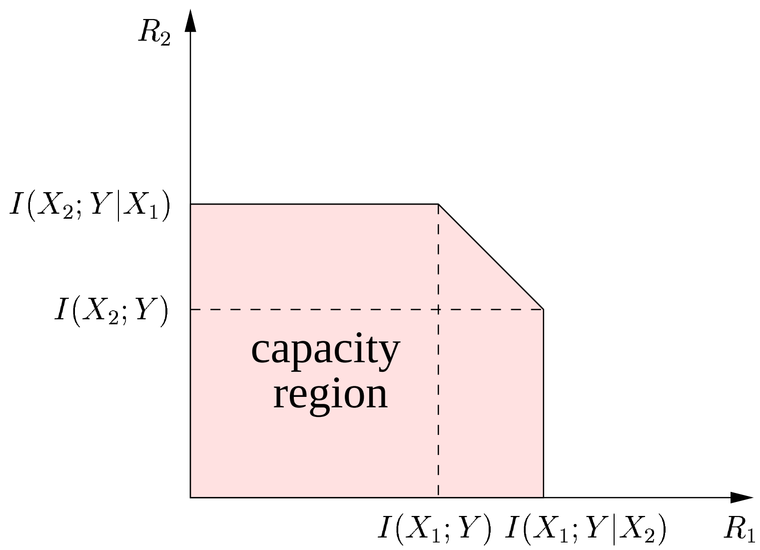


Figure 2.8: The capacity region of the multiple access channel.

The capacity region of the multiple access channel with independent sources can be increased by allowing the encoders to *cooperate* and exchange messages over communication links of limited capacity. This problem was studied by Willems in [93], where it is shown that the set of achievable rates is given by

$$\begin{aligned} R_1 &< I(X_1; Y|X_2 Z) + C_{12} \\ R_2 &< I(X_2; Y|X_1 Z) + C_{21} \\ R_1 + R_2 &< \min\{\, I(X_1 X_2; Y|Z) + C_{12} + C_{21},\ I(X_1 X_2; Y)\, \}, \end{aligned}$$

for some auxiliary random variable Z such that $|\mathcal{Z}| \leq \min(|\mathcal{X}_1| \cdot |\mathcal{X}_2| + 2, |\mathcal{Y}| + 3)$, and for a joint distribution $p(zx_1x_2y) = p(z)p(x_1|z)p(x_2|z)p(y|x_1x_2)$. Here, C_{12} and C_{21} denote the capacities of the links between encoder 1 and 2, and vice versa. As will be explained in detail in Chapter 3, the messages exchanged over these links are described by a suitable random variable Z. Somewhat informally, we can state that the capacity region grows by the amount of information exchanged between the encoders.

2.4.3 Multiple Access with Correlated Sources

When the sources are no longer independent, computing the capacity region becomes a more complicated matter. In a first step, one could consider compressing the data using Slepian-Wolf codes and then adding capacity-attaining channel codes. This coding strategy, which follows naturally from the separation principle, yields the following *sufficient* conditions for reliable communication

$$H(U_1|U_2) < I(X_1;Y|X_2) \tag{2.4}$$
$$H(U_2|U_1) < I(X_2;Y|X_1) \tag{2.5}$$
$$H(U_1U_2) < I(X_1X_2;Y). \tag{2.6}$$

These conditions basically state that the Slepian-Wolf region and the capacity region of the multiple access channel have a non-empty intersection. As unintuitive as it may seem, it is possible to show with a simple example that conditions (2.4)-(2.6) although sufficient are certainly not necessary.

***Example* 2.4 (From [26])** *Let U_1U_2 be two correlated binary sources distributed according to*

	$p(u_1u_2)$	U_1 = *0*	U_1 = *1*
U_2	*0*	*1/3*	*1/3*
	1	*1/3*	*0*

such that the joint entropy of U_1U_2 is given by $H(U_1U_2) = \log 3 \approx 1.58$. Assume now that U_1U_2 are separately encoded and transmitted to a common receiver over a multiple access channel for which $Y = X_1 + X_2$, i.e. the channel output results from the addition of its channel inputs. Given the characteristics of the channel, we can compute the maximum sum rate allowed by the channel yielding $R_1 + R_2 \approx I(X_1X_2;Y) = 1.5$ bits per channel use. Clearly, $H(U_1U_2) > I(X_1X_2;Y)$ and so we immediately conclude that reliable communication is not possible using a coding strategy based on Slepian-Wolf source codes and separate channel codes. However, inspite of the fact that the Slepian-Wolf region and the capacity region do not intersect, reliable communication is possible using the simplest conceivable coding strategy: we send the source symbols directly to the channel, i.e. $X_1 = U_1$ and $X_2 = U_2$ — since the symbol pair $(U_2, U_2) = (1, 1)$ never occurs, we can be sure that all source symbols will be correctly recovered by the decoder.

The reason why conditions (2.4)-(2.6) fail to give a complete characterization of the achievable rates with correlated sources has to do with the fact that the capacity region is computed under the assumption of independent inputs. In the same paper from which the previous example was taken [26], Cover, El Gamal and Salehi introduce a class of *correlated* joint source/channel codes, which enables them to increase the region of achievable rates to

$$H(U_1|U_2) < I(X_1;Y|X_2U_2) \quad (2.7)$$
$$H(U_2|U_1) < I(X_2;Y|X_1U_1) \quad (2.8)$$
$$H(U_1U_2) < I(X_1X_2;Y), \quad (2.9)$$

for some $p(u_1u_2x_1x_2y) = p(u_1u_2)p(x_1|u_1)p(x_2|u_2)p(y|x_1x_2)$. Also in [26], the authors generalize this set of sufficient conditions to sources (U_1U_2) with a common part $W = f(U_1) = g(U_2)$, but they were not able to prove a converse, i.e. , they were not able to show that their region is indeed the capacity region of the multiple access channel with correlated sources. Later, it was shown with a carefully constructed example by Dueck in [34] that indeed the region defined by (2.7)-(2.9) is not tight. To this date, the problem still remains open.

2.4.4 Related Problems

The aforementioned problems are the ones more closely related to the sensor reachback problem and the multiterminal source coding problem discussed in Chapters 3 and 4, respectively. In this section, we will briefly point at other topics in network information theory that are potentially relevant for wireless sensor networks.

The Broadcast Channel

While in the multiple access scenario we have multiple sources and one destination, in the *broadcast case* the information of one source is transmitted to multiple users. Thus, the classical model for the broadcast channel (proposed by Cover in [24]), has one input X and multiple outputs Y_i, $i = 1, 2, \ldots, M$, which are governed by the conditional probability distribution $p(y_1y_2\ldots y_M|x)$. Applications that fall under this system model include the downlink channel of a satellite or of a base station in a mobile communications network. In the context of wireless sensor networks, it is conceivable that a remote control center broadcasts messages to the sensor nodes on the field in order to coordinate their transmissions or change their configurations.

As in many other fundamental problems of network information theory, determining the capacity of the broadcast channel turns out to be a very difficult task. Consequently, a complete characterization of the achievable rates is only known for a few special cases, e.g. the physically *degraded* broadcast channel in which $p(y_1y_2|x)$ factors to $p(y_1|x)p(y_2|y_1)$ [28, Section 14.6] or, most recently, the multiple-input multiple-output Gaussian broadcast channel [92]. For a survey on other interesting results, we refer the reader to Cover's survey [25].

The Relay Channel

In wireless communications, fading of the signals transmitted due to multipath propagation is one of the major impairments that a communications system has to deal with. A natural way to deal with these impairments is by the use of *diversity*: redundant signals are transmitted over essentially independent channels and can then be combined at the receiver to average out distortion/noise effects induced by the independent channels [77]. If two or more transmitters are allowed to exchange information and coordinate their transmissions, they can exploit the resulting *spatial diversity* to improve the reliability and the efficiency of their communications. An information-theoretic abstraction of this user cooperation problem is the so called *relay channel*. At time i a relay node observes a noisy version $Y_R(i)$ of the symbol $X(i)$ transmitted by the sender and forms a symbol $X_R(i)$, which depends on all previously observed channel outputs $Y_R(1), \ldots, Y_R(i-1)$. The receiver observes the channel output $Y(i)$, whose relationship with $X(i)$, $X_R(i)$ and $Y_R(i)$ is characterized by the conditional $p(yy_R|xx_R)$. Once again, the capacity region is only known in special cases (see e.g. [28, Section 14.7] and [57]). Recently, several contributions appeared, which connect the insights gained from the classical relay problem with practical wireless communications, most notably the papers of Narula et al. [71], Laneman et al. [59], Sendonaris et al. [79, 80], and Dawy et al. [31].

The Two-Way Channel

The previous problems are instances of the general *two-way channel* proposed by Shannon in [83]. In its original formulation, two users both with transmitting and receiving capability, send information to each other over a common channel, as shown in *Figure 2.9*. The channel outputs Y_1 and Y_2, depend on the channel input symbols X_1 and X_2 according to $p(y_1y_2|x_1x_2)$ (see *Figure 2.9*). Since encoder 1 can decide on the next symbol X_1 to send based on the received channel symbol Y_2, the two-way channel introduces a new important aspect in the study of communications networks: *transmission feedback*. Unfortunately, the capacity of the two-way channel is only known in the Gaussian case, which decomposes into two independent channels [28, p.383].

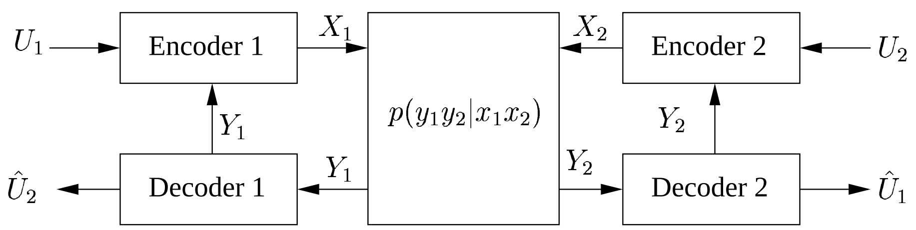


Figure 2.9: The two-way channel.

The previous examples show that network information theory offers a miriad of very challenging problems, some of which have been open for more than two decades. Nevertheless, in the past few years we have witnessed considerable progress in this field, partly motivated by the remarkable advancements of mobile communications systems and, more recently, wireless sensor networks. Although certainly not an easy task, the development of a comprehensive theory of information networks is likely to have a very strong impact on the design of contemporary communications systems.

3

The Sensor Reachback Problem

The beginning of knowledge is the discovery of something we do not understand.

FRANK HERBERT

3.1 Introduction

3.1.1 Reachback Communication in Wireless Sensor Networks

Wireless sensor networks made up of small, cheap, and mostly unreliable devices equipped with limited sensing, processing and transmission capabilities, have recently sparked a fair amount of interest in communications problems involving multiple correlated sources and large-scale wireless networks. As outlined in the introduction, it is envisioned that an important class of applications for such networks involves a dense deployment of a large number of sensors over a fixed area, in which some kind of physical process unfolds—the task of these sensors is then to collect measurements, encode them, and relay them over a noisy channel to some data collection point where this data is to be analyzed, and possibly acted upon. This scenario is illustrated in *Figure 3.1*.

There are several aspects that make this communications problem interesting:

1. *Correlated Observations:* If we have a large number of nodes sensing a physical process within a confined area, it is reasonable to assume that their measurements are correlated. This correlation may be exploited for efficient encoding/decoding.

2. *Cooperation among Nodes:* Before transmitting data to the remote receiver, the sensor nodes may establish a *conference* to exchange information over the wireless medium and increase their efficiency or flexibility through cooperation.

3. *Channel Interference:* If multiple sensor nodes use the wireless medium at the same time (either for conferencing or reachback), their signals will necessarily interfere with each

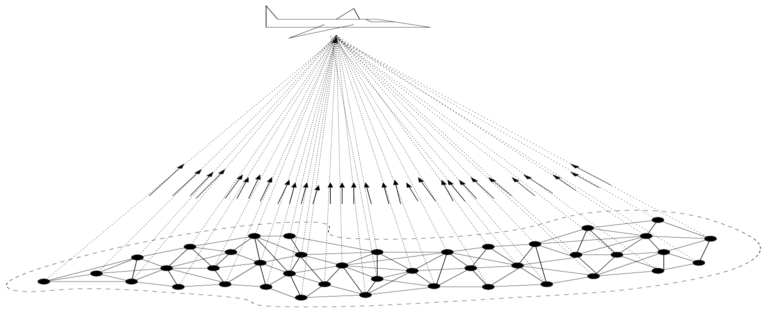


Figure 3.1: A large number of sensors is deployed over a target area. After collecting the data of interest, the sensors must *reach back* and transmit this information to a single receiver (e.g. , an overflying plane) for further processing.

other. Consequently, reliable communication in a reachback network requires a set of rules that control (or exploit) the interference in the wireless medium.

Based on the assumption of correlated measurements, cooperating sensor nodes and a medium access scheme that eliminates the interference, we formulate the sensor reachback problem as follows. Let $U_1U_2\ldots U_M$ be a set of correlated sources drawn i.i.d. from the joint distribution $p(u_1u_2\ldots u_M)$. The information generated by the M sources is separately encoded by M encoders, and transmitted to a remote receiver over an array of M independent channels, equivalent to a multiple access channel with $p(y|x_1x_2\ldots x_M) = \prod_{i=1}^{M} p(y_i|x_i)$. The encoders are interconnected by an underlying communications network, such that encoder i is able to send messages to encoder j reliably at rates $R_{ij} \leq C_{ij}$, with $i = 1,\ldots,M$, $j = 1,\ldots,M$ and $i \neq j$, before transmitting to the remote receiver. The solution to the problem is a complete characterization of the *reachback capacity*, i.e. , the exact set of conditions under which it is possible to reconstruct the values of $U_1U_2\ldots U_M$ at the far receiver with arbitrarily small probability of error. This problem setup is illustrated in *Figure 3.2* for $M = 2$ sources.

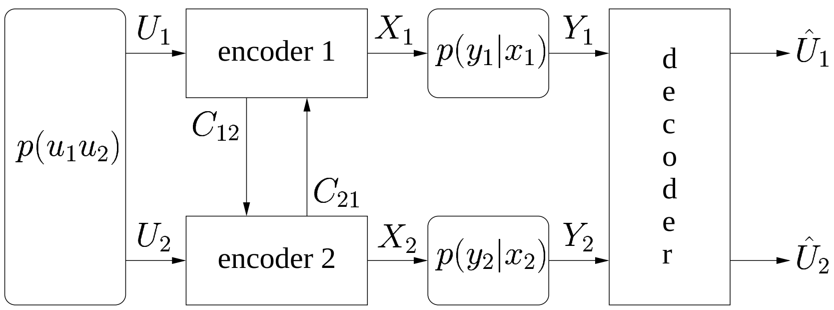


Figure 3.2: A system model for the sensor reachback problem for the case of two sources.

3.1.2 Modeling Assumptions

The previous problem setup forms what we deem to be a reasonable abstraction of the problem of sending the information picked up by a network of cooperating sensor nodes all the way back to a common receiver. In the spirit of George E. P. Box's maxim that *all models are wrong but* *some are useful*, we provide next some motivation for the most relevant aspects of the problem formulation and briefly discuss some of the alternatives.

Source Model

At each time instant the sources generate a random vector $U_1 U_2 \dots U_M$ drawn i.i.d. from discrete alphabets $\mathcal{U}_1 \times \mathcal{U}_2 \times \cdots \times \mathcal{U}_M$ according to $p(u_1 u_2 \dots u_M)$. For simplicity, we assume memoryless sources, and thus consider only the spatial correlation of the observed samples and not their temporal dependence (since the latter dependencies could be dealt with by simple extensions of our results to the case of ergodic sources). Furthermore, each sensor node i observes only one component U_i and must transmit enough information to enable the remote user to reconstruct the whole vector $U_1 U_2 \dots U_M$. This assumption is the most natural one to make for scenarios in which data is required at a remote location for fusion and further processing, but the data capture process is distributed, with sensors able to gather *local* measurements only, and deeply embedded in the environment.

A conceptually different approach would be to assume that all sensor nodes get to observe independently corrupted noisy versions of one and the same source of information U, and it is this source (and not the noisy measurements) that needs to be estimated at a remote location. This approach seems better suited for applications involving non-homogeneous sensors, where each one of the sensors gets to observe different characteristics of the same source (e.g. , multispectral imaging), and therefore leads to a conceptually very different type of sensing applications from those of interest in this thesis. Such an approach leads to the so called *CEO problem* studied by Berger, Zhang and Viswanathan in [20].

An Array of Independent Channels

Our motivation to consider reachback communication over an array of independent channels, instead of a general multiple access channel, is twofold.

From a pure information-theoretic perspective, an array of independent channels is interesting because, as we will see later in this chapter, the assumption of independence among channels gives rise to long Markov chains which play a central role in our ability to solve this problem: it is based on those chains that we are able to prove the converse part of our coding theorems, thus obtaining complete results in terms of capacity for the sensor reachback problem. Furthermore, said coding theorems provide solutions for special cases of the multiple access channel with correlated sources, cases for which no general converse is known.

From a more practical point of view, the assumption of independent channels is valid for any network that controls interference by means of a reservation-based medium-access control protocol (e.g. , TDMA or FDMA). Provided individual nodes have enough resources to reach directly the far receiver, this option is perfectly reasonable for sensor networking scenarios in which sensors collect data over extended periods of time, and then an agent collects data at certain time instants (like the plane of *Figure 3.1*). In such a scenario, all sensor nodes must transmit their accumulated measurements simultaneously, and therefore a key assumption in the design of standard random access techniques for multiaccess communication breaks down—the fact that individual nodes will transmit with low probability [22, Chapter 4]. As a result, classical random access would result in too many collisions and hence low throughput[1]. Alternatively, instead of *mitigating* interference, a MAC protocol could attempt to *exploit* it, in the form of using cooperation among nodes to generate waveforms that add up constructively at the receiver (cf. [46]). Providing an information theoretic analysis of such cooperation mechanisms would be very desirable, but since it entails dealing with correlated sources and a general multiple access channel, dealing with correlated sources and an array of independent channels appears a very reasonable first step working towards that goal, and also interesting in its own right, since it provides the ultimate performance limits for an important class of sensor networking problems.

Communication among Sensors

Before transmitting their data to the remote receiver, the sensors are allowed to exchange messages over a network. Note however that the problem statement only specifies that pairs of sensor nodes are able to communicate reliably below given rates C_{ij}, but it does not say anything about *how* this flow of information actually takes place. The latter would force us to consider classical networking topics like topology formation, routing, and flow control, which would only complicate matters for the goals of this thesis. In our context, knowing that there is some network that would allow nodes to exchange information at certain rates is enough to prove our main results.

Perfect Reconstruction at the Receiver

In our formulation of the sensor reachback problem, the far receiver is interested in reconstructing the entire field of sensor measurements with arbitrarily small probability of error. This formulation leads us to a natural *capacity* problem, in the classical sense of Shannon. Alternatively, one could relax the condition of perfect reconstruction, and tolerate some distortion in the reconstruction of the field of measurements at the far receiver. Natural extensions in this direction are discussed in Chapter 4.

3.1.3 Related Work

The sensor reachback problem is a close relative of (a) the multiple access channel with correlated sources considered by Cover et al. [26], and (b) of the multiple access channel with partially cooperating encoders[2] solved by Willems [93], both of which were discussed in detail in the previous chapter. A general problem subsuming these two problems would be a multiple access channel with correlated sources *and* partially cooperating encoders, which to the best of our knowledge has not been studied before. Our problem is also a special case of this more general problem, since we consider correlated sources, partially cooperating encoders, and a multiple access channel without interference.

3.1.4 Main Results

In this chapter we prove a number of coding theorems, which give a complete characterization of capacity in all relevant instances of the sensor reachback problem defined in Section 3.1.1.

[1] Recent work has considered the problem of random multiaccess communication in large-scale sensor networks—see [2]. A collection of papers on multiaccess communication was compiled by Massey —see [49].

[2] Recently, the concept of partially cooperating encoders appeared again in a framework for universal multiterminal source coding proposed by Jaggi and Effros in [51].

First, we study the case of non-cooperating encoders (meaning, the case in which encoders do not exchange any information at all prior to communicating with the far receiver). We show in this case that reliable communication is possible if and only if

$$H(U(S)|U(S^c)) < \sum_{i \in S} I(X_i; Y_i),$$

for all subsets $S \subseteq \{1, 2, \ldots, M\}$, where S^c denotes the complement of S and $U(S) = \{U_j : j \in S\}$. Our proof shows that, when the multiple access channel with correlated sources considered in [26] becomes an array of independent channels, distributed source coding (Slepian-Wolf), followed by separate channel coding, is an optimal coding strategy[3]. And although for the general case of [26] only a region of achievable rates is known (without a converse), for our problem setup we are able to give a complete solution, converse included.

Based on that result, we then proceed to analyze the differences between the two problems (ours and that in [26]), by noting that the crux of the proof in [26] is a class of *correlated* joint source/channel codes that preserve statistical dependencies of the sources in the transmitted channel codewords. In the context of sensor networks, this property is interesting, since source/channel codes that do not eliminate the source correlation completely are often simpler to implement than distributed source codes, and thus require less processing capabilities at the sensor nodes. For this case, we are able to give a region of achievable rates that is strictly contained in the reachback capacity region above, and we give an exact expression for the rate loss incurred by using correlated codes. In addition, we prove that this is not a trivial region, by showing that it is strictly larger than the one corresponding to independent encoders and decoders. This means that the coding technique proposed for the achievability result does lead to an improvement over the trivial solution based on M point-to-point problems, which in turn indicates that there is something to gain from exploiting at the decoder the fact that the data streams being uploaded are correlated.

Finally, we provide a complete characterization of reachback capacity for the general sensor reachback problem with an arbitrary number of partially cooperating encoders. In this case, reliable communication is possible if and only if

$$H(U(S)|U(S^c)) < \sum_{i \in S} I(X_i; Y_i) + I(U(S); Z(S^c)|U(S^c)),$$

for all subsets $S \subseteq \{1, 2, \ldots, M\}$, where $U(S) = \{U_j : j \in S\}$, $Z(S) = \{Z_{ij} : i \in S$ or $j \in S\}$, $I(Z_{ij}; U_i|U_j) < C_{ij}$ and $I(Z_{ij}; U_j|U_i) < C_{ji}$. Here, Z_{ij} denotes auxiliary random variables that represent the information exchanged by the encoders using a network of discrete memoryless independent channels.

For all instances of the sensor reachback problem considered in this work (with and without cooperation, and for any number of nodes), we are able to prove that natural generalizations of the joint source/channel coding theorem, commonly known as the *separation* theorem, hold [28, Chapter 8.13]. This observation motivates us to include at the end of the chapter an extra section, devoted to discussing the issue of optimality of separate source and channel coding in

[3]This in no way contradicts the results in [26], as explained in Section 3.2.5.

communication networks, which we argue is a question of “when” and not “if” it holds.

3.1.5 Chapter Outline

The rest of the chapter is organized as follows. In Section 3.2 we formulate the problem of transmitting *two* correlated sources over independent channels, without cooperation among encoders, and prove a coding theorem that gives an exact characterization of the conditions for reliable communication to be possible. We also investigate the implications of using correlated codes, and characterize the rate loss resulting in that case. Then, in Section 3.3, we give another characterization of reachback capacity, this time still for two nodes, but now under the assumption that the encoders are allowed to exchange information over network links of limited capacity. Based on this result, we discuss the impact of cooperation on the reachback capacity and again address the issue of encoding constraints. The results obtained for $M = 2$ sources are generalized for arbitrary number of sources M in Section 3.4, where we also give a Gaussian example and discuss the reachback capacity for a large class of sensor networks with different topologies. In Section 3.5, we revisit the issue of source and channel separation in communication networks, under the light of the results presented in earlier sections. The chapter is summarized in Section 3.6.

3.2 Reachback Capacity with Two Non-Cooperating Nodes

We begin our study of the sensor reachback problem by providing a solution for the case of $M = 2$ non-cooperating nodes ($C_{12} = C_{21} = 0$), illustrated in *Figure 3.3*. This simple problem setup provides valuable insights into the structure of the problem, allowing us to gain a firm understanding of the main issues involved in reachback communication over independent channels, before discussing the more subtle aspects of cooperation between encoders.

3.2.1 Definitions and Problem Statement

Consider two information sources generated by repeated independent drawings of a pair of discrete random variables U_1 and U_2 from a given joint distribution $p(u_1 u_2)$. We start by providing definitions of independent channels, source/channel block codes, probability of error and reliable communication.

Definition 3.1 A reachback channel consists of two independent discrete memoryless channels $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2)$, with input alphabets $\mathcal{X}_1$ and $\mathcal{X}_2$, output alphabets $\mathcal{Y}_1$

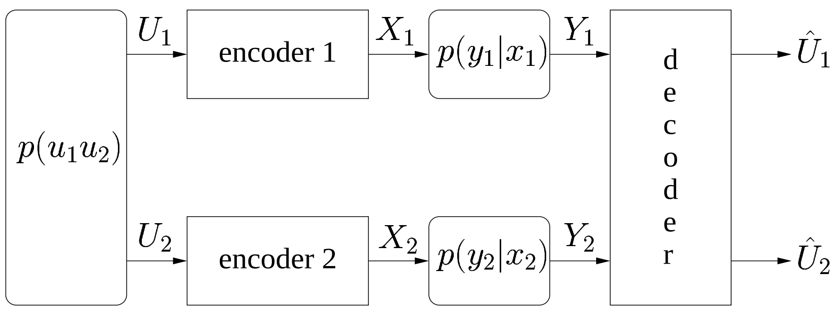


Figure 3.3: The sensor reachback problem with two non-cooperating nodes.

and $\mathcal{Y}_2$, and transition probability matrices $p(y_1|x_1)$ and $p(y_2|x_2)$.

Note that according to this definition, the reachback channel can be viewed as a multiple access channel with transition probability $p(y|x_1x_2) = p(y_1|x_1)p(y_2|x_2)$, and $Y = (Y_1Y_2)$. In other words, without interference, the multiple access channel becomes an array of independent channels.

Definition 3.2 A source/channel block code consists of an integer N, two encoding functions

$$f_1 : \mathcal{U}_1^N \to \mathcal{X}_1^N \quad \text{and} \quad f_2 : \mathcal{U}_2^N \to \mathcal{X}_2^N,$$

and a decoding function

$$g : \mathcal{Y}_1^N \times \mathcal{Y}_2^N \to \mathcal{U}_1^N \times \mathcal{U}_2^N.$$

Definition 3.3 The probability of error is given by

$$\begin{aligned} P_N &= p\{g(Y_1^N Y_2^N) \neq (U_1^N U_2^N)\} \\ &= \sum_{(u_1^N u_2^N)\in(\mathcal{U}_1^N \times \mathcal{U}_2^N)} p(u_1^N u_2^N) \cdot \; P\{g(Y_1^N Y_2^N) \neq (u_1^N u_2^N)|(U_1^N U_2^N) = (u_1^N u_2^N)\}, \end{aligned}$$

where, for a code assignment $x_1^N = f_1(u_1^N)$ and $x_2^N = f_2(u_2^N)$, the joint probability mass function is given by

$$p(u_1^N u_2^N y_1^N y_2^N) = \prod_{i=1}^{N} p(u_{1i}u_{2i})p(y_{1i}|x_{1i}(u_1^N))p(y_{2i}|x_{2i}(u_2^N)).$$

Definition 3.4 Reliable communication of the source $(U_1U_2) \sim p(u_1u_2)$ over independent channels $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2)$ is possible if there exists a sequence of block codes $\{x_1^N(u_1^N), x_2^N(u_2^N)\}$, with decoding function $g(y_1^N y_2^N)$ such that, as $N \to \infty$,

$$P_N = p\{g(Y_1^N Y_2^N) \neq (U_1^N U_2^N)\} \to 0.$$

In the following subsections we make use of the standard notions of jointly typical sequences and the AEP, explained in Chapter 2.

The main goal in this section is to characterize reachback capacity, by giving single-letter information-theoretic conditions for reliable communication.

3.2.2 Main Result

The following theorem gives necessary and sufficient conditions for reliable communication under this scenario.

Theorem 3.1 (Barros, Servetto [14]) A source (U_1U_2) drawn i.i.d. $\sim p(u_1u_2)$ can be communicated reliably over two independent channels denoted $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2)$, if and only if

$$\begin{aligned} H(U_1|U_2) &< I(X_1;Y_1) \\ H(U_2|U_1) &< I(X_2;Y_2) \\ H(U_1U_2) &< I(X_1;Y_1) + I(X_2;Y_2). \end{aligned}$$

3.2.3 Proof of Theorem 3.1

The proof begins with the converse and shows that the conditions of the theorem are necessary conditions for reliable communication to be possible. The forward part of the theorem then follows easily from the region of achievable rates defined by the Slepian-Wolf theorem.

Converse Proof

Consider a given code of block length N. The joint distribution on $\mathcal{U}_1^N \times \mathcal{U}_2^N \times \mathcal{X}_1^N \times \mathcal{X}_2^N \times \mathcal{Y}_1^N \times \mathcal{Y}_2^N$ is well defined as

$$p(u_1^N u_2^N x_1^N x_2^N y_1^N y_2^N) = \left(\prod_{i=1}^N p(u_{1i}u_{2i})\right) p(x_1^N|u_1^N)p(x_2^N|u_2^N)\left(\prod_{j=1}^N p(y_{1j}|x_{1j})\right)\left(\prod_{k=1}^N p(y_{2k}|x_{2k})\right)$$

By Fano's inequality, we can write:

$$\begin{aligned} \frac{1}{N}H(U_1^N U_2^N|Y_1^N Y_2^N) &\leq P_N \frac{1}{N}(\log|\mathcal{U}_1^N \times \mathcal{U}_2^N|) + \frac{1}{N} \\ &= \underbrace{P_N(\log|\mathcal{U}_1| + \log|\mathcal{U}_2|) + \frac{1}{N}}_{\lambda_N} \end{aligned} \tag{3.1}$$

where $|\mathcal{U}_1|$ and $|\mathcal{U}_2|$ are the alphabet sizes of U_1 and U_2, respectively. Notice that if $P_N \to 0$ as $N \to \infty$, λ_N must also go to zero. Plus, since $H(U_1^N U_2^N|Y_1^N Y_2^N) = H(U_1^N|Y_1^N Y_2^N) + H(U_2^N|U_1^N Y_1^N Y_2^N)$, we must also have $\frac{1}{N}H(U_1^N|Y_1^N Y_2^N) \leq \lambda_N$, and so we can write the following chain of inequalities:

$$\begin{aligned} NH(U_1) &= H(U_1^N) \\ &= I(U_1^N;Y_1^N Y_2^N) + H(U_1^N|Y_1^N Y_2^N) \\ &\overset{(a)}{\leq} I(U_1^N;Y_1^N Y_2^N) + N\lambda_N \\ &\overset{(b)}{\leq} I(U_1^N;Y_1^N U_2^N) + N\lambda_N \\ &= I(U_1^N;U_2^N) + I(U_1^N;Y_1^N|U_2^N) + N\lambda_N \\ &\overset{(c)}{\leq} I(U_1^N;U_2^N) + I(X_1^N;Y_1^N|U_2^N) + N\lambda_N, \end{aligned}$$

where (a) follows from (3.1), and (b) and (c) follow from the data processing inequality for the long Markov chain of the form $Y_1^N \to X_1^N \to U_1^N \to U_2^N \to X_2^N \to Y_2^N$. While the first term can be written as $I(U_1^N;U_2^N) = \sum_{i=1}^N I(U_{1i};U_{2i})$, the second term can be upper bounded by

$$\begin{aligned}
I(X_1^N;Y_1^N|U_2^N) &= H(Y_1^N|U_2^N) - H(Y_1^N|X_1^N U_2^N) \\
&\overset{(a)}{=} H(Y_1^N|U_2^N) - H(Y_1^N|X_1^N) \\
&= H(Y_1^N|U_2^N) - \sum_{i=1}^N H(Y_{1i}|Y_1^{i-1}X_1^N) \\
&\overset{(b)}{=} H(Y_1^N|U_2^N) - \sum_{i=1}^N H(Y_{1i}|X_{1i}) \\
&\overset{(c)}{\leq} H(Y_1^N) - \sum_{i=1}^N H(Y_{1i}|X_{1i}) \\
&= \sum_{i=1}^N H(Y_{1i}) - \sum_{i=1}^N H(Y_{1i}|X_{1i}) \\
&= \sum_{i=1}^N I(X_{1i};Y_{1i}),
\end{aligned}$$

since (a) Y_1^N and U_2^N are independent given X_1^N, (b) the channel is memoryless, and (c) conditioning reduces entropy. Thus,

$$\begin{aligned}
H(U_1) &\leq \frac{1}{N}\sum_{i=1}^N I(U_{1i};U_{2i}) + \frac{1}{N}\sum_{i=1}^N I(X_{1i};Y_{1i}) + \lambda_N \\
&= I(U_1;U_2) + I(X_1;Y_1) + \lambda_N,
\end{aligned}$$

because the U_{1i}'s and the U_{2i}'s are i.i.d. Now, going through identical derivations, we get

$$H(U_2) \leq I(U_1;U_2) + I(X_2;Y_2) + \lambda_N,$$

and

$$H(U_1U_2) \leq I(X_1;Y_1) + I(X_2;Y_2) + \lambda_N.$$

Taking the limit as $N \to \infty$, $P_N \to 0$, we have that

$$\begin{aligned}
H(U_1) &\leq I(U_1;U_2) + I(X_1;Y_1) \\
H(U_2) &\leq I(U_1;U_2) + I(X_2;Y_2) \\
H(U_1U_2) &\leq I(X_1;Y_1) + I(X_2;Y_2).
\end{aligned}$$

By subtracting $I(U_1;U_2)$ on both sides of the first two inequalities, we arrive at the conditions in the theorem, given by

$$\begin{aligned}
H(U_1|U_2) &\leq I(X_1;Y_1) \\
H(U_2|U_1) &\leq I(X_2;Y_2) \\
H(U_1U_2) &\leq I(X_1;Y_1) + I(X_2;Y_2),
\end{aligned}$$

thus concluding the proof of the converse. ■

Achievability Proof

We deal next with the achievability part of Theorem 3.1. Consider the following coding strategy. First, encoders 1 and 2 compress the input source blocks U_1^N and U_2^N separately using Slepian-Wolf codes. Then, the encoders add classical channel coding to transmit the compressed versions over the two independent channels. Since the Slepian-Wolf theorem guarantees that the rates

$$\begin{aligned} R_1 &\geq H(U_1|U_2) \\ R_2 &\geq H(U_2|U_1) \\ R_1 + R_2 &\geq H(U_1U_2) \end{aligned}$$

are achievable, and the channel coding theorem guarantees that the probability of error goes to zero for all rates $R_1 < I(X_1; Y_1)$ and $R_2 < I(X_2; Y_2)$, reliable communication is possible using this coding strategy if the conditions of the theorem are fulfilled.

3.2.4 A Simple Visualization of the Problem

To illustrate the issues involved in this problem consider the rate regions shown in *Figure 3.4*.

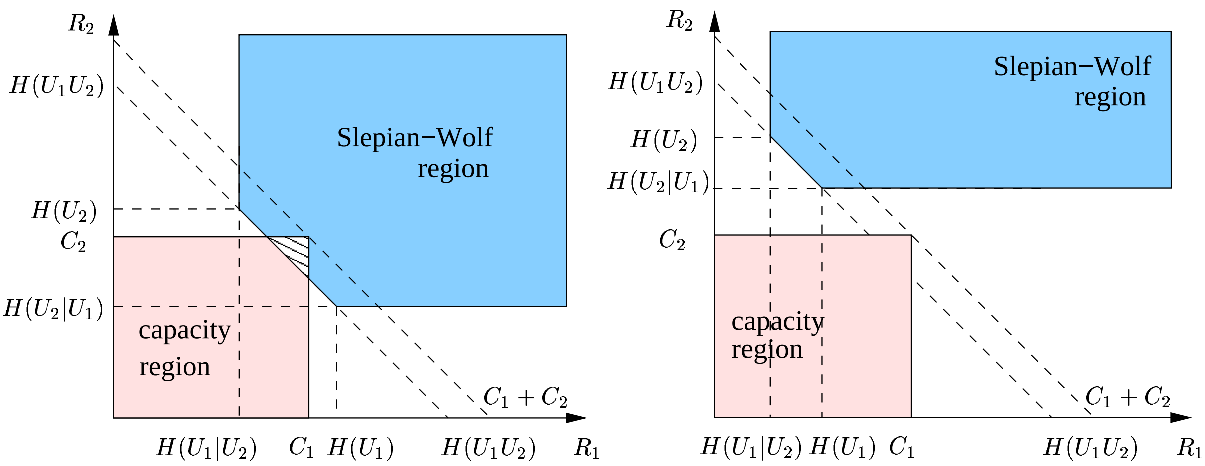


Figure 3.4: Relationship between the Slepian-Wolf region and the capacity region for two independent channels. In the left figure, as $H(U_1|U_2) < C_1$ and $H(U_2|U_1) < C_2$ the two regions intersect and therefore reliable communication is possible. The figure on the right shows the case in which $H(U_2|U_1) > C_2$ and there is no intersection between the two regions.

When the multiple access channel is reduced to two independent channels with capacities C_1 and C_2, its capacity region becomes a rectangle with side lengths C_1 and C_2 [28, Chapter 14.3]. Also shown is the Slepian-Wolf region of achievable rates for separate encoding of correlated sources, whose limits are defined by $H(U_1)$, $H(U_2)$ and $H(U_1U_2)$. Clearly, $H(U_1U_2) < C_1+C_2$ is a necessary condition for reliable communication as a consequence of Shannon's joint source and channel coding theorem for point-to-point communication. Assuming that this is the case, consider now the following possibilities:

- $H(U_1) < C_1$ and $H(U_2) < C_2$. The Slepian-Wolf region and the capacity region intersect, so any point (R_1, R_2) in this intersection makes reliable communication possible. Alternatively, we can argue that reliable transmission of U_1 and U_2 is possible even with independent decoders, therefore a joint decoder will also achieve an error-free reconstruction of the source.

- $H(U_1) > C_1$ and $H(U_2) > C_2$. Since $H(U_1U_2) < C_1 + C_2$ there is always at least one point of intersection between the Slepian-Wolf region and the capacity region, so reliable communication is possible.

- $H(U_1) < C_1$ and $H(U_2) > C_2$ (or vice versa). If $H(U_2|U_1) < C_2$ (or if $H(U_1|U_2) < C_1$) then the two regions will intersect. On the other hand, if $H(U_2|U_1) > C_2$ (or if $H(U_1|U_2) > C_1$), then there are no intersection points, but it is not immediately clear whether reliable communication is possible or not (see *Figure 3.4*), since examples are known in which the intersection between the capacity region of the multiple access channel and the Slepian-Wolf region of the correlated sources is empty and still reliable communication is possible [26].

Theorem 3.1 gives a definite answer to this last question: in the special case of correlated sources and independent channels an intersection between the capacity region and the Slepian-Wolf rate regions is not only sufficient, but also a necessary condition for reliable communication to be possible. From this observation we conclude that, in the case of independent channels, a two-stage encoder that uses Slepian-Wolf codes to compress the sources to their most efficient representation and then separately adds capacity attaining channel codes, indeed forms an optimal coding strategy—that is, for this reachback network, separation holds.

3.2.5 Rate Loss with Correlated Codes

The key ingredient of the achievability proof presented by Cover, El Gamal and Salehi for the multiple access channel with correlated sources is the generation of random codes, whose codewords X_i^N are statistically dependent on the source sequences U_i^N [26]. This property, which is achieved by drawing the codewords according to $\prod_{j=1}^{N} p(x_{ij}|u_{ij})$ with u_{ij} and x_{ij} denoting the j-th element of u_i^N and x_i^N, respectively, implies that U_i^N and X_i^N are jointly typical with high probability. Since the source sequences U_1^N and U_2^N are correlated, the codewords $X_1^N(U_1^N)$ and $X_2^N(U_2^N)$ are also correlated, and so we speak of *correlated codes*. This class of random codes, which is treated in more general terms in [4], can be viewed as joint source and channel codes that preserve the given correlation structure of the source sequences, which can then be exploited in the decoder to lower the probability of error.

Since correlated codes yield the best known characterization of achievable rates for the problem of transmitting correlated sources over the multiple access channel, it is only natural that we ask how this class of codes performs in the sensor reachback problem, for which we know that separate source and channel coding is optimal. This issue is also interesting from a practical point of view, since sensor nodes with limited processing capabilities may be forced to use very simple codes that do not eliminate correlations between measurements prior to transmission [16]. In this case, we are interested in knowing how far the remaining correlation in the codewords can still be used by the receiver to improve the decoding result. The following result gives a characterization of the reachback capacity under this scenario.

Theorem 3.2 (Barros, Servetto [14]) A source $(U_1^N U_2^N) \sim \prod_{i=1}^N p(u_{1i}u_{2i})$ can be sent with arbitrarily small probability of error over two independent channels $\{\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1\}$ and $\{\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2\}$, with *correlated codes* $\{X_1^N(U_1^N), X_2^N(U_2^N)\}$ if

$$H(U_1|U_2) < I(X_1;Y_1|U_2) \tag{3.2}$$
$$H(U_2|U_1) < I(X_2;Y_2|U_1) \tag{3.3}$$
$$H(U_1U_2) < I(X_1X_2;Y_1Y_2), \tag{3.4}$$

for some $p(u_1u_2)p(x_1|u_1)p(x_2|u_2)p(y_1|x_1)p(y_2|x_2)$.

The proof is based on the joint source-channel codes of [26]. We repeat the description of that code construction here to highlight the property that is most relevant to us: the codewords are generated in statistical dependence to the source sequences, and are therefore correlated. *Proof:* Fix $p(x_1|u_1)$ and $p(x_2|u_2)$. For each $u_1^N \in \mathcal{U}_1^N$, independently generate one x_1^N sequence according to $\prod_{i=1}^N p(x_{1i}|u_{1i})$. Index the x_1^N sequences by $x_1^N(u_1^N)$, $u_1^N \in \mathcal{U}_1^N$. Similarly, for each $u_2^N \in \mathcal{U}_2^N$, independently generate one x_2^N sequence according to $\prod_{i=1}^N p(x_{2i}|u_{2i})$. Index the x_2^N sequences by $x_2^N(u_2^N)$, $u_2^N \in \mathcal{U}_2^N$. Notice that each random code is generated according to a conditional probability on the source observed by the corresponding encoder.

Encoding

To send sequence u_1^N, transmitter 1 sends the codeword $x_1^N(u_1^N)$. Similarly, to send sequence u_2^N, transmitter 2 sends codeword $x_2^N(u_2^N)$.

Decoding

Upon observing the received sequences y_1^N and y_2^N, the decoder declares $(\hat{u}_1^N\hat{u}_2^N)$ to be the transmitted source sequence pair if $(\hat{u}_1^N\hat{u}_2^N)$ is the unique pair $(u_1^N u_2^N)$ such that

$$(u_1^N, u_2^N, x_1^N(u_1^N), x_2^N(u_2^N), y_1^N, y_2^N) \in A_\epsilon^N,$$

where A_ϵ^N is the appropriate set of jointly typical sequences according to the definition in [28]. In [26], the code construction described above is used to characterize the conditions for reliable communication of two correlated sources drawn i.i.d. $\sim p(u_1u_2)$ over a multiple access channel $p(y|x_1x_2)$. The conditions obtained in that more general setup are given by (2.7)-(2.9). Thus, it suffices to specialize this result to the case of independent channels. Let $y = (y_1y_2)$ and let

$p(y|x_1x_2) = p(y_1y_2|x_1x_2) = p(y_1|x_1)p(y_2|x_2)$. Based on (2.7) we can write

$$\begin{aligned}
H(U_1|U_2) &< I(X_1;Y|U_2X_2) \\
&= I(X_1;Y_1Y_2|U_2X_2) \\
&= H(Y_1Y_2|U_2X_2) - H(Y_1Y_2|U_2X_2X_1) \\
&= H(Y_1|U_2X_2) + H(Y_2|U_2X_2Y_1) - H(Y_1|U_2X_2X_1) - H(Y_2|U_2X_2X_1Y_1) \\
&\overset{(a)}{=} H(Y_1|U_2) + H(Y_2|X_2) - H(Y_1|U_2X_1) - H(Y_2|X_2) \\
&= H(Y_1|U_2) - H(Y_1|U_2X_1) \\
&= I(X_1;Y_1|U_2),
\end{aligned}$$

where (a) follows from the long Markov chain condition $Y_1 \rightarrow X_1 \rightarrow U_1 \rightarrow U_2 \rightarrow X_2 \rightarrow Y_2$, which implies that Y_1 is independent of X_2 given U_2, Y_2 is independent of U_2, X_1 and Y_1 given X_2, and Y_1 is independent of X_2 given X_1. Similarly, for conditions (2.8) and (2.9) we get $H(U_2|U_1) < I(X_2;Y_2|U_1)$ and $H(U_1U_2) < I(X_1X_2;Y_1Y_2)$ thus concluding the proof. ■

It is interesting to observe that in this instance of the problem, we not only have a long Markov chain $Y_1^N \rightarrow X_1^N \rightarrow U_1^N \rightarrow U_2^N \rightarrow X_2^N \rightarrow Y_2^N$ at the block level (due to the functional dependencies introduced by the encoders and the channels), but we also have a *single-letter* Markov chain $Y_1 \rightarrow X_1 \rightarrow U_1 \rightarrow U_2 \rightarrow X_2 \rightarrow Y_2$ that comes from using correlated codes and fixing the conditional probability distributions $p(x_1|u_1)$ and $p(x_2|u_2)$.

More importantly, the proof shows that the result of Cover, El Gamal and Salehi for the multiple access channel with correlated sources in [26] does not immediately specialize to Theorem 3.1, when we assume a multiple access channel with conditional probability distribution $p(y|x_1x_2) = p(y_1y_2|x_1x_2) = p(y_1|x_1)p(y_2|x_2)$. Taking a closer look at the first condition in the theorem, we observe that

$$\begin{aligned}
I(X_1;Y_1|U_2) &= H(Y_1|U_2) - H(Y_1|X_1U_2) \\
&= H(Y_1|U_2) - H(Y_1|X_1) \\
&\leq H(Y_1) - H(Y_1|X_1) \\
&= I(X_1;Y_1),
\end{aligned}$$

with equality for $p(u_1u_2x_1x_2) = p(u_1u_2)p(x_1)p(x_2)$. This means that we must choose independent codewords to obtain the achievable region of Theorem 3.1, and so we conclude that when transmitting correlated sources over independent channels correlated codes are in general suboptimal. At first glance, this observation is somewhat surprising, since the sensor reachback problem with non-cooperating nodes is a special case of the multiple access channel with correlated sources considered in [26], where it is shown that in the general case correlated codes outperform Slepian-Wolf codes (independent codewords). The crucial difference between the two problems is the presence (or absence) of interference in the channel. Albeit somewhat informally, we can state that correlated codes are advantageous when the transmitted codewords are combined in the channel through interference, which is clearly not the case in our formulation of the sensor reachback problem.

We now give a characterization of the rate loss incurred into by using correlated codes in our problem setup. Comparing the conditions of Theorems 3.1 and 3.2, we can describe the gap between the two regions of achievables rates. Focusing on the first of the three conditions, the extent Δ_1 of this gap can be written as:

$$\begin{aligned}
\Delta_1 &= I(X_1;Y_1) - I(X_1;Y_1|U_2) \\
&= I(X_1;Y_1) - (H(Y_1|U_2) - H(Y_1|X_1U_2)) \\
&= I(X_1;Y_1) - (H(Y_1|U_2) - H(Y_1|X_1)) \\
&= I(X_1;Y_1) - (H(Y_1) - H(Y_1) + H(Y_1|U_2) - H(Y_1|X_1)) \\
&= I(X_1;Y_1) - I(X_1;Y_1) + I(Y_1;U_2) \\
&= I(Y_1;U_2). \qquad (3.5)
\end{aligned}$$

Similarly, for the second of the three conditions, we get a gap $\Delta_2 = I(Y_2;U_1)$ and for the sum rate condition

$$\begin{aligned}
\Delta_0 &= I(X_1;Y_1) + I(X_2;Y_2) - I(X_1X_2;Y_1Y_2) \\
&= I(X_1;Y_1) + I(X_2;Y_2) - (I(X_1;Y_1Y_2) + I(X_2;Y_1Y_2|X_1)) \\
&= I(X_1;Y_1) + I(X_2;Y_2) - (I(X_1;Y_1) + I(X_1;Y_2|Y_1) + I(X_2;Y_2|X_1) + I(X_2;Y_1|X_1Y_2)) \\
&= I(X_2;Y_2) - (I(X_1;Y_2|Y_1) + I(X_2;Y_2|X_1)) \\
&= I(X_2;Y_2) - (H(Y_2|Y_1) - H(Y_2|Y_1X_1) + H(Y_2|X_1) - H(Y_2|X_1X_2)) \\
&= I(X_2;Y_2) - (H(Y_2|Y_1) - H(Y_2|X_1X_2)) \\
&= I(X_2;Y_2) - (H(Y_2) - H(Y_2) + H(Y_2|Y_1) - H(Y_2|X_2))
\end{aligned}$$

$$\begin{aligned}
&= I(X_2;Y_2) - (I(X_2;Y_2) - I(Y_1;Y_2)) \\
&= I(Y_1;Y_2)
\end{aligned}$$

Since $\Delta_i \geq 0$, $i \in \{0,1,2\}$ (mutual information is always nonnegative), we conclude that the region of achievable rates given by Theorem 3.2 is contained in the region defined by Theorem 3.1. Furthermore, we find that the rate loss terms have a simple, intuitive interpretation: Δ_0 is the loss in sum rate due to the dependencies between the outputs of different channels, and Δ_1 (or Δ_2) represent the rate loss due to the dependencies between the outputs of channel 1 (or 2) and the source transmitted over channel 2 (or 1). All these terms become zero if, instead of using correlated codes, we fix $p(x_1)p(x_2)$ and remove the correlation between the source blocks before transmission over the channels.

Note also that the region defined by Theorem 3.2 is not trivial, in that it contains *more* points than those that can be achieved by two pairs of independent encoders/decoders. From (3.2) and (3.5) it follows that

$$\begin{aligned}
H(U_1) &< I(U_1;U_2) + I(X_1;Y_1|U_2) \\
&= I(U_1;U_2) + I(X_1;Y_1) - I(Y_1;U_2),
\end{aligned}$$

and similarly, using (3.3) we get

$$\begin{aligned} H(U_2) &< I(U_1;U_2) + I(X_2;Y_2|U_1) \\ &= I(U_1;U_2) + I(X_2;Y_2) - I(Y_2;U_1). \end{aligned}$$

Applying the data processing inequality based on the long Markov chain $Y_1 \to X_1 \to U_1 \to U_2 \to X_2 \to Y_2$, we observe that $I(U_1;U_2) - I(Y_1;U_2) \geq 0$, and similarly, $I(U_1;U_2) - I(Y_2;U_1) \geq 0$, and thus conclude that the region of Theorem 3.2 is in general larger than the region obtained by solving two independent point-to-point problems. The difference between these two regions is what we can expect to gain from exploiting the correlation between codewords in the decoding step.

3.3 Reachback Capacity with $M = 2$ Partially Cooperating Nodes

We now extend the results in Section 3.2 to allow for partial cooperation between the encoders. Once again, we start with some definitions and a formal statement of the problem. The latter differs from the previous problem statement in the definition of the encoders, and in their ability to establish a *conference* prior to transmission over the reachback channel.

We start with some discussion about the conferencing mechanism, inspired by [93]. Assume that encoder 1 can send messages to encoder 2 over a channel with capacity C_{12}, and vice versa (encoder 2 to encoder 1 over a channel with capacity C_{21}). These messages could represent, for example, synchronization information: "In transmission 12 I will send $X_1 = 3$", "I will transmit zeros in transmissions 22, 24 and 26", etc. They could also represent quantized versions of the observed source values: "My sample is positive", "I will send an index between 128 and 132", etc. The simplest form of conference can be characterized as two *simultaneous monologues*: encoder 1 sends a block of messages to encoder 2, and encoder 2 sends a block of messages to encoder 1. In [93], Willems presents a more general definition of a conference, which is closer to a *dialogue*. Let V_{ik} be the message sent by encoder i at the k-th transmission; encoder 1 sends the first message V_{11}, then encoder 2 sends its first message V_{21}, after which encoder 1 sends another message V_{12}, then encoder 2 sends V_{22}, and so on. This type of conference is more general not only because it admits multiple messages to be exchanged between the encoders, but, more interestingly, because it allows the next message V_{1k} (or V_{2k}) to be sent by encoder 1 (or encoder 2) to depend on all previously received messages V_2^{k-1} (or V_1^{k-1}). It turns out that both in the capacity problem considered by Willems in [93] and in the sensor reachback problem, two *simultaneous monologues* are sufficient to achieve all points in the capacity region.

3.3.1 Definitions and Problem Statement

A *reachback network* consists of two sender nodes and one receiver node. Sender 1 is connected to the receiver via a discrete memoryless channel $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1)$ and sender 2 via $(\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2)$. Senders 1 and 2 are joined by network links of capacity C_{12} and C_{21} with information being exchanged in opposite directions. This setup was illustrated in *Figure 3.2*.

A *conference* among encoders is specified by a set of $2K$ functions

$$\begin{aligned} h_{1k} : \mathcal{U}_1^N \times \mathcal{V}_{21} \times ... \times \mathcal{V}_{2(k-1)} &\rightarrow \mathcal{V}_{1k} \\ h_{2k} : \mathcal{U}_2^N \times \mathcal{V}_{11} \times ... \times \mathcal{V}_{1(k-1)} &\rightarrow \mathcal{V}_{2k}, \end{aligned}$$

such that the conference message $V_{1k} \in \mathcal{V}_{1k}$ (or $V_{2k} \in \mathcal{V}_{2k}$) of encoder 1 (or encoder 2) at time k depends on the previously received messages V_2^{k-1} (or V_1^{k-1}) and the corresponding source message. The conference rates[4] are given by

$$R_{12} = (1/K) \sum_{k=1}^{K} \log_2 |\mathcal{V}_{1k}| \quad \text{and} \quad R_{21} = (1/K) \sum_{k=1}^{K} \log_2 |\mathcal{V}_{2k}|.$$

A conference is said to be (C_{12}, C_{21})-admissible if and only if

$$KR_{12} \leq NC_{12} \quad \text{and} \quad KR_{21} \leq NC_{21}.$$

The *encoders* are two functions:

$$\begin{aligned} f_1 : \mathcal{U}_1^N \times \mathcal{V}_{21} \times ... \times \mathcal{V}_{2K} &\rightarrow \mathcal{X}_1^N \\ f_2 : \mathcal{U}_2^N \times \mathcal{V}_{11} \times ... \times \mathcal{V}_{1K} &\rightarrow \mathcal{X}_2^N \end{aligned}$$

These functions map a block of N source symbols observed by each encoder, and a block of K messages received from the other encoder, to a block of N channel inputs. The *decoder* is a function

$$g : \mathcal{Y}_1^N \times \mathcal{Y}_2^N \rightarrow \hat{\mathcal{U}}_1^N \times \hat{\mathcal{U}}_2^N.$$

g maps two blocks of channel outputs (one from each channel) into two blocks of reconstructed source sequences.

An $(R_1, R_2, R_{12}, R_{21}, N, K, P_N)$ *code* for this problem is defined by:

- Two encoders f_1,f_2, with $|f_1| = 2^{NR_1}$, $|f_2| = 2^{NR_2}$,
- A decoder g for the two encoders f_1 and f_2,
- A (C_{12}, C_{21})-admissible conference of length K and rates R_{12} and R_{21},
- $p(\hat{U}_1^N \hat{U}_2^N \neq U_1^N U_2^N) \leq P_N$.

Finally, we say that *reliable communication* is possible, meaning that the sources $(U_1 U_2)$ can be sent over this network with arbitrarily small probability of error, if, for sufficiently large blocklength N, there exists a (C_{12}, C_{21})-admissible conference of length K and a $(R_1, R_2, R_{12}, R_{21}, N, K, \epsilon)$ code, for all $\epsilon > 0$.

[4]At first glance, it might seem puzzling that the conference rates are defined in terms of the size of the alphabets as in [93], because in our problem the conference messages sent by one encoder and the source values observed by the other encoder are dependent. Note, however, that the definition of the $2K$ encoding functions that characterize the conference is general enough to admit a random binning mechanism that eliminates said statistical dependence. The present definition of the conference rates is therefore perfectly reasonable.

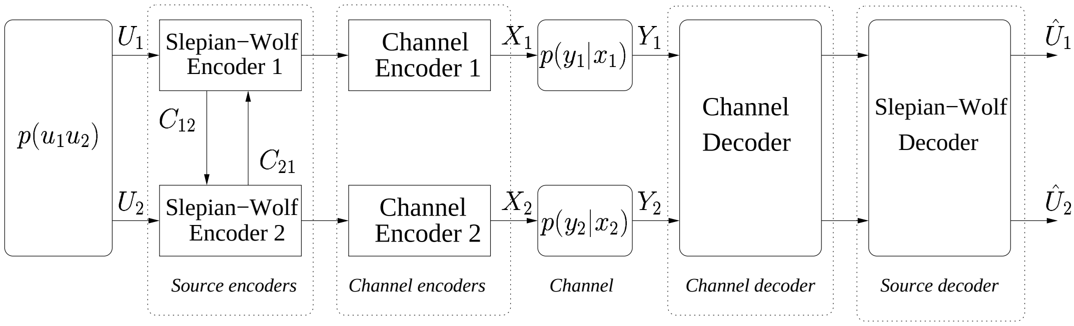


Figure 3.5: Coding strategy for the achievability proof of Theorem 3.3: cooperative Slepian-Wolf source codes followed by classical channel codes.

The goal of the problem is to characterize the *reachback capacity* of the network by giving single-letter information-theoretic conditions for reliable communication.

3.3.2 Statement of Main Result

Theorem 3.3 (Barros, Servetto [9]) Reliable communication is possible if and only if

$$H(U_1|U_2) < I(X_1;Y_1) + I(U_1;Z|U_2) \quad (3.6)$$
$$H(U_2|U_1) < I(X_2;Y_2) + I(U_2;Z|U_1) \quad (3.7)$$
$$H(U_1U_2) < I(X_1;Y_1) + I(X_2;Y_2), \quad (3.8)$$

for some auxiliary variable Z such that $I(U_1;Z|U_2) < C_{12}$, $I(U_2;Z|U_1) < C_{21}$, $|\mathcal{Z}| \leq |\mathcal{U}_1||\mathcal{U}_2|$.

3.3.3 Achievability Proof based on Cooperative Source Coding

The achievability part of the proof is based on separate source and channel coding. First, we describe the conferencing mechanism, then we give the rate region for distributed source coding with partial cooperation between encoders. The conditions in the theorem then follow from the intersection of this rate region with the capacity region of the channels. The resulting system architecture is illustrated in *Figure 3.5*.

Proof: Partition the set $\mathcal{U}_1$ in M_1 cells, indexed by $v_1 \in \{1, 2, \ldots, M_1\}$, such that $v_1(u_1) = c_1$ if u_1 is inside cell c_1. Similarly, partition the set $\mathcal{U}_2$ in M_2 cells, indexed by $v_2 \in \{1, 2, \ldots, M_2\}$, such that $v_2(u_2) = c_2$ if u_2 is inside cell c_2.

Upon observing a block u_1^N of source outputs, encoder 1 determines v_1 for each observed value u_1. Similarly, encoder 2 determines v_2 for each observed value u_2 of the source output block u_2^N. Using the conference mechanism encoder 1 can send a block v_1^N to encoder 2 at rate $R_{12} < C_{12}$, and encoder 2 can send a block v_2^N to encoder 1 at rate $R_{21} < C_{21}$. We will now show that the rates $R_{12} = H(V_1|U_2)$ and $R_{21} = H(V_2|U_1)$ are sufficient for V_1^N and V_2^N to be exchanged between the encoders with arbitrarily small probability of error. Since (U_1U_2) are random and (V_1V_2) are functions of (U_1U_2), (V_1V_2) are random as well. The encoders are assumed to have knowledge of the joint distribution $p(u_1u_2v_1v_2)$, from which they can obtain the marginals $p(u_1v_2)$ and $p(u_2v_1)$. Notice that these two distributions can be viewed as two pairs of correlated sources (U_1V_2) and (U_2V_1). Since U_2^N is known at encoder 2, it follows from the Slepian-Wolf theorem for (U_2V_1) that V_1^N can be compressed at rates $R_{12} \geq H(V_1|U_2)$ and still be reconstructed perfectly at encoder 2. Similarly, V_2^N can be compressed at rates $R_{21} \geq H(V_2|U_1)$ and still be reconstructed perfectly at encoder 1. Thus, using separate source and channel coding, V_1^N and V_2^N can be transmitted over the conference links at rates

$$R_{12} = H(V_1|U_2) \quad < \quad C_{12} \tag{3.9}$$

$$R_{21} = H(V_2|U_1) \quad < \quad C_{21}, \tag{3.10}$$

with arbitrarily small probability of error.

Let $Z = (V_1V_2)$. Since (V_1V_2) are functions of the source random variables (U_1U_2), Z is also a random variable and a function of (U_1U_2), which in turn means that $p(u_1u_2z) = p(u_1u_2)p(z|u_1u_2)$ is a well-defined probability distribution. Instead of (3.9), we can now write

$$\begin{aligned}
C_{12} &> H(V_1|U_2) \\
&= H(V_1V_2|U_2) \\
&= H(V_1V_2|U_2) - H(V_1V_2|U_1U_2) \\
&= I(U_1;V_1V_2|U_2) \\
&= I(U_1;Z|U_2).
\end{aligned}$$

Similarly, (3.10) yields $C_{21} > I(U_2;Z|U_1)$.

After conferencing, the encoders compress their data using distributed source codes. Let $U_1' = (U_1Z)$ and $U_2' = (U_2Z)$. Since U_1 and U_2 are i.i.d. sources, $Z = f(U_1U_2)$ is also i.i.d. U_1' and U_2' can be viewed as two i.i.d. sources $\sim p(u_1'u_2') = p(u_1u_2z)$. Then, according to the Slepian-Wolf theorem, the following compression rates are achievable:

$$\begin{aligned}
R_1 &> H(U_1'|U_2') \\
R_2 &> H(U_2'|U_1') \\
R_1 + R_2 &> H(U_1'U_2').
\end{aligned}$$

Substituting $U_1' = (U_1Z)$ and $U_2' = (U_2Z)$, we get

$$\begin{aligned}
R_1 &> H(U_1Z|U_2Z) = H(U_1|U_2Z) \\
R_2 &> H(U_2Z|U_1Z) = H(U_2|U_1Z) \\
R_1 + R_2 &> H(U_1U_2Z) = H(U_1U_2).
\end{aligned}$$

Adding channel codes separately, we conclude that reliable communication is possible if this rate region intersects the capacity region of the channels. We can write this as

$$\begin{aligned}
H(U_1|U_2Z) &< I(X_1;Y_1) \\
H(U_2|U_1Z) &< I(X_2;Y_2) \\
H(U_1U_2) &< I(X_1;Y_1) + I(X_2;Y_2),
\end{aligned}$$

which is equivalent to

$$\begin{aligned} H(U_1|U_2) &< I(X_1;Y_1) + I(Z;U_1|U_2) \\ H(U_2|U_1) &< I(X_2;Y_2) + I(Z;U_2|U_1) \\ H(U_1U_2) &< I(X_1;Y_1) + I(X_2;Y_2), \end{aligned}$$

thus concluding the proof of achievability. ■

Notice that the conference mechanism described in the proof relies on two deterministic partitions, which can be chosen arbitrarily. These two partitions determine $p(v_1v_2)$, which can easily be obtained from $p(u_1u_2)$ by summing over all (u_1u_2) in each pair of partition cells indexed by (v_1v_2). Since $p(v_1v_2) = p(z)$, the choice of partition determines the auxiliary random variable Z, which together with the source and channel encoders define the operation point in the reachback capacity region. In other words, for an arbitrary choice of Z (or equivalently of partitions) for which there exists an admissible conference such that $I(U_1;Z|U_2) < C_{12}$ and $I(U_2;Z|U_1) < C_{21}$, Theorem 3.3 gives the conditions for reliable communication, i.e. , the exact reachback capacity with partially cooperating encoders. The latter includes all achievable points for an arbitrary choice of Z, and so it is not necessary to specify the partitions any further.

Instead of exchanging conference messages first and then performing separate source and channel coding, one could start by compressing the sources using Slepian-Wolf codes, and then allowing the channel encoders to exchange messages as proposed by Willems in [93]. We address this issue in Appendix A.1, by giving an alternative achievability proof based on the coding strategy shown in *Figure 3.6*. It turns out that there is nothing to lose from moving the conference mechanism to the channel encoders.

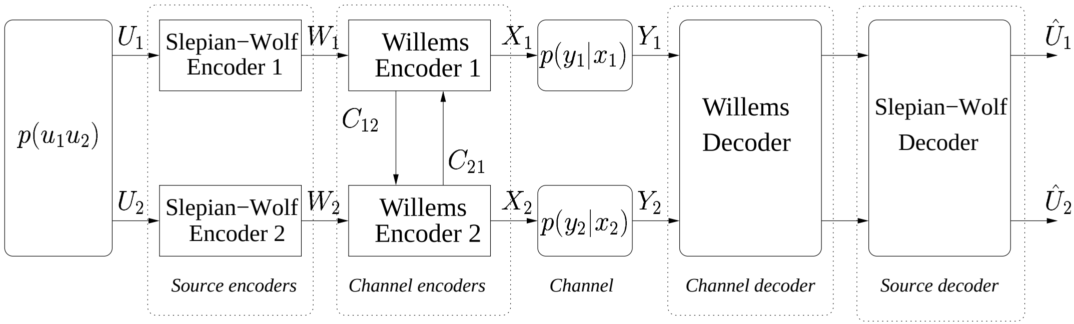


Figure 3.6: Coding strategy for the alternative achievability proof of Theorem 3.3: classical Slepian-Wolf source codes followed by Willems' cooperative channel codes.

3.3.4 Converse of Theorem 3.3

The converse part of Theorem 3.3 can be proved, similarly to Theorem 3.1, using Fano's inequality and standard techniques. By exploiting two long Markov chains, in this case $Y_2^N \to X_2^N \to (Z^N U_1^N) \to X_1^N \to Y_1^N$ and $Y_2^N \to X_2^N \to (U_2^N Z^N) \to X_1^N \to Y_1^N$, we can show that the conditions obtained in the previous subsection are not only sufficient but also necessary for reliable communication to be possible. Since the complete proof is rather technical and lengthy (but conceptually straightforward), details are only provided in Appendix A.2.

3.3.5 Cooperative versus Non-Cooperative Reachback

We now take a closer look at the implications of Theorem 3.3. The first thing to note is that Z is a variable that can be interpreted as the information exchanged by the two encoders. Therefore, by explicitly solving the channel capacity problems for communication between the two encoders and with the far receiver, a simpler (and more intuitive) version of the theorem is obtained:

$$\begin{aligned} H(U_1|U_2) &< C_1 + C_{12} \\ H(U_2|U_1) &< C_2 + C_{21} \\ H(U_1U_2) &< C_1 + C_2, \end{aligned}$$

where C_1, C_2, C_{12}, C_{21} are the capacities of the corresponding channels. Now, from this simpler version, we can see clearly that there is indeed a strict improvement over the conditions given in Theorem 3.1 for the case of no cooperation. For example, it is easy to see that if $C_{12} > H(U_1|U_2)$, and $C_{21} > H(U_2|U_1)$, then these conditions reduce to the one for the classical point-to-point problem: $H(U_1U_2) < C_1 + C_2$. Any point on the surface of the sum-rate face of the region can be achieved by having encoders send their realization to each other using Slepian-Wolf codes—in this way, both can reconstruct both sources, generate a joint encoding, and then split this encoding in whatever way they choose to. Some of these points are clearly not achievable without cooperation among encoders: e.g. , $(R_1, R_2) = (0, H(U_1U_2))$ and $H(U_1|U_2) > 0$, even if $H(U_1U_2) < I(X_2; Y_2)$. This is illustrated in *Figure 3.7*.

It is also interesting to observe in *Figure 3.7* that, contrary to what happens with general multiple access channels, in the case of independent channels considered in this work cooperation does *not* lead to an increase in the achievable sum-rate. This suggests a *routing* interpretation for Theorem 3.3. If, say, encoder 1 has too much data to send and a channel not good enough (that is, $H(U_1|U_2) > C_1$), and encoder 2 has enough idle capacity (that is, $H(U_2|U_1) < C_2$), then encoder 1 can use Z to route some of its data to the far receiver via encoder 2. The total number of bits that can be sent over the reachback network is still bounded by $I(X_1; Y_1) + I(X_2; Y_2)$. But, provided this constraint is not violated, each encoder can also act as a relay for the other encoder, in this way relaxing the conditions on the minimum amount of data required from each encoder. And the total amount of information that can be exchanged among nodes is given by the capacity of the interconnection network between encoders.

Incidentally, note also that the widely accepted view that "according to the Slepian-Wolf theorem separate encoders can achieve the same compression performance of a joint encoder" is only partially accurate. Indeed, the *total* number of bits required with separate and with joint encoders remains the same. However, for the informal statement above to be an accurate description of things, the joint decoder needs to receive a minimum amount of information from each encoder, so not any point in the sum-rate region $R_1 + R_2 > H(U_1U_2)$ is achievable without cooperation. The net effect of cooperation is to relax this requirement, to the extent supported by the interconnection network between encoders.

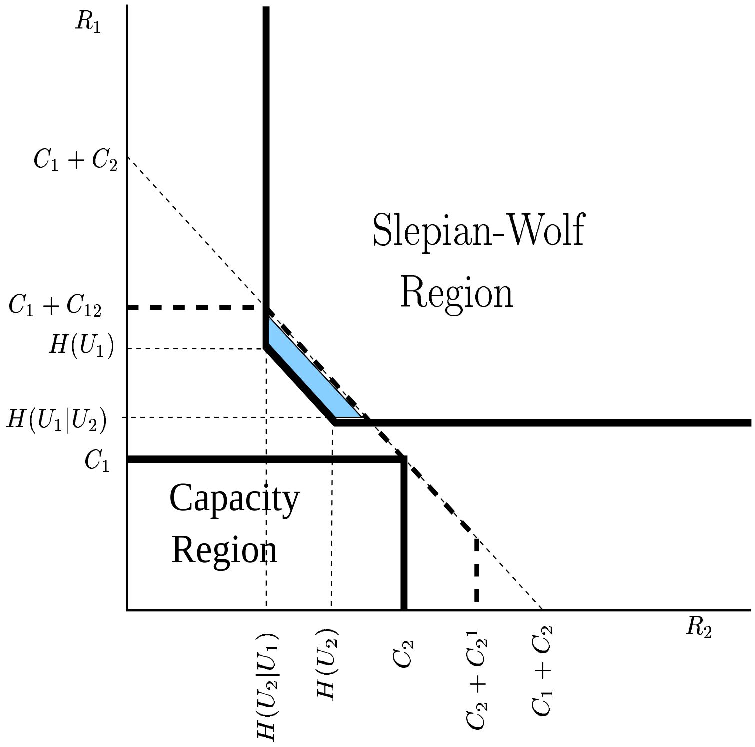


Figure 3.7: An example to illustrate the effect of cooperation among encoders on the feasibility of reliable communication between the reachback network and the far receiver. Since the capacity region of the pair of independent channels and the Slepian-Wolf region do not intersect, it follows from Theorem 3.1 that reliable communication is not possible. However, with cooperation we can enlarge side faces of the capacity region by the capacity of the conference channels until there are points of intersection with the Slepian-Wolf region (the shaded portion of the picture). To achieve such points, it is necessary for the encoder with a bad channel to route some of its data to the joint decoder via the good channel available to the other encoder.

3.3.6 Partially Cooperating Nodes with Constrained Encoders

In Section 3.2.5 we discussed a reachback scenario in which the sensor nodes send correlated codewords instead of using Slepian-Wolf source coding. As argued there, this constraint is interesting in part because it models the case in which the sensor nodes have limited complexity and are not capable of encoding their data optimally to remove correlations. In this subsection, we address the same issue, now in the presence of partial cooperation among encoders.

Going back to the achievability proof of Theorem 3.3, we observe that Slepian-Wolf codes are used for two different tasks: (1) to compress the source messages prior to transmission over the reachback channels, and (2) to compress the conference messages prior to transmission over the conference links:

- The first task is identical to the case with non-cooperative encoders, and so it is reasonable to assume that the use of correlated codes will lead to a rate loss relative to the region given by Theorem 3.3 similar to that shown in Theorem 3.2.

- The second task imposes one additional requirement on the encoders – to be able to reconstruct the Slepian-Wolf encoded conference messages the encoders must have full knowledge of the joint probability distribution $p(u_1u_2)$.

From a practical point of view, this requirement could pose some difficulties, since it may be hard for the sensor nodes to obtain or estimate the dependencies among all observed variables. Part of the reason that makes Slepian-Wolf codes attractive for practical sensor networking applications is the fact that, by not requiring knowledge of typical sets at the encoders, complexity is moved to the decoder. Yet in this conferencing mechanism, we do require knowledge of those typical sets at the encoders, to decode conference messages. We are therefore interested in obtaining a set of conditions for reliable communication in which the encoders do not exploit the joint distribution for conferencing.

The following theorem gives sufficient conditions for reliable communication under these two encoding constraints.

Theorem 3.4 (Barros, Servetto [10]) Let (U_1U_2) be two correlated sources drawn i.i.d. $\sim p(u_1u_2)$, which are to be transmitted over two independent channels $\{\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1\}$ and $\{\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2\}$. Assuming that the encoders are connected by communication links of capacities C_{12} and C_{21}, do not have knowledge of $p(u_1u_2)$, and use *correlated* codes denoted $\{x_1^n(u_1^n), x_2^n(u_2^n)\}$ then reliable communication is possible if

$$\begin{aligned} H(U_1|U_2) &< I(X_1;Y_1|U_2Z) + I(U_1;Z|U_2) \\ H(U_2|U_1) &< I(X_2;Y_2|U_1Z) + I(U_2;Z|U_1) \\ H(U_1U_2) &< I(X_1X_2;Y_1Y_2), \end{aligned}$$

for some $p(u_1u_2)p(v_1|u_1)p(v_2|u_2)p(x_1|u_1)p(x_2|u_2)p(y_1|x_1)p(y_2|x_2)$ and $Z = (V_1V_2)$, such that $H(Z|V_2) < C_{12}$, $H(Z|V_1) < C_{21}$, and $|\mathcal{Z}| \leq |\mathcal{U}_1||\mathcal{U}_2|$.

Proof: We start with the conferencing mechanism and then obtain the conditions for reliable communication by generalizing Theorem 3.2. The conference messages are generated using the same partitions as in the proof of Theorem 3.3. The pair of conference channels is then equivalent to a two-way channel without interference [86, pp. 351-352], yet with correlated inputs. Since the encoders cannot exploit the joint probability distribution, we assume they compress the conference messages to their marginal entropies and then add channel coding to transmit them reliably over the conference channels.[5] It follows then from the source coding theorem and from the channel coding theorem that a sufficient condition for reliable communication of the conference messages to be possible is $H(V_1) = H(Z|V_2) < C_{12}$ and $H(V_2) = H(Z|V_1) < C_{21}$, with $Z = (V_1V_2)$.

Now, let $U_1' = (U_1Z)$ and $U_2' = (U_2Z)$. Since U_1 and U_2 are i.i.d. sources, $Z = f(U_1U_2)$ is also i.i.d., and U_1' and U_2' can be viewed as two i.i.d. sources $\sim p(u_1'u_2') = p(u_1u_2z)$. Then,

[5] Note that we do *not* claim separate source and channel coding to be an optimal coding strategy for sending correlated sources over a two-way channel without interference and without knowledge of the joint probability distribution at the encoders. This is because, without proof, we cannot rule out the existence of a *universal* coding strategy (meaning, without a priori knowledge of source statistics), leading to less restrictive conditions for reliable communication over the conference channels than those stated above.

according to Theorem 3.2, reliable communication is possible if

$$\begin{aligned} H(U_1'|U_2') &< I(X_1;Y_1|U_2') \\ H(U_2'|U_1') &< I(X_2;Y_2|U_1') \\ H(U_1'U_2') &< I(X_1X_2;Y_1Y_2). \end{aligned}$$

Substituting $U_1' = (U_1Z)$ and $U_2' = (U_2Z)$, we get

$$\begin{aligned} H(U_1|U_2Z) &< I(X_1;Y_1|U_2Z) \\ H(U_2|U_1Z) &< I(X_2;Y_2|U_1Z) \\ H(U_1U_2Z) &< I(X_1X_2;Y_1Y_2). \end{aligned}$$

The conditions in the theorem then follow from standard identities. ∎

Comparing the expressions in Theorems 3.2, 3.3, and 3.4, we conclude that the rate loss under the given encoding constraints is twofold. First, the use of correlated codes leads to similar rate loss terms as in Section 3.2.5. Secondly, the restriction on the choice of codes for conferencing implies a restriction on the choice of auxiliary variable Z, possibly leading to smaller values of $I(U_1;Z|U_2)$ and $I(U_2;Z|U_1)$ and thus to a potential reduction in reachback capacity.

3.4 Reachback Capacity with an Arbitrary Number of Nodes

3.4.1 $M \geq 2$ Non-Cooperating Nodes

Having established the reachback capacity region for the case of two non-cooperating nodes, we now generalize this result to the transmission of M correlated sources over M independent channels, for arbitrary $M \geq 2$. The following theorem gives the conditions for reliable communication for this case, illustrated in *Figure 3.8*.

Theorem 3.5 (Barros, Servetto [14]) A set of correlated sources $U^M = \{U_1, U_2, \ldots, U_M\}$ can be communicated reliably over multiple independent channels denoted $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1) \ldots (\mathcal{X}_M, p(y_M|x_M), \mathcal{Y}_M)$ if and only if

$$H(U(S)|U(S^c)) < \sum_{i\in S} I(X_i;Y_i), \tag{3.11}$$

for all subsets $S \subseteq \{1, 2, \ldots, M\}$ and $U(S) = \{U_j : j \in S\}$.

Proof: The converse can be proved exactly as in Theorem 3.1, this time dealing with $2^M - 1$ inequalities. To prove the forward part of the theorem consider the Slepian-Wolf region of achievable rates for multiple sources, given by

$$R(S) > H(U(S)|U(S^c)), \tag{3.12}$$

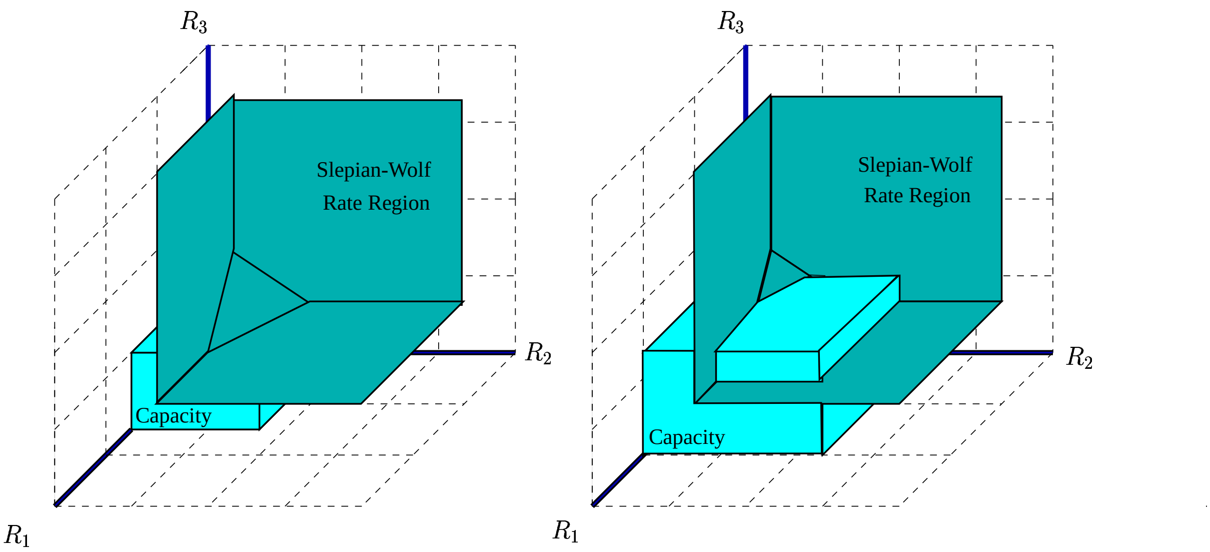


Figure 3.8: A sketch of the regions involved in Theorem 3.5 for $M = 3$ sources. Once again, when the capacity region of the independent reachback channels does not intersect the Slepian-Wolf rate region (left plot) reliable communication is not possible. If the two regions do intersect (right plot) than all points in the intersection are achievable.

for all $S \subseteq \{1, 2, \ldots, M\}$ where $R(S) = \sum_{i \in S} R_i$ and $U(S) = \{U_j : j \in S\}$. If the conditions in the theorem are fulfilled for all subsets $S \subseteq \{1, 2, \ldots, M\}$, then the Slepian-Wolf region defined by (3.12) intersects the capacity region given by $R_i < I(X_i; Y_i)$, $i = 1 \ldots M$, which means that the compressed source blocks can be transmitted over the array of M independent channels with arbitrarily small probability of error by adding separate channel codes. ■

Generalizing Theorem 3.2 in a similar way, we obtain the following result for correlated codes.

Theorem 3.6 (Barros, Servetto [11]) A set of correlated sources $\{U_1, U_2, \ldots, U_M\}$ can be communicated reliably over independent channels $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1) \ldots (\mathcal{X}_M, p(y_M|x_M), \mathcal{Y}_M)$ with correlated codes if

$$H(U(S)|U(S^c)) < \sum_{i \in S} I(X_i; Y_i|U(S^c)),$$

for all subsets $S \subseteq \{1, 2, \ldots, M\}$.

Proof: The proof is similar to the proof of Theorem 3.2, using the more general version of the theorem by Cover, El Gamal and Salehi for $M > 2$ sources [26]. ■

3.4.2 $M \geq 2$ Cooperating Nodes

The sensor reachback problem with $M \geq 2$ cooperating nodes is a network problem in which M encoders observe M correlated sources, then exchange messages over an interconnection network of limited capacity, and finally send the information to a far receiver over an array of

M independent channels, as illustrated in *Figure 3.9*.

We define a general *network conference* as follows. At each time k, encoder i chooses a message V_{ij} for each of the other $M-1$ encoders in the network. The $M-1$ messages generated by encoder i at time k may depend on the observed source sequence U_1^N and all previously received messages from all other nodes. Hence, the network conference is specified by $K(M-1)$ functions

$$h_{ij}(k) : \mathcal{U}_i^N \times \bigotimes_{m \in \mathcal{M} \setminus \{i\}} \bigotimes_{l=1}^{k-1} \mathcal{V}_{mi}(l) \to \mathcal{V}_{ij}(k),$$

where K is the number of conferencing steps, $\mathcal{M} = \{1, \ldots, M\}$ and $\bigotimes$ denotes the cartesian product operator. The set of conference rates that must be supported by the network for each pair of nodes is thus given by

$$R_{ij} = (1/K) \sum_{k=1}^{K} \log_2 |\mathcal{V}_{ij}(k)|,$$

with $i, j = 1, \ldots, M$. As in the two source case, a conference is said to be admissible if and only if

$$KR_{ij} \leq NC_{ij} \qquad \text{for all } (i,j) \text{ pairs.}$$

Notice that this definition does *not* imply that there is a direct connection between every node i and every node j in the network. The correct interpretation of the previous conditions is that the network is capable of supporting a communication rate between nodes i and j up to the given capacity C_{ij} — how this communication actually takes place (topology control, routing, etc.) is of no concern for the problem at hand.

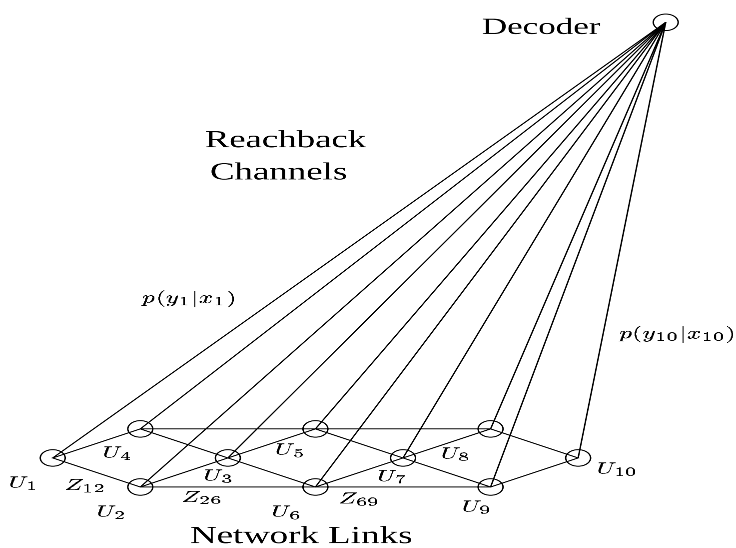


Figure 3.9: Reachback communication with $M \geq 2$ partially cooperating nodes.

The reachback *encoders* are functions

$$f_i : \mathcal{U}_i^N \times \bigotimes_{m\in\mathcal{M}\setminus\{i\}} \bigotimes_{l=1}^{K} \mathcal{V}_{mi}(l) \ \rightarrow \ \mathcal{X}_i^N,$$

that map the observed source vector U_i^N and the $K(M-1)$ messages received by encoder i over the network to a single codeword X_i^N that is sent to the central receiver. These functions map a block of N source symbols observed by each encoder, and a set of blocks of K messages received from the other encoders, to a block of N channel inputs. Finally, the *decoder* is a function

$$g : \mathcal{Y}_1^N \times \mathcal{Y}_2^N \times \cdots \times \mathcal{Y}_M^N \rightarrow \hat{\mathcal{U}}_1^N \times \hat{\mathcal{U}}_2^N \times \cdots \times \hat{\mathcal{U}}_M^N$$

that maps M blocks of channel outputs (one from each channel) into M blocks of reconstructed source sequences.

Based on the previous definitions we can generalize Theorem 3.3 and get the following necessary and sufficient conditions for reliable communication.

Theorem 3.7 (Barros, Servetto [10]) A set of correlated sources $U^M = \{U_1U_2\dots U_M\}$ can be communicated reliably with partially cooperating encoders over multiple independent channels $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1)\dots(\mathcal{X}_M, p(y_M|x_M), \mathcal{Y}_M)$, if and only if there exist random variables Z_{ij} with $i,j = 1,\dots,M$ and $i \neq j$, such that (a) Z_{ij} depend arbitrarily on $U_1\dots U_M$, (b) the maximum rates provided by the network are not exceeded, i.e.

$$I(U_1\dots U_{j-1}U_{j+1}\dots U_M; Z_{ij}|U_j) < C_{ij},$$

and (c) the conditions

$$H(U(S)|U(S^c)) < \sum_{i\in S} I(X_i;Y_i) + I(U(S); Z(S^c)|U(S^c)),$$

hold true for all subsets $S \subseteq \{1,2,\dots,M\}$, where $U(S) = \{U_j : j \in S\}$ and $Z(S) = \{Z_{ij} : i \in S \text{ or } j \in S\}$.

Proof: The proof is very similar to Theorem 3.3. We start with the achievability part. First, the encoders establish a conference over the network. We define M^2 auxiliary random variables V_{ij}, $i,j = 1,\dots,M$, which correspond to the messages sent by encoder i to encoder j at the rate C_{ij} supported by the network. Since no node sends a message to itself we set $V_{ii} = 1$ for all $i = 1,\dots,M$.

Upon observing a block u_i^N of source outputs, encoder i determines a set of message values v_{ij}, $j = 1,\dots,M$ for each observed value u_i. Said messages are coarse versions of the observed sample, as explained in the proof of Theorem 3.3. Using the conference mechanism encoder i can send a block v_{ij}^N to encoder j at rate $R_{ij} < C_{ij}$. We will now show that the rates $R_{ij} = H(V_{ij}|U_j)$ are sufficient for a block of messages V_{ij}^N to be transmitted reliably. Since (U_iU_j) are random and V_{ij} is a function of U_i, V_{ij} is random as well. Nodes i and j know the joint distribution $p(u_iu_jv_{ij})$, from which they can obtain the marginal $p(u_jv_{ij})$. Since this distribution can be viewed as a pair of correlated sources $(U_j V_{ij})$ and U_j^N is known at encoder j, it follows from the Slepian-Wolf theorem for $(U_j V_{ij})$ that V_{ij}^N can be compressed at rates $R_{ij} \geq H(V_{ij}|U_j)$ and still be reconstructed perfectly at encoder 2. Thus, using separate source and channel coding, V_{ij}^N can be transmitted over the network at rates

$$R_{ij} = H(V_{ij}|U_j) \quad < \quad C_{ij} \tag{3.13}$$

with arbitrarily small probability of error. Since V_{ij} is a function of U_i, we can write (3.13) as

$$\begin{aligned} C_{ij} \quad &> \quad H(V_{ij}|U_j) - H(V_{ij}|U_1 \ldots U_M) \\ &> \quad I(U_1 \ldots U_{j-1} U_{j+1} \ldots U_M; V_{ij}|U_j) \\ &> \quad I(U_1 \ldots U_{j-1} U_{j+1} \ldots U_M; Z_{ij}|U_j), \end{aligned}$$

where we set $Z_{ij} = V_{ij}$, thus defining the auxiliary random variables declared in the statement of the theorem[6].

After conferencing the encoders compress the source blocks and the received/sent message blocks using Slepian-Wolf codes. Let

$$Z_{i\mathcal{M}} = \left\{ Z_{kl} : k = i \text{ and } l \in \mathcal{M} \right\}$$

be the messages sent by node i and let

$$Z_{\mathcal{M}i} = \left\{ Z_{kl} : k \in \mathcal{M} \text{ and } l = i \right\}$$

be the messages received by node i. Moreover, we define $U_i' = (U_i Z_{i\mathcal{M}} Z_{\mathcal{M}i})$ with $i = 1, \ldots, M$, such that $p(u_1' \ldots u_M') = p(u_1 \ldots u_M z_{11} \ldots z_{MM})$. The M-source version of the Slepian-Wolf theorem [28, Theorem 14.4.2] guarantees that the rates

$$\begin{aligned} R(S) \quad &> \quad H(U'(S)|U'(S^c)) && (3.14) \\ &= \quad H(U(S)Z(S)|U(S^c)Z(S^c)) \\ &= \quad H(U(S)|U(S^c)Z(S^c)) \\ &= \quad H(U(S)|U(S^c)) - I(U(S); Z(S^c)|U(S^c)) && (3.15) \end{aligned}$$

for all subsets $S \subseteq \{1, 2, \ldots, M\}$ with $R(S) = \sum_{i \in S} R_i$, $U'(S) = \{U_k : k \in S\}$, $U(S) = \{U_j : j \in S\}$, and $Z(S) = \{Z_{ij} : i \in S \text{ or } j \in S\}$, are achievable. If conditions

$$H(U(S)|U(S^c)) < \sum_{i \in S} I(X_i; Y_i) + I(U(S); Z(S^c)|U(S^c)),$$

are fulfilled for all subsets $S \subseteq \{1, 2, \ldots, M\}$, then the Slepian-Wolf region defined by (3.15) intersects the capacity region given by $R_i < I(X_i; Y_i)$, $i = 1, 2, \ldots, M$, which is true whenever the conditions in the theorem are fulfilled, then the compressed source blocks can be transmitted

[6]Although we could continue the proof using V_{ij} as auxiliary random variables, this would lead to some confusion with respect to the converse part of the proof.

over the array of M independent channels with arbitrarily small probability of error by adding separate channel codes.

The converse part is similar to the converse proof of Theorem 3.3 with $2^M - 1$ inequalities. Details are provided in Appendix A.3. ■

3.4.3 Examples

With a few concrete examples, we illustrate the usefulness of Theorems 3.5 and 3.7.

A) Reachback Communication over Gaussian Channels with Orthogonal Multiple Access

The capacity of the Gaussian multiple access channel with M independent sources is given by

$$\sum_{i=1}^{m} R_i \leq \frac{1}{2}\log\left(1 + \frac{\sum_{i=1}^{m} P_i}{\sigma^2}\right),$$

for all $m \in \{1, 2, \ldots, M\}$, where σ^2 and P_i are the noise power and the power of the i-th user respectively [28, pp. 378-379]. If we use orthogonal accessing (e.g. TDMA), and assign different time slots to each of the transmitters, then the Gaussian multiple access channel is reduced to an array of M independent single-user Gaussian channels with capacity

$$C_i = \tau_i \cdot \frac{1}{2}\log\left(1 + \frac{P_i}{N\tau_i}\right), \qquad 1 \leq i \leq M,$$

where τ_i is the time fraction allocated to source user i.

Applying Theorem 3.5, we obtain the reachback capacity of the Gaussian channel with orthogonal accessing[7]. Reliable communication is possible if and only if

$$H(U(S)|U(S^c)) \leq \sum_{i \in S} \tau_i \cdot \frac{1}{2}\log\left(1 + \frac{P_i}{N\tau_i}\right),$$

for all subsets $S \subseteq \{1, 2, \ldots, M\}$.

B) Reachback Networks with Different Topologies

Theorem 3.7 can be applied to a great variety of network topologies, depending on the capacities of the links between nodes. Three such examples are shown in *Figure 3.10*.

Three interesting cases are illustrated in *Figure 3.10*:

- In (a) the sensor nodes build a linear array and conference communication occurs along the vertical axis. To obtain the reachback capacity we must set all capacities to zero, except those corresponding to the reachback channels $C_1 \ldots C_4$ and those corresponding to the active conference links C_{12}, C_{21}, C_{23}, C_{32}, C_{34} ,C_{43}. The auxiliary random variables become Z_{12}, Z_{23} and Z_{34} and the application of Theorem 3.7 follows easily.

- In (b) not all sensor nodes are connected to the remote location, and hence the information needs to be relayed to those nodes which do have a reachback channel available. Here,

[7] The generalization of Theorem 3.5 for channels with real-valued output alphabets can be easily obtained using the techniques in [28, Section 9.2 and Chapter 10].

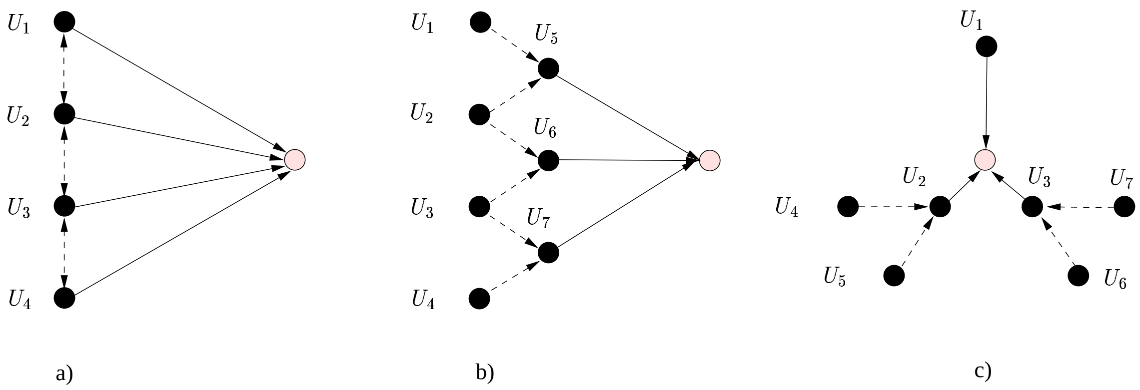


Figure 3.10: Three sensor network topologies for which Theorem 3.7 gives the reachback capacity. Dark circles denote the sensor nodes and a light circle represents the remote receiver. Conference channels and reachback channels are depicted as dashed and solid arrows, respectively, with the arrow head indicating the direction of the flow of information. Conferencing is done before transmission over the reachback channels.

we must consider the capacities $C_5 \dots C_7$ for reachback, and $C_{15}, C_{25}, C_{26}, C_{36}, C_{37}, C_{47}$ for conferencing, such that the auxiliary random variables become Z_{15}, Z_{25}, Z_{26}, Z_{36}, Z_{37}, Z_{47}.

- Finally, (c) shows a tree topology, in which each parent node has a reachback channel available ($C_5 \dots C_3$), and the children can relay their information over conference channels with capacities C_{42}, C_{52}, C_{63}, C_{73}. To compute the reachback capacity we must define the auxiliary random variables Z_{24}, Z_{25}, Z_{36}, Z_{37}. Once again, the application of Theorem 3.7 is straightforward.

In all cases, it is interesting to see how reliable communication is possible *if and only if* the network of interconnections among encoders has enough capacity to redistribute sensed data, so as to match the amount of data to upload to the capacity of the channels from each node to the far receiver, and further expanding on the routing interpretation developed in Section 3.3.5.

3.5 On the Separation of Source and Channel Coding in Communication Networks

3.5.1 Separation is Optimal for the Sensor Reachback Problem

In the context of point to point communication, given a source U (from a finite alphabet and satisfying the AEP) and a channel of capacity C, it is well known that the condition $H(U) < C$ is both necessary and sufficient for sending the source over the channel with arbitrarily small probability of error [28, Chapter 8.13]. From this result, commonly known as the *separation* principle, it follows that there is nothing to lose in using a two-stage encoder, which first compresses the source to its most efficient representation (at a rate close to $H(U)$), and then separately adds channel codes which can deal with the errors caused by the channel.

In the context of the sensor reachback problem with non-cooperating nodes, the proof of Theorem 3.5 gives necessary and sufficient conditions for reliable communication that, after solving the capacity problems, can be written as:

$$H(U(S)|U(S^c)) < \sum_{i \in S} C_i,$$

for all subsets $S \subseteq \{1, 2, \ldots, M\}$. These conditions show that a generalization of the previous statement also holds for the transmission of multiple correlated sources $(U_1, U_2, \ldots, U_M)$ over independent channels of capacities $(C_1, C_2, \ldots, C_M)$ – there is nothing to lose by compressing the sources to their most efficient representation (Slepian-Wolf coding) and separately adding channel codes. Furthermore, for the sensor reachback problem with partially cooperating encoders, Theorem 3.3 also shows that an optimal coding strategy for this problem consists once again of a cascade of a cooperative version of Slepian-Wolf source codes, followed by classical channel codes.

Therefore, these results identify an important class of non-trivial communication networks and information theory problems, in which the classical notion of separation between sources and channels holds.

3.5.2 Is Separation Relevant for Networks?

Based on the observations above it is only natural that we revisit the issue of optimality of separate source and channel coding in communication networks. This question is certainly not trivial, and we are not yet in a position to provide a definite answer. However, we feel it is only appropriate to discuss some of the intuition we derive about this most relevant issue from the results presented in this chapter.

We observe first that both in the converse proof of the separation theorem, as well as in the converse proofs for the different instances of the sensor reachback problem addressed in this chapter, the key ingredient that renders separation optimal is the data processing inequality [28, Section 2.8]. Application of this inequality requires Markov dependencies among random variables used to model sources and channel inputs/outputs. And as shown in this work, this property arises not only in point-to-point problems, but also in various non-trivial networks. Now, it is well known that this Markov property does not hold for a general multiple access channel with correlated sources, as established by the simple example of a binary adder channel and two binary sources with joint probability $(1/3, 1/3, 0, 1/3)$ in [26], and this has been the basis so far for arguing that separation does not hold in networks. However, after looking at *all* the evidence available, concluding from that simple example that the separation principle is not useful in the context of communication networks does appear to us to be too hasty a step:

- *Separation holds in other networks.* The team of Effros, Médard, Koetter, et al., showed that separation is optimal for a large class of networks [35], the crux being that all operations are carried out over a common finite field. A most remarkable aspect of their result is that, with this simple and natural restriction, it is shown in [35] that separation is optimal even for the example of the binary adder channel used in [26] to motivate the need for joint source/channel codes in networks. Also, in [67], Merhav and Shamai give an example of a point-to-point problem with side information for which separation holds. Also, in an unpublished manuscript that we recently became aware of, Yeung had also established some initial separation results in some simple networks [98]. So, for one network example where separation fails, there are multiple other network examples where separation holds, and one could easily argue more relevant, too.

- *The performance gap between a separation based approach and an unconstrained approach is not known in general*. In those cases for which separate source and channel coding has been found to be a suboptimal strategy, it seems pertinent to establish the extent to which separation leads to a loss of performance. To the best of our knowledge, no conclusive piece of evidence has been provided establish beyond a reasonable doubt the need for joint source/channel codes in a network setup.[8]

- *Reservation-based accessing schemes are common practice, and practical distributed source codes exist*. Because of their simplicity and low complexity, multiple access schemes that deal with the interference issue by dividing the medium into independent channels are widely used in many communication networks, and are of particular appeal for highly resource constrained nodes in sensor networks [6]. One important advantage is the reduction of the very challenging multi-user channel coding problem (e.g. , [7] and [66]), to multiple instances of the point-to-point problem, which is well understood both in theory and in practice. Similarly, wireless sensor networks have led a considerable amount of research in the area of distributed source coding, yielding practical schemes that come close to the theoretical limits obtained by Slepian and Wolf [85], and Wyner and Ziv [95] (e.g. [1], [38], [69], [75], [81]). It is therefore interesting to know the ultimate performance limits of communications systems with distributed source coding and separate channel coding even in networks where the separation principle does not hold.

The separation of source and channel coding is one of the cornerstones of digital communications. By representing information in terms of bits, Shannon provided an architecture for point-to-point communication systems in which the task of data compression and the task of channel noise mitigation are carried out by separate modules of the system, without any performance degradation due to this split of tasks. In the context of communication networks, we argue that joint source/channel codes are not always the only viable approach and that, whenever possible we should take advantage of over 50 years of experience in the design of communication systems based on separate source and channel coding.

Even in networks for which the separation principle does not hold, separate design still gives us the practical advantages of a system with multiple reusable components. In a real application, performance is a most important factor, but is not the only one: for example, the ability to quickly assemble working systems out of off-the-shelf components might justify some performance loss in highly dynamic environments such as a battlefield, or a marketplace. Therefore,

[8]Gastpar and Vetterli have presented some preliminary work along these lines [39], related to a sensor networking application derived from Berger's CEO problem [20], and built on top of Berger's results on uncoded transmission [17, pg. 162]. Now, while that preliminary result does hint at a potentially large gap in the extreme case of *no* cooperation at all, in another extreme case of the same setup (*full* cooperation), the gap vanishes. Hence, it does seem to us that the work of [39] needs to be further developed (e.g. , to address *partial* cooperation), before valid inferences can be made about the utility of separation in that setup.

even for networks in which separation turns out to be suboptimal, it is still of great interest to know what are the performance limits when *enforcing* separation constraints.

We end this section quoting Ahlswede and Han on the issue of separation in networks, from an early paper on multiterminal source coding [4]:

> *Another way of coming closer to a real communication situation with our models consists of enforcing the separation principle (in spite of its suboptimality in an ideal situation) and investigating what can be done (also optimally) if source and channel coding are carried out separately*.

3.6 Summary and Conclusions

In this chapter we have considered the sensor reachback problem. We formulated this problem as one of communication of multiple correlated sources over an array of independent channels, with partial cooperation among encoders. We defined the notion of reachback capacity, and gave exact characterizations for this capacity in a number of scenarios (nodes cooperating or not, constraints on the encoders, and numbers of nodes in the network). Having found in all cases that a natural network generalization of the classical joint source/channel coding theorem holds, we revisited the issue of source/channel separation in networks, where we argued that it may be too soon to dismiss separation as irrelevant in networks.

4

The Multiterminal Source Coding Problem

It takes two to know one.

GREGORY BATESON

4.1 Introduction

In the previous chapter, we gave a complete solution for the sensor reachback problem with perfect reconstruction at the receiver, in other words we assumed that the sent messages were to be recovered with arbitrarily small probability of error. Having determined the set of necessary and sufficient conditions for reliable communication under this scenario, there is another fundamental question that naturally comes to mind: what if, in a given scenario, it turns out to be impossible to match the rates of the Slepian-Wolf encoders to the capacities of the channels, as required for arbitrarily small probability of error? In that case, the best we can hope for is to reconstruct an *approximation* to the original source message. Recall that in the point-to-point problem, this is equivalent to having a source U with entropy $H(U) > C$, where C is the capacity of the channel. As mentioned in Chapter 2, the answer to this question is provided by classical rate-distortion theory. Consequently, it is only natural for us to consider a rate-distortion version of the sensor reachback problem — in the case of non-cooperating encoders we encounter the well-known *multiterminal source coding* problem [18].

4.1.1 Problem Statement

Before establishing a precise connection between the sensor reachback problem and the multiterminal source coding problem, we start with a rigorous statement of the latter.

Assume two sources U_1 and U_2, which are drawn i.i.d. $\sim p(u_1 u_2)$ from two finite alphabets, denoted $\mathcal{U}_1$ and $\mathcal{U}_2$. The two sources are processed by two separate encoders and decoded jointly, as shown in *Figure 4.1*. We require the following definitions:

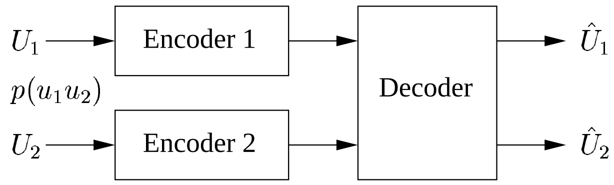


Figure 4.1: System setup for the multiterminal source coding problem.

1. Let the distortion measures be $d_1 : \mathcal{U}_1 \times \hat{\mathcal{U}}_1 \to \mathbb{R}^+$ and $d_2 : \mathcal{U}_2 \times \hat{\mathcal{U}}_2 \to \mathbb{R}^+$, where $\hat{U}_1$ and $\hat{U}_2$ are the reconstruction values with alphabets $\hat{\mathcal{U}}_1$ and $\hat{\mathcal{U}}_2$, respectively.

2. An $(N, 2^{NR_1}, 2^{NR_2}, \overline{D}_1, \overline{D}_2)$ code is defined by two encoding functions,

$$f_1 : \mathcal{U}_1^N \to \{1, 2, \ldots, 2^{NR_1}\},$$

$$f_2 : \mathcal{U}_2^N \to \{1, 2, \ldots, 2^{NR_2}\},$$

a decoding function,

$$g : \{1, 2, \ldots, 2^{NR_1}\} \times \{1, 2, \ldots, 2^{NR_2}\} \to \hat{\mathcal{U}}_1^N \times \hat{\mathcal{U}}_2^N$$

and an average distortion pair

$$\overline{D}_1 = \mathrm{E}\Big[\frac{1}{N}\sum_{i=1}^{N} d_1(U_{1i}, \hat{U}_{1i})\Big],$$

$$\overline{D}_2 = \mathrm{E}\Big[\frac{1}{N}\sum_{i=1}^{N} d_2(U_{2i}, \hat{U}_{2i})\Big],$$

where $(\hat{U}_1^N, \hat{U}_2^N) = g(f_1(U_1^N), f_2(U_2^N))$.

3. The rate-distortion tuple (R_1, R_2, D_1, D_2) is *achievable* if for any $\epsilon > 0$, for sufficiently large N an $(N, 2^{NR_1}, 2^{NR_2}, \overline{D}_1, \overline{D}_2)$ code exists such that $\overline{D}_1 \le D_1 + \epsilon$ and $\overline{D}_2 \le D_2 + \epsilon$.

4. The *rate-distortion region* $\mathcal{R}(D_1, D_2)$ of the two sources is the closure of the set of achievable rate and distortion tuples (R_1, R_2, D_1, D_2).

The goal of the problem is to give a complete characterization of the rate-distortion region $\mathcal{R}(D_1, D_2)$ in terms of single letter information-theoretic quantities.

4.1.2 A Converse for the Sensor Reachback Problem with Distortions

Having made a precise statement of the multiterminal source coding problem, we will now show how this formulation is a natural generalization of the sensor reachback problem, when the conditions for perfect reconstruction of the source messages at the receiver are not fulfilled. We recall once again that in the point-to-point case, discussed in detail in Section 2.1, if the entropy of the source U exceeds the capacity C of the channel, i.e. $H(U) > C$, the best we can hope for is to reconstruct an approximation of the source messages in the spirit of the rate-distortion theorem. Consider now the rate-distortion formulation of the sensor reachback problem with $M = 2$ sources illustrated in *Figure 4.2*. Let U_1 and U_2 be two correlated sources drawn i.i.d. $\sim p(u_1 u_2)$

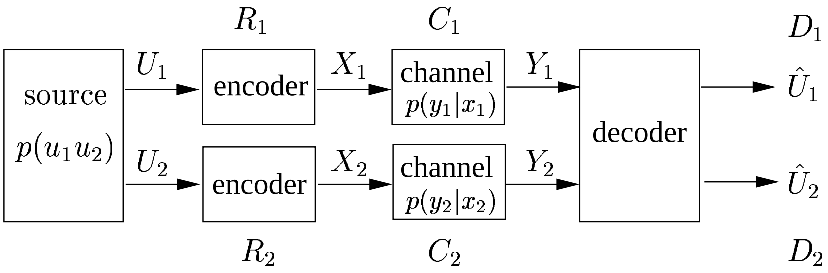


Figure 4.2: A rate-distortion version of the sensor reachback problem.

that are to be encoded separately at rates R_1 and R_2, and transmitted over two orthogonal channels $(\mathcal{X}_1, p(y_1|x_1), \mathcal{Y}_1)$ and $(\mathcal{X}_2, p(y_2|x_2), \mathcal{Y}_2)$ of capacities C_1 and C_2, respectively. The receiver is expected to reproduce U_1 with average distortion $\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} d_1(U_{1i}, \hat{U}_{1i})\right] \leq D_1$ and $\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} d_2(U_{2i}, \hat{U}_{2i})\right] \leq D_2$. The following result, whose proof is given in the appendix, relates the rate-distortion region of the sources with the capacities of the channels.

Theorem 4.1 (Barros) The sources U_1 and U_2 can be reconstructed at the receiver with distortions $\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} d_1(U_{1i}, \hat{U}_{1i})\right] \leq D_1$ and $\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} d_2(U_{2i}, \hat{U}_{2i})\right] \leq D_2$ if and only if the intersection between the capacity region $\{R_1 \leq C_1, R_2 \leq C_2\}$ and the rate-distortion region $\mathcal{R}(D_1, D_2)$ is non-empty.

Proof: See the appendix, Section A.4. ■

Given that the capacity region for two independent channels is trivial to obtain, we conclude from the previous theorem that solving the sensor reachback problem with distortions is equivalent to finding the rate-distortion region for the multiterminal source coding problem. From a coding point of view, *Theorem 4.1* establishes the optimality of separate source and channel coding for this problem — all achievable rate-distortion pairs can be obtained by cascading multiterminal source codes with capacity attaining channel codes. Consequently, we can ignore the channels and focus on the underlying source coding problem.

4.1.3 Previous Work

Separate encoding of correlated sources was first studied by Slepian and Wolf [85], who assumed no cooperation between encoders and solved the case of perfect reconstruction at the receiver, i.e. $D_1 = D_2 = 0$, as explained in Chapter 3. Shortly thereafter, Wyner and Ziv gave a solution for the case in which the receiver is provided with a perfect copy of U_2, i.e. $R_2 \geq H(U_2)$ and $D_2 = 0$ [95]. This result was further generalized by Berger and Yeung, to include all rates $R_2 < H(U_2)$ and $D_2 = 0$ [19].

To this date, the most significant contribution on the non-cooperative case with $D_1 \geq 0$ and $D_2 \geq 0$ stems from Berger and Tung's work ([18], [90]), who derived an inner and an outer bound for the rate-distortion region of generic (R_1, R_2, D_1, D_2) tuples. Both bounds are given in terms of *identical* mutual information expressions, involving two auxiliary random variables denoted as W_1 and W_2. The only difference between these two bounds is due to the fact that, while the outer bound assumes two Markov chain conditions $W_1 \rightarrow U_1 \rightarrow U_2$ and $U_1 \rightarrow U_2 \rightarrow W_2$ on W_1 and W_2, the inner bound requires them to obey an additional long chain condition of the form $W_1 \rightarrow U_1 \rightarrow U_2 \rightarrow W_2$. This latter condition poses a non-trivial restriction on the set of auxiliary random variables over which we can perform the minimization of the mutual information terms giving the boundaries of the rate-distortion region. Consequently, the inner bound thus obtained contains only a subset of the rate-distortion tuples suggested by the outer bound.

Due to the difficulty in solving the multiterminal source coding problem, subsequent contributions have focused on a few special cases that have practical relevance: Gaussian sources with squared distortions were considered by Oohama in [73], and a high-resolution version of the problem was studied by Zamir and Berger in [100]. Recent interest on the duality of source and channel coding in the context of multi-user communications has led several authors to investigate the relationship between the Berger-Tung inner rate-distortion region and the Marton capacity region for the broadcast channel (e.g. [91, 99, 76]).

With respect to user cooperation, in [53] Berger and Kaspi proposed a multiterminal source coding problem with partially cooperating encoders and gave a full characterization of the rate-distortion region for the case in which *one* of the two encoders can observe not only its source, but also the codeword generated by the other encoder. A similar setup, but with *both* encoders observing the codewords sent to the decoder, was used by Oohama to study the Slepian-Wolf problem [72] from a universal source coding perspective. Here, the encoders cooperate by processing each pair of input source blocks in multiple stages. In each stage, the codeword generated by each encoder is sent both to the decoder and the other encoder, such that the codewords at stage k depend not only on the source blocks but also on the $k-1$ previous pairs of codewords. For this setup, Oohama showed that the Slepian-Wolf rate region does not increase through cooperation.

Somewhat informally, we can state that in the previous form of cooperation, the encoders process the data and exchange information *at the same time*. A different kind of cooperation between encoders was proposed by Willems in [93], in which the encoders exchange information *before* they encode the data. The procedure is simple and conceptually pleasing: first the encoders establish a *conference* to exchange K messages and then they encode the data based on the messages exchanged and on the observed source blocks. This assumption is perfectly reasonable for sensor networks where a fusion center collects data at certain times, and the sensors can share information between transmissions [9]. It was this observation that led us to opt for this type of cooperation in the statement of the sensor reachback problem in Section 3.3, and the latter inspired us to formulate a cooperative source coding problem [15]. Recently, this form of cooperation was also used by Jaggi and Effros [50], to re-formulate the universal source coding problem proposed by Oohama [72], and show that a conference with asymptotically negligible rate is sufficient to guarantee the existence of universal Slepian-Wolf source codes.

4.1.4 Main Contributions

We offer several original contributions in this chapter. First, we give a simple proof of the Berger and Tung achievable rate-distortion region for multiterminal source coding. Based on a generalization of this proof we are able to obtain another inner bound, whose characterization does not require the long chain condition $W_1 \to U_1 \to U_2 \to W_2$, thus extending the family of random variables which can be considered as candidates for the minimization of the mutual informations defining the boundaries of the rate-distortion region. For this purpose we break the encoding process into two separate stages: a rate-distortion encoder (lossy compression), which guarantees the prescribed distortion, and random binning (lossless distributed compression), which removes the remaining correlation thus delivering the required rates for the prescribed distortions. This combined encoder can be viewed as the cascade of a vector quantizer and a Slepian-Wolf encoder, a coding strategy which makes perfect sense from a practical point of view and is therefore intuitively pleasing.

Then, we present a second inner bound for the sought rate-distortion region, whose characterization does *not* require the auxiliary random variables W_1 and W_2, which help describe the achievable rate-distortion tuples to obey the aforementioned long chain condition $W_1 \to U_1 \to U_2 \to W_2$. The coding strategy we use to prove this result is based on time-sharing of two complementary classes of rate-distortion codes, which can reach all points of the Berger-Yeung region, i.e. all rate-distortion tuples of the form $(R_1, R_2, D_1, 0)$ and, similarly, $(R_1, R_2, 0, D_2)$. Although this coding strategy does not achieve all the rate-distortion tuples (R_1, R_2, D_1, D_2) promised by the outer bound of Berger and Tung, the rate-distortion region we obtain does give the fundamental performance limits of a family of rate-distortion codes, which is likely to be simpler to implement than multiterminal rate-distortion codes and therefore relevant from a practical point of view[1]. A third bound is also obtained based on a specific collection of auxiliary random variables. In addition, we discuss the limitations of typical sequence decoding and the Markov lemma when seeking a solution for the multiterminal source coding problem, and give an achievable rate-distortion region for a special case of practical importance — correlated binary sources with Hamming distortions.

In the second part of this chapter, we formulate a general cooperative source coding problem with two distortion criteria, including a thorough definition of the conferencing mechanism. We first focus on the lossless case ($D_1 = D_2 = 0$), for which we give an exact characterization of the corresponding rate region. Our approach differs from the work of Jaggi and Effros [50, Thm. 4] in three important aspects: (1) in our case the encoders can exploit the joint statistics of the sources (we do not address universality issues), (2) we give a very simple proof based on a combination of deterministic binning and Slepian-Wolf random binning, and (3) our theorem delivers the classical Slepian-Wolf theorem for $R_{12} = R_{21} = 0$, whereas [50] does *not*: as the authors themselves explain, the universal encoders they use are not capable of learning the joint statistics of the sources unless they are allowed a minimal amount of cooperation.

Finally, we apply the same ideas to the multiterminal rate-distortion problem, where we obtain an inner and an outer bound for the cooperative rate-distortion tuples $(R_1, R_2, R_{12}, R_{12}, D_1, D_2)$ based on Berger and Tung's results for the classical multiterminal source coding

[1]Successful design of practical codes along these lines was recently presented in [96].

problem. In this case, we find that the rate expressions defining the regions for the two problems are identical, the only difference being given by the class of auxiliary random variables used to describe the encodings. But still we are able to prove that partial cooperation between encoders does in general lead to an increase of the achievable rate-distortion region.

4.1.5 Chapter Outline

The rest of this chapter is divided in three different parts. Section 4.2 discusses the classical version of the multiterminal source coding problem, providing a simple proof for the Berger-Tung rate-distortion region and two new inner bounds for this problem. The ideas developed in Section 4.2 are then applied to the binary case with Hamming distortion, addressed in Section 4.3. The third part, contained in Section 4.4, is dedicated entirely to a cooperative version of the multiterminal source coding problem, with and without perfect reconstruction at the receiver. The main results in this chapter are summarized in Section 4.5.

4.2 The Classical Multiterminal Source Coding Problem

A formal problem statement for the classical multiterminal source coding problem was given in the previous section. Based on this formulation, we will now discuss the best known rate-distortion bounds for this problem and prove some additional results.

4.2.1 A Simple Proof for the Berger-Tung Inner Bound

In [18] and [90], Berger and Tung include two theorems defining an inner and an outer bound for the rate-distortion region of the multiterminal source coding problem. The characterization of these bounds requires the use of two auxiliary random variables W_1 and W_2, which have special properties. In our own notation their inner bound can be written as follows..

Theorem 4.2 (Berger-Tung Inner Bound [18, 90]) Let (U_1, U_2) be drawn i.i.d. $\sim p(u_1 u_2)$. For a given distortion pair (D_1, D_2) an achievable rate region is given by

$$\begin{aligned} R_1 &\geq I(U_1U_2; W_1|W_2), &(4.1)\\ R_2 &\geq I(U_1U_2; W_2|W_1), &(4.2)\\ R_1 + R_2 &\geq I(U_1U_2; W_1W_2), &(4.3) \end{aligned}$$

if $W_1 \to U_1 \to U_2 \to W_2$ form a Markov chain and there exist $\hat{U}_1(W_1, W_2)$ and $\hat{U}_2(W_1, W_2)$, such that $D_1 \geq \mathrm{E}\left[d_1(U_1, \hat{U}_1)\right]$ and $D_2 \geq \mathrm{E}\left[d_2(U_2, \hat{U}_2)\right]$.

In the next section we will see that the mutual information expressions describing the Berger-Tung inner and outer bounds are identical. However, the long Markov chain $W_1 \to U_1 \to U_2 \to W_2$, which is only present in the description of the inner bound, allows for a few minor simplifications. Specifically, we can write

$$I(U_1U_2;W_1|W_2) = I(U_1;W_1|W_2) + I(U_2;W_1|W_2U_1) = I(U_1;W_1|W_2),$$

and, similarly, $I(U_1U_2;W_2|W_1) = I(U_2;W_2|W_1)$. A somewhat more intuitive form is given by the following identities:

$$I(U_1;W_1|W_2) = I(U_1;W_1) - I(W_1;W_2),$$

$$I(U_2;W_2|W_1) = I(U_2;W_2) - I(W_1;W_2),$$

$$I(U_1U_2;W_1W_2) = I(U_1;W_1) + I(U_2;W_2) - I(W_1;W_2),$$

where it becomes obvious that the minimum rates required for multiterminal source coding are lower than the rates for separate rate-distortion by the amount of mutual information between the auxiliary random variables W_1 and W_2, which are generally correlated.

The original proof of Berger and Tung [90] is based on strongly typical sequences and somewhat complicated combinatorial arguments. We will now give a simpler proof based on a two-stage encoder, which combines rate-distortion encoding with Slepian-Wolf style random binning, as shown in *Figure 4.3*.

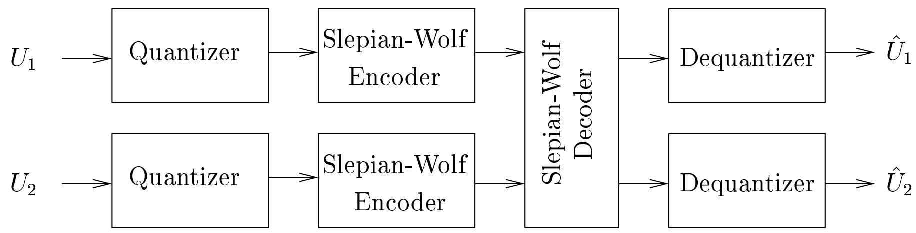


Figure 4.3: Coding strategy used in the proof of *Theorem 4.2*.

In a first step, each rate-distortion encoder (quantizer) takes blocks of source symbols and uses its codebook to generate quantization indices guaranteeing that the reconstructed sequences obey the prescribed average distortions. Subsequently the Slepian-Wolf encoders take blocks of quantization indices and perform lossless compression down to the achievable rates. On the decoding side, a perfect reconstruction of the quantization indices is given by the Slepian-Wolf decoder and the rate-distortion decoders (dequantizers) look up the codebook to produce the final output sequences. A similar coding strategy based on separate quantization and distributed compression was used by Wyner and Ziv to prove their result for the rate-distortion problem with side information [95]. More recently, Zamir and Berger [100] showed that for Gaussian sources and in the high resolution regime (i.e. very low distortions) said two step encoding strategy is optimal, i.e. separate quantizers and Slepian Wolf compression of the quantization indices yields an encoder which can operate at all achievables points of the rate-distortion region for this special case. Whether this is true for other sources and distortions remains to be shown. We proceed with our proof of the Berger and Tung inner bound.

Proof: First, we define the generation of the codebooks and the encoding and decoding pro-

cedures. Then, by analyzing the resulting distortion, we will show that, for an adequate choice of auxiliary random variables, there exists at least one $(N, 2^{NR_1}, 2^{NR_2}, \overline{D}_1, \overline{D}_2)$ rate-distortion code for any (R_1, R_2, D_1, D_2) which lies in the rate-distortion region given by *Theorem 4.2*. For this purpose, we will use the concepts and properties of strongly typical sequences as defined in Chapter 2.
We start this proof by fixing $p(w_1 w_2|u_1 u_2)$, such that

1. $p(u_1 u_2 w_1) = p(u_1 u_2)p(w_1|u_1)$ and $p(u_1, u_2, w_2) = p(u_1 u_2)p(w_2|u_2)$.
2. $I(U_1; W_1) \geq R_1(D_1)$ and $I(U_2; W_2) \geq R_2(D_2)$, where $R_1(\cdot)$ and $R_2(\cdot)$ denote the rate-distortion functions of U_1 and U_2, respectively.

Calculate $p(w_1) = \sum_{u_1 u_2 w_2} p(u_1 u_2)p(w_1 w_2|u_1 u_2)$ and $p(w_2) = \sum_{u_1 u_2 w_1} p(u_1 u_2)p(w_1 w_2|u_1 u_2)$. We will prove the existence of a rate-distortion code with rates R_1 and R_2 and distortions less than or equal to $D_1 + \delta_1$ and $D_2 + \delta_2$, with arbitrary $\delta_1 > 0$ and $\delta_2 > 0$.

Random code generation

Randomly generate a rate-distortion codebook $\mathcal{C}_1$ consisting of $2^{NI(U_1;W_1)}$ sequences W_1^N drawn i.i.d. $\sim \prod_{i=1}^N p(w_{1i})$. Index these codewords by $i \in \{1, 2, \ldots, 2^{NI(U_1;W_1)}\}$. Generate a second rate-distortion codebook $\mathcal{C}_2$ consisting of $2^{NI(U_2;W_2)}$ sequences W_2^n drawn i.i.d. $\sim \prod_{i=1}^N p(w_{2i})$. Index the codewords of the second codebook by $j \in \{1, 2, \ldots, 2^{NI(U_2;W_2)}\}$.

Encoding rule

Encoder 1 determines if there exists a i such that $(U_1^N, W_1^N(i)) \in T_\delta^N(U_1 W_1)$, the strongly typical set. If there is more than one such i, encoder 1 chooses the least. If there is no such i, encoder 1 sets $i = 1$. Similarly, encoder 2 determines if there exists a j such that $(U_2^N, W_2^N(j)) \in T_\delta^N(U_2 W_2)$. If there is more than one such j, encoder 1 chooses the least. If there is no such j, encoder 1 sets $j = 1$. Let g_1 and g_2 denote the assignment of blocks of samples U_1^N to index letters I and the assignment of blocks of samples U_2^N to index letters J, respectively.

Calculation of the probability distribution of the indices I and J

It follows from the encoding rule that each possible $(U_1^N, U_2^N) \in \mathcal{U}_1^N \times \mathcal{U}_2^N$ is assigned a unique pair of letters (i, j). Therefore, we can compute the probability distribution $p(ij)$ according to

$$p(ij) = \sum_{(U_1^N U_2^N):i=g_1(U_1^N)\wedge\ j=g_2(U_2^N)} p(U_1^N U_2^N),$$

where $p(U_1^N U_2^N) = \prod_{k=1}^N p(U_{1k} U_{2k})$. The decoder is informed about $p(ij)$.

Random binning of blocks of indices I^K and J^K

Independently assign every $i^K \in \mathcal{I}^K$ to one of 2^{KNR_1} bins according to a uniform distribution on $\{1, 2, \ldots, 2^{KNR_1}\}$. Index the bins by $b_1 \in \{1, 2, \ldots, 2^{KNR_1}\}$. Similarly, randomly assign every $j^K \in \mathcal{J}^K$ to one of 2^{KNR_2} bins indexed by $b_2 \in \{1, 2, \ldots, 2^{KNR_2}\}$. KNR_1 bits are sufficient to encode the bin index b_1 and, similarly, KNR_2 bits are sufficient to encode the bin index b_2. The bin assignments f_1 and f_2 are revealed to the decoder.

Encoding procedure

In a first step, encoder 1 takes K blocks U_1^N (denoted $\{U_1^N\}^K$) and applies the encoding rule to each of the blocks, thus obtaining a block of index letters $i^K = i_1, \ldots, i_K$. Subsequently, encoder 1 outputs the index b_1 of the bin which contains i^K. Similarly, encoder 2 takes K blocks U_2^N (denoted $\{U_2^N\}^K$) and applies the encoding rule to each of the blocks, thus obtaining a block of index letters $j^K = j_1, \ldots, j_K$. Finally, encoder 2 sends the index b_2 of the bin which contains j^K.

Decoding

Assume that the decoder receives the index pair $(\hat{b}_1, \hat{b}_2)$. If there is one and only one pair of jointly typical sequences (i^K, j^K), such that $f_1(i^K) = \hat{b}_1$, $f_2(j^K) = \hat{b}_2$, and $(i^K, j^K) \in T_\epsilon^K(IJ)$, the K reproduced sequences are $\{w_1^N\}^K = \{w_1^N(i_1), \ldots, w_1^N(i_K)\}$ and $\{w_2^N\}^K = \{w_2^N(j_1), \ldots, w_2^N(j_K)\}$. Otherwise choose $(\hat{\imath}^K, \hat{\jmath}^K) = (i^K, j^K)$ randomly and generate the reproduced sequences accordingly. Notice that in order to determine the typical set $T_\epsilon^K(IJ)$, the decoder must know $p(ij)$, which must be shared a priori.

Calculation of distortion

We set $\hat{U}_i^N = W_i^N$, $i = 1, 2$, and calculate the expected distortions over the random choice of codebooks $\mathcal{C}_1$ and $\mathcal{C}_2$ as given by $\overline{D_1} = E_{U_1^N, \mathcal{C}_1}\{d(U_1^N, \hat{U}_1^N)\}$ and $\overline{D_2} = E_{U_2^N, \mathcal{C}_2}\{d(U_2^N, \hat{U}_2^N)\}$. For a fixed codebook $\mathcal{C}_1$ and $\epsilon_1 > 0$ we divide the sequences $u_1^N \in \mathcal{U}_1^N$ into two groups:

- Sequences u_1^N such that $\hat{u}_1^N$ is strongly typical with u_1^N, i.e. $d_1(u_1^N, \hat{u}_1^N) < D_1 + \epsilon_1$. Since the total probability of these sequences is at most 1, these sequences contribute at most $D_1 + \epsilon_1$ to the expected distortion.

- Sequences u_1^N such that $\hat{u}_1^N$ is not strongly typical with u_1^N. Let P_1^e be the total probability of these sequences. Since the distortion for any individual sequence is bounded by $d'_{\max}$, these sequences contribute at most $P_1^e d'_{\max}$ to the expected distortion.

Similarly, for a fixed codebook $\mathcal{C}_2$ and $\epsilon_2 > 0$ we divide the sequences $u_2^N \in \mathcal{U}_2^N$ into two groups:

- Sequences u_2^N such that $\hat{u}_2^N$ is strongly typical with u_2^N, i.e. $d(u_2^N, \hat{u}_2^N) < D_2 + \epsilon_2$. These sequences contribute at most $D_2 + \epsilon_2$ to the expected distortion.

- Sequences u_2^N such that $\hat{u}_2^N$ is not strongly typical with u_2^N. Let P_2^e be the total probability of these sequences. As the distortion for any individual sequence is bounded by $d''_{\max}$, these sequences contribute at most $P_2^e d''_{\max}$ to the expected distortion.

Hence, the expected distortions are bounded by

$$\mathrm{E}\left[d(U_1^N, \hat{U}_1^N)\right] \leq D_1 + \epsilon_1 + P_1^e d'_{\max}$$

and

$$\mathrm{E}\left[d(U_2^N, \hat{U}_2^N)\right] \leq D_2 + \epsilon_2 + P_2^e d''_{\max},$$

which can be made less than $D_1 + \delta_1$ and $D_1 + \delta_2$, respectively, for appropriately chosen ϵ_1 and ϵ_2, and small enough P_1^e and P_2^e. It is thus sufficient to show that P_1^e and P_2^e are small, in order to prove that the expected distortions are close to D_1 and D_2.

Calculation of P_1^e and P_2^e

The probability P_1^e that the decoder produces a sequence $\hat{u}_1^N$ which is not strongly typical with u_1^N results from the union of the following events:

$$\begin{aligned}
E_0 &= \{(u_1^N, \hat{u}_1^N) \notin \mathcal{T}_\delta^N(U_1)\}, \\
E_1 &= \{(i^K, j^K) \notin \mathcal{T}_\epsilon^K(IJ)\}, \\
E_2 &= \{\exists i'^K \neq i^K : f_1(i'^K) = f_1(i^K) \text{ and } (i'^K, j^K) \in \mathcal{T}_\epsilon^K(IJ)\}, \\
E_3 &= \{\exists j^K \neq j'^K : f_2(j'^K) = f_2(j^K) \text{ and } (i^K, j'^K) \in \mathcal{T}_\epsilon^K(IJ)\}, \\
E_4 &= \{\exists (i'^K, j'^K) \neq (i^K, j^K) : f_1(i'^K) = f_1(i^K), f_2(j'^K) = f_2(j^K) \text{ and } (i'^K, j'^K) \in \mathcal{T}_\epsilon^K(IJ)\}.
\end{aligned}$$

Similarly, P_2^e results from the union of events E_1, E_2, E_3, E_4 and event $E_5 = \{(u_2^N, \hat{u}_2^N) \notin \mathcal{T}_\delta^N(U_1U_2)\}$. Notice that (U_1^N, U_2^N) are random and so are (i^K, j^K). By the union of events bound, we can write:

$$P_1^e = P(E_0 \cup E_1 \cup E_2 \cup E_3 \cup E_4) \leq P(E_0) + P(E_1 \cup E_2 \cup E_3 \cup E_4), \quad (4.4)$$

$$P_2^e = P(E_5 \cup E_1 \cup E_2 \cup E_3 \cup E_4) \leq P(E_5) + P(E_1 \cup E_2 \cup E_3 \cup E_4). \quad (4.5)$$

Now, since by definition $I(U_1; \hat{U}_1) \geq R_1(D_1)$ the rate-distortion theorem applies, and we can safely assume that when averaged over all randomly chosen codebooks $P(E_0)$ goes to zero for N sufficiently large (see [28, Section 13.5]) and thus there exists at least one codebook $\mathcal{C}_1$ for which $P(E_0)$ goes to zero for N sufficiently large. Using a similar argument, $I(U_2; \hat{U}_2) \geq R_2(D_2)$ implies that there exists at least one codebook $\mathcal{C}_2$ for which $P(E_5)$ also goes to zero for N sufficiently large.

Independently of the choice of codebooks, the probability of the remaining error events goes to zero if the conditions of the Slepian-Wolf theorem (see [28, Section 14.4]) are fulfilled, i.e. if

$$\begin{aligned}
NR_1 &\geq H(I|J), & (4.6) \\
NR_2 &\geq H(J|I), & (4.7) \\
N(R_1 + R_2) &\geq H(IJ). & (4.8)
\end{aligned}$$

Notice that R_1 and R_2 are given in bits per symbol and $H(IJ)$ is the joint entropy of I and J which correspond to blocks of N symbols. We also recall that by our own choice of decoding functions we have $\hat{U}_1 = \hat{U}_1(W_1, W_2) = W_1$ and $\hat{U}_2 = \hat{U}_2(W_1, W_2) = W_2$. To show that the rate-distortion region given by the theorem is achievable it suffices to prove that

$$NR_1 \geq NI(U_1;W_1|W_2) \geq H(I|J), \quad (4.9)$$

$$NR_2 \geq NI(U_2;W_2|W_1) \geq H(J|I), \quad (4.10)$$

$$N(R_1+R_2) \geq NI(U_1U_2;W_1W_2) \geq H(IJ). \quad (4.11)$$

We start with the last inequation. since the rate-distortion encodings are deterministic one-to-one mappings, we have

$$\begin{aligned}
H(IJ) &\leq H(IJW_1^N W_2^N) \\
&= H(W_1^N W_2^N) \\
&\overset{(a)}{=} H(W_1^N W_2^N) - H(W_1^N W_2^N|U_1^N U_2^N) \\
&\leq NH(W_1W_2) - (H(W_1^N|U_1^N U_2^N) + H(W_2^N|U_1^N U_2^N W_1^N)) \\
&\overset{(b)}{=} NH(W_1W_2) - (H(W_1^N|U_1^N) + H(W_2^N|U_2^N)) \\
&\overset{(c)}{=} NH(W_1W_2) - NH(W_1|U_1) - NH(W_2|U_2) \\
&\overset{(d)}{=} N(H(W_1W_2) - H(W_1|U_1U_2) - H(W_2|U_2U_1W_1) \\
&= NI(U_1U_2;W_1W_2),
\end{aligned}$$

where use the following arguments:

(a) follows from the fact that W_1^N and W_2^N are functions of U_1^N and U_2^N,

(b) follows from the long Markov chain on blocks $W_1^N \to U_1^N \to U_2^N \to W_2^N$,

(c) follows from the fact that W_1^N and W_2^N are drawn i.i.d. from the conditional probability distributions $p(w_1|u_1)$ and $p(w_2|u_2)$,

(b) follows from the long Markov chain on single letters $W_1 \to U_1 \to U_2 \to W_2$.

Taking (4.9) we use similar arguments to obtain

$$\begin{aligned}
H(I|J) &= H(IJ) - H(J) \\
&\leq NI(U_1U_2;W_1W_2) - H(J) \\
&= NI(U_1U_2;W_1W_2) - H(W_2^N) \\
&= NI(U_1U_2;W_1W_2) - (H(W_2^N) - H(W_2^N|U_2^N)) \\
&= NI(U_1U_2;W_1W_2) - I(U_2^N;W_2^N) \\
&= NI(U_1U_2;W_1W_2) - NI(U_2;W_2) \\
&= N(I(U_2;W_2) + I(U_1;W_2|U_2) + I(U_1;W_1|W_2) + I(U_2;W_1|W_2U_1) - I(U_2;W_2)) \\
&= NI(U_1;W_1|W_2),
\end{aligned}$$

and, similarly, $H(J|I) \leq NI(U_2;W_2|W_1)$, thus concluding the proof. ■

Remark 2 It is worth noting that the conditions $I(U_1;W_1) \geq R_1(D_1)$ and $I(U_2;W_2) \geq R_2(D_2)$, where $R_1(D_1)$ and $R_2(D_2)$ denote the rate-distortion functions given by

$$R_1(D_1) = \min_{p(w_1|u_1):\sum_{(u_1,w_1)} p(u_1)p(w_1|u_1)d(u_1,w_1)\leq D_1} I(U_1;W_1),$$

$$R_2(D_2) = \min_{p(w_2|u_2):\sum_{(u_2,w_2)} p(u_2)p(w_2|u_2)d(u_2,w_2)\leq D_2} I(U_2;W_2),$$

are *not* in contradiction with the conditions in the theorem, i.e. we do not have an additional restriction on the choice of $p(w_1w_2|u_1u_2)$. Remember that to compute the rate-distortion functions $R_i(D_i)$, $i = 1, 2$, we must minimize the corresponding mutual information expression over the set $\mathcal{P}_{W_i|U_i}$ of all possible $p(w_i|u_i)$ that fulfill the corresponding condition on the average distortion. On the other hand, to obtain the boundaries of the multiterminal rate-distortion region, we must minimize the expressions on the right side of the three inequalities in *Theorem 4.2* by choosing the appropriate distribution $p(w_1w_2|u_1u_2)$ among all distributions that satisfy $W_1 \to U_1 \to U_2 \to W_2$. Since the set of distributions $p(w_1|u_1)$ and $p(w_2|u_2)$ that obey the conditions of *Theorem 4.2* cannot be larger than $\mathcal{P}_{W_1|U_1} \cup \mathcal{P}_{W_2|U_2}$, we conclude that any choice of $p(w_1w_2|u_1u_2)$ that maximizes the rate-distortion region defined by *Theorem 4.2* guarantees the conditions $I(U_1;W_1) \geq R_1(D_1)$ and $I(U_2;W_2) \geq R_2(D_2)$, as required in the proof.

Beyond its conceptual simplicity, the previous proof is interesting, because it clearly shows that all rate-distortion pairs in the Berger-Tung inner bound can be achieved by cascading classical vector quantizers with Slepian-Wolf entropy codes. This suggests an informal notion of duality between this problem and the sensor reachback problem, for which we have shown that cascading Slepian-Wolf codes and capacity attaining point-to-point channel codes is an optimal coding strategy.

4.2.2 The Berger-Tung Outer Bound

The outer bound of Berger and Tung is given by the following theorem.

Theorem 4.3 (Berger-Tung Outer Bound [18, 90]) Let (U_1, U_2) be drawn i.i.d. $\sim p(u_1u_2)$. For a given distortion pair (D_1, D_2) all achievable rates (R_1, R_2) must satisfy the following conditions:

$$R_1 \geq I(U_1U_2;W_1|W_2), \tag{4.12}$$

$$R_2 \geq I(U_1U_2;W_2|W_1), \tag{4.13}$$

$$R_1 + R_2 \geq I(U_1U_2;W_1W_2), \tag{4.14}$$

where W_1 and W_2 are two auxiliary random variables, such that $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$ form two Markov chains, and there exist $\hat{U}_1(W_1, W_2)$ and $\hat{U}_2(W_1, W_2)$, such that $D_1 \geq \mathrm{E}\left[d_1(U_1, \hat{U}_1)\right]$ and $D_2 \geq \mathrm{E}\left[d_2(U_2, \hat{U}_2)\right]$.

Proof: The proof uses Fano's inequality and standard information-theoretic arguments. For details, see [18]. ■

Notice that the expressions in the Berger-Tung inner bound and the outer bound are in fact identical, except for a minor simplification made possible by the long Markov chain condition in *Theorem 4.2*. Specifically, if $W_1 \to U_1 \to U_2 \to W_2$ form a Markov chain we can write

$$I(U_1U_2; W_1|W_2) = I(U_1; W_1|W_2) + I(U_2; W_1|W_2U_1) = I(U_1; W_1|W_2),$$

and, similarly, $I(U_1U_2; W_2|W_1) = I(U_2; W_2|W_1)$. Since the auxiliary random variables that characterize the outer bound obey the short chain conditions $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$ but not necessarily the long chain condition, it is not possible to carry out the same manipulation on the rate expressions of *Theorem 4.3*. The only difference between the two bounds lies thus in the class of probability distributions over which the optimization of the mutual information terms is carried out. The implications of this discord are discussed in the next section.

4.2.3 The Limitations of the Markov Lemma

The original proof of Berger and Tung for *Theorem 4.2* (see [90, 18]) is based on the *Markov Lemma*, which, as explained in Section 2.2.3, is valid for *strongly typical sequences*. For convenience, we restate this important result in the notation we adopted for the multiterminal source coding problem.

Lemma 4.4 (Markov Lemma) Let $(U_1U_2W_1)$ form a Markov chain $W_1 \to U_1 \to U_2$. If for a given $(u_1^N, u_2^N) \in \mathcal{T}_\delta^N(U_1U_2)$, W_1^N is drawn $\sim \prod_{i=1}^N p(w_{1i}|u_{1i})$, then $P\{(U_1^N, U_2^N, W_1^N) \in \mathcal{T}_\delta^N(U_1U_2W_1)\} \to 1$ as $n \to \infty$.

In more intuitive terms, the Markov property of $(U_1U_2W_2)$ implies that if we take two sequences u_1^N and u_2^N which are jointly typical and generate a third sequence w_1^N according to the conditional probability $p(w_1|u_1)$, then with high probability u_1^N, u_2^N and w_1^N are jointly typical. This in turn implies that with high probability u_2^N and w_2^N are jointly typical, a property which leads to elegant solutions for source coding problems with side information (e.g. [95] and [19]).

Unfortunately, in the case of multiterminal source coding the Markov lemma falls short of yielding a complete characterization of the rate-distortion region. By randomly generating the rate-distortion codewords W_1^N and W_2^N according to $\sim \prod_{i=1}^N p(w_{1i}|u_{1i})$ and $\sim \prod_{i=1}^N p(w_{2i}|u_{2i})$, respectively, and applying the Markov lemma to guarantee successful typical sequence decoding, Berger and Tung's coding strategy only works if the auxiliary random variables W_1 and W_2 obey the long Markov chain condition $W_1 \to U_1 \to U_2 \to W_2$, whereas the converse based on Fano's inequality admits all random variables W_1 and W_2 that obey the short chains $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$.

As Berger points out in [18], the former condition, which implies

$$p(w_1w_2|u_1u_2) = p(w_1|u_1)p(w_2|u_2)$$

is much weaker than the latter, because it does not admit joint conditional probability distributions in the form of a mixture of product conditionals, i.e.

$$p(w_1 w_2|u_1 u_2) = \sum_{k=1}^{K} \lambda_k p_k(w_1|u_1) p_k(w_2|u_2), \text{ with } \sum_{k=1}^{K} \lambda_k = 1.$$

Thus, by forcing the auxiliary random variables $(W_1 W_2)$ to form a *long* Markov chain with U_1 and U_2, the inner bound poses an additional restriction on the set of auxiliary random variables over which we can perform the search for the minimal values of the mutual information, which in turn yield the exact boundaries of the sought rate-distortion region.

4.2.4 An Inner Bound based on Time-sharing of Berger-Yeung Codes

It follows from the previous discussion that to solve the multiterminal source coding problem, one has either to (a) tighten the converse, (b) to prove the existence of codes which guarantee the achievability of the rate-distortion region defined by *Theorem 4.3* for an arbitrary choice of $(W_1 W_2)$ among all random variables which obey the short chain conditions, but not necessarily the long chain condition, or (c) prove that the two classes of probability distributions yield the same rate-distortion region. In the following, we will pursue strategy (b) and derive a new inner bound for the multiterminal source coding problem.

We recall that in [19] Berger and Yeung give a complete characterization for the rate-distortion region of one important particular case of our problem, namely $D_2 = 0$. Their main result can be formulated as follows.

Theorem 4.5 (Berger-Yeung [19]) Let $(U_1 U_2)$ be drawn i.i.d. $\sim p(u_1 u_2)$. For a given distortion pair $(D_1, 0)$ the rate pair (R_1, R_2) is achievable if and only if

$$\begin{aligned} R_1 &\geq I(U_1; W_1|U_2) \\ R_2 &\geq H(U_2|W_1) \\ R_1 + R_2 &\geq H(U_2) + I(U_1; W_1|U_2) \end{aligned} \tag{4.15}$$

where W_1 is an auxiliary random variable, which obeys the following conditions:

1. $W_1 \to U_1 \to U_2$ forms a Markov chain,
2. $\hat{U}_1(U_2, W_1)$ exists such that $\mathrm{E}\left[d_1(U_1, \hat{U}_1)\right] \leq D_1$, and
3. $|\mathcal{W}_1| \leq |\mathcal{U}_1| + 2$,

Naturally, their result is also valid for the symmetric case, i.e. $D_1 = 0$, leading to a second version of theorem.

Theorem 4.6 (Berger-Yeung, second version) Let $(U_1 U_2)$ be drawn i.i.d. $\sim p(u_1 u_2)$. For a given distortion pair $(0, D_2)$ the rate pair (R_1, R_2) is achievable if and only if

$$\begin{aligned} R_1 &\geq H(U_1|W_2) \\ R_2 &\geq I(U_2; W_2|U_1) \\ R_1 + R_2 &\geq H(U_1) + I(U_2; W_2|U_1) \end{aligned} \tag{4.16}$$

where W_2 is an auxiliary random variable, which obeys the following conditions:

1. $U_1 \to U_2 \to W_2$ forms a Markov chain,
2. $\hat{U}_2(U_1, W_2)$ exists such that $\mathrm{E}\left[d_2(U_2, \hat{U}_2)\right] \leq D_2$, and
3. $|\mathcal{W}_2| \leq |\mathcal{U}_2| + 2$.

Notice that the proof of achievability of *Theorem 4.5* guarantees the existence of $(N, 2^{NR_1}, 2^{NR_2}, \overline{D}_1, \overline{D}_2)$ codes, such that $\overline{D}_1 \leq D_1 + \epsilon$ and $\overline{D}_2 \leq \epsilon$ [19]. Similarly, *Theorem 4.6* implies the existence of $(N, 2^{NR_1}, 2^{NR_2}, \overline{D}_1, \overline{D}_2)$ codes, such that $\overline{D}_1 \leq \epsilon$ and $\overline{D}_2 \leq D_2 + \epsilon$.

A fact that is not mentioned in the original contribution of [19], is that the achievability part of the Berger-Yeung rate-distortion regions can be easily proved using the Berger-Tung inner bound in *Theorem 4.2*. For the first version, we can set $W_2 = U_2$ and $\hat{U}_2(W_1, W_2) = W_2 = U_2$ thus guaranteeing $\mathrm{E}\left[d_2(U_2, \hat{U}_2)\right] \leq \epsilon$. For the rate expressions we get

$$\begin{aligned} R_1 &\geq I(U_1U_2; W_1|W_2) = I(U_1; W_1|W_2) = I(U_1; W_1|U_2) \\ R_2 &\geq I(U_1U_2; W_2|W_1) = (U_2; W_2|W_1) = I(U_2; U_2|W_1) = H(U_2|W_1) \\ R_1 + R_2 &\geq I(U_1U_2; W_1W_2) \\ &= I(U_2; W_1W_2) + I(U_1; W_1W_2|U_2) \\ &= I(U_2; W_1U_2) + I(U_1; W_1U_2|U_2) \\ &= H(U_2) + I(U_1; W_1|U_2). \end{aligned}$$

The achievability proof for the second version follows analogously by setting $W_1 = U_1$ and $\hat{U}_1(W_1, W_2) = W_1 = U_1$.

The most salient aspect of Berger-Yeung codes is that they require only one short Markov chain to work. Based on this property, we will now show how a time-sharing combination of the family of Berger-Yeung codes that give all points of the form $(R_1, R_2, D_1, 0)$ with the second family of Berger-Yeung codes that achieve all points of the form $(R_1, R_2, 0, D_2)$, enables us to construct a family of codes for the multiterminal rate-distortion problem, whose region of achievable rates for arbitrary (D_1, D_2) depends on two auxiliary random variables which obey the two short chain conditions, but not necessarily the long chain condition.

Theorem 4.7 (Barros, Servetto [13]) Let (U_1U_2) be drawn i.i.d. $\sim p(u_1u_2)$. For a given distortion pair (D_1, D_2) the rate pair (R_1, R_2) is achievable if

$$\begin{aligned} R_1 &\geq I(U_1; U_2 W_1 | W_2) && (4.17)\\ R_2 &\geq I(U_2; U_1 W_2 | W_1) && (4.18)\\ R_1 + R_2 &\geq H(U_1 U_2) - H(U_1 | U_2 W_1) - H(U_2 | U_1 W_1) && (4.19) \end{aligned}$$

where W_1 and W_2 are two auxiliary random variables, which obey the following conditions:

1. $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$ form two Markov chains,

2. $|\mathcal{W}_1| \leq |\mathcal{U}_1| + 2$ and $|\mathcal{W}_2| \leq |\mathcal{U}_2| + 2$, and there exist $\hat{U}_1(U_2, W_1)$ and $\hat{U}_2(U_1, W_2)$, such that

3. $D_1 \geq \mathrm{E}\left[d_1(U_1, \hat{U}_1)\right]$ and $D_2 \geq \mathrm{E}\left[d_2(U_2, \hat{U}_2)\right]$.

Proof: Let C_1 and C_2 be two Berger-Yeung codes at different rates, such that $\mathbf{R}' = (R_1', R_2')$ yields $\mathbf{D}' = (0, D_2)$ and $\mathbf{R}'' = (R_1'', R_2'')$ yields $\mathbf{D}'' = (D_1, 0)$. If we use the first pair of codebooks for the first $\lambda\ N$ symbols and the second codebook for the last $(1-\lambda)N$ symbols, with $0 \leq \lambda \leq 1$, we can construct a new pair of codes, whose number of codewords is $2^{N(\lambda R_1' + (1-\lambda) R_1'')}$ for source U_1 and $2^{N(\lambda R_2' + (1-\lambda) R_2'')}$ for source U_2. Consequently, this new pair of codes will have rates $\lambda \mathbf{R}' + (1-\lambda)\mathbf{R}''$ and induce the distortions $\lambda \mathbf{D}' + (1-\lambda)\mathbf{D}''$, i.e. $(1-\lambda)D_1$ for source U and $\lambda\ D_2$ for source U_2.

Thus, we can state that the rate-distortion tuple $(R_1, R_2, D_1, D_2) = (\lambda R_1' + (1-\lambda) R_1'', \lambda R_2' + (1-\lambda) R_2'', (1-\lambda) D_1, \lambda\ D_2)$ is achievable if (R_1, R_2) obey the time-sharing combination of the two versions of the Berger-Yeung conditions, i.e.

$$\begin{aligned} R_1 &\geq \lambda H(U_1 | W_2') + (1-\lambda) I(U_1; W_1'' | U_2) && (4.20)\\ R_2 &\geq \lambda I(U_2; W_2' | U_1) + (1-\lambda) H(U_2 | W_1'') && (4.21)\\ R_1 + R_2 &\geq \lambda (H(U_1) + I(U_2; W_2' | U_1)) + (1-\lambda)(H(U_2) + I(U_1; W_1'' | U_2)), && (4.22) \end{aligned}$$

where W_2' and W_1'' are two auxiliary random variables, obeying the short chain conditions $W_1'' \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2'$, which correspond to the two families of Berger-Yeung codes. Furthermore we introduce a time-sharing random variable Q, which is independent of all other random variables.

Define $W_1 = (Q, W_{1Q})$, such that $W_1 = W_1' = U_1$ with probability λ and $W_1 = W_1''$ with probability $1-\lambda$. Similarly, define $W_2 = (Q, W_{2Q})$, such that $W_2 = W_2'$ with probability λ and $W_2 = W_2'' = U_2$ with probability $1-\lambda$. Notice that W_1 and W_2 also obey the short chain conditions $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$.

For the sum rate condition (4.22), we can write

$$\begin{aligned}
R_1 + R_2 &\geq \lambda(H(U_1) + I(U_2; W_2'|U_1)) + (1-\lambda)(H(U_2) + I(U_1; W_1''|U_2)) \\
&= H(U_1U_2) - (\lambda H(U_2|U_1W_2') + (1-\lambda)H(U_1|U_2W_1'')) \\
&= H(U_1U_2) - (\lambda H(U_2|U_1W_2') + (1-\lambda)H(U_2|U_1W_2'') \\
&\quad +\lambda H(U_1|U_2W_1') + (1-\lambda)H(U_1|U_2W_2'')) \qquad (4.23) \\
&= H(U_1U_2) - (H(U_2|U_1W_{2Q}, Q) + H(U_1|U_2W_{1Q}Q)) \\
&= H(U_1U_2) - H(U_2|U_1W_2) - H(U_1|U_2W_1),
\end{aligned}$$

where (4.23) stems from the fact that $H(U_1|U_2W_1') = H(U_1|U_2U_1) = 0$ and $H(U_2|U_1W_2'') = H(U_2|U_1U_2) = 0$. Applying similar steps to the condition in (4.20) we get

$$\begin{aligned}
R_1 &\geq \lambda H(U_1|W_2') + (1-\lambda)I(U_1; W_1''|U_2) \\
&= \lambda H(U_1|W_2') + (1-\lambda)H(U_1|W_2'') - (1-\lambda)H(U_1|W_2'') + \lambda I(U_1; W_1'|U_2) \\
&\quad -\lambda I(U_1; W_1'|U_2) + (1-\lambda)I(U_1; W_1''|U_2) \\
&= H(U_1|W_{2Q}Q) + I(U_1; W_{1Q}Q|U_2) - ((1-\lambda)H(U_1|W_2'') + \lambda I(U_1; W_1'|U_2)) \\
&= H(U_1|W_2) + I(U_1; W_1|U_2) - ((1-\lambda)H(U_1|U_2) + \lambda I(U_1; U_1|U_2)) \\
&= H(U_1|W_2) + I(U_1; W_1|U_2) - H(U_1|U_2) \\
&= H(U_1|W_2) + H(U_1|U_2) - H(U_1|U_2W_1) - H(U_1|U_2) \\
&= H(U_1|W_2) - H(U_1|U_2W_1W_2) \\
&= I(U_1; U_2W_1|W_2),
\end{aligned}$$

and, similarly, for the remaining inequation we get $R_2 \geq I(U_2; U_1W_2|W_1)$, thus completing the proof. ■

It is easy to prove that the first version of the Berger-Yeung result ($D_2 = 0$) can be obtained from *Theorem 4.7* by setting $W_2 = U_2$ and that, similarly, our result specializes to the second version ($D_1 = 0$) of the Berger-Yeung theorem $W_1 = U_1$. This is not surprising, as the code construction which achieves all the rate-distortion tuples in our inner bound, is in fact based on the families of codes which achieve all points in the two versions of the rate-distortion region of Berger and Yeung.

More interesting is a comparison between our result and the Berger-Tung outer bound. For this purpose, we must compare the inequalities (4.17)-(4.19) to the inequalities (4.12)-(4.14). Starting once again with the sum rate condition, we get

$$\begin{aligned}
H(U_1U_2) - H(U_2|U_1W_2) - H(U_1|U_2W_1) &= H(U_1U_2) - H(U_2|U_1W_1W_2) \\
&\quad -H(U_1|U_2W_1W_2) \qquad (4.24) \\
&\geq H(U_1U_2) - H(U_1U_2|W_1W_2) \\
&\geq I(U_1U_2; W_1W_2),
\end{aligned}$$

where (4.24) follows from the two short chain conditions $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$.

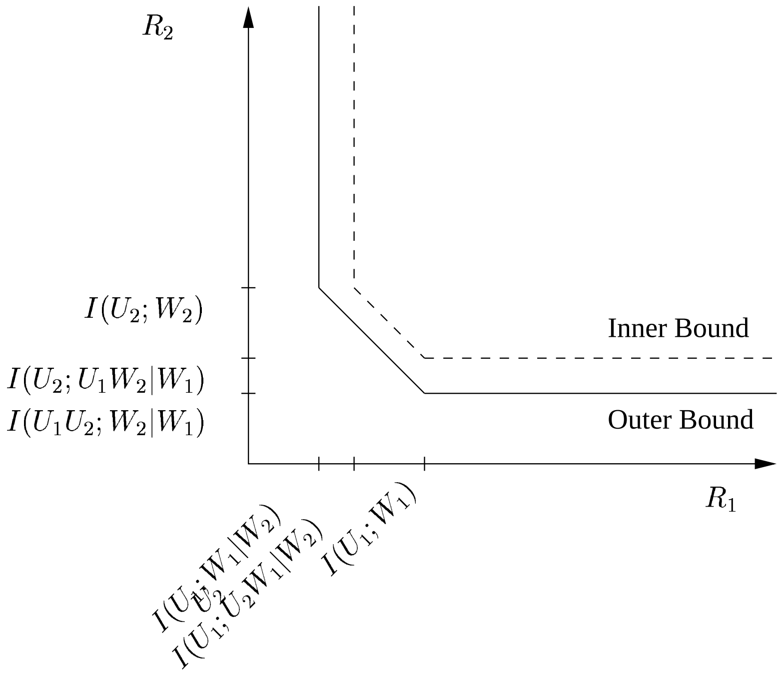


Figure 4.4: Our rate-distortion inner bound in comparison to the Berger-Tung outer bound for a fixed value of (D_1, D_2).

For the other two conditions, we get

$$\begin{aligned} I(U_1U_2;W_1|W_2) &= H(U_1U_2|W_2) - H(U_1U_2|W_1W_2) \\ &= H(U_1|W_2) + H(U_2|U_1W_2) - H(U_2|W_1W_2) - H(U_1|U_2W_1W_2) \\ &= H(U_1|W_2) - (H(U_1|U_2W_1W_2) + H(U_2|U_1W_2) - H(U_2|W_1W_2)) \\ &\leq H(U_1|W_2) - H(U_1|U_2W_1W_2) \qquad (4.25) \\ &\leq I(U_1;U_2W_1|W_2), \end{aligned}$$

and, similarly, $I(U_1U_2;W_2|W_1) \leq I(U_1;U_2W_2|W_1)$. Notice that (4.25) stems from the fact that $W_1 \to U_1 \to U_2$ forms a Markov chain and therefore $H(U_2|U_1W_2) - H(U_2|W_1W_2) \geq 0$. Clearly, our region of achievable rates is strictly contained in the Berger-Tung outer bound, as illustrated in *Figure 4.4*. As the two bounds are not tight, the problem still requires a different solution. Nevertheless, our rate-distortion inner bound does allow a direct comparison with the outer bound and a quantification of the distance between the two bounds, which the Berger-Tung inner bound clearly does not.

Finally, we point out that since both Berger-Yeung rate-distortion regions are contained in the Berger-Tung inner region and the latter is convex (see proof in the appendix, Section A.7), the new time-sharing inner bound is strictly contained in the Berger-Tung inner bound. Thus, we conclude that the gap between the Berger-Tung outer bound and our time-sharing inner bound, which we can quantify exactly, is an upper bound to the gap between the two Berger-Tung bounds, a gap which is difficult to compute due to the difference in the classes of probability distributions that define them.

4.2.5 An Alternative Inner Bound

Consider once again the proof of the Berger-Tung inner bound given by *Theorem 4.2*. The coding strategy based on separate quantization and distributed source coding is guaranteed to work if the Slepian-Wolf conditions (4.6)-(4.8) are fulfilled with respect to the quantization indices I and J. To prove that the conditions in *Theorem 4.2* are sufficient for (4.6)-(4.8) to be valid, we used the Markov lemma and showed that these conditions are indeed valid for all W_1 and W_2 that obey the long Markov chain condition $W_1 \to U_1 \to U_2 \to W_2$. Next, we will show that this coding strategy also works for a collection of auxiliary random variables that obey the two short chain conditions $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$ but not necessarily the long one. The result is given by the following theorem, which is proved in the appendix.

Theorem 4.8 (Barros, Servetto [12]) Let (U_1, U_2) be drawn i.i.d. $\sim p(u_1 u_2)$ and let (W_1, W_2) be a pair of random variables for which there exist functions $\hat{U}_1(W_1, W_2)$ and $\hat{U}_2(W_1, W_2)$ such that $D_1 \geq \mathrm{E}\left[d_1(U_1, \hat{U}_1)\right]$ and $D_2 \geq \mathrm{E}\left[d_2(U_2, \hat{U}_2)\right]$. Then, the pair of rates (R_1, R_2) is achievable with distortions (D_1, D_2) if

$$
\begin{aligned}
R_1 &\geq I(U_1 U_2; W_1 | W_2), &(4.26)\\
R_2 &\geq I(U_1 U_2; W_2 | W_1), &(4.27)\\
R_1 + R_2 &\geq I(U_1 U_2; W_1 W_2), &(4.28)
\end{aligned}
$$

for all the pairs (W_1, W_2) in a collection $\mathcal{S}$ having the property that $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$ form two Markov chains.

Notice that not only are expressions (4.26)-(4.28), which define the rate-distortion region, the same as in *Theorem 4.3*, but in contrast to *Theorem 4.2* (where the auxiliary random variables W_1 and W_2 are required to obey the long chain condition $W_1 \to U_1 \to U_2 \to W_2$, our auxiliary variables are only required to satisfy the two short chain conditions $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$. This theorem does not yet guarantee that all the (R_1, R_2, D_1, D_2) points promised by the converse *Theorem 4.3* and the outer bound are achievable: *Theorem 4.8* does *not* state that $\mathcal{S}$ contains *all* pairs satisfying a short chain condition, it only says that the pairs contained in there do satisfy that condition—whether all such pairs are in there or not remains to be seen. However, this result does suggest that the long chain condition $W_1 \to U_1 \to U_2 \to W_2$ is not a necessary condition: we prove that there exist *some* random variables (W_1, W_2) which, obeying only the short chain conditions $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$, still yield achievable rate-distortion points.

4.3 The Binary Case with Hamming Distortion

In this section, we want to illustrate the applicability of the previous ideas by considering a specific case of practical importance: separate encoding of correlated binary sources subject to the Hamming distortion criterion. Wyner and Ziv considered a similar problem setup in [95] and derived the rate-distortion function of a binary source with binary side information [95]. Interestingly enough, although Berger and Yeung suceeded at characterizing the rate-distortion region for the generic case in which one of two memoryless sources must be reconstructed almost perfectly [19], they were not able to specialize their result to the (apparently simpler) binary case with Hamming distortion. To the best of our knowledge, the binary case with two distortion criteria discussed in this section, was not addressed in previous publications on multiterminal source coding, inspite of its undeniable relevance.

4.3.1 Problem Statement

Let U_1 and U_2 be two correlated binary symmetric sources, such that $p(U_1 = 0) = p(U_2 = 0) = 1/2$ and $p(U_2 \neq U_1) = p_0$ with $0 < p_0 \leq 1/2$. U_1 and U_2 (or vice versa) can be viewed as the input and the output, respectively, of a binary symmetric channel with crossover probability p_0. Let the distortion measure be

$$d(u, \hat{u}) = \begin{cases} 0 & \text{if } u = \hat{u}, \\ 1 & \text{otherwise.} \end{cases}$$

For prescribed distortions $\mathrm{E}\left[\sum_{k=1}^{N} d(U_{1k}, \hat{U}_{1k})\right] \leq D_1 + \epsilon$ and $\mathrm{E}\left[\sum_{k=1}^{N} d(U_{2k}, \hat{U}_{2k})\right] \leq D_2 + \epsilon$, our goal is to characterize the achievable rates R_1 and R_2 for source U_1 and U_2, respectively.

4.3.2 Achievable Rates

We begin by applying *Theorem 4.2,* the Berger-Tung inner bound, to the specific case of binary sources and Hamming distortions. For simplicity, we introduce the following definitions:

$$r_1^* \equiv I(U_1; W_1|W_2) = I(U_1; W_1) - I(W_1; W_2), \tag{4.29}$$

$$r_2^* \equiv I(U_2; W_2|W_1) = I(U_2; W_2) - I(W_1; W_2), \tag{4.30}$$

$$r_0^* \equiv I(U_1 U_2; W_1 W_2) = I(U_1; W_1) + I(U_2; W_2) - I(W_1; W_2). \tag{4.31}$$

These definitions allow us to rewrite the conditions in *Theorem 4.2* as $R_1 \geq r_1^*$, $R_2 \geq r_2^*$ and $R_1 + R_2 \geq r_0^*$. Moreover, we will make use of the binary entropy function $H_b(a) = -a \log a - (1 - a) \log(1 - a)$, and the notation $a * b$ to denote the operation

$$a * b = (1 - a)b + (1 - b)a,$$

as proposed by Wyner and Ziv in [94]. This notation is convenient for cascaded binary symmetric channels: if the first channel has crossover probability a and the second channel has crossover probability b, then the crossover probability from the input of the first channel to the output of the second channel is $a * b$.

Next, we define a pair of *test channels*, as illustrated in *Figure 4.5*. Let W_1 be the output of binary symmetric channel with input U_1 and crossover probability $0 \leq \delta_1 \leq 1/2$. Similarly, let W_2 be the output of binary symmetric channel with input U_2 and crossover probability $0 \leq \delta_2 \leq 1/2$. Clearly, we have a long Markov chain of the form $W_1 \rightarrow U_1 \rightarrow U_2 \rightarrow W_2$.

Consider the following instances of this setup:

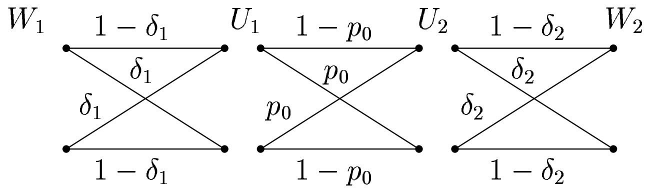


Figure 4.5: Binary symmetric test channels. We can view the correlated binary sources U_1 and U_2 as the input and the output of a binary symmetric channel with crossover probability p_0. The auxiliary random variables W_1 and W_2 can be obtained likewise using two test channels with crossover probabilities δ_1 and δ_2, respectively.

A. Non-Degenerate W_1 and W_2:

Setting $\hat{U}_1 = \hat{U}_1(W_1, W_2) = W_1$ and $\hat{U}_2 = \hat{U}_2(W_1, W_2) = W_2$, we have the average distortions $\mathrm{E}\left[\sum_{k=1}^{n} d(U_{1k}, \hat{U}_{1k})\right] = \delta_1$ and $\mathrm{E}\left[\sum_{k=1}^{n} d(U_{2k}, \hat{U}_{2k})\right] = \delta_2$. For the minimum coding rates we get

$$\begin{aligned} r_1^* &= I(U_1; W_1) - I(W_1; W_2) \\ &= (1 - H_b(\delta_1)) - (1 - H_b(\delta_1 * \delta_2 * p_0)), \\ &= H_b(\delta_1 * \delta_2 * p_0) - H_b(\delta_1), \end{aligned}$$

and, similarly,

$$\begin{aligned} r_2^* &= I(U_2; W_2) - I(W_1; W_2) \\ &= H_b(\delta_1 * \delta_2 * p_0) - H_b(\delta_2). \end{aligned}$$

Finally, the required sum rate becomes

$$\begin{aligned} r_0^* &= I(U_1; W_1) + I(U_2; W_2) - I(W_1; W_2) \\ &= 1 - H_b(\delta_1) - H_b(\delta_2) + H_b(\delta_1 * \delta_2 * p_0). \end{aligned}$$

We note that for all possible values of δ_1, δ_2 and p_0, we have $\delta_1 * \delta_2 * p_0 \geq \delta_1 * p_0 \geq \delta_1$, and similarly $\delta_1 * \delta_2 * p_0 \geq \delta_2 * p_0 \geq \delta_2$.

B. Degenerate $W_1 (W_1 = 0)$

Setting $\hat{U}_1 = \hat{U}_1(W_1, W_2) = W_2$ and $\hat{U}_2 = \hat{U}_2(W_1, W_2) = W_2$, we have the average distortions $\mathrm{E}\left[\sum_{k=1}^{n} d(U_{1k}, \hat{U}_{1k})\right] = \delta_2 * p_0$ and $\mathrm{E}\left[\sum_{k=1}^{n} d(U_{2k}, \hat{U}_{2k})\right] = \delta_2$. For the minimum coding rates we get

$$\begin{aligned} r_1^* &= 0 \\ r_2^* &= r_0^* = 1 - H_b(\delta_2). \end{aligned}$$

C. Degenerate W_2($W_2 = 0$)

Setting $\hat{U}_1 = \hat{U}_1(W_1, W_2) = W_1$ and $\hat{U}_2 = \hat{U}_2(W_1, W_2) = W_1$, we have E $\left[\sum_{k=1}^{n} d(U_{1k}, \hat{U}_{1k})\right] = \delta_1$ and E $\left[\sum_{k=1}^{n} d(U_{2k}, \hat{U}_{2k})\right] = \delta_1 * p_0$, and similarly,

$$\begin{aligned} r_1^* &= r_0^* = 1 - H_b(\delta_2) \\ r_2^* &= 0. \end{aligned}$$

Now, by taking the lower convex hull over the rate-distortion regions defined by $R_1 \geq r_1^*$, $R_2 \geq r_2^*$ and $R_1 + R_2 \geq r_0^*$ for each of the three cases[2] presented above, we obtain the inner bound that we seek. For prescribed distortion values $0 \leq D_1 \leq 1/2$ and $0 \leq D_2 \leq 1/2$ for sources U_1 and U_2, respectively, we must consider all $0 \leq \delta_1, \delta_2 \leq 1/2$ and $0 \leq \lambda_1, \lambda_2 \leq 1$ such that

$$\begin{aligned} D_1 &= \lambda_1\delta_1 + \lambda_2(\delta_2 * p_0) + (1-\lambda_1-\lambda_2)\delta_1 = \lambda_2(\delta_2 * p_0) + (1-\lambda_2)\delta_1 \\ D_2 &= \lambda_1(\delta_1 * p_0) + \lambda_2\delta_2 + (1-\lambda_1-\lambda_2)\delta_2 = \lambda_1(\delta_1 * p_0) + (1-\lambda_1)\delta_2. \end{aligned}$$

Since the functions r_i^*, $i = 0, 1, 2$ are convex (see the appendix, Section A.7), we can obtain the corresponding minimum rates R_i^* according to

$$\begin{aligned} R_1^* &= \lambda_1 \cdot (1 - H_b(\delta_1)) + \lambda_2 \cdot 0 + (1-\lambda_1-\lambda_2) \cdot (H_b(\delta_1 * \delta_2 * p_0) - H_b(\delta_1)) \\ R_2^* &= \lambda_1 \cdot 0 + \lambda_2 \cdot (1 - H_b(\delta_2)) + (1-\lambda_1-\lambda_2) \cdot (H_b(\delta_1 * \delta_2 * p_0) - H_b(\delta_2)) \\ R_0^* &= \lambda_1 \cdot (1 - H_b(\delta_1)) + \lambda_2 \cdot (1 - H_b(\delta_1) + (1-\lambda_1-\lambda_2) \cdot (H_b(\delta_1 * \delta_2 * p_0)) \\ &\quad - H_b(\delta_1) - H_b(\delta_2)). \end{aligned}$$

4.4 Cooperative Multiterminal Source Coding

Having studied the classical multiterminal source coding problem in detail, we will now take into consideration the effects of partial cooperation between encoders. For this purpose, we re-formulate the problem using the conference mechanism explained in Chapter 3.

4.4.1 Problem Statement

Assume two sources U_1 and U_2, which are drawn i.i.d. $\sim p(u_1 u_2)$ from two finite alphabets, denoted $\mathcal{U}_1$ and $\mathcal{U}_2$. The two sources are processed by two separate encoders, which can exchange messages over two independent links at rates R_{12} and R_{21} prior to encoding, with information flowing in opposite directions, as illustrated in *Figure 4.6*.

A *conference* among encoders is specified by a set of $2K$ functions

$$\begin{aligned} h_{1k} : \mathcal{U}_1^N \times \mathcal{V}_{21} \times ... \times \mathcal{V}_{2(k-1)} &\rightarrow \mathcal{V}_{1k} \\ h_{2k} : \mathcal{U}_2^N \times \mathcal{V}_{11} \times ... \times \mathcal{V}_{1(k-1)} &\rightarrow \mathcal{V}_{2k}, \end{aligned}$$

[2]The case when both W_1 and W_2 are degenerate is not relevant, because it corresponds to the maximum distortions $D_1 = D_2 = 1/2$ and $R_1 = R_2 = 0$.

We say that the rate-distortion tuple $(R_1, R_2, R_{12}, R_{21}, D_1, D_2)$ is *achievable* if for any $\epsilon > 0$ and sufficiently large N and K a $(R_1, R_2, R_{12}, R_{21}, N, K, \overline{D}_1, \overline{D}_2)$ code exists such that $\overline{D}_1 \leq D_1 + \epsilon$ and $\overline{D}_2 \leq D_2 + \epsilon$. Finally, the *rate-distortion region* $\mathcal{R}(\mathcal{D})$ of the two sources is the closure of the set of achievable rate distortion tuples $(R_1, R_2, R_{12}, R_{21}, D_1, D_2)$. The goal of the problem is to give a complete characterization of the rate-distortion region in terms of single letter information theoretic quantities.

4.4.2 Lossless Cooperative Source Coding (CSC)

A. The CSC Rate Region with Two Sources

We start by solving the cooperative source coding problem formulated in the previous section, for the case of perfect reconstruction at the receiver (lossless compression), i.e. $D_1 = D_2 = 0$. The corresponding rate region, which generalizes the result of Slepian and Wolf, is given by the following coding theorem.

Theorem 4.9 (Barros, Servetto [15]) The sources $(U_1U_2) \sim p(u_1u_2)$ can be encoded at rates (R_1, R_2), respectively, and perfectly reconstructed at the decoder if and only if

$$\begin{aligned} R_1 &\geq H(U_1|U_2Z) \\ R_2 &\geq H(U_2|U_1Z) \\ R_1 + R_2 &\geq H(U_1U_2) \\ R_{12} &\geq I(U_1;Z|U_2) \\ R_{21} &\geq I(U_2;Z|U_1) \end{aligned}$$

where Z is an auxiliary random variable, such that $p(u_1u_2z) = p(u_1u_2)p(z|u_1u_2)$.

Proof: We now prove the existence of codes that satisfy the conditions of the theorem. Partition the set $\mathcal{U}_1$ in M_1 cells, indexed by $v_1 \in \{1, \ldots, M_1\}$, such that $v_1(u_1) = c_1$ if u_1 is inside cell c_1. Similarly, partition the set $\mathcal{U}_2$ in M_2 cells, indexed by $v_2 \in \{1, \ldots, M_2\}$, such that $v_2(u_2) = c_2$ if u_2 is inside cell c_2.

Prior to transmission encoder 1 sends v_1 for each observed value u_1 to encoder 2, and the latter sends to the former the index v_2 for each observed value u_2. Let $Z = (V_1V_2)$. Since (V_1V_2) are functions of the source random variables (U_1U_2), Z is also a random variable and a function of (U_1U_2), which in turn means that $p(u_1u_2z) = p(u_1u_2)p(z|u_1u_2)$ is a well-defined probability distribution.

We will now show that the rates $R_{12} \geq H(V_1|U_2)$ and $R_{21} \geq H(V_2|U_1)$ are sufficient for V_1^N and V_2^N to be exchanged between the encoders with arbitrarily small probability of error. The encoders are informed about the joint probability distribution $p(u_1u_2v_1v_2)$, from which they can obtain the marginals $p(u_1v_2)$ and $p(u_2v_1)$. Notice that these two distributions can be viewed as two pairs of correlated sources (U_1V_2) and (U_2V_1). Since U_2^N is perfectly known at encoder 2, it follows from the Slepian-Wolf theorem for (U_2V_1) that V_1^N can be compressed at rates $R_{12} \geq H(V_1|U_2)$ and still be reconstructed perfectly at encoder 2. Similarly, V_2^N can be compressed at rates $R_{21} \geq H(V_2|U_1)$ and still be reconstructed perfectly at encoder 1. Thus, we can write

$$\begin{aligned} R_{12} &\geq H(V_1|U_2) \\ &= H(V_1V_2|U_2) \\ &= H(Z|U_2) - H(Z|U_1U_2) \\ &= I(U_1; Z|U_2) \end{aligned}$$

and similarly $R_{21} \geq I(U_2; Z|U_1)$.

Let $U_1' = (U_1Z)$ and $U_2' = (U_2Z)$. Since U_1 and U_2 are i.i.d. sources, then $Z = f(U_1U_2)$ is also i.i.d. and U_1' and U_2' can be viewed as two i.i.d. sources $\sim p(u_1'u_2') = p(u_1u_2z)$. Then according to the Slepian-Wolf theorem the following rates are achievable:

$$\begin{aligned} R_1 &> H(U_1'|U_2'), \\ R_2 &> H(U_2'|U_1'), \\ R_1 + R_2 &> H(U_1'U_2'). \end{aligned}$$

Substituting $U_1' = (U_1Z)$ and $U_2' = (U_2Z)$, we get

$$\begin{aligned} R_1 &> H(U_1Z|U_2Z) = H(U_1|U_2Z), \\ R_2 &> H(U_2Z|U_1Z) = H(U_2|U_1Z), \\ R_1 + R_2 &> H(U_1U_2Z) = H(U_1U_2). \end{aligned}$$

The converse part of the proof is based on Fano's inequality and standard techniques. Details can be found in Appendix A.6. ■

B. Lossless CSC with an Arbitrary Number of Sources $M > 2$

The arguments used to prove *Theorem 4.9* can be easily extended to the case of $M > 2$ sources, yielding the following result.

Theorem 4.10 (Barros, Servetto [15]) A set of correlated sources $U^M = \{U_1U_2 \dots U_M\}$ can be encoded with partially cooperating encoders if and only if there exist random variables Z_{ij}, $1 \leq i < j \leq M$, such that

$$R(S) > H(U(S)|U(S^c)Z(S^c)), \quad (4.32)$$

for all subsets $S \subseteq \{1, 2, \dots, M\}$, where S^c denotes the complement of S, $U(S) = \{U_j : j \in S\}$, $Z(S) = \{Z_{ij} : i \in S \text{ or } j \in S\}$ and $I(Z_{ij}; U_i|U_j) < R_{ij}$.

Proof: The achievability proof is based on the M-source version of the Slepian-Wolf theorem [28, Theorem 14.4.2], whereas the converse proof uses the same arguments as the converse proof of *Theorem 4.9* but with $2^M - 1$ inequalities. Since the proof contains no new ideas we omit the details. ■

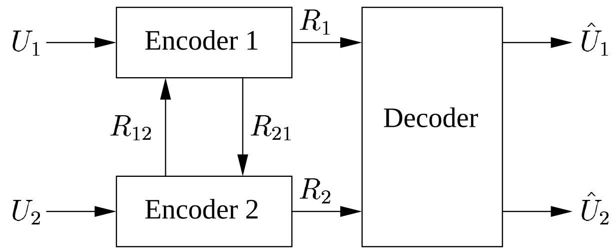


Figure 4.6: A cooperative version of the multiterminal source coding problem.

such that the conference message $V_{1k} \in \mathcal{V}_{1k}$ (or $V_{2k} \in \mathcal{V}_{2k}$) of encoder 1 (or encoder 2) at time k depends on the previously received messages V_2^{k-1} (or V_1^{k-1}) and the corresponding source message. A conference is said to be (R_{12}, R_{21})-admissible if and only if

$$\sum_{k=1}^{K} \log_2 |\mathcal{V}_{1k}| \leq NR_{12},$$

and

$$\sum_{k=1}^{K} \log_2 |\mathcal{V}_{2k}| \leq NR_{21}.$$

The *encoders* are two functions:

$$\begin{aligned} f_1 : \mathcal{U}_1^N \times \mathcal{V}_{21} \times ... \times \mathcal{V}_{2K} &\rightarrow \{1, 2, \ldots, 2^{NR_1}\}, \\ f_2 : \mathcal{U}_2^N \times \mathcal{V}_{11} \times ... \times \mathcal{V}_{1K} &\rightarrow \{1, 2, \ldots, 2^{NR_2}\}. \end{aligned}$$

These encoding functions map a block of N source symbols U_i^N, $i \in \{1, 2\}$, observed by each encoder, and a block of K messages received from the other encoder, to a discrete index $f_i(U_i^N)$.

The *decoder* is a function

$$g : \{1, 2, \ldots, 2^{NR_1}\} \times \{1, 2, \ldots, 2^{NR_2}\} \rightarrow \hat{\mathcal{U}}_1^N \times \hat{\mathcal{U}}_2^N,$$

which maps a pair of indices into two blocks of reconstructed source sequences.

Let the *distortion* measures be

$$d_1 : \mathcal{U}_1 \times \hat{\mathcal{U}}_1 \rightarrow \mathcal{R}^+ \quad \text{and} \quad d_2 : \mathcal{U}_2 \times \hat{\mathcal{U}}_2 \rightarrow \mathcal{R}^+.$$

An $(R_1, R_2, R_{12}, R_{21}, N, K, \overline{D}_1, \overline{D}_2)$ code is defined by

- two encoding functions f_1 and f_2,
- a decoding function g,
- a (R_{12}, R_{21})-admissible conference of length K,
- a distortion pair $\overline{D}_1 = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} d_1(U_{1i}, \hat{U}_{1i})\right]$, and $\overline{D}_2 = \mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N} d_2(U_{2i}, \hat{U}_{2i})\right]$, where $(\hat{U}_1^N \hat{U}_2^N) = g(f_1(U_1^N), f_2(U_2^N))$.

C. A Network Flow Interpretation of the Rate Region

The net effect of cooperation on the classical Slepian-Wolf rate region is to relax the conditions on the minimum amount of data required from each encoder, to the extent allowed by the conference network. This is illustrated in *Figure 4.7*.

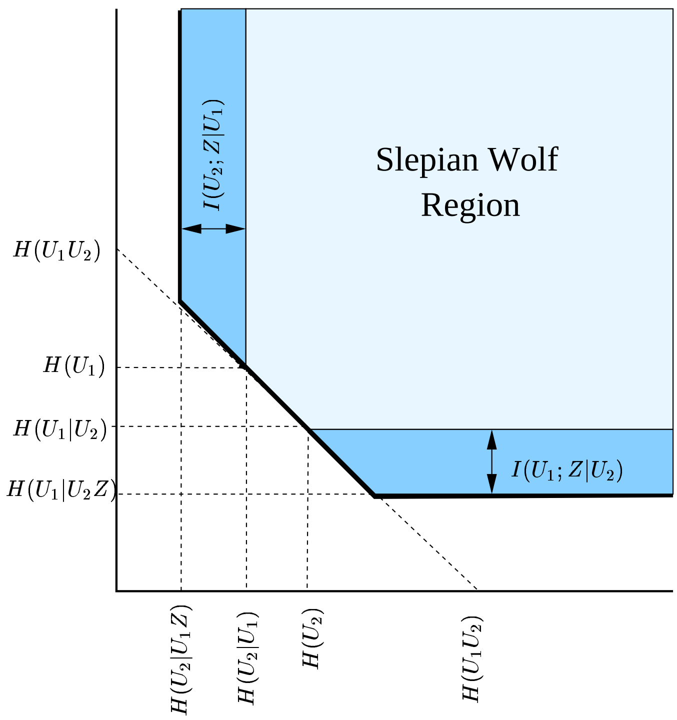


Figure 4.7: An example to illustrate the effect of cooperation among encoders. With cooperation we can enlarge the Slepian-Wolf rate region by the exact amount information exchanged by the encoders over the conference links (the dark shaded portion of the picture).

For $M = 2$ sources, we can offer an intuitive interpretation in terms of flow networks [88]: the two sources of correlated information (U_1U_2) can be viewed as two distinct commodities that must be transported to the sink. The first commodity U_1 can either flow directly from encoder 1 to the decoder, or indirectly via encoder 2, and similarly for the second commodity U_2. What the converse part of our cooperative source coding theorem shows is that, analogously to the classical multi-commodity flow problem, rates (R_1, R_2) (i.e. the amounts of flow for each commodity) are feasible *if and only if* there exists a way to split $H(U_1U_2)$ information bits among the two sources, and a way to route all the information for each source partly over the direct path to the decoder and partly through the conference links—this interpretation is further developed in [8].

4.4.3 Rate-Distortion Bounds for Cooperative Source Coding

In this section, we apply the conferencing techniques explained in the previous section, to obtain inner and outer bounds for the rate-distortion version of the problem. Our proofs are based on the work of Berger and Tung for the classical multiterminal source coding problem, explained in detail in Section 4.2.

A. Inner Bound for Cooperative Source Coding

Based on the conference mechanism presented in Section 4.4, and the forward part of Berger and Tung's theorem, we are able to obtain sufficient conditions for the existence of rate-distortion codes for cooperative source coding.

Theorem 4.11 (Barros, Servetto [15]) Let (U_1U_2) be drawn i.i.d. $\sim p(u_1u_2)$. For a given distortion pair (D_1, D_2), there exist codes with rates (R_1, R_2), where

$$\begin{aligned} R_1 &\geq I(U_1U_2; W_1|W_2) \\ R_2 &\geq I(U_1U_2; W_2|W_1) \\ R_1 + R_2 &\geq I(U_1U_2; W_1W_2) \\ R_{12} &\geq I(U_1; Z|U_2) \\ R_{21} &\geq I(U_2; Z|U_1), \end{aligned}$$

if there exist auxiliary random variables W_1, W_2, Z, such that the joint distribution of all variables is of the form $p(u_1u_2)p(z|u_1u_2)p(w_1|u_1z)p(w_2|u_2z)$, and for which there exist reconstruction functions $\hat{U}_1(W_1, W_2)$ and $\hat{U}_2(W_1, W_2)$ such that $D_1 \geq \mathrm{E}\left[d_1(U_1, \hat{U}_1)\right]$ and $D_2 \geq \mathrm{E}\left[d_2(U_2, \hat{U}_2)\right]$.

Proof: We start with the conferencing mechanism and then obtain achievability conditions by applying *Theorem 4.2*. The conference messages (V_1V_2) are generated using the partitions explained in the proof of *Theorem 4.9*, such that by defining the auxiliary random variable $Z = (V_1V_2)$ we obtain a well-defined probability distribution $p(u_1u_2z) = p(u_1u_2)p(z|u_1u_2)$. It follows from the proof of *Theorem 4.9* that conditions $R_{12} \geq I(U_1; Z|U_2)$ and $R_{21} \geq I(U_2; Z|U_1)$ are sufficient to guarantee perfect recontruction of V_2 at encoder 1 and V_1 at encoder 2.

Let $U_1' = (U_1Z)$ and $U_2' = (U_2Z)$. Since U_1 and U_2 are i.i.d. sources, then $Z = f(U_1U_2)$ is also i.i.d. and U_1' and U_2' can be viewed as two i.i.d. sources $\sim p(u_1'u_2') = p(u_1u_2z)$. Then, according to *Theorem 4.2*, there exist codes at rates

$$R_1 \geq I(U_1'U_2'; W_1|W_2) = I(U_1U_2Z; W_1|W_2) \tag{4.33}$$
$$R_2 \geq I(U_2'U_2'; W_2|W_1) = I(U_1U_2Z; W_1|W_2) \tag{4.34}$$
$$R_1 + R_2 \geq I(U_1'U_2'; W_1W_2) = I(U_1U_2Z; W_1W_2), \tag{4.35}$$

where W_1, W_2 are random variables such that $p(u_1'u_2'w_1w_2) = p(u_1'u_2')p(w_1|u_1')\ p(w_2|u_2')$, or similarly $p(u_1u_2zw_1w_2) = p(u_1u_2)p(z|u_1u_2)p(w_1|u_1z)p(w_2|u_2z)$. Since $Z = (V_1V_2)$ and (V_1V_2) are functions of (U_1U_2) we can omit Z in (4.33)-(4.35), thus obtaining the same conditions as in *Theorem 4.2*. ■

B. Outer Bound for Cooperative Source Coding

We are also able to obtain necessary conditions for the existence of rate-distortion codes for the cooperative source coding problem:

Theorem 4.12 (Barros, Servetto [15]) For a given distortion pair (D_1, D_2) all achievable rates (R_1, R_2) must satisfy the following conditions:

$$\begin{aligned} R_1 &\geq I(U_1U_2; W_1|W_2) \\ R_2 &\geq I(U_1U_2; W_2|W_1) \\ R_1 + R_2 &\geq I(U_1U_2; W_1W_2) \\ R_{12} &\geq I(U_1; Z|U_2) \\ R_{21} &\geq I(U_2; Z|U_1), \end{aligned}$$

where (i) the joint probability distribution $p(u_1u_2)p(z|u_1u_2)p(w_1w_2|u_1u_2z)$ is such that $W_1 \to U_1Z \to U_2$ and $U_1 \to U_2Z \to W_2$ form two Markov chains, and (ii) there exist $\hat{U}_1(W_1, W_2)$ and $\hat{U}_2(W_1, W_2)$, such that $D_1 \geq \mathrm{E}\left[d_1(U_1, \hat{U}_1)\right]$ and $D_2 \geq \mathrm{E}\left[d_2(U_2, \hat{U}_2)\right]$.

Proof: The proof uses the same arguments as the proof of the Berger-Tung outer bound (cf. [90]), and the proof of *Theorem 4.9*, therefore we omit it here. ■

C. Two Important Remarks

1) The inner and outer bounds for cooperative source coding may or may not be tight.

As in the non-cooperative case, the descriptions of the inner and outer bounds are based on different sets of auxiliary random variables. The former imposes a long chain Markov condition $W_1 \to U_1Z \to U_2Z \to W_2$, whereas the latter only requires two short chains $W_1 \to U_1Z \to U_2$ and $U_1 \to U_2Z \to W_2$. This situation is entirely analogous to that of the classical problem without cooperation (same rate expressions, description differs only in that the inner bound requires $W_1 \to U_1 \to U_2 \to W_2$, whereas the outer bound requires $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$, and therefore, cooperation does not lead to any new insights: solving our problem with cooperation is unfortunately exactly as hard as solving the classical problem.

2) Cooperation does generate new achievable rate pairs.

Although all four rate expressions (inner and outer bounds, with and without cooperation) are identical, this does not imply that the rate-distortion region with and without cooperation are identical. The subtle difference lies again in the set of auxiliary random variables W_1 and W_2, which do not have the same constraints in both cases. In the non-cooperative inner bound, they must form a long Markov chain of the form $W_1 \to U_1 \to U_2 \to W_2$, whereas in the cooperative inner bound, they depend on the two sources and the conference messages Z leading to the long chain $W_1 \to U_1Z \to U_2Z \to W_2$. As argued in the paragraph above, like in the case without cooperation, we do not know whether or not there are rate pairs generated by distributions of the form $W_1 \to U_1Z \to U_2$ and $U_1 \to U_2Z \to W_2$ that are not already contained among those rate pairs generated by distributions of the form $W_1 \to U_1Z \to U_2Z \to W_2$. However, we *do* know that there are rate pairs generated by distributions of the form $W_1 \to U_1Z \to U_2Z \to W_2$ (for the inner bound with cooperation) which cannot be generated by distributions of the form $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$ (for the outer bound without cooperation): consider the case in which $Z = U_1$, such that by communicating at rates $R_{12} > I(U_1; Z|U_2) = H(U_1|U_2)$

encoder 2 is able to reconstruct the source block U_1^N perfectly prior to encoding. Then, the encoders can operate for example at rates $(R_1, R_2) = (0, I(U_1U_2; W_1W_2))$, corresponding to a rate-distortion tuple which in the general case is clearly outside of the region defined by the non-cooperative outer bound.

4.5 Summary and Conclusions

In this chapter, we showed that a natural extension of the sensor reachback problem from a rate-distortion perspective leads directly to the long-standing multiterminal source coding problem. After showing through a new proof that separate quantization followed by Slepian-Wolf compression achieves all the rate-distortion tuples given by the classical inner bound of Berger and Tung, we presented two new rate-distortion regions which overcome one of the major obstacles towards a complete solution for the problem: the long Markov chain $W_1 \to U_1 \to U_2 \to W_2$. An important feature of the new inner bound based on time-sharing of Berger-Yeung codes is that it allows us to determine the gap between known achievable rates and the outer bound of Berger and Tung. Turning our attention to a practical example — binary sources with Hamming distortions — we showed how to apply the presented techniques in order to obtain a partial caracterization of the rate-distortion region.

Another key contribution in this chapter was the extension of the multiterminal source coding problem to the case of partial cooperation between encoders. Although we encountered the same technical difficulties as in the non-cooperative case, it is worth pointing out that once the classical multiterminal source coding problem without cooperation is solved, our results in the previous section automatically allow us to obtain a solution for the problem with cooperation.

5

Scalable Decoding for Sensor Reachback

Todos os caminhos vão dar a Roma.[1]
PORTUGUESE PROVERB

In the previous two chapters, we studied the fundamental limits of reachback communication in sensor networks using some of the mathematical tools provided by network information theory. Said information-theoretic approach allows us to make very general and durable statements about communications systems, albeit at the cost of a somewhat idealistic assumption — the encoding and decoding algorithms have *unbounded* complexity. Our goal in this chapter, is to re-investigate the sensor reachback problem from a more practical point of view, using a basic system setup that allows us to identify a practical complexity bottleneck and find realizable solutions based on contemporary estimation techniques.

5.1 Introduction

5.1.1 Problem Background

Consider once again a large-scale sensor network, in which hundreds of sensor nodes pick up samples from a physical process in a field, encode their observations and transmit the data back to a remote location over an array of *reachback* channels. The task of the decoder at the remote location is then to produce the best possible estimates of the data sent by all the nodes.

In Chapter 3 we showed that the correlation between the measurements of different sensors — a condition that must be necessarily true whenever we have a large number of nodes sensing a physical process within a confined area — can in general be exploited to improve the decoding result and thus increase the reachback capacity of the network. This principle holds, even when the sensor nodes themselves are not capable of eliminating the redundancy in the data prior

[1] *All roads lead to Rome.*

to transmission, as explained in Section 3.2.5. To fulfil this data compression task each node would have to use complex Slepian-Wolf source codes, a requirement that may well turn out to be impractical for large-scale sensor networks. In that case, the decoder can still take advantage of the remaining correlation to produce a more accurate estimate of the sent information.

The previous observation motivates us to consider now a reachback communications model, in which the system complexity is shifted from the sensor nodes to the receiver, in other words a reachback network with very simple encoders (e.g. a scalar quantizer, a bit mapping and a modulator) and a decoder of increased yet manageable complexity. Our goal is then to devise a practical decoding algorithm for this instance of the sensor reachback problem.

5.1.2 Main Contributions

Assuming a large-scale sensor network with hundreds of nodes, we argue that the optimal decoder based on minimum mean square error (MMSE) estimation is unfeasible, because its complexity grows exponentially with the number of sensors in the network. To guarantee the *scalability* of the decoding algorithm, we propose the following approach. First, we construct a factor graph [58] that models the correlation between the sensor signals in a flexible way depending on the targeted decoding complexity and the desired reconstruction fidelity. Then, based on this factor graph, we use the *belief propagation algorithm* (often called the *sum-product algorithm* [58]) to estimate the transmitted data. We are able to show that by choosing the factor graph in an appropriate way we can make the overall decoding complexity grow linearly with the number of nodes.

Naturally, the performance of the decoding algorithm depends heavily on the accuracy of the chosen factor graph as a model for the correlation in the sensor data. For large-scale sensor networks with arbitrary topology, we show that factor trees are particularly well suited for this application, because (a) they can be easily optimized, (b) they have no cycles thus allowing the belief propagation algorithm to yield an accurate solution, and (c) they provide a simple way to compute an exact complexity count for the associated decoder. Using the Kullback-Leibler distance as a measure of the fidelity of the approximated correlation model, we give a detailed mathematical treatment of factorization of multivariate Gaussian sources and a set of optimization algorithms for different classes of graphs. Moreover, we investigate the impact of degree constraints on the function nodes of the tree and find optimal trees for several instances of the problem.

Finally, we add a number of examples and numerical results that underline the performance and scalability of the proposed approach. It turns out that under reasonable assumptions on the spatial correlation of the sensor data, the performance of our decoder is very close to the optimal MMSE solution.

5.1.3 Related Work

The idea of exploiting the remaining correlation in the encoded data to enhance the decoding result was already present in Shannon's landmark paper [82]. This principle was put effectively into practice by Hagenauer in [43], triggering many contributions that exploit the redundancy left by suboptimal quantizers in combination with convolutional codes or turbo codes [21] and powerful iterative decoding schemes [44]. More recently, this approach has also been success-fully implemented using low-density parity-check codes (see [56] and the references therein).

5.1.4 Chapter Outline

The rest of the chapter is organized as follows. Section 5.2 sets the stage for the main decoding problem by describing the system setup and elaborating on the drawbacks of the optimal decoder. Then, Section 5.3 describes our approach based on factor graphs and iterative decoding, and presents a few examples. The crux of this chapter is the set of optimization tools presented in Section 5.4. This chapter concludes with some numerical results in Section 5.5 and some comments in Section 5.6.

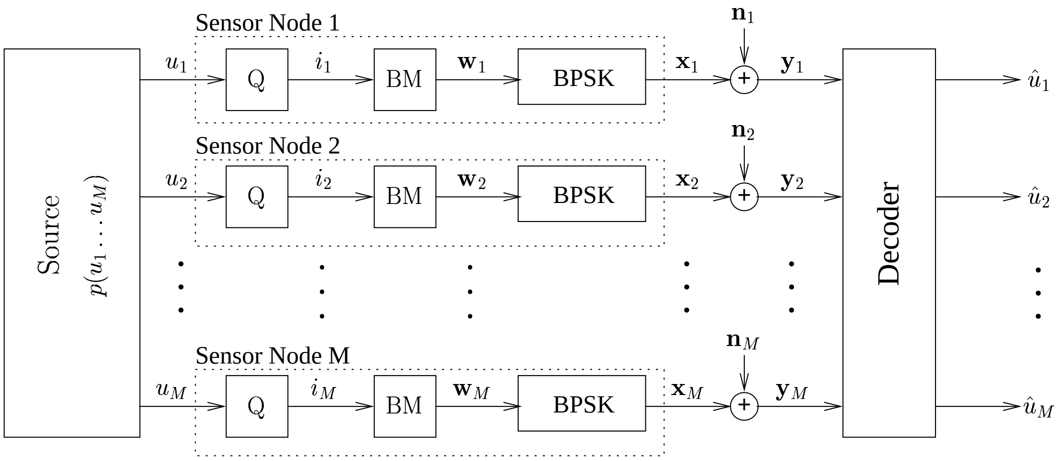


Figure 5.1: System model. M correlated samples are separately encoded by M sensor nodes, consisting of a scalar quantizer, a bit mapper and a modulator. The data is then transmitted over an array of independent AWGN channels and decoded jointly at the receiver.

5.2 Problem Setup

5.2.1 System Model

The basic system model that accompanies us throughout this chapter is illustrated in *Figure 5.1*. We begin with a brief explanation of useful notation and a precise description of the source model, the encoding procedure and the reachback channel.

Notation

In the following, we consider all vectors to be column vectors and denote them with small bold letters. On the other hand, matrices are denoted with *capital* bold letters, unless otherwise stated. The expression $\mathbf{0}_N$ corresponds to the length-N all-zero column vector, $\mathbf{I}_N$ is the $N \times N$ identity matrix, and $|\mathbf{A}|$ denotes the determinant of $\mathbf{A}$. We will often refer to the covariance defined as

$$\mathrm{Cov}\{\mathbf{x}, \mathbf{y}\} \triangleq \mathrm{E}\left[\mathbf{x}\mathbf{y}^T\right] - \mathrm{E}\left[\mathbf{x}\right] \mathrm{E}\left[\mathbf{y}\right]^T,$$

where $\mathrm{E}\left[\cdot\right]$ denotes again the expected value. An N-dimensional random variable with realizations $\mathbf{u} = (u_1\, u_2\, ...\, u_N)^T$, $u_i \in \mathbb{R}$, is Gaussian distributed with mean $\boldsymbol{\mu} = \mathrm{E}\left[\mathbf{u}\right]$ and covariance matrix $\mathbf{R} = \mathrm{Cov}\{\mathbf{u}, \mathbf{u}\}$, when its probability density function (PDF) $p(\mathbf{u})$ is given by

$$p(\mathbf{u}) = \exp(-\frac{1}{2}(\mathbf{u} - \boldsymbol{\mu})^T \mathbf{R}^{-1}(\mathbf{u} - \boldsymbol{\mu}))/(2\pi|\mathbf{R}|)^{1/2}. \tag{5.1}$$

Such a PDF is simply denoted as $\mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$.

Source Model

Each sensor k observes at time t continuous real-valued data samples $u_k(t)$ with $k = 1, 2, \ldots, M$. For simplicity, we assume that the M sensor nodes are placed randomly on the unit square and focus on the spatial correlation of measurements and not their temporal dependence. Thus, we drop the time variable t and consider only one time step. Nevertheless, we do point out that the discussed techniques can be easily extended to account for sources with memory. The sample vector $\mathbf{u} = (u_1\, u_2 \ldots u_M)^T$ at any given time t is assumed to be the realization of an N-dimensional Gaussian random variable, whose PDF $p(\mathbf{u})$ is given by $\mathcal{N}(\mathbf{0}_M, \mathbf{R})$ with

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,M} \\ \rho_{2,1} & 1 & \cdots & \rho_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{M,1} & \rho_{M,2} & \cdots & 1 \end{bmatrix}.$$

It follows that the samples u_k have zero mean $\mathrm{E}[u_k]$ and unit variance $Cov\{u_k, u_k\}$. Gaussian models for capturing the spatial correlation between sensors at different locations are discussed in [78], whereas examples of reasonable models for the correlation coefficients $\rho_{k,k'} = \mathrm{E}[u_k u_{k'}]$ of physical processes unfolding in a field can be found in [32]. In the following, we make two simplifying assumptions:

1. the sensors are randomly placed in a unit square according to a uniform distribution
2. the correlation between the measurements of any two sensors decays exponentially with the Euclidean distance between them.

Denoting the position of sensor m as $\mathbf{z}_m = [z_m(1), z_m(2)]^T$, we can write the distance between two sensors i and j as $l_{i,j} = \|\mathbf{z}_i - \mathbf{z}_j\|$, such that the correlation between their measurements is given by $\rho_{i,j} = \exp(-\beta \cdot l_{i,j})$, where β is a positive constant.

Encoding

We assume that the sensors are "cheap" devices consisting of a scalar quantizer, a bit mapper and a modulator[2]. Sensor k quantizes the sample u_k to output the index $i_k \in \mathcal{L} = \{1, 2, \ldots, 2^Q\}$, such that i_k is represented by Q bits. There are 2^Q reconstruction values $\tilde{u}(i_k) \in \mathbb{R}$ also indexed by $i_k \in \mathcal{L}$. The modulator maps i_k to a tuple $\mathbf{x}_k$ of channel symbols, which are transmitted to the remote receiver. In our examples we use binary phase shift keying (BPSK), such that in a discrete-time baseband description of our transmission scheme i_k is mapped first to a binary codeword $\mathbf{w}_k$ and then to Q channel symbols $\mathbf{x}_k = (x_{k,1} \ldots x_{k,Q})$, $x_{k,q} \in \{+1, -1\}$.

[2]The chosen models for the encoder and the channel may seem too simple, yet they allows us to focus on the essential aspects of the problem and highlight the key features of our decoding algorithm. The latter can be easily extended to include, for example, more sophisticated channel coding.

Reachback Channel

As we argued in Chapter 3, since all sensors must transmit some data to the central receiver *all* the time, reservation based medium access protocols (e.g. TDMA or FDMA) are a very reasonable choice for this type of reachback networks. Thus, we assume that the reachback channel is virtually interference-free, i.e., the joint PDF $p(\mathbf{y}_1 \ldots \mathbf{y}_M|\mathbf{x}_1 \ldots \mathbf{x}_M)$ factors into $\prod_{k=1}^{M} p(\mathbf{y}_k|\mathbf{x}_k)$. In addition, we model the reachback channel as an array of additive white Gaussian noise (AWGN) channels with noise variance σ^2. Assuming coherent demodulation, we write the channel outputs as $\mathbf{y}_k = \mathbf{x}_k + \mathbf{n}_k$, $k = 1, 2, \ldots, M$, where $\mathbf{n}_k$ is distributed according to $\mathcal{N}(\mathbf{0}_Q, \sigma^2 \mathbf{I}_Q)$.

Having defined the main building blocks on the transmitting side of our reachback network, we now turn our attention to the decoding algorithm at the receiving end.

5.2.2 Optimal MMSE Decoding

After obtaining $\mathbf{y}_k$ from each sensor, the decoder uses $\mathbf{y} = (\mathbf{y}_1\, \mathbf{y}_2, \ldots, \mathbf{y}_M)^T$ and the available *a priori* knowledge[3] of the source correlation $\mathbf{R}$ to produce the estimates $\hat{u}_k$ of the measurements u_k. Since we are interested in obtaining reconstructed samples $\hat{u}_k$ with high fidelity, the chosen figure of merit to be minimized by the decoder is the mean square error (MSE) $\mathrm{E}\left[(\hat{u}_k - \tilde{u}(i_k))^2\right]$, which in this case is achieved by the conditional mean estimator (CME, [74]):

$$\hat{u}_k = \mathrm{E}\left[\tilde{u}(i_k)|\mathbf{y}\right] = \sum_{\forall i \in \mathcal{L}} \tilde{u}(i) \cdot p(i_k = i|\mathbf{y}). \tag{5.2}$$

Notice that for PDF optimized quantizers this estimator also minimizes the MSE $\mathrm{E}\left[(\hat{u}_k - u_k)^2\right]$ between $\hat{u}_k$ and u_k [52]. Clearly, in order to perform optimal decoding, we require the posterior probabilities $p(i_k|\mathbf{y})$, which in this case are given by

$$p(i_k = i|\mathbf{y}) = \gamma \cdot \sum_{\forall \mathbf{i} \in \mathcal{L}^M : i_k = i} p(\mathbf{y}|\mathbf{i}) p(\mathbf{i}), \tag{5.3}$$

where $\mathbf{i} = (i_1\, i_2 \ldots i_M)^T$ and $\gamma = 1/p(\mathbf{y})$ is a constant normalizing the sum over the product of probabilities to one. Since the AWGN channels are independent, $p(\mathbf{y}|\mathbf{i})$ factors into $\prod_{k=1}^{M} p(\mathbf{y}_k|i_k)$, with $p(\mathbf{y}_k|i_k)$ given by $\mathcal{N}(\mathbf{x}_k(i_k), \sigma^2 \mathbf{I}_Q)$. In addition, we require the probability mass function (PMF) $p(\mathbf{i})$ of the index vector $\mathbf{i}$, which can be obtained by numerically integrating the source PDF $p(\mathbf{u})$ characterized by $\mathbf{R}$ over the quantization region indexed by $\mathbf{i}$. Alternatively, one can resort to Monte Carlo simulations in order to estimate $p(\mathbf{i})$, a task which needs to be carried out only once and can therefore be performed offline.

To measure the computational complexity of the decoding process, we count the number of additions and multiplications required to compute the estimates $\hat{u}_k$ for all k. The decoding operation can be split into 3 steps:

1. Calculate $p(\mathbf{y}_k|i_k)$ for all k.

[3] It is reasonable to assume that for a large class of sensor applications, such as environmental monitoring or precision farming, the correlation matrix $\mathbf{R}$ can be obtained from past observations of the targeted physical process.

2. Marginalize i_k in $p(\mathbf{i}) \cdot \prod_{k=1}^{M} p(\mathbf{y}_k|i_k)$ for all k:

$$m(i_k) = \sum_{\forall \mathbf{i} \in \mathcal{L}^M : i_k = i} p(\mathbf{i}) \cdot \prod_{k=1}^{M} p(\mathbf{y}_k|i_k). \tag{5.4}$$

3. Calculate $\hat{u}_k$ for all k using $p(i_k|\mathbf{y}) = \gamma \cdot m(i_k)$.

Steps 1 and 3 require a number of additions and multiplications which is linear in the number M of sensors. In contrast, step 2 requires $M(2^{Q(M-1)}-1)$ additions and $M^2 2^{QM}$ multiplications to compute the marginals $m(i_k)$ for all k, which becomes unfeasible for sensor networks with a large number M of nodes. For example, for a reachback network with 100 sensors and $Q=1$, i.e. one-bit quantization, we would have to perform 10^{32} multiplications and 10^{30} additions!

5.2.3 Problem Statement

From the previous observation we conclude that the MMSE-optimal decoder is unfeasible with step 2 as the major bottleneck — its computational complexity grows exponentially with the number of nodes in the network. Our goal is thus to find a *scalable* decoding algorithm yielding the best possible trade-off between complexity and estimation error.

5.3 Scalable Decoding using Factor Graphs

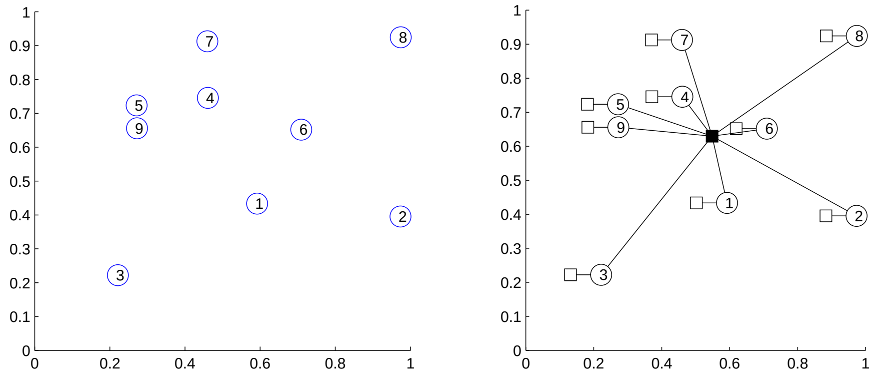
In this section, we present a scalable decoding solution based on factor graphs and the belief propagation (BP) algorithm [58]. These tools enable us to control the computational complexity of the decoding algorithm using the following two-step approach. First, we define a simplified approximate model of the dependencies between the samples u_k by constructing a suitable factor graph. Then, we perform belief propagation decoding on the factor graph defined in the first step, in order to produce the desired data estimates with complexity growing linearly with the size M of the reachback network.

5.3.1 Factor Graphs and the Belief Propagation Algorithm

A factor graph depicts a function of typically many variables, which factors with respect to a suitable operation such as multiplication. There are two types of nodes in a factor graph: *variable nodes* representing the variables and *function nodes* representing the factors. The dependencies between variables and factors are indicated by *edges* connecting some of the nodes. We use the term *degree of a node* to refer to the number of edges connected to a certain node.

Going back to our problem, the function that needs to be factorized is $p(\mathbf{i}) \cdot \prod_{k=1}^{M} p(\mathbf{y}_k|i_k)$, as stated in step 2 of the decoding operation explained in the previous section. The corresponding factor graph, illustrated in *Figure 5.2* for $M=9$ sensors, consists of M variable nodes (one for each index i_k in the index vector $\mathbf{i}$) and $M+1$ function nodes (one degree-1 node for each of the M PDFs $p(\mathbf{y}_k|i_k)$ and one degree-M node for $p(\mathbf{i})$).

Although factor graphs are certainly not the only way to design a decoder for the reachback network, there are several properties that make this tool particularly appealing for the problem at hand:



(a) Sensor Positions. (b) Factor Graph.

Figure 5.2: The factor graph of the function $p(\mathbf{i}) \cdot \prod_{k=1}^{9} p(\mathbf{y}_k|i_k)$ for the sensor network depicted in (a), consists of 9 variable nodes for each index i_k (circles), 9 function nodes for each factor $p(\mathbf{y}_k|i_k)$ (empty boxes), and a function node for the factor $p(\mathbf{i})$ (filled box), as shown in (b). Notice that the positions of the variable nodes conveniently correspond to the actual sensor positions.

A) The factor graph reflects the network topology.

Assuming the positions of the sensor on the unit square are known, their dependencies can be easily depicted by connecting the corresponding variable nodes, as illustrated in *Figure 5.2*.

B) In combination with the belief propagation decoding algorithm the factor graph yields the desired estimates directly.

Based on the factor graph we can compute the marginals $m(i_k)$ very efficiently using the well-known belief propagation (BP) algorithm (for a detailed description, see [58]). The main idea is to let the nodes pass "messages" to their neighbours along the edges of the graph. As long as the factor graph is cycle-free, the BP algorithm yields the correct marginals $m(i_k)$ in the M variable nodes. Otherwise, the BP algorithm becomes iterative (the messages circulate forever) and the desired marginals $m(i_k)$ cannot be computed exactly[4].

C) The resulting algorithmic complexity can be easily computed directly from the factor graph.

Cycle-free factor graphs have another very useful property: the number of additions and multiplications required during message passing can be derived directly from the degrees of the variable nodes and the function nodes [5]. For our factor graph, in which the variables are drawn from the alphabet $\mathcal{L}$ of size 2^Q, these numbers are as follows:

1. A variable node of degree d_v requires $d_v(d_v - 2)2^Q$ multiplications (a message consists of 2^Q values, d_v messages must be computed, $d_v - 2$ multiplications per message) for the messages sent to other function nodes and $(d_v - 1)2^Q$ multiplications to compute the marginal $m(\cdot)$.

[4] Surprisingly, in many applications, the *loopy* version of the belief propagation does yield sufficiently accurate results [58].

2. A function node of degree d_f requires $d_f(d_f-1)2^{Qd_f}$ multiplications and $d_f(2^{Q(d_f-1)}-1)$ additions [5].

These complexity counts hold for graphs with cycles as well, but the number of operations scales with the number of iterations performed during message passing. From these formulas we can obtain the complexity count for message passing on the factor graph that represents the function $p(\mathbf{i}) \cdot \prod_{k=1}^{M} p(\mathbf{y}_k|i_k)$, corresponding to the MMSE-optimal decoder outlined in Section 5.2.2, as shown in *Table 5.1*.

Table 5.1: Complexity count for the optimal MMSE decoder.

	additions	multiplications
M variable nodes	0	$M2^Q$
M func. nodes for $p(\mathbf{y}_k|i_k)$	0	0
function node for $p(\mathbf{i})$	$M(2^{Q(M-1)}-1)$	$M(M-1)2^{QM}$

5.3.2 Scalable Decoding on Factor Trees

The large number of operations required to compute $m(i_k)$ for all k can decrease tremendously if $p(\mathbf{i})$ factors into functions with *small* numbers of arguments (indices i_k) yielding function nodes with small degree. There are general ways to factorize a joint PMF such as the chain rule, e.g.

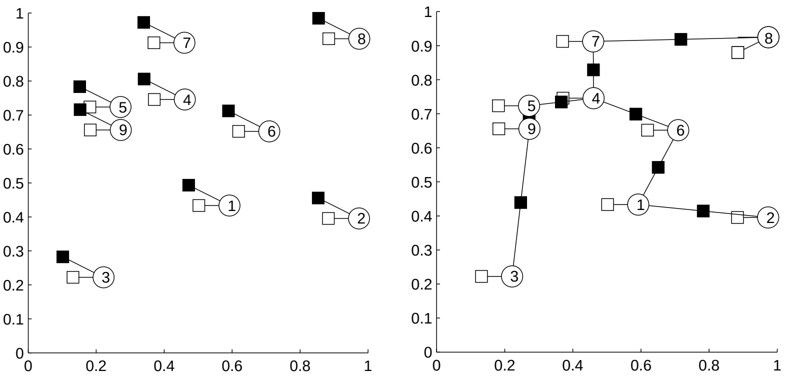
$$p(\mathbf{i})=p(i_1)p(i_2|i_1)...p(i_M|i_1 \dots i_{M-1}),$$

where each factor is again a PMF. However, some factors in this factorization still evidence a large degree (up to degree M) and the factor graph contains cycles, so that the BP algorithm cannot be exact. To overcome this drawback, we propose factorizations which yield a *factor tree*. As explained above, for this class of factor graphs the message passing algorithm is exact (i.e., it computes $m(i_k)$ correctly). Moreover, we can restrict the connectivity of the factor tree by limiting the function nodes to have a prescribed degree. For example, if we factorize according to $p(\mathbf{i})=g_1(i_1)g_2(i_2,i_1)...g_M(i_M,i_{M-1})$ for some functions $g_k(\cdot)$, we get a chain-like factor tree, whose function nodes have degree at most two.

In most cases the PMF $p(\mathbf{i})$ derived from $p(\mathbf{u})$ will not have a structure leading to such a factorization, and we must seek an approximate source distribution $\hat{p}(\mathbf{u})$ that does lead to a PMF $\hat{p}(\mathbf{i})$ with the desired properties. It is important to note that since a particular index i_k depends only on the source symbol u_k (scalar quantization), any factorization of $\hat{p}(\mathbf{i})$ implies the same factorization of $\hat{p}(\mathbf{u})$ and vice-versa. Consider the following factorizations of $\hat{p}(\mathbf{u})$ and $\hat{p}(\mathbf{i})$, which will be addressed in more detail in Section 5.4:

Scalar decoding: Suppose that $\hat{p}(\mathbf{u})$ and $\hat{p}(\mathbf{i})$ factor into M functions with a single argument:

$$\hat{p}(\mathbf{u})=\prod_{k=1}^{M} f_k(u_k) \quad \text{and} \quad \hat{p}(\mathbf{i})=\prod_{k=1}^{M} g_k(i_k).$$



(a) Degree 1 (b) Degree 1 and 2 (c) Degree 1, 2 and 3

Figure 5.3: Factor trees similar to that in *Figure 5.2* for the functions (a) $\prod_{k=1}^{M} p(\mathbf{y}_k|i_k)g_k(i_k)$ (degree 1 function nodes), (b) $\prod_{k=1}^{M} p(\mathbf{y}_k|i_k)g_k(i_{p_k}, i_{q_k})$ (degree 1 and 2 function nodes), and (c) $\prod_{k=1}^{M} p(\mathbf{y}_k|i_k)g_k(i_{p_k}, i_{q_k}, i_{r_k})$ (degree 1,2, and 3 function nodes).

The corresponding approximate marginals $\hat{m}(i_k)$ are given by $\prod_{k=1}^{M} p(\mathbf{y}_k|i_k)g_k(i_k)$, which yields the factor tree in *Figure 5.3*(a) for our example sensor network with $M = 9$ sensors. Obviously, no information about the correlations between sensors is considered. The complexity count is shown in *Table 5.2*.

Table 5.2: Complexity count for the scalar decoder.

	additions	multiplications
M variable nodes	0	$M2^Q$
M function nodes for $p(\mathbf{y}_k|i_k)$	0	0
M function nodes for $g_k(i_k)$	0	0

Decoding with degree-2 factor tree: Suppose that $\hat{p}(\mathbf{u})$ and $\hat{p}(\mathbf{i})$ factor into $M-1$ functions with 2 arguments:

$$\hat{p}(\mathbf{u}) = \prod_{k=1}^{M-1} f_k(u_{p_k}, u_{q_k}) \quad \text{and} \quad \hat{p}(\mathbf{i}) = \prod_{k=1}^{M-1} g_k(i_{p_k}, i_{q_k}),$$

where p_k indicates the index of the main variable and q_k specifies the index of the conditioning variable, e.g. with $(p_k, q_k) = (3, 1)$ we get $f_k(u_{p_k}, u_{q_k}) = p(u_3|u_1)$. Exactly $M-1$ such factors are required to connect all M sensors in a tree. The corresponding approximate marginals $\hat{m}(i_k)$ are given by $\prod_{k=1}^{M-1} p(\mathbf{y}_k|i_k)g_k(i_{p_k}, i_{q_k})$. A possible factor tree for this factorization as shown in *Figure 5.3*(b) for the index pairs $(1, 2)$, $(1, 6)$, $(4, 6)$, $(3, 9)$, $(5, 9)$, $(4, 5)$, $(7, 8)$, and $(4, 7)$. *Table 5.3* shows the corresponding complexity count.

There are numerous other factorizations yielding different complexity counts. One possibility is to increase the admissible degree of the function nodes. As an example, consider the factor

Table 5.3: Complexity count for the degree-2 factor graph decoder.

	additions	multiplications
M variable nodes	0	$M2^Q$
M function nodes for $p(\mathbf{y}_k \vert i_k)$	0	0
$M-1$ function nodes for $g_k(i_{p_k}, i_{q_k})$	$2(M-1)(2^Q-1)$	$2(M-1)2^{2Q}$

tree shown in *Figure 5.3*(c), where $\hat{p}(\mathbf{i})$ factors into

$$g_1(i_1, i_2)g_2(i_3, i_9)g_3(i_1, i_4, i_6)g_4(i_4, i_5, i_9)g_5(i_4, i_7, i_8),$$

i.e. the largest admissible degree is 3. If we do allow graphs with cycles or clusters of variable nodes, we end up with a very large class of factors graphs, which admit an iterative BP algorithm [16, 60].

5.4 Model Optimization

The performance of the scalable decoder proposed in the previous section naturally depends on how well the factor graph model is able to approximate the given source distribution. In this section, we will provide a set of optimization tools that allow us to find adequate factorizations of the joint probability distribution $\hat{p}(\mathbf{u})$. We begin with a brief justification for the chosen optimization criterion.

5.4.1 Optimization Criterion

According to the problem statement in Section 5.2, the ultimate goal of our decoder is to minimize the mean square error between the transmitted samples and the reconstructed values. For this purpose, we require the approximated distribution used by the decoder to be as *close* as possible to the original distribution, while observing the imposed complexity constraint. As we have seen in Section 5.3.2, the number of operations required to produce the desired estimates can be computed directly from the chosen factor graph — thus, we can meet the target complexity by selecting an adequate factor graph or, equivalently, an adequate *constrained factorization*.

Thus, to solve the problem we must determine the functions $f_k(\cdot)$ of said factorization that yield the best possible approximation $\hat{p}(\mathbf{u})$ of the source distribution $p(\mathbf{u})$. A useful measure for this task is the Kullback-Leibler distance (KLD) between $p(\mathbf{u})$ and $\hat{p}(\mathbf{u})$ given by

$$D(p(\mathbf{u})||\hat{p}(\mathbf{u})) = \int p(\mathbf{u}) \log \frac{p(\mathbf{u})}{\hat{p}(\mathbf{u})}\, d\mathbf{u}, \tag{5.5}$$

measured in bits [28, Section 9.5].

Our motivation for using the KLD as the optimization criterion for this instance of the problem comes from previous work by Li, Chaddha, and Gray on fixed-rate[5] vector quantization with mismatched codebooks [62, Section 6]. Their main result can be summarized as follows:

[5]This result was later partially extended to entropy-constrained vector quantization by Gray and Linder [41].

if the quantizer is optimized for a model probability distribution $\hat{p}(\mathbf{u})$ instead of the true source distribution $p(\mathbf{u})$, the resulting excess quadratic distortion in decibels is proportional to the KLD between $\hat{p}(\mathbf{u})$ and $p(\mathbf{u})$ given by (5.5). As a rule of thumb, the authors indicate that every one bit difference in the KLD leads asymptotically to a 6 dB loss in performance. Evidently, this result does not apply directly to our reachback system, since in our case the source coding is done by an array of scalar quantizers that process each sample u_k individually and *not* by a vector quantizer operating on the whole block of samples $\mathbf{u}$. However, it is worth pointing out that the vector quantizer approach does correspond to the case of *full cooperation* between the sensors – i.e. every node knows *all* the source realizations observed by *all* the nodes. Therefore, it is perfectly reasonable to view the performance of a vector quantizer processing $\mathbf{u}$ as an upper bound to the fidelity achieved by our coding scheme, and for this upper bound we know from [62] that the loss in mean square error is proportional to the KLD between $\hat{p}(\mathbf{u})$ and $p(\mathbf{u})$.

Indeed, our numerical results (discussed in detail in Section 5.5) sustain a similar useful connection between the KLD for $p(\mathbf{u})$ and $\hat{p}(\mathbf{u})$ and the mean square error of our decoder. Naturally, we would like to have a mathematical proof that quantifies this relationship, but unfortunately this has proved to be a difficult task due to the non-linearity of the array of scalar quantizers. Therefore, we must leave it as a challenging problem for future work.

5.4.2 Constrained Chain Rule Expansions

Before proceeding with a detailed description of our optimization algorithms, we must introduce a few mathematical tools that prove very useful for the problem at hand. Recall that our goal is to find a set of factors $f_k(\cdot)$ for $p(\mathbf{u})$ that obeys certain constraints and ultimately yields the approximated distribution $\hat{p}(\mathbf{u}) = \prod_k f_k(\cdot)$ used by the decoder. The next example introduces a very useful concept for this task.

***Example* 5.1** *Let* $p(\mathbf{u}) = p(u_1 u_2 \dots u_5)$ *admitting the chain rule expansion given by*

$$p(u_1 \dots u_5) = p(u_1)p(u_2|u_1)p(u_3u_4|u_1u_2)p(u_5|u_1u_2u_3u_4).$$

We can obtain an approximate PDF $\hat{p}(\mathbf{u}) = \hat{p}(u_1 \dots u_5)$ *of* $p(\mathbf{u})$ *by taking the factors of the chain rule expansion and removing some of the conditioning variables. For instance, the expansion*

$$\hat{p}(u_1 \dots u_5) = p(u_1)p(u_2|u_1)p(u_3u_4|u_2)p(u_5|u_4) \tag{5.6}$$

is a constrained chain rule expansion with at most one conditioning variable.

The next definition makes this concept more precise.

Definition 5.1 Consider a PDF $\hat{p}(\mathbf{u})$, which factors into N_f PDFs $p(\mathbf{a}_k|\mathbf{b}_k)$ according to

$$\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} p(\mathbf{a}_k|\mathbf{b}_k), \tag{5.7}$$

where $\mathbf{a}_k$ and $\mathbf{b}_k$ are subsets of the elements in $\mathbf{u}$. This PDF is a constrained chain rule expansion (CCRE) of the source distribution $p(\mathbf{u})$, if the following constraints are met:

1. Two vectors $\mathbf{a}_k$ and $\mathbf{a}_l$, $k \neq l$, are disjoint: $\mathbf{a}_k \cap \mathbf{a}_l = \emptyset$.

2. The elements in $\mathbf{b}_k$ are connected: $\mathbf{b}_k \subseteq \bigcup_{l=1}^{k-1} \mathbf{a}_l$.

3. All elements u_k of $\mathbf{u}$ are connected: $\bigcup_{k=1}^{N} \mathbf{a}_k = \mathbf{u}$.

Thus, the set $\mathbf{b}_1$ is always empty. A special case is the usual chain rule expansion, where $\mathbf{b}_k = \bigcup_{l=1}^{k-1} \mathbf{a}_l$ holds.

In the previous example, the CCRE $p(u_1)p(u_2|u_1)p(u_3, u_4|u_2)p(u_5|u_4)$ of the PDF $p(u_1, ..., u_5)$ is specified by $\mathbf{a}_1 = u_1$, $\mathbf{b}_1 = \emptyset$, $\mathbf{a}_2 = u_2$, $\mathbf{b}_2 = u_1$, $\mathbf{a}_3 = (u_3, u_4)$, $\mathbf{b}_3 = u_2$, $\mathbf{a}_4 = u_5$, and $\mathbf{b}_4 = u_4$. The next definition introduces another useful property, which we also illustrate with an example.

Definition 5.2 The CCRE $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} p(\mathbf{a}_k|\mathbf{b}_k)$ is said to be *symmetric*, if any $\mathbf{b}_k$, $k = 2, 3, ..., N$, is a subset of $(\mathbf{a}_l, \mathbf{b}_l)$ for some $l < k$.

***Example* 5.2** *The CCRE given by*

$$p(u_1)p(u_2|u_1)p(u_3, u_4|u_2)p(u_5|u_4)$$

of the PDF $p(u_1, ..., u_5)$ *is symmetric because* $\mathbf{b}_2 \subset \mathbf{a}_1$, $\mathbf{b}_3 \subset \mathbf{a}_2$, *and* $\mathbf{b}_4 \subset \mathbf{a}_3$ *holds. The CCRE corresponding to*

$$p(u_1, u_2)p(u_3, u_4|u_2)p(u_5|u_4, u_1)$$

of $p(u_1, ..., u_5)$ *is not symmetric, because* $\mathbf{b}_3 = (u_1, u_4)$ *is not contained in* $(\mathbf{a}_1, \mathbf{b}_1) = (u_1, u_2)$ *or* $(\mathbf{a}_2, \mathbf{b}_2) = (u_2, u_3, u_4)$.

It turns out that symmetric CCREs of the source distribution $p(\mathbf{u})$ yield the type of factor trees that we are interested in:

Lemma 5.1 If a CCRE $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} p(\mathbf{a}_k|\mathbf{b}_k)$ for the source distribution $p(\mathbf{u})$ has at most one conditioning variable in every factor, i.e., all $\mathbf{b}_k$ are either empty or contain a single element, then (1) the CCRE is symmetric, and (2) the factor graph corresponding to $\hat{p}(\mathbf{u})$ is a tree.

Proof: See the appendix, Section A.8.

From this lemma follows, for example, that a CCRE $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} p(\mathbf{a}_k|\mathbf{b}_k)$ in which both the $\mathbf{a}_k$ and the $\mathbf{b}_k$ consist of single elements yield a factor tree where all function nodes have degree 2. The next example illustrates the relationship between the different classes of chain rule expansions and their corresponding factor graphs.

***Example* 5.3** *Consider the probability distribution* $p(\mathbf{u}) = p(u_1 u_2 u_3 u_4)$ *admitting the standard chain rule expansion*

$$p(\mathbf{u}) = p(u_1 u_2 u_3 u_4) = p(u_1 u_2)p(u_3|u_1 u_2)p(u_4|u_1 u_2 u_3) \tag{5.8}$$

The corresponding factor graph, depicted in Figure 5.4, is inconvenient for decoding — it contains many cycles, for instance $(u_1, p(u_3|u_1 u_2), u_2, p(u_1 u_2), u_1)$. *A better alternative would be*

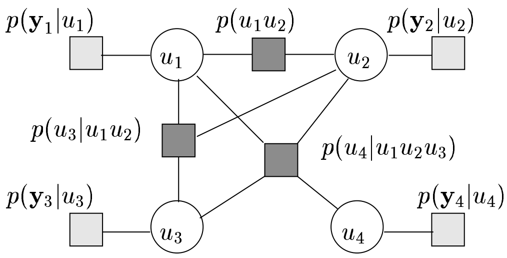


Figure 5.4: Corresponding factor graph for $p(\mathbf{u})$ in (5.8).

the factor graph consisting of one function node $p(u_1u_2u_3u_4)$, which is equivalent yet has less function nodes and no cycles at all.
Next, consider the symmetric CCRE of $p(\mathbf{u})$ given by

$$\hat{p}(\mathbf{u}) = p(u_1u_2)p(u_3|u_1)p(u_4|u_3) \tag{5.9}$$

Figure 5.5 shows the corresponding factor graph, which is now symmetric. Notice that it is possible to find equivalent CCRE containing $p(u_1u_3)$ or $p(u_3u_4)$.

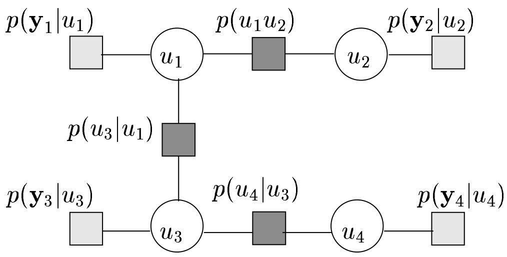


Figure 5.5: Corresponding factor graph for $\hat{p}_{\text{s}}(\mathbf{u})$ in (5.9).

Finally, if we have a non-symmetric CCRE of $p(\mathbf{u})$, such as

$$\hat{p}(\mathbf{u}) = p(u_1u_2)p(u_3|u_1)p(u_4|u_2u_3), \tag{5.10}$$

we get a non-symmetric factor graph, as shown in Figure 5.6. Clearly, it is not possible to find an equivalent CCRE containing $p(u_2u_3u_4)$.

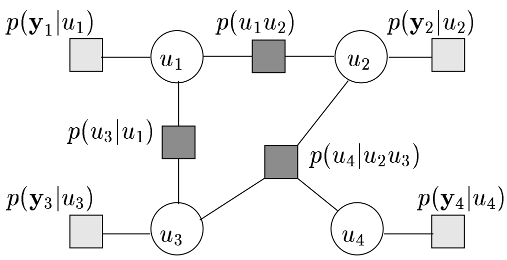


Figure 5.6: Corresponding factor graph for $\hat{p}_{\text{ns}}(\mathbf{u})$ in (5.10).

5.4.3 Constrained Factorization of Gaussian Distributions

The definitions and properties discussed in the previous section are very general and can in fact be applied to any given PDF. Given the nature of our source model, we are of course particularly interested in CCREs of multivariate Gaussian distributions given by $\mathcal{N}(\mathbf{0}_M, \mathbf{R})$. Ultimately, we would like to compute the approximate covariance matrix $\hat{\mathbf{R}}$ induced by a given CCRE. The next lemma presents a few very useful properties for this purpose.

Lemma 5.2 Let $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} p(\mathbf{a}_k|\mathbf{b}_k)$ be a CCRE of a Gaussian PDF $p(\mathbf{u})$ that is given by $\mathcal{N}(\mathbf{0}_M, \mathbf{R})$. Let $\mathbf{P}$ be an $M \times M$ indicator matrix, whose entry in the l-th row and l'-th column is 1 if both u_l and $u_{l'}$ are contained in one of the N_f factors $p(\mathbf{a}_k|\mathbf{b}_k)$ and 0 otherwise. For example, for the CCRE $p(u_1)p(u_2|u_1)p(u_3, u_4|u_2)p(u_5|u_4)$ of $p(u_1, ..., u_5)$ we find

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

The following holds:

1. The PDF $\hat{p}(\mathbf{u})$ is a zero-mean Gaussian PDF with covariance matrix $\hat{\mathbf{R}}$, i.e., it is given by $\mathcal{N}(\mathbf{0}_M, \hat{\mathbf{R}})$.

2. The entries of $\hat{\mathbf{R}}^{-1}$ are zero for all zero-positions in $\mathbf{P}$.

3. The trace of $\mathbf{R}\hat{\mathbf{R}}^{-1}$ equals M, i.e., $\mathrm{tr}(\mathbf{R}\hat{\mathbf{R}}^{-1}) = M$.

4. If the CCRE is symmetric, then the entries of $\hat{\mathbf{R}}$ are equal those in $\mathbf{R}$ for all one-positions in $\mathbf{P}$.

Proof: See the appendix, Section A.9.

In the appendix, Section A.10, we present a numerical example to illustrate how $\hat{\mathbf{R}}$ can be computed for a given CCRE.

Based on Lemma 5.2, we can prove the following connection between symmetric CCREs and the KLD-optimal functions $f_k(\cdot)$ of the factorization $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} f_k(\cdot)$ that minimize the KLD $D(p(\mathbf{u})||\hat{p}(\mathbf{u}))$:

Theorem 5.3 Consider the Gaussian source distribution $p(\mathbf{u})$ given by $\mathcal{N}(\mathbf{0}_M, \mathbf{R})$ and the PDF $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} f_k(\mathbf{u}_k)$, which factors into N_f functions $f_k(\mathbf{u}_k)$ with subsets $\mathbf{u}_k$ of $\mathbf{u}$ as argument. If the latter factorization admits a symmetric CCRE, i.e., all $\mathbf{u}_k$ can be split into pairs $(\mathbf{a}_k, \mathbf{b}_k)$ satisfying the constraints in Definitions 5.1 and 5.2, then the KLD-optimal functions $f_k(\mathbf{u}_k)$ minimizing the $D(p(\mathbf{u})||\hat{p}(\mathbf{u}))$ are equal to the Gaussian PDFs

$$p(\mathbf{a}_k|\mathbf{b}_k) = p(\mathbf{a}_k, \mathbf{b}_k)/p(\mathbf{b}_k),$$

and the corresponding minimal KLD is given by

$$D(p(\mathbf{u})||\hat{p}(\mathbf{u})) = -\frac{1}{2}\log_2|\mathbf{R}| + \frac{1}{2}\sum_{k=1}^{N_f}\log_2\frac{|\mathbf{R}_{\mathbf{a}_k,\mathbf{b}_k}|}{|\mathbf{R}_{\mathbf{b}_k}|}, \tag{5.11}$$

where $\mathbf{R}_{\mathbf{a}_k,\mathbf{b}_k}$ and $\mathbf{R}_{\mathbf{b}_k}$ are the covariance matrices of the zero-mean Gaussian PDFs $p(\mathbf{a}_k, \mathbf{b}_k)$ and $p(\mathbf{b}_k)$, respectively.

Proof: See the appendix, Section A.11.

This theorem considerably simplifies our search for KLD-optimal approximate source distributions $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} f_k(\mathbf{u}_k)$ that yield factor trees with function nodes of degree at most 1, 2, or 3, by allowing us to restrict our attention to the set of symmetric CCREs and determine step by step the factor arguments $\mathbf{a}_k$ and $\mathbf{b}_k$ that minimize the KLD. Moreover, it follows from (5.11) that each factor $p(\mathbf{a}_k|\mathbf{b}_k)$ of $\hat{p}(\mathbf{u})$ reduces the KLD $D(p(\mathbf{u})||\hat{p}(\mathbf{u}))$ by the amount $\log_2|\mathbf{R}_{\mathbf{a}_k,\mathbf{b}_k}|/|\mathbf{R}_{\mathbf{b}_k}|$, which is strictly negative, because in general $|\mathbf{R}_{\mathbf{a}_k,\mathbf{b}_k}| < |\mathbf{R}_{\mathbf{b}_k}|$ holds. Moreover, this amount depends only on the determinants of the matrices $\mathbf{R}_{\mathbf{u}_k}$ and $\mathbf{R}_{\mathbf{b}_k}$ and is thus independent of all other factors. This important property is used extensively by the optimization algorithms presented in the next section. We conclude this part with a simple numerical example.

***Example* 5.4** *Suppose we are given a Gaussian source distribution $p(\mathbf{u}) = p(u_1u_2u_3u_4)$ with zero mean and the covariance matrix*

$$\mathbf{R} = \begin{bmatrix} 1 & 0.7 & 0.5 & 0.2 \\ 0.7 & 1 & 0.6 & 0.6 \\ 0.5 & 0.6 & 1 & 0.3 \\ 0.2 & 0.6 & 0.3 & 1 \end{bmatrix}. \tag{5.12}$$

For the standard chain rule expansion

$$p(\mathbf{u}) = p(u_1u_2u_3u_4) = p(u_1u_2)p(u_3|u_1u_2)p(u_4|u_1u_2u_3),$$

we can compute the KLD according to

$$\begin{aligned} D(p(\mathbf{u})||p(\mathbf{u})) &= -\frac{1}{2}\log|\mathbf{R}| + \frac{1}{2}\sum_{k=1}^{3}\log\frac{|\mathbf{R}_{\mathbf{u}_k}|}{|\mathbf{R}_{\mathbf{b}_k}|} \\ &= 1.26 - 0.49 - 0.33 - 0.44 \\ &= 0 \textit{ bits.} \end{aligned}$$

As expected, every factor decreases the KLD by a certain amount, and the overall KLD turns out to be zero, because the standard chain rule expansion is equivalent to the given distribution $p(\mathbf{u})$.

Assume now that we have instead the symmetric CCRE

$$\hat{p}(\mathbf{u}) = p(u_1u_2)p(u_3|u_1)p(u_4|u_3).$$

If we carry out the previous KLD computation, we now get

$$\begin{aligned} D(p(\mathbf{u})||\hat{p}(\mathbf{u})) &= -\frac{1}{2}\log|\mathbf{R}| + \frac{1}{2}\sum_{k=1}^{3}\log\frac{|\mathbf{R}_{\mathbf{u}_k}|}{|\mathbf{R}_{\mathbf{b}_k}|} \\ &= 1.26 - 0.49 - 0.20 - 0.07 \\ &= 0.50 \textit{ bits.} \end{aligned}$$

As expected, every factor decreases the KLD by a certain amount, however since $\hat{p}(\mathbf{u})$ is an approximation of $p(\mathbf{u})$, the KLD is greater than zero, giving us a measure for the accuracy of the approximation.

5.4.4 Optimization Algorithms

In the previous section, we proved for Gaussian sources that symmetric CCREs of the source distribution $p(\mathbf{u})$ yield the KLD-optimal functions $f_k(\cdot)$ of the factorization $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} f_k(\mathbf{u}_k)$ provided that the arguments $\mathbf{u}_k$ admit a symmetric CCRE. We also showed that factorizations yielding a factor tree with function nodes of degree 1, 2, or 3 always admit symmetric CCREs. Nevertheless, there exist many factor trees that connect the M variable nodes for each sensor in the network and so the problem becomes finding the factor tree for which the underlying symmetric CCRE $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} p(\mathbf{a}_k|\mathbf{b}_k)$ yields the smallest KLD $D(p(\mathbf{u})||\hat{p}(\mathbf{u}))$.

Let l_a and l_b denote the allowed maximal number of elements in the sets $\mathbf{a}_k$ and $\mathbf{b}_k$ of the CCRE $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} p(\mathbf{a}_k|\mathbf{b}_k)$, respectively. Recall that the algorithmic complexity for scalable decoding based on sum-product decoding on a factor tree grows exponentially with the degree $d_f = l_a + l_b$ of the function nodes, which is why we consider factor trees with $d_f \leq 3$ only, as specified in Section 5.3.2.

Besides the trivial scalar decoder corresponding to the symmetric CCRE $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} p(u_k)$, i.e., $(l_a, l_b) = (1, 0)$, we consider decoders based on the choice $(l_a, l_b) = (1, 1)$ or $(l_a, l_b) = (2, 1)$, which are based on factor trees with function node degrees of at most 2 or 3. From Lemma 5.1 follows that symmetric CCREs generate such factor trees when $l_b = 1$, i.e. when the factors $p(\mathbf{a}_k|\mathbf{b}_k)$ of the CCRE contain only a single conditioning variable.

Next, we provide optimization algorithms for these two classes of scalable decoders.

1) Factor Tree with Degree-2 Function Nodes

A symmetric CCRE $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} p(\mathbf{a}_k|\mathbf{b}_k)$ yields a degree-2 factor tree if $(l_a, l_b) = (1, 1)$ with $N = M - 1$. Starting with the trivial factorization $\hat{p}(\mathbf{u}) = \prod_{k=0}^{M-1} p(u_{r_k})$, where $\{r_0, ..., r_{M-1}\}$ is a permutation of the index set $\{1, ..., M\}$, admissible CCREs are constructed by adding conditioning variables to $M-1$ of these factors, i.e.,

$$\hat{p}(\mathbf{u}) = p(u_{r_0}) \prod_{k=1}^{M-1} p(u_{r_k}|u_{s_k}),$$

where the index s_1 is necessarily equal to r_0 and all other indices s_k are chosen from the set $\{1, ..., M\}$. Combining the PDFs $p(u_{r_0})=p(u_{s_1})$ and $p(u_{r_1}|u_{s_1})$ yields the CCRE

$$\hat{p}(\mathbf{u}) = p(u_{r_1}, u_{s_1}) \prod_{k=2}^{M-1} p(u_{r_k}|u_{s_k}),$$

consisting of $N = M-1$ factors with two arguments. The corresponding index factorization $\hat{p}(\mathbf{i})=\prod_{k=1}^{N_f} g_k(\cdot)$ is given by

$$\hat{p}(\mathbf{i}) = p(i_{r_1}, i_{s_1}) \prod_{k=2}^{M-1} p(i_{r_k}|i_{s_k}).$$

The calculation of the KLD $D(p(\mathbf{u})||\hat{p}(\mathbf{u}))$ via (5.11) requires the *local* covariance matrices

$$\mathbf{R}_{u_{r_k},u_{s_k}} = \begin{bmatrix} 1 & \rho_{r_k,s_k} \\ \rho_{r_k,s_k} & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{R}_{u_{s_k}} = 1, \tag{5.13}$$

which follow from the entries $\rho_{k,k'}$ of the covariance matrix $\mathbf{R}$ of the source distribution $p(\mathbf{u})$, such that

$$\begin{aligned} D(p(\mathbf{u})||\hat{p}(\mathbf{u})) &= -\frac{1}{2}\log_2|\mathbf{R}| + \frac{1}{2}\log_2|\mathbf{R}_{u_{r_1,s_1}}| + \frac{1}{2}\sum_{k=2}^{M-1}\log_2\frac{|\mathbf{R}_{u_{r_k},u_{s_k}}|}{|\mathbf{R}_{u_{s_k}}|} \\ &= -\frac{1}{2}\log_2|\mathbf{R}| + \frac{1}{2}\sum_{k=1}^{M-1}\log_2(1-\rho^2_{r_k,s_k}). \end{aligned}$$

Notice that a function node connecting the variable nodes i_{r_k} and i_{s_k} decreases the KLD by

$$\Delta D_{1|1} = \frac{1}{2}\log_2(1-\rho^2_{r_k,s_k}), \tag{5.14}$$

corresponding to the factor $p(u_{r_k}|u_{s_k})$. We denote the first decrease due to the factor $p(u_{r_1}, u_{s_1})$ as $\Delta D_0 = \frac{1}{2}\log_2(1-\rho^2_{r_1,s_1})$.

The function nodes corresponding to $p(u_{r_1}, u_{s_1})$ or $p(u_{r_k}|u_{s_k})$ can be regarded as vertices in a classical graph connecting the (variable) nodes i_{r_k} and i_{s_k}, which have the undirected weight $\frac{1}{2}\log_2(1-\rho^2_{r_k,s_k})$. Our optimization task — finding the factor tree arguments $\mathbf{a}_k$ and $\mathbf{b}_k$ for the factors $p(\mathbf{a}_k|\mathbf{b}_k)$ yielding a minimal KLD — can thus be formulated as a minimum weight spanning tree problem where the undirected weight of an edge between two nodes i_{r_k} and i_{s_k} is given by $\frac{1}{2}\log_2(1-\rho^2_{r_k,s_k})$. To find this tree, we applied Algorithm 1 (see Table 5.4), which is an adaptation of Prim's minimum weight spanning tree algorithm taken from [40]. The algorithm finds the optimal tree in the M^{M-2} possible trees [89] with a very low complexity. *Figure 5.7* shows the outcomes of the proposed algorithm for sensor networks with $M = 9$ and $M = 100$ nodes using the source model outlined in Section 5.2.1.

Table 5.4: Algorithm 1: Find the optimal degree-2 factor tree

Initialization

Function node counter: $k \leftarrow 1$

Find the two variable nodes $[i_{r_1} \quad i_{s_1}]$

- which yield the smallest $\Delta D_0 = \frac{1}{2}\log_2(1 - \rho^2_{r_1,s_1})$

Connect i_{r_1} and i_{s_1} by a new function node with function $g_1 = p(i_{r_1}, i_{s_1})$

Set of connected variable nodes: $\mathcal{S} \leftarrow \{i_{r_1}, i_{s_1}\}$

Main Loop

repeat

$k \leftarrow k + 1$

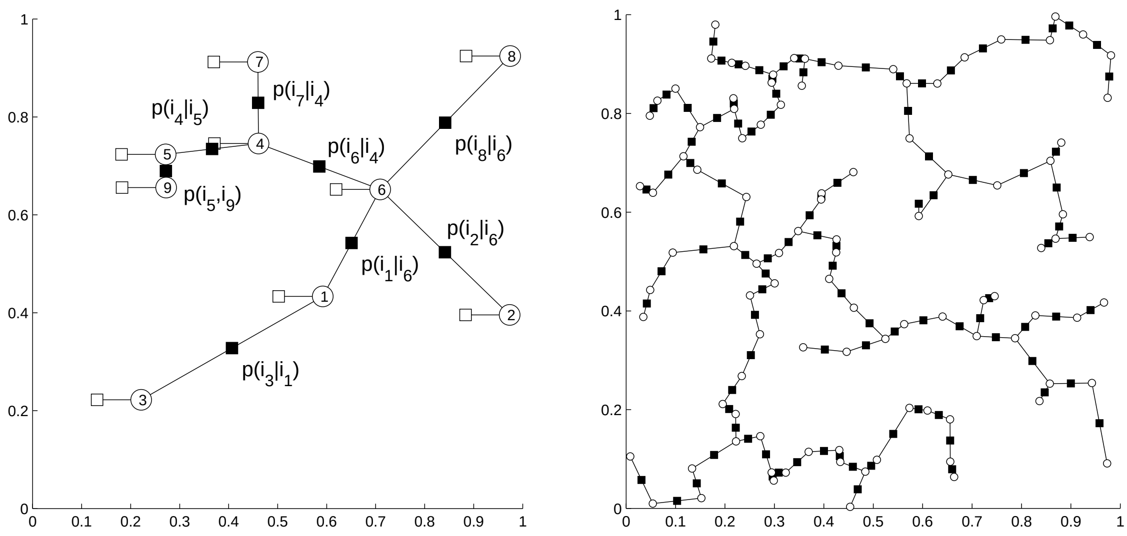
Find the two variable nodes $[i_{r_k} \quad i_{s_k}]$

- where $i_{s_k} \in \mathcal{S}$
- where $i_{r_k} \notin \mathcal{S}$
- which yield the smallest $\Delta D_{1|1} = \frac{1}{2}\log_2(1 - \rho^2_{r_k,s_k})$

Connect i_{r_k} and i_{s_k} by a new function node with function $g_k = p(i_{r_k}|i_{s_k})$

$\mathcal{S} \leftarrow \mathcal{S} + \{i_{r_k}\}$

until all variable nodes are connected



(a) $M = 9$ sensors (b) $M = 100$ sensors

Figure 5.7: Degree-2 factor trees for M sensors placed randomly on the unit square, according to the source model described in Section 5.2.1.

2) Degree-3 Factor Trees

The optimization procedure for the previous case turned out to be relatively simple, because degree-2 factor trees can be interpreted as classical graphs and we could exploit well-established graph-theoretic techniques. Unfortunately, this is not true for degree-3 factor trees, forcing us to seek an alternative solution.

In analogy with the previous case, we begin by rewriting (5.7) specifically for degree-3 factor trees according to

$$\hat{p}(\mathbf{u}) = p(u_{r_0}) \prod_{k=2}^{(M-1)/2} p(u_{r_k}, u_{s_k}|u_{t_k}),$$

where $\mathbf{a}_1 = u_{r_0}$, $\mathbf{a}_k = [u_{r_{k-1}} \quad u_{s_{k-1}}]$ (for $k > 1$) and $\mathbf{b}_k = u_{r_{k-1}}$. In practice, it is not always possible or useful to construct a degree-3 factor tree that consists solely of degree-3 function nodes, however to simplify the explanation we will neglect the additional degree-2 function nodes and assume that $(M-1)/2$ is a natural number.

Once again, we require the local covariance matrices

$$\mathbf{R}_{\mathbf{u}_{k+1}} = \mathbf{R}_{r_k,s_k,t_k} = \begin{bmatrix} 1 & \rho_{r_k,s_k} & \rho_{r_k,t_k} \\ \rho_{r_k,s_k} & 1 & \rho_{s_k,t_k} \\ \rho_{r_k,t_k} & \rho_{s_k,t_k} & 1 \end{bmatrix} \tag{5.15}$$

and $\mathbf{R}_{\mathbf{b}_{k+1}} = \mathbf{R}_{s_k} = 1$, where $\rho_{r_k,s_k} = \mathbf{R}(r_k, s_k)$ denotes the covariance between u_{r_k} and u_{s_k}. Now, we can calculate the KLD using (5.11), which results in

$$\begin{aligned} D(p(\mathbf{u})||\hat{p}(\mathbf{u})) &= -\frac{1}{2}\log_2|\mathbf{R}| + \frac{1}{2}\sum_{k=1}^{(M-1)/2}\log_2\frac{|\mathbf{R}_{r_k,s_k,t_k}|}{|\mathbf{R}_{s_k}|} \\ &= -\frac{1}{2}\log_2|\mathbf{R}| + \frac{1}{2}\sum_{k=1}^{(M-1)/2}\log_2|\mathbf{R}_{r_k,s_k,t_k}|. \end{aligned}$$

Since the degree-3 factor tree cannot be described as a classical graph, we cannot apply a minimum weight spanning tree algorithm. Moreover, the search space is not a matroid, so that we cannot rely on a *greedy algorithm* [89] to deliver an optimal solution. Instead, we propose a suboptimal algorithm that constructs a degree-3 factor tree based on the optimal degree-2 factor tree: First, we try to replace a pair of degree-2 function nodes with one degree-3 function node that reduces the KLD without changing the original structure of the tree (as illustrated in *Figure 5.8*); then, we repeat this procedure over and over again until it is no longer possible to replace any function nodes.

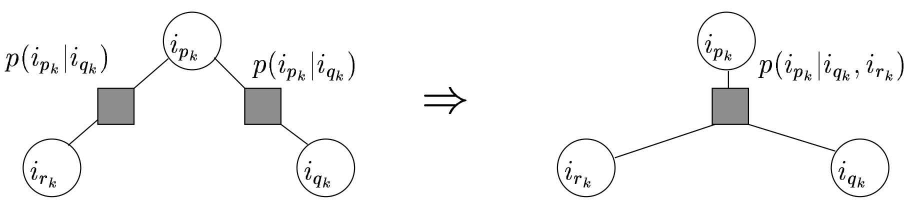


Figure 5.8: The basic procedure in Algorithm 5.5 consists of replacing two degree-2 function nodes by one degree-3 function node. The replacements are chosen according to the associated reduction of the KLD.

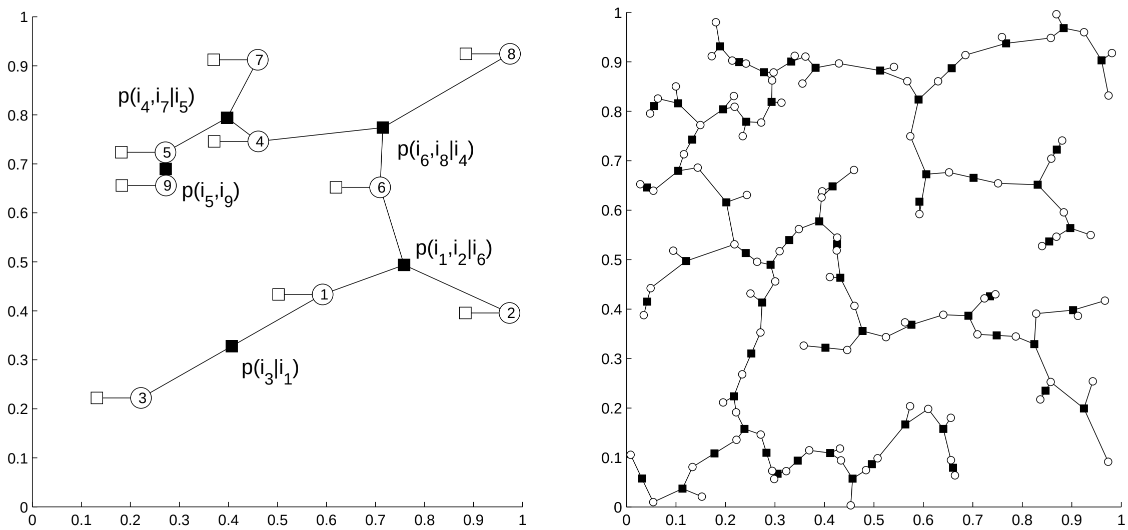
Table 5.5: Algorithm 2: Find the optimized degree-3 factor tree

Initialization

Construct the optimal degree-2 factor tree $\mathcal{T}_2$ with Algorithm 1
Make a list of all combinations of three variable nodes which are neighbours in $\mathcal{T}_2$
Calculate $\Delta D_{2|1} - \Delta D_{2\text{x}1|1} = \frac{1}{2}\log_2\left(\frac{|\mathbf{R}_{s_k,r_k,t_k}|}{|\mathbf{R}_{s_k,r_k}||\mathbf{R}_{r_k,t_k}|}\right)$ for every list entry
sort the list in order of increasing $\Delta D_{2|1} - \Delta D_{2\text{x}1|1}$
Function node counter: $k \leftarrow 0$

Main Loop

repeat
read next row $[i_{r_k} \quad i_{s_k} \quad i_{t_k} \quad \Delta D_{2|1} - \Delta D_{2\text{x}1|1}]$ of list
if connection of i_{r_k}, i_{s_k} and i_{t_k} does not form a cycle **then**
$k \leftarrow k + 1$
remove the two function nodes connecting i_{r_k}, i_{s_k} and i_{t_k}
connect i_{r_k}, i_{s_k} and i_{t_k} by a new function node with function $g_k = p(i_{r_k}, i_{s_k}|i_{t_k})$, where i_{t_k} represents the only conditioning argument in the previous functions
end if
until end of list



(a) $M = 9$ sensors (b) $M = 100$ sensors

Figure 5.9: Degree-3 factor trees for M sensors placed randomly on the unit square, according to the source model described in Section 5.2.1.

In terms of the underlying CCRE, this step is equivalent to replacing the factors $p(u_{s_k}|u_{r_k})$ and $p(u_{r_k}|u_{t_k})$ with the factor $p(u_{s_k}, u_{r_k}|u_{t_k})$. Denoting the correlations between the sensors as ρ_{r_k,s_k}, ρ_{r_k,t_k} and ρ_{s_k,t_k}, we can write the KLD decrement $\Delta D_{2|1}$ associated with the degree-3 function node representing the factor $p(u_{s_k}, u_{r_k}, u_{t_k})$ or $p(u_{s_k}, u_{r_k}|u_{t_k})$ as

$$\Delta D_{2|1} \;=\; \frac{1}{2}\log_2|\mathbf{R}_{s_k,r_k,t_k}| = \frac{1}{2}\log_2(1 + 2\cdot\rho_{r_k,s_k}\cdot\rho_{r_k,t_k}\cdot\rho_{s_k,t_k} - \rho^2_{r_k,s_k} - \rho^2_{r_k,t_k} - \rho^2_{s_k,t_k})$$

On the other hand, the KLD decrement $\Delta D_{2\times 1|1}$ associated with the pair of degree-2 function nodes representing the factors $p(u_{s_k}|u_{r_k})$ and $p(u_{r_k}|u_{t_k})$ is

$$\Delta D_{2\times 1|1} = \frac{1}{2}\log_2|\mathbf{R}_{s_k,r_k}| + \frac{1}{2}\log_2|\mathbf{R}_{r_k,t_k}| = \frac{1}{2}\log_2(1+\rho^2_{r_k,s_k}\cdot\rho^2_{r_k,t_k} - \rho^2_{r_k,s_k} - \rho^2_{r_k,t_k})$$

Consequently, the overall reduction in KLD that results from the substitution is given by

$$\Delta D_{2|1} - \Delta D_{2\times 1|1} \;=\; \frac{1}{2}\log_2\left(\frac{1+2\cdot\rho_{r_k,s_k}\cdot\rho_{r_k,t_k}\cdot\rho_{s_k,t_k} - \rho^2_{r_k,s_k} - \rho^2_{r_k,t_k} - \rho^2_{s_k,t_k}}{1+\rho^2_{r_k,s_k}\cdot\rho^2_{r_k,t_k} - \rho^2_{r_k,s_k} - \rho^2_{r_k,t_k}}\right)$$

This quantity, which is used by the algorithm to choose the appropriate substitutions, has the property that

$$\Delta D_{2|1} - \Delta D_{2\times 1|1} \leq 0 \qquad \text{if} \qquad \frac{\rho_{r_k,t_k}\cdot\rho_{r_k,s_k}}{\rho_{s_k,t_k}} \geq 0 \tag{5.16}$$

with equality when $\rho_{r_k,t_k}\cdot\rho_{r_k,s_k} = \rho_{s_k,t_k}$. It follows that a degree-3 function node always leads to a smaller KLD, except when the variables i_{t_k}, i_{r_k} and i_{s_k} form a Markov chain and $\rho_{t_k,r_k}\rho_{r_k,s_k} = \rho_{t_k,s_k}$. In this case, the two degree-2 factors translate the connection between the variables in an optimal way [45].

The resulting degree-3 factor trees for the previous sensor networks with $M = 9$ and $M = 100$ are shown in *Figure 5.9*, where again we used the source model outlined in Section 5.2.1. Other special cases, including unsymmetric constrained chain rule expansions and factor graphs with cycles, are discussed in detailed in the diploma thesis of Hausl [45].

5.5 Numerical Examples

In this section, we present some numerical results that underline the effectiveness and the scalability of the proposed decoding approach. In order to evaluate the performance of our scalable decoders, we measure the output signal-to-noise ratio (SNR) in decibel (dB) given by

$$\text{Output SNR} = 10\cdot\log_{10}\left(\frac{\|\mathbf{u}\|^2}{\|\mathbf{u}-\hat{\mathbf{u}}\|^2}\right) \text{ in dB} \tag{5.17}$$

versus the channel SNR E_S/N_0 averaged over a sufficient amount of sample transmissions. We consider two cases: $M = 9$ sensors and $M = 100$ sensors. The first case is interesting, because for nine sensors we can still simulate the optimal MMSE decoder and compare it with our proposed scalable decoders. The second case illustrates well the scalability of our approach, whose complexity grows linearly in the number of sensors.

Naturally, the results are highly dependent on the chosen source model. For example, if the source samples are independent there is nothing to gain from trying to exploit the correlations in order to improve the decoding result. On the other hand, if the sensor measurements are highly correlated, we can expect high gains in terms of output SNR. Therefore, we make the following reasonable assumptions, which were justified already in Section 5.2.1: (1) the sensors are placed randomly on the unit square, and (2) the correlation between sensor i and j decays exponentially with the distance between them. Specifically, given the position $\mathbf{z}_m = [z_m(1), z_m(2)]^T$ of each

Table 5.6: KLD in bits per sensor for different approximations and sensor scenarios. The degree-2 factor tree, the degree-3 factor tree and, for reference, the degree-3 factor graph (with cycles) correspond to a first order, a mixed first and second order and a second order Markov approximation, respectively. The scalar decoder uses a zero order Markov approximation.

	$M=9$	$M=100$
scalar decoder, KLD	$6.81 \cdot 10^{-1}$	$7.48 \cdot 10^{-1}$
degree-2 factor tree, KLD	$5.25 \cdot 10^{-2}$	$7.24 \cdot 10^{-2}$
degree-3 factor tree, KLD	$3.93 \cdot 10^{-2}$	$5.78 \cdot 10^{-2}$
degree-3 factor graph, KLD	$9.20 \cdot 10^{-3}$	$1.66 \cdot 10^{-2}$
optimal decoder, KLD	0.00	0.00

sensor m we can compute the distance between two sensors i and j from $l_{i,j} = \|\mathbf{z}_i - \mathbf{z}_j\|$. Based on this distance, we calculate the correlation between their measurements according to

$$\rho_{i,j} = \exp(-\beta \cdot l_{i,j}).$$

The key parameter in this formula is β, a positive constant that determines the speed of the exponential decay.

Notice that if we keep increasing the number of sensors in the unit square without altering the value of β, the sensor measurements would become increasingly correlated. Therefore, to obtain a fair result, we set $\beta = 1.05$ and $\beta = 4.2$ for the simulations with $M = 9$ sensors and $M = 100$ sensors, respectively. Finally, each sensor node uses a Lloyd-Max quantizer to map u_k to i_k, which is then transmitted in accordance with the system setup described in Section 5.2.1.

For each sensor scenario and each class of factor graph models, we carried out the optimization algorithms described in the previous section. The resulting KLD values are summarized in Table 5.6, where for reference we also include the outcomes for a scalar decoder, a degree-3 factor graph with cycles and the optimal decoder. As expected, the higher the complexity of the model and the associated decoding algorithm, the better the approximation of the correlation structure of the sensor data and consequently the lower the resulting KLD.

The overall system performance is illustrated in *Figure 5.10*, which depicts the simulation results for the network with $M = 9$ sensors depicted in *Figure 5.2* and *Figure 5.3*. We used $Q = 1$-bit quantization and the correlation parameter was set to $\beta = 1.05$, as explained above. Clearly, the factor-tree-based decoders (degree-2 and degree-3 tree) are nearly as good as the CME, since the KLD is small. Also, the improvement of the degree-3 tree over the degree-2 tree is barely noticeable. The scalar decoder loses a lot of performance, since it does not exploit any information about the source correlations.

Figure 5.11 depicts the performance results for a network with $M = 100$ sensors with multiple quantizers. The correlation parameter is $\beta = 4.2$ (the sensors are closer and, thus, more correlated). The KLD-optimal degree-2 factor tree is depicted in *Figure 5.9*(b). Again, the KLD of the degree-2 tree is nearly as good that of the degree-3 tree, which also applied to their SNR performance. The gains in output SNR provided by the proposed class of scalable decoders are highest (up to 2 dB) for low channel SNR values (0 dB) and $Q = 3$-bit quantization. Recall that, in this case ($M = 100$), the optimal decoder is unfeasible.

5.6 Summary and Conclusions

We studied the problem of jointly decoding the correlated measurements picked up by a sensor reachback network. First, we showed that the complexity of the optimal MMSE decoder grows exponentially with the number of nodes in the network, thus motivating the search for scalable solutions offering a trade-off between complexity and end-to-end distortion. Then, we presented a scalable decoding scheme for the sensor reachback problem, which uses a simplified factor graph model of the dependencies between the sensor measurements such that a belief propagation decoder can produce the required estimates efficiently.

Focusing on factor trees — for which we know that the BP algorithm delivers optimal estimates — we introduced the concept of constrained chain rule expansions and provided two optimization algorithms for the Gaussian case. The analysis tools we presented can be equally applied to many other factorization models yielding decoders with various complexities.

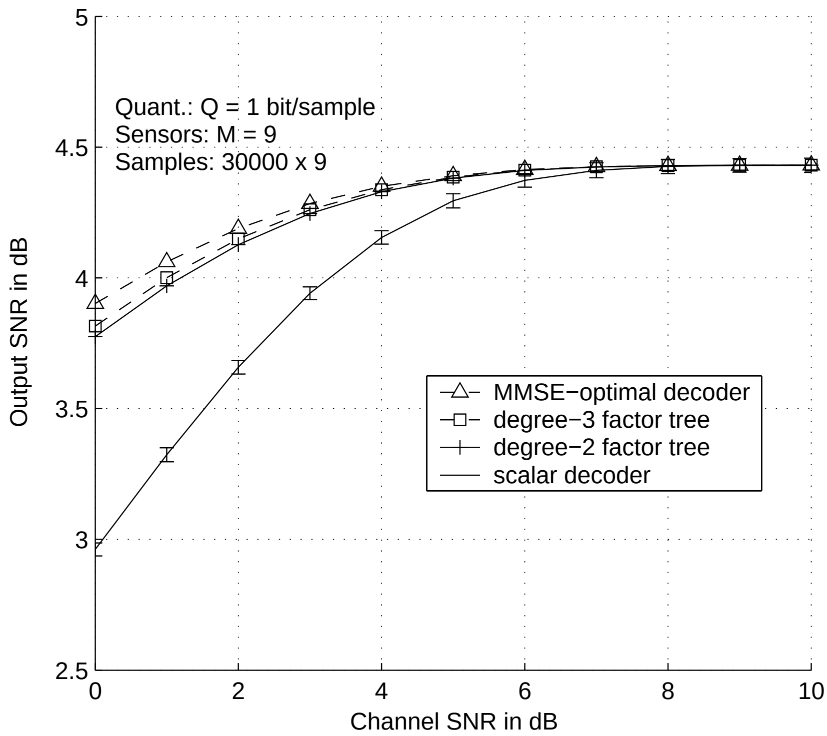


Figure 5.10: Performance of the MMSE-optimal CME and three decoders applying the BP algorithm on the factor graphs in *Figure 5.3* for a network with $M = 9$ sensors and $Q = 1$-bit quantization. The correlation factor between any two sensor measurements varies between $\rho = 0.217$ and $\rho = 0.930$. We consider the following cases: (1) scalar decoder (cf. *Figure 5.3*(a), $D(p(\mathbf{u})||\hat{p}(\mathbf{u})) = 3.92$ bits), (2) optimal degree-2 factor tree (cf. *Figure 5.3*(b), $D(p(\mathbf{u})||\hat{p}(\mathbf{u})) = 0.43$ bits), (3) optimized degree-3 factor tree (cf. *Figure 5.3*(c), $D(p(\mathbf{u})||\hat{p}(\mathbf{u})) = 0.40$ bits), (4) optimal MMSE Decoder. The bottom curve includes the standard deviation for 200 experiments thus supporting the reliability of our simulation results.

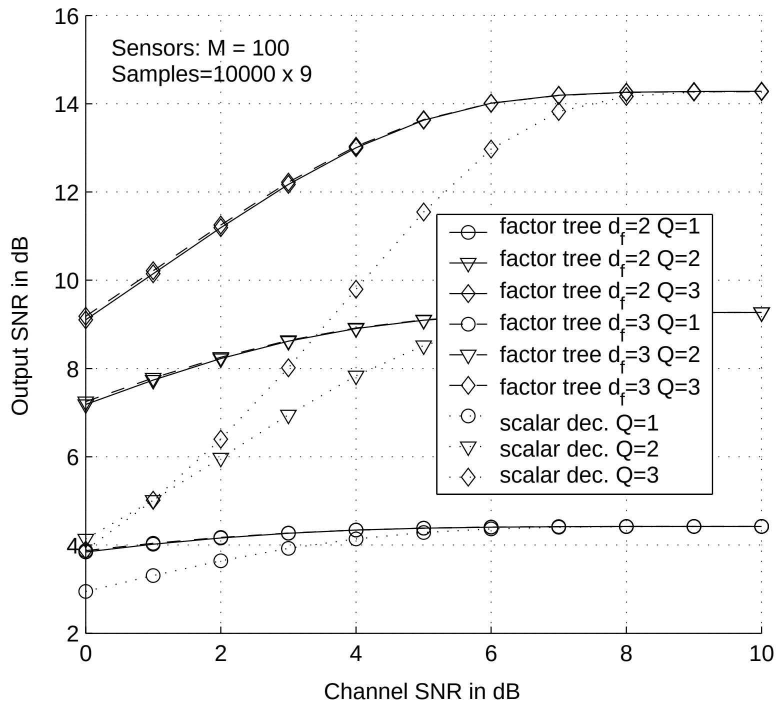


Figure 5.11: Performance of 3 decoders based on optimized factor graphs for a network with $M = 100$ sensors using various quantizers (1, 2, or 3-bit quantization). The correlation factor between any two sensor measurements varies between $\rho = 0$ and $\rho = 0.945$. We consider the following cases: (1) scalar decoder (trivial factor graph, $D(p(\mathbf{u})||\hat{p}(\mathbf{u})) = 45.37$ bits), (2) KLD-optimal degree-2 factor tree ($D(p(\mathbf{u})||\hat{p}(\mathbf{u})) = 6.13$ bits), (3) optimized degree-3 factor tree ($D(p(\mathbf{u})||\hat{p}(\mathbf{u})) = 5.40$ bits).

Our analyses and simulation results indicate that the proposed approach is well suited for large-scale sensor networks. Natural extensions could include (a) extending the factor graph to account for sensor nodes that have more complex features, such as entropy coding, channel coding or higher modulations, and (b) reducing the complexity further by running linear message updates in the nodes of the factor graph based on a Gaussian approximation of the message distributions [65].

6

Main Contributions and Future Work

It is better to know some of the questions than all of the answers.
JAMES THURBER

Main Contributions

This dissertation presented a number of original contributions towards a fundamental understanding of reachback communication in wireless sensor networks — both in theory and in practice. The first part dealt with the problem of reliable communication in sensor networks from the point of view of network information theory, whereas the second part looked at the underlying distributed data compression problem in terms of rate-distortion trade-offs. The third and more practical part was devoted to the development of decoding algorithms that meet the complexity and scalability requirements of large-scale sensor networks. The next paragraphs give an overview of our main contributions in each of these parts.

1. Fundamental Limits of Reachback Communication

We proposed an information-theoretic treatment of the communications scenario in which a large number of sensors deployed on a field measure the state of some physical process, and cooperate to send this information to a remote receiver. Formulating the problem as one of communicating multiple correlated sources over an array of independent channels, and with partial cooperation among encoders, we proved a set of coding theorems that give a complete characterization of the *reachback capacity*, i.e., of the exact conditions on the sources and the channels under which reliable communication with perfect reconstruction of the source messages at the far receiver is possible. Although in the general problem of sending correlated sources over multiple access channels considered by Cover et al. [26] only a region of achievable rates is known (without a converse), for our model we were able to give a *complete* characterization, converse included.

Moreover, for all instances of the sensor reachback problem considered in this thesis (with and without cooperation, and for any number of nodes), we were able to prove that natural generalizations of the joint source/channel coding theorem, commonly known as the *separation* theorem, hold [28, Ch. 8.13]. To the best of our knowledge, these are the first non-trivial examples involving multiple sources and multiple channels for which the separation principle does provide an optimal system architecture. This observation motivated us to revisit the issue of optimality of separate source and channel coding in communication networks, which we argued is a question of "when" and not "if" it holds.

2. Rate-Distortion Bounds for Multiterminal Source Coding

If the conditions for perfect reconstruction at the far receiver cannot be met, the best we can hope for is to reconstruct an *approximation* of the original source message — in this case, the fundamental communication limits are provided by rate-distortion theory [17]. The rate-distortion generalization of the sensor reachback problem lead to the well-known multiterminal source coding problem [18], in which correlated sources are compressed separately with respect to a fidelity criterion and we ask for a complete characterization of the achievable compression rates and distortion levels (the so called *rate-distortion region*).

Our contributions here were as follows. First, we gave a simple proof for the best known achievable rate-distortion region for general discrete sources and distortion measures (the Berger-Tung inner bound [90]). The proof shows for the first time that all points of this rate-distortion region can be achieved through a combination of independent quantizers followed by distributed compression (i.e. *Slepian-Wolf coding*) of the quantization indices. Secondly, we derived a new inner bound based on time-sharing of two complementary families of codes, which were initially used by Berger and Yeung [19] to solve the special case where one of the distortions goes to zero. A key feature of our codes is that they work for a large family of auxiliary random variables (W_1, W_2), whose dependence on the random variables (U_1, U_2) describing the correlated sources satisfies only two short Markov chain conditions, $W_1 \to U_1 \to U_2$ and $U_1 \to U_2 \to W_2$, but not the more restrictive long chain condition $W_1 \to U_1 \to U_2 \to W_2$ typical of all previously known codes. For the more practical case of two binary sources with Hamming distortion, we were able to give a new partial characterization of the corresponding rate-distortion region.

In addition, we re-formulated the multiterminal source coding problem to account for partial cooperation between encoders. In the lossless case, we gave a simple proof for a coding theorem that extends the result of Slepian and Wolf [84] for non-cooperating encoders, resulting in a complete characterization of the rate region under this scenario. Then we extended this setup to consider an arbitrary pair of distortions (D_1, D_2), where we extended the results of Berger and Tung [90] to obtain an inner and an outer bound for the region of achievable rates (R_1, R_2). Interestingly enough, we found that the rate expressions for the Berger-Tung inner and outer bounds, and for our inner and outer bounds with cooperation, are all identical—the only differences among all four regions lie in the class of probability distributions over which each of these bounds is defined. A close inspection of these classes of distributions reveals two important facts: (a) the uncertainty on whether the Berger-Tung inner and outer bounds are tight or not carries over to our inner and outer bounds with cooperation; (b) cooperation does produce a strict enlargement of the rate-distortion region.

3. Scalable Decoding Algorithms for Large-Scale Sensor Networks

Taking yet a different, more practical approach to reachback communication in wireless sensor networks, we assumed that each sensor node uses a very simple encoder (a scalar quantizer and a modulator) and focused on decoding algorithms that exploit the correlation structure of the sensor data to produce the best possible estimates under the minimum mean square error (MMSE) criterion. Our analysis showed that the optimal MMSE decoder is unfeasible for large scale sensor networks, because its complexity grows exponentially with the number of nodes in the network. Seeking a *scalable* alternative, we used factor graphs to obtain a simplified model for the correlation structure of the sensor data. This model allowed us to use a practical decoding algorithm (the so called *belief propagation* algorithm) whose complexity can be made to grow linearly with the size of the network. Considering large sensor networks with arbitrary topologies, we focused on factor trees and gave an exact characterization of the decoding complexity, as well as mathematical tools for factorizing Gaussian sources and optimization algorithms for finding optimal factor trees under the Kullback-Leibler criterion.

In short, our main contributions are as follows:

- We gave an information-theoretic characterization of the fundamental performance limits of a general class of sensor networks, consisting of an arbitrary number of partially cooperating nodes that use orthogonal accessing to transmit the picked up data back to a remote receiver. For this class of communication networks, our results establish the optimality of separate source and channel coding.

- Seeking a solution for the multiterminal source coding problem, we contributed with two inner bounds for the corresponding rate-distortion region and showed that a two-stage coding strategy (independent quantization followed by Slepian Wolf compression) yields all points of the Berger-Tung inner rate-distortion region.

- We formulated a cooperative source coding problem, in which encoders are allowed to establish a conference to exchange information before compressing the data. In the lossless version of the problem, we obtained the exact rate region, and for the lossy case, we proved that partial cooperation increases the achievable rate-distortion region.

- Addressing the issues of scalability and complexity in large-scale reachback networks, we introduced a class of scalable decoders and showed how to optimize factor trees for efficient decoding of correlated sensor data. The analysis tools we presented can be equally applied to other factorization models yielding decoders with various trade-offs in terms of complexity and end-to-end distortion.

Some of these results have recently been cited by different authors, e.g. in the context of coding strategies for sensor networks [102, 33], cooperation between sensor nodes [70], and fundamental limits of wireless sensor networks [37, 48, 47].

Future Work

In this final section, we would like to share some of the intuition gained from our work on reachback communication in wireless sensor networks and present four classes of research problems that we find particularly interesting and promising for future research activities: (a) stochastic models for sensor data, (b) fundamental limits with constrained resources, (c) source/channel coding for sensor networks and (d) joint estimation and data fusion of encoded sensor data.

A) Stochastic Models for Sensor Data

The design of contemporary communications systems relies strongly on finding mathematical models for the sources of information. There are many successful examples: in text compression we use the probabilities of the letters, in speech coding we use a linear model for the vocal tract, in audio processing we exploit the imperfections of the human ear, and for images we use transforms that yield simple Gaussian parameters. In the context of sensor networks, source modelling is still at a very infant stage. Most of the envisioned applications, are characterized by parameters (e.g. temperature, humidity and wind speed) that are dynamic by nature — the sensors measure realizations of physical processes that unfold continuously in time and space. Exploiting the available information from real measurements taken over long periods of time(e.g. seismic data or farming charts), we can hope to capture the random properties of the target parameters, which include the noisy deviations introduced by the measuring sensors and the statistical dependencies between measurements taken at different times, as well as those between neighboring sensor nodes. These correlations are interesting to the communications engineer because, as we showed in Chapter 5, they can be used to leverage the overall system efficiency. Therefore, general models that capture the random properties of large classes of sensor data are likely to have a strong impact on the design of practical sensor networks.

B) Communication Limits with Practical Constraints

This dissertation offered several contributions towards a characterization of the fundamental limits of sensor networks. Based on classical tools of network information theory, we were able to characterize the *reachback* capacity of a large class of sensor networks and identify network architectures that yield optimal performance. Said information-theoretic results are valid for discrete memoryless sources and channels. The next step towards a thorough understanding of the underlying principles of sensor networks, would be a mathematical characterization of their performance limits taking into consideration two fundamental types of constraints: transmit power and computational complexity.

It is common practice to constraint the average transmit power when computing the capacity of point-to-point wireless channels [28]. Similarly, we can adapt some of the classical multi-user capacity problems, most notably the multiple access channel, the broadcast channel and the relay channel, to sensor network scenario for example by introducing the source models described in the previous section, fading channels with Gaussian noise, some form of cooperation between users and, most importantly, average power constraints. Computing the capacity of sensor networks based on this model requires us to come up with coding strategies and power allocation schemes that sustain the achievability of said capacity region, leading to theoretical constructions that often give important hints on which designs make most sense in practical applications.

Computational complexity, on the other hand, is a system aspect that is more difficult to model, since it is highly dependent on the actual implementation (semiconductor technology, processor architecture, memory access, parallel algorithms). One way to capture at least part of the sensor nodes' processing constraints is to restrict their communications in terms of coding rates and codeword size ([68], [55]). Introducing this type of practical constraints we can evaluate their impact on the information-theoretic capacity of the classical multi-user channels mentioned above and characterize the trade-offs between decoding complexity and error probability (e.g. in terms of error exponents [36]).

In this context, some attention should be directed to the effects of feedback between the sensor nodes and the fusion center. Although feedback does not affect the capacity of a point-to-point discrete memoryless channel [28], it does increase the capacity of a generic multiple access channel [27]. Going back to our example, between the sensor nodes on the field and the fusion center controlling the irrigation arm we have a noisy multiple access channel on the uplink and a noisy broadcast channel on the downlink. The latter can be used by the fusion center to coordinate the transmissions of the sensor nodes, yielding a multiple access channel with noisy broadcast feedback. Since feedback is known to increase the capacity of multiple access channels, we believe this to be a communications aspect that deserves some more investigation in the context of system architectures for sensor networks.

C) Source/Channel Coding for Sensor Data

Based on accurate models for the sensor data and the wireless channel, we can find practical codes that come close to the communications limits described in the previous section. There are several possibilities for the data observed by the sensors and the desired reconstruction at the decoder, for example:

1. the sensors observe different sets of realizations of a multi-dimensional random process, the decoder reconstructs each of the realizations (*multiterminal* or *distributed* source coding, [85, 18]),

2. the sensors observe different sets of realizations of a multi-dimensional random process, the decoder estimates a function of the random variables describing the process,

3. the sensors observe multiple noisy versions of the same realization, the decoder estimates said realization (the so called CEO problem [20]).

Practical code constructions for the first scenario have recently appeared in several publications (e.g. [1, 38, 75]), most of which rely on dual versions of capacity attaining channel codes (e.g. turbo codes and low-density parity-check codes). Allthough equally important, the other two scenarios are less well understood and require further investigation. The same observation applies to the development of distributed joint source/channel codes that are robust to channel errors, as well as cooperative source/channel codes for sensor networks.

D) Joint Decoding, Estimation and Data Fusion

The receiver of the information sent by a sensor network has three different tasks: (1) to decode the received signals, (2) to perform some form of data fusion, (3) to make a decision based on the estimated data. These data processing steps are usually done separately, which means that the data fusion procedure only takes into consideration the statistics of the source and neglects the characteristics of the channel. In Chapter 5, we proposed to change this paradigm and perform joint decoding and data fusion, exploiting both the correlation in the data and the available channel state information to aid the decision making process.

The basic decoding and estimation operations in the receiver require in general the factorization of a multi-dimensional probability distribution that reflects the *a priori* source statistics and the likelihood values of the channel outputs. Assuming that the nodes use very simple encoders, for example a scalar quantizer and a modulator, it is straightforward to derive the optimal decoder based on minimum mean square estimation (MMSE). However, the simplicity of said derivation is in fact misleading — the complexity of the optimal MMSE decoder grows exponentially with the number of nodes in the network. The basic intuition to be gained from our work on scalable solutions for this decoding problem is that for practical purposes the two-dimensional statistics of the data collected by a sensor network can be well approximated by embedding a factor tree, thus reducing the complexity of the decoding, estimation or data fusion algorithms that process this data. Future directions include (a) extending this method to sensor networks with cooperation, M-ary modulations and channel interference, and (b) obtaining performance bounds based on random networks and random graphs.

From the conceptual point of view, this approach is also interesting in applications where the topology of the network can be defined by the system designer. The underlying research problem could then be formulated as follows: Given the probability density function of the sources observed by the sensors, a model for the communications channel and the target complexity of the joint decoding/data fusion algorithm what is the optimal placement for the sensor nodes?

In conclusion, when we take into consideration the large breadth of sensor applications and the many research challenges they entail, it becomes quite clear that wireless sensor networks will remain a very exciting topic in the years to come. If the theoretical and practical insights gained from our thesis contribute to a fundamental understanding of said networks, thus nurturing their technological development, then the ultimate goals we set for our personal research path will no doubt have been accomplished.

A

Proofs

Mathematatik ist die perfekte Methode, sich selbst an der Nase herumzuführen.
ALBERT EINSTEIN

A.1 An Alternative Achievability Proof for Theorem 3.3 based on Cooperative Channel Coding

In [93], Willems obtained the capacity region of the multiple access channel with partially cooperating encoders. He did this by introducing a class of channel codes with cooperation, that build on a construction due to Slepian and Wolf for a multiple access channel with two encoders, two independent sources and a third common source observed by both encoders [84]. We now show that in the sensor reachback problem there is no performance loss associated with an alternative architecture to the one presented in the main text (illustrated in *Figure 3.5*). In this new architecture, we use classical Slepian-Wolf codes to remove the correlation between the sources, and then apply Willems' cooperative channel coding approach (as illustrated in *Figure 3.6*). Besides the historical interest (we developed this proof first), this alternative proof also serves the purpose of showing that there is nothing to lose in terms of performance by moving cooperation from the source coders to the channel coders.

Proof: Each source encoder takes an input block U_i^N, $i \in \{1, 2\}$ and outputs a bin index W_i from the alphabet $\mathcal{W}_i = \{1, 2, ..., 2^{NR_i}\}$. Prior to transmission the channel encoders exchange messages over the conference channels. The conference messages are obtained as follows. First, we partition the set of messages $\mathcal{W}_1 = \{1, 2, ..., 2^{NR_1}\}$ in $2^{NR_{12}}$ cells, indexed by $i_1 \in \{1, 2, ..., 2^{NR_{12}}\}$, such that $i_1(w_1) = c_1$ if w_1 is inside cell c_1. Similarly, we partition the set of messages $\mathcal{W}_2 = \{1, 2, ..., 2^{NR_2}\}$ in $2^{NR_{21}}$ cells, indexed by $i_2 \in \{1, 2, ..., 2^{NR_{21}}\}$, such that $i_2(w_2) = c_2$ if w_2 is inside cell c_2. All messages inside each cell c_1 are indexed by $j_1 \in \{1, 2, ..., 2^{N(R_1 - R_{12})}\}$, and all messages inside each cell c_2 are indexed by $j_2 \in \{1, 2, ..., 2^{N(R_2 - R_{21})}\}$. During the conference, encoder 1 sends index i_1 to encoder 2, and encoder 2 sends index i_2 to encoder 1. Since U_1^N and U_2^N are random variables, W_1, W_2, I_1 and I_2 are also random.

The conditions for reliable communication under this conference scenario are given in [93] and can be written as

$$\begin{aligned} R_1 - R_{12} &\leq I(X_1;Y_1Y_2|X_2Z) \\ R_2 - R_{21} &\leq I(X_2;Y_1Y_2|X_1Z) \\ R_1 + R_2 - R_{12} - R_{21} &\leq I(X_1X_2;Y_1Y_2|Z) \\ R_1 + R_2 &\leq I(X_1X_2;Y_1Y_2), \end{aligned}$$

where Z is an auxiliary random variable such that $Z = (I_1I_2)$, $p(w_1w_2z) = p(w_1)p(w_2)p(z|w_1w_2)$. Since $R_1 = \frac{1}{N}H(W_1)$, and $R_1 - R_{12} = \frac{1}{N}H(W_1|Z)$, we have that $R_{12} = \frac{1}{N}(H(W_1) - H(W_1|Z)) = \frac{1}{N}I(W_1;Z)$. Similarly, $R_{21} = \frac{1}{N}I(W_2;Z)$. Using these identities and the fact that the channels are independent, we get

$$R_1 \leq I(X_1;Y_1|Z) + \frac{1}{N}I(W_1;Z) \tag{A.1}$$

$$R_2 \leq I(X_2;Y_2|Z) + \frac{1}{N}I(W_2;Z) \tag{A.2}$$

$$R_1 + R_2 \leq I(X_1;Y_1|Z) + I(X_2;Y_2|Z) + \frac{1}{N}I(W_1W_2;Z) \tag{A.3}$$

$$R_1 + R_2 \leq I(X_1;Y_1) + I(X_2;Y_2), \tag{A.4}$$

where we used the fact that W_1 and W_2 are independent and therefore $I(W_1;Z) + I(W_2;Z) = I(W_1W_2;Z)$.

As in the other proof, we know that reliable communication is possible if the capacity region given by (A.1)-(A.4) intersects the Slepian-Wolf rate region for (U_1U_2). This is the case if and only if

$$H(U_1|U_2) \leq I(X_1;Y_1|Z) + \frac{1}{N}I(W_1;Z) \tag{A.5}$$

$$H(U_2|U_1) \leq I(X_2;Y_2|Z) + \frac{1}{N}I(W_2;Z) \tag{A.6}$$

$$H(U_1U_2) \leq I(X_1;Y_1|Z) + I(X_2;Y_2|Z) + \frac{1}{N}I(W_1W_2;Z) \tag{A.7}$$

$$H(U_1U_2) \leq I(X_1;Y_1) + I(X_2;Y_2). \tag{A.8}$$

We now develop the sum rate condition (A.7). First, we note that, since W_1W_2 are a function of $U_1^NU_2^N$ (encoding property), and $U_1^NU_2^N$ are a function of W_1W_2 (decoding property), *both* Markov chains $W_1W_2 \rightarrow U_1^N \rightarrow U_2^N \rightarrow Z$ and $U_1^NU_2^N \rightarrow W_1W_2 \rightarrow Z$ hold, and so it follows from the data processing inequality that $I(W_1W_2;Z) = I(U_1^NU_2^N;Z)$. Noting that $\frac{1}{N}I(U_1^NU_2^N;Z) = I(U_1U_2;Z)$ (the sources are i.i.d.), we can rewrite (A.7) as

$$H(U_1U_2) \leq I(X_1;Y_1|Z) + I(X_2;Y_2|Z) + I(U_1U_2;Z). \tag{A.9}$$

To develop the side conditions (A.5) and (A.6) accordingly, we use a simple time-sharing argument. Assume the two Slepian-Wolf source encoders operate at rates $R_1 = H(U)$ and $R_2 = H(U_2|U_1)$, such that $H(W_1) = H(U_1^N)$, $H(U_1^N|W_1) = 0$ and, consequently,

$$I(W_1;Z) = I(U_1^NW_1;Z) = I(U_1^N;Z) + I(W_1;Z|U_1^N) = I(U_1^N;Z).$$

Substituting $I(W_1; Z) = I(U_1^N; Z)$ and $R_1 = H(U)$ in condition (A.1), we get

$$H(U) + \epsilon = I(X_1; Y_1|Z) + I(U_1; Z), \tag{A.10}$$

for some $\epsilon > 0$ arbitrarily small. Since (A.10) follows from (A.1), and (A.7) follows from (A.3), the source/channel coding theorem by Slepian and Wolf [84] guarantees that there exists a code satisfying both (A.7) and (A.10). We can now combine these two conditions by subtracting $H(U_1)$ from both sides of the modified condition (A.9), so that

$$\begin{aligned} H(U_2|U_1) &\leq I(X_1; Y_1|Z) + I(X_2; Y_2|Z) + I(U_1U_2; Z) - H(U_1) \\ &\leq I(X_2; Y_2|Z) + I(U_2; Z|U_1) + I(X_1; Y_1|Z) + I(U_1; Z) - H(U) \\ &\leq I(X_2; Y_2|Z) + I(U_2; Z|U_1) + \epsilon, \end{aligned} \tag{A.11}$$

where ϵ can be made arbitrarily small to yield the second condition in the theorem. The first condition can be obtained by a symmetric argument, with Slepian-Wolf encoders operating at rates $R_1 = H(U_2)$ and $R_2 = H(U_1|U_2)$, so that

$$H(U_2) + \epsilon = I(X_2; Y_2|Z) + I(U_2; Z),$$

and

$$H(U_1|U_2) \leq I(X_1; Y_1|Z) + I(U_1; Z|U_2) + \epsilon. \tag{A.12}$$

By time-sharing between the code construction for $(R_1, R_2) = (H(U_1), H(U_2|U_1))$ and $(R_1, R_2) = (H(U_1|U_2), H(U_2))$, we conclude that conditions (A.12), (A.11), (A.9) and (A.8) are sufficient for reliable communication. Looking at the first two terms of the right-hand side of (A.9), we can write

$$\begin{aligned} I(X_1; Y_1|Z) + I(X_2; Y_2|Z) &= H(Y_1|Z) - H(Y_1|X_1Z) + H(Y_2|Z) - H(Y_2|X_2Z) \\ &= H(Y_1|Z) - H(Y_1|X_1) + H(Y_2|Z) - H(Y_2|X_2) \\ &\leq H(Y_1) - H(Y_1|X_1) + H(Y_2) - H(Y_2|X_2) \\ &= I(X_1; Y_1) + I(X_2; Y_2), \end{aligned}$$

with equality for $p(x_1x_2z) = p(x_1)p(x_2)p(z)$. Since this choice of Z maximizes the right-handside of (A.9), we can modify this condition to

$$H(U_1U_2) \leq I(X_1; Y_1) + I(X_2; Y_2) + I(U_1U_2; Z).$$

We conclude that (A.9) is always satisfied when (A.8) is satisfied, and so we omit the former.

■

A.2 Proof of Converse for Theorem 3.3

A.2.1 Preliminaries

To develop the converse we start with Fano's inequality. Assuming there exists a suitable code with parameters $(R_1, R_2, R_{12}, R_{21}, N, K, P_e)$, then we must have

$$H(U_1^N U_2^N|\hat{U}_1^N \hat{U}_2^N) \leq P_e \log\left(|\mathcal{U}_1^N \times \mathcal{U}_2^N|\right) + H_b(P_e), \tag{A.13}$$

where $H_b(P_e)$ is the binary entropy function. For convenience, define also

$$\delta(P_e) = \left(P_e \log\left(|\mathcal{U}_1^N \times \mathcal{U}_2^N|\right) + H_b(P_e)\right)/N.$$

It follows from (A.13) that

$$\begin{aligned} H(U_1^N U_2^N|Y_1^N Y_2^N) &= H(U_1^N U_2^N|Y_1^N Y_2^N g(Y_1^N Y_2^N)) \\ &= H(U_1^N U_2^N|Y_1^N Y_2^N \hat{U}_1^N \hat{U}_2^N) \\ &\leq H(U_1^N U_2^N|\hat{U}_1^N \hat{U}_2^N) \\ &\leq N\delta(P_e), \end{aligned}$$

and therefore,

$$H(U_1^N U_2^N|Y_1^N Y_2^N V_1^K V_2^K) \leq H(U_1^N U_2^N|Y_1^N Y_2^N) \leq N\delta(P_e),$$

and also

$$\begin{aligned} H(U_1^N|Y_1^N Y_2^N V_1^K V_2^K) &\leq N\delta(P_e) \\ H(U_2^N|Y_1^N Y_2^N V_1^K V_2^K) &\leq N\delta(P_e) \\ H(U_1^N|Y_1^N Y_2^N V_1^K V_2^K U_2^N) &\leq N\delta(P_e) \\ H(U_2^N|Y_1^N Y_2^N V_1^K V_2^K U_1^N) &\leq N\delta(P_e). \end{aligned}$$

According to the problem statement, we have two long Markov chains in place: $Y_2^N \to X_2^N \to (V_1^K V_2^K U_1^N) \to X_1^N \to Y_1^N$ and $Y_2^N \to X_2^N \to (U_2^N V_1^K V_2^K) \to X_1^N \to Y_1^N$. These chains (informally referred to as *the long chains* in this section) will prove quite useful in our derivations.

A.2.2 The Side Faces: Necessity of Equations (3.6) and (3.7)

We start by bounding $H(U_1^N)$:

$$\begin{aligned} H(U_1^N) &= I(U_1^N; Y_1^N Y_2^N V_1^K V_2^K U_2^N) + H(U_1^N|Y_1^N Y_2^N V_1^K V_2^K U_2^N) \\ &\leq I(U_1^N; Y_1^N Y_2^N V_1^K V_2^K U_2^N) + N\delta(P_e) \\ &= I(U_1^N; U_2^N) + I(U_1^N; Y_1^N Y_2^N V_1^K V_2^K|U_2^N) + N\delta(P_e) \\ &= I(U_1^N; U_2^N) + I(U_1^N; V_1^K V_2^K|U_2^N) + I(U_1^N; Y_1^N Y_2^N|U_2^N V_1^K V_2^K) + N\delta(P_e) \end{aligned}$$

Now, $I(U_1^N;U_2^N)$ and $N\delta(P_e)$ stay the same, and we start with $I(U_1^N;V_1^KV_2^K|U_2^N)$:

$$\begin{aligned}
I(U_1^N;V_1^KV_2^K|U_2^N) &= \sum_{n=1}^N I(U_{1n};V_1^KV_2^K|U_2^NU_1^{n-1}) \\
&= \sum_{n=1}^N H(U_{1n}|U_2^NU_1^{n-1}) - H(U_{1n}|V_1^KV_2^KU_2^NU_1^{n-1}) \\
&= \sum_{n=1}^N H(U_{1n}|U_{2n}) - H(U_{1n}|V_1^KV_2^KU_{2n}) \\
&= \sum_{n=1}^N I(U_{1n};V_1^KV_2^K|U_{2n}) \\
&= \sum_{n=1}^N I(U_{1n};Z_n|U_{2n}),
\end{aligned}$$

where we set $Z_n = V_1^KV_2^K$. Now, we simplify $I(U_1^N;Y_1^NY_2^N|U_2^NV_1^KV_2^K)$:

$$\begin{aligned}
&I(U_1^N;Y_1^NY_2^N|U_2^NV_1^KV_2^K) \\
&= \sum_{n=1}^N I(U_1^N;Y_{1n}Y_{2n}|U_2^NV_1^KV_2^KY_1^{n-1}Y_2^{n-1}) \\
&\overset{(a)}{=} \sum_{n=1}^N I(U_1^NX_{1n};Y_{1n}Y_{2n}|U_2^NV_1^KV_2^KY_1^{n-1}Y_2^{n-1}X_{2n}) \\
&= \sum_{n=1}^N H(Y_{1n}Y_{2n}|U_2^NV_1^KV_2^KY_1^{n-1}Y_2^{n-1}X_{2n}) \\
&\qquad - H(Y_{1n}Y_{2n}|U_2^NV_1^KV_2^KY_1^{n-1}Y_2^{n-1}X_{2n}U_1^NX_{1n}) \\
&\leq \sum_{n=1}^N H(Y_{1n}Y_{2n}|U_2^NV_1^KV_2^KX_{2n}) \\
&\qquad - H(Y_{1n}Y_{2n}|U_2^NV_1^KV_2^KY_1^{n-1}Y_2^{n-1}X_{2n}U_1^NX_{1n}) \\
&\overset{(b)}{=} \sum_{n=1}^N H(Y_{1n}Y_{2n}|U_2^NV_1^KV_2^KX_{2n}) - H(Y_{1n}Y_{2n}|U_2^NV_1^KV_2^KX_{2n}U_1^NX_{1n}) \\
&= \sum_{n=1}^N I(U_1^NX_{1n};Y_{1n}Y_{2n}|U_2^NV_1^KV_2^KX_{2n}) \\
&= \sum_{n=1}^N I(U_1^NX_{1n};Y_{1n}|U_2^NV_1^KV_2^KX_{2n}) + \underbrace{I(U_1^NX_{1n};Y_{2n}|U_2^NV_1^KV_2^KX_{2n}Y_{1n})}_{=0} \\
&\overset{(c)}{=} \sum_{n=1}^N I(U_1^NX_{1n};Y_{1n}|U_2^NV_1^KV_2^KX_{2n})
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=1}^{N} H(Y_{1n}|U_2^N V_1^K V_2^K X_{2n}) - H(Y_{1n}|U_2^N V_1^K V_2^K X_{2n} U_1^N X_{1n}) \\
&= \sum_{n=1}^{N} H(Y_{1n}|U_2^N V_1^K V_2^K X_{2n}) - H(Y_{1n}|V_1^K V_2^K X_{1n}) \\
&\leq \sum_{n=1}^{N} H(Y_{1n}) - H(Y_{1n}|V_1^K V_2^K X_{1n}) \\
&\stackrel{(d)}{=} \sum_{n=1}^{N} H(Y_{1n}) - H(Y_{1n}|X_{1n}) \\
&= \sum_{n=1}^{N} I(X_{1n}; Y_{1n}) \qquad \text{(A.14)}
\end{aligned}
$$

where: (a) follows from the fact that $(U_1^N, V_2^K) \rightsquigarrow X_1^N \rightsquigarrow X_{1n}$ and $(U_2^N, V_1^K) \rightsquigarrow X_2^N \rightsquigarrow X_{2n}$; (b) follows from from the fact that the channels are discrete and memoryless, so given $X_{1n}X_{2n}$, $Y_{1n}Y_{2n}$ are independent of anything else, so we can drop conditioning terms without changing the entropy; (c) follows from the long chains; and (d) follows from the fact that (V_1^K, V_2^K) are independent of Y_{1n} given X_{1n}.

Combining all of the above, we get that

$$H(U_1^N) \leq I(U_1^N; U_2^N) + \sum_{n=1}^{N} I(X_{1n}; Y_{1n}) + \sum_{n=1}^{N} I(U_{1n}; Z_n|U_{2n}) + N\delta(P_e),$$

or equivalently,

$$\frac{1}{N} H(U_1^N|U_2^N) \leq \frac{1}{N} \sum_{n=1}^{N} I(X_{1n}; Y_{1n}) + \frac{1}{N} \sum_{n=1}^{N} I(U_{1n}; Z_n|U_{2n}) + \delta(P_e).$$

Symmetric arguments yield

$$\frac{1}{N} H(U_2^N|U_1^N) \leq \frac{1}{N} \sum_{n=1}^{N} I(X_{2n}; Y_{2n}) + \frac{1}{N} \sum_{n=1}^{N} I(U_{2n}; Z_n|U_{1n}) + \delta(P_e).$$

A.2.3 The Sum-Rate Face — Necessity of Equation (3.8)

Again, we start by bounding $H(U_1^N U_2^N)$:

$$\begin{aligned} H(U_1^N U_2^N) &= I(U_1^N U_2^N; Y_1^N Y_2^N) + H(U_1^N U_2^N | Y_1^N Y_2^N) \\ &\leq I(U_1^N U_2^N; Y_1^N Y_2^N) + N\delta(P_e). \end{aligned}$$

Now we need to simplify $I(U_1^N U_2^N; Y_1^N Y_2^N)$. Here we make use of the fact that, from the long chains, it follows that $U_1^N U_2^N \to X_1^N X_2^N \to X_{1n} X_{2n} \to Y_{1n} Y_{2n}$ also forms a Markov chain. So,

$$\begin{aligned} & I(U_1^N U_2^N; Y_1^N Y_2^N) \\ &= \sum_{n=1}^N I(U_1^N U_2^N; Y_{1n} Y_{2n} | Y_1^{n-1} Y_2^{n-1}) \\ &\leq \sum_{n=1}^N I(U_1^N U_2^N; Y_{1n} Y_{2n} | Y_1^{n-1} Y_2^{n-1}) + I(X_{1n} X_{2n}; Y_{1n} Y_{2n} | Y_1^{n-1} Y_2^{n-1} U_1^N U_2^N) \\ &= \sum_{n=1}^N I(U_1^N U_2^N X_{1n} X_{2n}; Y_{1n} Y_{2n} | Y_1^{n-1} Y_2^{n-1}) \\ &= \sum_{n=1}^N I(X_{1n} X_{2n}; Y_{1n} Y_{2n} | Y_1^{n-1} Y_2^{n-1}) + \underbrace{I(U_1^N U_2^N; Y_{1n} Y_{2n} | Y_1^{n-1} Y_2^{n-1} X_{1n} X_{2n})}_{=0} \\ &\stackrel{(a)}{=} \sum_{n=1}^N I(X_{1n} X_{2n}; Y_{1n} Y_{2n} | Y_1^{n-1} Y_2^{n-1}) \\ &= \sum_{n=1}^N H(Y_{1n} Y_{2n} | Y_1^{n-1} Y_2^{n-1}) - H(Y_{1n} Y_{2n} | Y_1^{n-1} Y_2^{n-1} X_{1n} X_{2n}) \\ &\stackrel{(b)}{=} \sum_{n=1}^N H(Y_{1n} Y_{2n} | Y_1^{n-1} Y_2^{n-1}) - H(Y_{1n} Y_{2n} | X_{1n} X_{2n}) \\ &\leq \sum_{n=1}^N H(Y_{1n} Y_{2n}) - H(Y_{1n} Y_{2n} | X_{1n} X_{2n}) \\ &\leq \sum_{n=1}^N H(Y_{1n}) + H(Y_{2n}) - H(Y_{1n} Y_{2n} | X_{1n} X_{2n}) \\ &= \sum_{n=1}^N H(Y_{1n}) + H(Y_{2n}) - H(Y_{1n} | X_{1n}) - H(Y_{2n} | X_{2n}) \\ &= \sum_{n=1}^N I(X_{1n}; Y_{1n}) + \sum_{n=1}^N I(X_{2n}; Y_{2n}). \end{aligned} \tag{A.15}$$

where: (a) follows from the chain above, and (b) follows from the DMC property. Therefore, we get that

$$\frac{1}{N}H(U_1^N U_2^N) \leq \frac{1}{N}\sum_{n=1}^{N} I(X_{1n};Y_{1n}) + \frac{1}{N}\sum_{n=1}^{N} I(X_{2n};Y_{2n}) + \delta(P_e).$$

A.2.4 Conditions on the Conference Rates

We now obtain necessary conditions for an admissable conference in terms of the auxiliary random variable Z:

$$\begin{aligned}
NC_{12} &\geq \sum_{k=1}^{K} \log|\mathcal{V}_{1k}| \\
&\geq H(V_1^K) \\
&\geq H(V_1^K|U_2^N) \\
&= H(V_1^K V_2^K|U_2^N) \\
&= H(V_1^K V_2^K|U_2^N) - H(V_1^K V_2^K|U_1^N U_2^N) \\
&= I(V_1^K V_2^K;U_1^N|U_2^N) \\
&= \sum_{n=1}^{N} I(V_1^K V_2^K;U_{1n}|U_1^{n-1}U_2^N) \\
&= \sum_{n=1}^{N} H(U_{1n}|U_1^{n-1}U_2^N) - H(U_{1n}|V_1^K V_2^K U_1^{n-1} U_2^N) \\
&= \sum_{n=1}^{N} H(U_{1n}|U_{2n}) - H(U_{1n}|V_1^K V_2^K U_{2n}) \\
&= \sum_{n=1}^{N} I(U_{1n};V_1^K V_2^K|U_{2n}) \\
&= \sum_{n=1}^{N} I(U_{1n};Z_n|U_{2n}),
\end{aligned}$$

where the first inequality is due to the admissibility condition for conferences, and the rest are standard information theoretic manipulations. Taking the second inequality, a similar argument yields

$$NC_{21} \geq \sum_{n=1}^{N} I(U_{2n};Z_n|U_{1n}).$$

Thus, the conditions on the conference rates become

$$\frac{1}{N}\sum_{n=1}^{N} I(U_{1n};Z_n|U_{2n}) \leq C_{12} \quad \text{and} \quad \frac{1}{N}\sum_{n=1}^{N} I(U_{2n};Z_n|U_{1n}) \leq C_{21}$$

A.2.5 Final Remarks

So far, we have established that

$$\begin{aligned}
\frac{1}{N}H(U_1^N|U_2^N) &\leq \frac{1}{N}\sum_{n=1}^{N} I(X_{1n};Y_{1n}) + \frac{1}{N}\sum_{n=1}^{N} I(U_{1n};Z_n|U_{2n}) + \delta(P_e)\\
\frac{1}{N}H(U_2^N|U_1^N) &\leq \frac{1}{N}\sum_{n=1}^{N} I(X_{2n};Y_{2n}) + \frac{1}{N}\sum_{n=1}^{N} I(U_{2n};Z_n|U_{1n}) + \delta(P_e)\\
\frac{1}{N}H(U_1^N U_2^N) &\leq \frac{1}{N}\sum_{n=1}^{N} I(X_{1n};Y_{1n}) + \frac{1}{N}\sum_{n=1}^{N} I(X_{2n};Y_{2n}) + \delta(P_e),
\end{aligned}$$

and also that

$$\frac{1}{N}\sum_{n=1}^{N} I(U_{1n};Z_n|U_{2n}) \leq C_{12} \qquad \text{and} \qquad \frac{1}{N}\sum_{n=1}^{N} I(U_{2n};Z_n|U_{1n}) \leq C_{21}.$$

Now, from here to the exact form of the conditions in the theorem there is a very short way. First, note that using the standard technique of introducing time-sharing variables (see, e.g. , [28, pg. 435]), we can replace the averages above by variables with the exact same distribution as prescribed by Theorem 3.3. Note also that by its own definition, $\delta(P_e) \to 0$ as $P_e \to 0$. Finally, note from the achievability proof that $|\mathcal{Z}| \leq |\mathcal{U}_1| \cdot |\mathcal{U}_2| < \infty$ (since Z is made up of partitions of $\mathcal{U}_1$ and $\mathcal{U}_2$). This concludes the proof of Theorem 3.3. ■

A.3 Converse Proof for Theorem 3.7

The proof uses the same arguments as the converse proof of Theorem 3.3, therefore we include here only the main steps.

A.3.1 Preliminaries

Assume that there exist codes with parameters $(R_1, \ldots, R_M, R_{12}, \ldots, R_{M-1,M}, N, K, P_e)$. Let $Z_{ij} = V_{ij}^K$, where V_{ij}^K denotes the block of messages sent by encoder i to encoder j. Based on Fano's inequality, we can write:

$$H(U_1^N \ldots U_M^N|\hat{U}_1^N \ldots \hat{U}_M^N) \leq P_e \log\left(|\mathcal{U}_1^N \times \cdots \times \mathcal{U}_M^N|\right) + H_b(P_e), \qquad \text{(A.16)}$$

where $H_b(P_e)$ is the binary entropy function. Define

$$\delta(P_e) = \left(P_e \log\left(|\mathcal{U}_1^N \times \cdots \times \mathcal{U}_M^N|\right) + H_b(P_e)\right)/N.$$

It follows from eqn. (A.16) that

$$H(U_1^N \ldots U_M^N|Y_1^N \ldots Y_M^N) \leq N\delta(P_e),$$

and consequently,

$$H(U^N(S)|Y_1^N \dots Y_M^N U^N(S^c)Z(S)) \leq N\delta(P_e) \tag{A.17}$$

for all subsets $S \subseteq \{1, 2, \dots, M\}$ with $U^N(S) = \{U_j^N : j \in S\}$, and $Z(S) = \{Z_{ij} : i \in S \text{ or } j \in S\}$.

A.3.2 Main Arguments

In order to obtain necessary conditions for an admissable conference, we write

$$\begin{aligned}
NC_{ij} &\geq \sum_{k=1}^{K} \log |\mathcal{V}_{ij}(k)| \\
&\geq H(V_{ij}^K) \\
&\geq H(V_{ij}^K|U_j^N) \\
&= H(V_{ij}^K|U_j^N) - H(V_{ij}^K|U_1^N \dots U_M^N) \\
&= I(U_1^N \dots U_{j-1}^N U_{j+1}^N \dots U_M^N; V_{ij}^K|U_j^N) \\
&= \sum_{n=1}^{N} I(U_{1n} \dots U_{j-1,n} U_{j+1,n} \dots U_{M,n}; V_{ij}^K|U_1^{n-1} \dots U_{j-1}^{n-1} U_{j+1}^{n-1} \dots U_M^{n-1} U_j^N) \\
&= \sum_{n=1}^{N} H(U_{1n} \dots U_{j-1,n} U_{j+1,n} \dots U_{M,n}|U_1^{n-1} \dots U_{j-1}^{n-1} U_{j+1}^{n-1} \dots U_M^{n-1} U_j^N) \\
&\qquad - H(U_{1n} \dots U_{j-1,n} U_{j+1,n} \dots U_{M,n}|U_1^{n-1} \dots U_{j-1}^{n-1} U_{j+1}^{n-1} \dots U_M^{n-1} U_j^N V_{ij}^K) \\
&= \sum_{n=1}^{N} H(U_{1n} \dots U_{j-1,n} U_{j+1,n} \dots U_{M,n}|U_{j,n}) - H(U_{1n} \dots U_{j-1,n} U_{j+1,n} \dots U_{M,n}|U_{j,n} V_{ij}^K) \\
&= \sum_{n=1}^{N} I(U_{1n} \dots U_{j-1,n} U_{j+1,n} \dots U_{M,n}; V_{ij}^K|U_{j,n}) \\
&= \sum_{n=1}^{N} I(U_{1n} \dots U_{j-1,n} U_{j+1,n} \dots U_{M,n}; Z_{ij}(n)|U_{j,n}),
\end{aligned}$$

where (a) the first inequality is due to the admissibility condition for conferences, (b) we set $Z_{ij}(n) = V_{ij}^K$ and (c) the rest follows from standard information theoretic manipulations.

Necessary conditions for reliable communication can be obtained from the following inequality:

$$\begin{aligned}
H(U^N(S)) &= I(U^N(S); Y_1^N \dots Y_M^N Z(S) U^N(S^c)) + H(U^N(S)|Y_1^N \dots Y_M^N Z(S) U^N(S^c))) \\
&\leq I(U^N(S); U^N(S^c)) + I(U^N(S); Z(S)|U^N(S^c)) \\
&\qquad + I(U^N(S); Y_1^N \dots Y_M^N|Z(S) U^N(S^c)) + N\delta.
\end{aligned} \tag{A.18}$$

We can now develop each of the mutual information terms on the right-hand side of this inequality to obtain single-letter expressions. The first term can be subtracted on both sides, yielding $H(U^N(S)|U^N(S^c))$ on the left-hand side of (A.18), which can be shown to be equal to $NH(U(S)|U(S^c))$ by arguing that the sources are memoryless and using a standard time-sharing argument. Similarly, using standard information-theoretic identities and inequalities to develop the second term we get

$$I(U^N(S); Z(S)|U^N(S^c)) \leq NI(U(S); Z(S)|U(S^c)).$$

Finally, for the third term we obtain

$$I(U^N(S); Y_1^N \dots Y_M^N|Z(S)U^N(S^c)) \leq N \sum_{i \in S} I(X_i; Y_i)$$

repeating the steps of (A.14) and (A.15), and using the aforementioned time-sharing argument.

■

A.4 Proof of Theorem 4.1

Assume that the rate-distortion region $\mathcal{R}(D_1, D_2)$ is known. It follows from the definition of $\mathcal{R}(D_1, D_2)$ that

$$\begin{aligned} R_1 &\geq r_1(D_1, D_2) \\ R_2 &\geq r_2(D_1, D_2) \\ R_1 + R_2 &\geq r_0(D_1, D_2), \end{aligned}$$

where $r_i(\cdot)$, $i = 0, 1, 2$, indicate real numbers that depend on the prescribed distortions D_1 and D_2. From the joint source/channel coding theorem, commonly known as the *separation* theorem [28, Ch. 8.13], and the definition of channel capacity, we must have

$$r_1(D_1, D_2) \leq I(X_1; Y_1) \leq C_1, \tag{A.19}$$
$$r_2(D_1, D_2) \leq I(X_1; Y_1) \leq C_2. \tag{A.20}$$

Since according to the definition of the rate-distortion region for the multiterminal source coding problem $r_0(D_1, D_2)$ is the minimum sum rate at which the reconstruction blocks $(\hat{U}_1^N, \hat{U}_2^N)$ approximate the source blocks (U_1^N, U_2^N) within distortion levels (D_1, D_2), any admissable code for this problem must fulfil $I(U_1^N U_2^N; \hat{U}_1^N \hat{U}_2^N) \geq r_0(D_1, D_2)$. We obtain the following chain of inequalities:

$$\begin{aligned} r_0(D_1, D_2) &\leq I(U_1^N U_2^N; \hat{U}_1^N \hat{U}_2^N) \\ &\overset{(a)}{\leq} I(U_1^N U_2^N; Y_1^N Y_2^N) \\ &= I(U_1^N U_2^N; Y_1^N) + I(U_1^N U_2^N; Y_2^N|Y_1^N) \end{aligned}$$

$$
\begin{aligned}
&\stackrel{(b)}{\leq} I(U_1^N U_2^N; Y_1^N) + I(U_1^N U_2^N; Y_2^N) \\
&= I(U_1^N; Y_1^N) + I(U_2^N; Y_1^N | U_1^N) + I(U_2^N; Y_2^N) + I(U_1^N; Y_2^N | U_2^N) \\
&\stackrel{(c)}{=} I(U_1^N; Y_1^N) + I(U_2^N; Y_2^N) \\
&\stackrel{(d)}{=} I(X_1^N; Y_1^N) + I(X_2^N; Y_2^N) \\
&\stackrel{(e)}{=} C_1 + C_2 \qquad \text{(A.21)}
\end{aligned}
$$

where we use the following the arguments:

(a) follows from the data processing inequality,

(b) follows from

$$
\begin{aligned}
I(U_1^N U_2^N; Y_2^N | Y_1^N) &= H(Y_2^N | Y_1^N) - H(Y_2^N | U_1^N U_2^N Y_1^N) \\
&= H(Y_2^N | Y_1^N) - H(Y_2^N | U_1^N U_2^N) \\
&\leq H(Y_2^N) - H(Y_2^N | U_1^N U_2^N) \\
&= I(U_1^N U_2^N; Y_2^N)
\end{aligned}
$$

(c) follows from the Markov chain of the form $Y_1^N \to X_1^N \to U_1^N \to U_2^N \to X_2^N \to Y_2^N$,

(d) follows from the data processing inequality,

(e) follows from the definition of channel capacity.

Since (A.19)-(A.21) were obtained from necessary conditions for any code that guarantees the prescribed distortion levels, we conclude that a non-empty intersection between the rate-distortion region and the capacity region is a necessary condition for any code that solves this problem. ■

A.5 Proof of Theorem 4.8

The proof is identical to the proof of the Berger-Tung inner bound up to the inequalities (4.6)-(4.8), which we rewrite here for clarity.

$$
\begin{aligned}
NR_1 &\geq H(I|J), \\
NR_2 &\geq H(J|I), \\
N(R_1 + R_2) &\geq H(IJ)
\end{aligned}
$$

Standard information theoretical identities and inequalities yield the following chain of long expressions:

$$\begin{aligned}
NKR_1 &\geq H(I^K|J^K) \\
&= H(\{U_1^N\}^K, \{U_2^N\}^K, I^K|J^K) - H(\{U_1^N\}^K, \{U_2^N\}^K|I^K, J^K) \\
&= H(\{U_1^N\}^K, \{U_2^N\}^K|J^K) - H(\{U_1^N\}^K, \{U_2^N\}^K|I^K, J^K) \qquad \text{(A.22)} \\
&= \sum_{k=1}^{K} H(\{U_1^N\}_k, \{U_2^N\}_k|J^K, \{U_1^N\}^{k-1}, \{U_2^N\}^{k-1}) \\
&\quad -H(\{U_1^N\}_k, \{U_2^N\}_k|I^K, J^K, \{U_1^N\}^{k-1}, \{U_2^N\}^{k-1}) \\
&= \sum_{k=1}^{K}\sum_{l=1}^{N} H(U_{1kl}, U_{2kl}|J^K, \{U_1^N\}^{k-1}, \{U_2^N\}^{k-1}, \{U_1^{l-1}\}_k, \{U_2^{l-1}\}_k) \\
&\quad -H(U_{1kl}, U_{2kl}|I^K, J^K, \{U_1^N\}^{k-1}, \{U_2^N\}^{k-1}, \{U_1^{l-1}\}_k, \{U_2^{l-1}\}_k), \qquad \text{(A.23)}
\end{aligned}$$

Here (A.22) stems from the fact that for a fixed code $\mathcal{C}_1$ the codeword I^K is a function of the block $\{U^N\}^K$. Defining

$$W_{1kl} = (I^K, \{U_1^N\}^{k-1}, \{U_2^N\}^{k-1}, \{U_1^{l-1}\}_k, \{U_2^{l-1}\}_k)$$

and

$$W_{2kl} = (J^K, \{U_1^N\}^{k-1}, \{U_2^N\}^{k-1}, \{U_1^{l-1}\}_k, \{U_2^{l-1}\}_k)$$

, we get

$$\begin{aligned}
NKR_1 &\geq \sum_{k}^{K}\sum_{l=1}^{n} H(U_{1kl}, U_{2kl}|W_{2kl}) - H(U_{1kl}, U_{2kl}|W_{2kl}, W_{1kl}) \\
&= \sum_{k}^{K}\sum_{l=1}^{N} I(U_{1kl}, U_{2kl}; W_{1kl}|W_{2kl}).
\end{aligned}$$

Using symmetric arguments, we get

$$NKR_2 \geq \sum_{k=1}^{K}\sum_{l=1}^{N} I(U_{1kl}, U_{2kl}; W_{1kl}|W_{2kl}).$$

Finally, for the sum rate condition we can write:

$$\begin{aligned}
NK(R_1 + R_2) &\geq H(I^K, J^K) \\
&= H(I^K, J^K|\{U_1^N\}^K, \{U_2^N\}^K) + I(\{U_1^N\}^K, \{U_2^N\}^K; I^K, J^K) \\
&= I(\{U_1^N\}^K, \{U_2^N\}^K; I^K, J^K)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k}^{K} H(\{U_1^N\}_k, \{U_2^N\}_k | \{U_1^N\}^{k-1}, \{U_2^N\}^{k-1}) \\
&\quad - H(\{U_1^N\}_k, \{U_2^N\}_k | I^K, J^K, \{U_1^N\}^{k-1}, \{U_2^N\}^{k-1}) \\
&= \sum_{k}^{K} \sum_{l}^{N} H(U_{1kl}, U_{2kl}) - H(U_{1k}, U_{2k} | W_{1kl}, W_{2kl}) \\
&= \sum_{k}^{K} \sum_{l}^{N} I(U_{1kl}, U_{2kl}; W_{1kl}, W_{2kl}),
\end{aligned}
$$

where once again we have used that fact that for fixed codes $\mathcal{C}_1$ and $\mathcal{C}_2$, the codewords I^K and J^K are functions of $\{U_1^N\}^K$ and $\{U_2^N\}^K$, respectively. Because $\{U_2^N\}^K - \{U_1^N\}^K - I^K$ form a Markov chain and U_1, U_2 are i.i.d random variables, so does $U_{2kl} \rightarrow U_{1kl} \rightarrow I^K$ and therefore $U_{2kl} \rightarrow U_{1kl} \rightarrow W_{1kl}$ forms a Markov chain. Similarly, it follows from $\{U_1^N\}^K \rightarrow \{U_2^N\}^K \rightarrow J^N$ and the fact that U_1, U_2 are drawn i.i.d that $U_{1k} \rightarrow U_{2k} \rightarrow J^K$ which in turn implies that $U_{1kl} \rightarrow U_{2kl} \rightarrow W_{2kl}$ forms a Markov chain as well. Notice that $U_{2kl} \rightarrow U_{1kl} \rightarrow W_{1kl}$ and $U_{1kl} \rightarrow U_{2kl} \rightarrow W_{2kl}$ do not imply $W_{1kl} \rightarrow U_{1kl} \rightarrow U_{2kl} \rightarrow W_{2kl}$ as W_{1kl} and W_{2kl} are not necessarily independent given U_{1kl} and U_{2kl}. Rewriting the previous inequalities as

$$
\begin{aligned}
R_1 &\geq \frac{1}{nN} \sum_{m=1}^{K \cdot N} I(U_{1m}, U_{2m}; W_{1m} | W_{2m}), \\
R_2 &\geq \frac{1}{nN} \sum_{m=1}^{K \cdot N} I(U_{1m}, U_{2m}; W_{1m} | W_{2m}), \\
R_1 + R_2 &\geq \frac{1}{nN} \sum_{m=1}^{K \cdot N} I(U_{1m}, U_{2m}; W_{1m}, W_{2m}),
\end{aligned}
$$

and introducing a time-sharing random variable Q, we get

$$
\begin{aligned}
R_1 &\geq \frac{1}{nN} \sum_{m=1}^{K \cdot N} I(U_{1m}, U_{2m}; W_{1m} | W_{2m}, Q = m) \\
&= I(U_{1Q}, U_{2Q}; W_{1Q} | W_{2Q}, Q) \\
R_2 &\geq \frac{1}{nN} \sum_{m=1}^{K \cdot N} I(U_{1m}, U_{2m}; W_{2m} | W_{1m}, Q = m) \\
&= I(U_{1Q}, U_{2Q}; W_{2Q} | W_{1Q}, Q)
\end{aligned}
$$

$$
\begin{aligned}
R_1 + R_2 &\geq \frac{1}{nN} \sum_{m=1}^{K \cdot N} I(U_{1m}, U_{2m}; W_{1m}, W_{2m} | Q = m) \\
&= I(U_{1Q}, U_{2Q}; W_{1Q}, W_{2Q} | Q).
\end{aligned}
$$

Since Q is independent of U_{1Q} and U_{2Q}, we have

$$\begin{aligned}
I(U_{1Q},U_{2Q};W_{1Q}|W_{2Q},Q) &= I(U_{1Q},U_{2Q};W_{1Q},Q|W_{2Q},Q) \\
I(U_{1Q},U_{2Q};W_{2Q}|W_{1Q},Q) &= I(U_{1Q},U_{2Q};W_{2Q},Q|W_{1Q},Q); \\
I(U_{1Q},U_{2Q};W_{1Q},W_{2Q}|Q) &= I(U_{1Q},U_{2Q};W_{1Q},W_{2Q},Q) \\
&\quad -I(U_{1Q},U_{2Q};Q) \\
&= I(U_{1Q},U_{2Q};W_{1Q},W_{2Q},Q)
\end{aligned} \tag{A.24}$$

Since U_{1Q} and U_{2Q} have the joint distribution $p(u_1u_2)$ in the theorem and the time-sharing variable does not alter the Markov structure of the random variables, defining $W_1 = (W_{1Q}, Q)$ and $W_2 = (W_{2Q}, Q)$, we have shown the existence of two random variables such that

$$R_1 \geq I(U_1U_2;W_1|W_2), \tag{A.25}$$

$$R_2 \geq I(U_1U_2;W_2|W_1), \tag{A.26}$$

$$R_1 + R_2 \geq I(U_1U_2;W_1W_2), \tag{A.27}$$

thus concluding the proof. ■

A.6 Proof of Converse for Theorem 4.9

Fix the encoders and the decoder. Let $I_0 = f_1(U_1^N, V_2^K)$ and $J_0 = f_2(U_2^N, V_1^K)$. It follows from Fano's inequality that

$$\begin{aligned}\frac{1}{N}H(U_1^N U_2^N|I_0 J_0) &\leq P_N \frac{1}{N}(\log\|\mathcal{U}_1^N \times \mathcal{U}_2^N\|) + \frac{1}{N} \\ &= \underbrace{P_N(\log\|\mathcal{U}_1\| + \log\|\mathcal{U}_2\|) + \frac{1}{N}}_{\lambda_N}\end{aligned} \tag{A.28}$$

where $\|\mathcal{U}_1\|$ and $\|\mathcal{U}_2\|$ are the alphabet sizes of U_1 and U_2, respectively. Notice that if $P_N \to 0$, λ_N must also converge to zero. Furthermore, we have

$$\frac{1}{N}H(U_1^N|I_0 J_0) \leq \lambda_N, \tag{A.29}$$

$$\frac{1}{N}H(U_2^N|I_0 J_0) \leq \lambda_N, \tag{A.30}$$

$$\frac{1}{N}H(U_1^N|I_0 J_0 U_2^N V_1^K V_2^K) \leq \lambda_N, \tag{A.31}$$

$$\frac{1}{N}H(U_2^N|I_0 J_0 U_1^N V_1^K V_2^K) \leq \lambda_N, \tag{A.32}$$

and so we can write the following chain of inequalities:

$$\begin{aligned}NR_1 &\geq H(I_0) \\ &\geq H(I_0|U_2^N V_1^K V_2^K J_0) \\ &\geq H(I_0|U_2^N V_1^K V_2^K U_1^N J_0) + I(I_0; U_1^N|U_2^N V_1^K V_2^K J_0) \\ &\overset{(a)}{=} I(I_0; U_1^N|U_2^N V_1^K V_2^K J_0) \\ &= H(U_1^N|U_2^N V_1^K V_2^K J_0) - H(U_1^N|I_0 U_2^N V_1^K V_2^K J_0) \\ &\geq H(U_1^N|U_2^N V_1^K V_2^K J_0) - \lambda_N \\ &\overset{(b)}{=} NH(U_1|U_2 Z) - \lambda_N,\end{aligned}$$

where (a) follows from the fact that I_0 is a function of U_1^N and V_2^K, and (b) results from setting $Z = (V_1^K V_2^K)$ and using the i.i.d. property of the sources and the Markov chain $U_1^N \to U_2^N V_1^K V_2^K \to J_0$. Similar arguments yield

$$NR_2 \geq NH(U_2|U_1 Z) - \lambda_N.$$

For the sum rate condition, we write

$$\begin{aligned}N(R_1 + R_2) &\geq H(I_0, J_0) \\ &= H(I_0 J_0|U_1^N U_2^N) + I(I_0 J_0; U_1^N U_2^N)\end{aligned}$$

$$
\begin{aligned}
&= I(I_0J_0;U_1^NU_2^N)\\
&= H(U_1^NU_2^N) - H(U_1^NU_2^N|I_0J_0)\\
&\geq H(U_1^NU_2^N) - \lambda_N\\
&= NH(U_1U_2) - \lambda_N.
\end{aligned}
$$

Finally, for the conference rates we have

$$
\begin{aligned}
NR_{12} &\geq \sum_{k=1}^{K} \log|\mathcal{V}_{1k}|\\
&\geq H(V_1^K)\\
&\geq H(V_1^K|U_2^N)\\
&= H(V_1^KV_2^K|U_2^N)\\
&= H(V_1^KV_2^K|U_2^N) - H(V_1^KV_2^K|U_1^NU_2^N)\\
&= I(V_1^KV_2^K;U_1^N|U_2^N)\\
&= \sum_{n=1}^{N} I(V_1^KV_2^K;U_{1n}|U_1^{N-1}U_2^N)\\
&= \sum_{n=1}^{N} H(U_{1n}|U_1^{N-1}U_2^N) - H(U_{1n}|V_1^KV_2^KU_1^{N-1}U_2^N)\\
&= \sum_{n=1}^{N} H(U_{1n}|U_{2n}) - H(U_{1n}|V_1^KV_2^KU_{2n})\\
&= \sum_{n=1}^{N} I(U_{1n};V_1^KV_2^K|U_{2n})\\
&= \sum_{n=1}^{N} I(U_{1n};Z_n|U_{2n}),
\end{aligned}
$$

where the first inequality is due to the admissibility condition for conferences, and the rest are standard information theoretic manipulations. Symmetric arguments yield

$$
NR_{21} \geq \sum_{n=1}^{N} I(U_{2n};Z_n|U_{1n}).
$$

Using a standard time-sharing argument (see, e.g. , [28, pg. 435]) we obtain the necessary conditions for distributed source coding with partial cooperation between the encoders. ■

A.7 Convexity of the Berger-Tung Inner Region

Let $(\mathbf{R}', \mathbf{D}') = (R_1', R_2', D_1', D_2')$ and $(\mathbf{R}'', \mathbf{D}'') = (R_1'', R_2'', D_1'', D_2'')$. We want to show that the rate-distortion region defined by (4.1)-(4.3)— the Berger-Tung inner bound, denoted $\mathcal{R}_{\mathcal{BT}}(\mathcal{D})$ is convex in the sense that if $(\mathbf{R}', \mathbf{D}') \in \mathcal{R}_{\mathcal{BT}}(\mathcal{D})$ and $(\mathbf{R}'', \mathbf{D}'') \in \mathcal{R}_{\mathcal{BT}}(\mathcal{D})$ then $(\lambda\, \mathbf{R}' + (1 -$

$\lambda)\mathbf{R}'', \lambda\, \mathbf{D}' + (1-\lambda)\mathbf{D}'') \in \mathcal{R}_{\mathcal{BT}}(\mathcal{D})$.

Assume that (D_1', D_2') is the distortion pair achieved by the rate pair (R_1', R_2') with the auxiliary random variable W_1', W_2' and the decoding function g'. Similarly, assume that (D_1'', D_2'') is the distortion pair achieved by (R_1'', R_2'') with W_1'', W_2'' and g''. Let Q be a random variable independent of $U_1, U_2, W_1', W_2', W_1''$, and W_2'' which takes on the value 1 with probability λ and 2 with probability $1-\lambda$. Furthermore let $S_1 = f_1(U_1)$ and $S_2 = f_2(U_2)$ denote the encoded versions of U_1 and U_2, respectively.

Define $W_1 = (W_{1Q}, Q)$ and $W_2 = (W_{2Q}, Q)$, such that $g(S_1, S_2, W_1, W_2) = g'(S_1, S_2, W_1', W_2')$ with probability λ and $g(S_1, S_2, W_1, W_2) = g''(S_1, S_2, W_1'', W_2'')$ with probability $1-\lambda$. This results in the following distortions:

$$\begin{aligned}
D_1 &= Ed(U_1, \hat{U}_1) \\
&= \lambda Ed(U_1, g_1'(S_1, S_2, W_1', W_2')) + (1-\lambda) Ed(U_1, g_1''(S_1, S_2, W_1'', W_2'') \\
&= \lambda D_1' + (1-\lambda) D_2'',
\end{aligned}$$

where g_1 denotes the first component of g. Similarly,

$$\begin{aligned}
D_2 &= Ed(U_2, \hat{U}_2) \\
&= \lambda Ed(U_2, g_2'(S_1, S_2, W_1', W_2')) + (1-\lambda) Ed(U_2, g_2''(S_1, S_2, W_1'', W_2'') \\
&= \lambda D_1' + (1-\lambda) D_2'',
\end{aligned}$$

where g_2 denotes the second component of g. On the other hand, the expression in (4.1) becomes

$$\begin{aligned}
I(U_1U_2; W_1|W_2) &= H(U_1U_2|W_2) - H(U_1U_2|W_1W_2) \\
&= H(U_1U_2|W_{2Q}Q) - H(U_1U_2|W_{1Q}W_{2Q}Q) \\
&= \lambda H(U_1U_2|W_2') + (1-\lambda)(U_1U_2|W_2'') \\
&\quad -\lambda H(U_1U_2|W_1'W_2') - (1-\lambda) H(U_1U_2|W_1''W_2'') \\
&= \lambda\, I(U_1U_2; W_1'|W_2') + (1-\lambda) I(U_1U_2; W_1''|W_2''),
\end{aligned}$$

and, similarly,

$$I(U_1U_2; W_2|W_1) = \lambda\, I(U_1U_2; W_2'|W_1') + (1-\lambda) I(U_1U_2; W_2''|W_1'').$$

For the sum rate condition in (4.3) we get

$$\begin{aligned}
I(U_1U_2; W_1W_2) &= H(U_1U_2) - H(U_1U_2|W_1W_2) \\
&= H(U_1U_2) - H(U_1U_2|W_{1Q}W_{2Q}Q) \\
&= H(U_1U_2) - \lambda H(U_1U_2|W_1'W_2') - (1-\lambda) H(U_1U_2|W_1''W_2'') \\
&= \lambda I(U_1U_2; W_1'W_2') + (1-\lambda) I(U_1U_2; W_1''W_2''),
\end{aligned}$$

thus proving the convexity of $\mathcal{R}_{\mathcal{BT}}(\mathcal{D})$. ■

A.8 Proof of Lemma 5.1

The proof of part 1) is straightforward: if $\mathbf{b}_k$ consists of at most a single element, this element must be contained in $\mathbf{a}_l$ for some $l < k$ according to definition 1.

To prove part 2), we start with the CCRE $\hat{p}'(\mathbf{u}) = \prod_{k=1}^{N_f} p(\mathbf{a}_k)$, which is derived from $\hat{p}(\mathbf{u})$ by removing all conditions $\mathbf{b}_k$. The factor graph of this CCRE is a tree (more precisely, a forest), since the subsets $\mathbf{a}_k$ are pairwise disjoint again according to definition 1. The N_f subtrees corresponding to the factors $p(\mathbf{a}_k)$ are connected to a complete tree by adding exactly $N-1$ extra edges to the graph, such that each edge starts in the function node of a $p(\mathbf{a}_k)$. This results from adding $\mathbf{b}_k$ conditions to the $p(\mathbf{a}_k)$ factors for all $k = 1, ..., N$, such that $\mathbf{b}_1$ is empty and all other $\mathbf{b}_k$ consist of exactly one element, as stated in the lemma. This construction also serves to prove part 2) of Theorem 5.1.

A.9 Proof of Lemma 5.2

Let $[\mathbf{R}_{\mathbf{a}_k}]_K$ denote the expansion of $\mathbf{R}_{\mathbf{a}_k}$ to a $K \times K$ matrix, where the non-zero entries correspond to the positions of the $\mathbf{a}_k$ elements in $\mathbf{u}$, e.g.,

$$\mathbf{R}_{u_1,u_3} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \to [\mathbf{R}_{u_1,u_3}]_5 = \begin{bmatrix} a & 0 & b & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ c & 0 & d & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Using this notation, the inverse covariance matrix $\hat{\mathbf{R}}^{-1}$ of the PDF $\hat{p}(\mathbf{u}) = \prod_{k=1}^{N_f} p(\mathbf{a}_k|\mathbf{b}_k)$ can be written as

$$\hat{\mathbf{R}}^{-1} = \sum_{k=1}^{N_f} [\mathbf{R}_{\mathbf{u}_k}^{-1}]_K - [\mathbf{R}_{\mathbf{b}_k}^{-1}]_K, \tag{A.33}$$

where $\mathbf{u}_k = (\mathbf{a}_k, \mathbf{b}_k)$ and $\mathbf{R}_{\mathbf{u}_k}$ and $\mathbf{R}_{\mathbf{b}_k}$ are the covariance matrices of the zero-mean Gaussian PDFs $p(\mathbf{u}_k) = p(\mathbf{a}_k, \mathbf{b}_k)$ and $p(\mathbf{b}_k)$, respectively. This follows from the equivalence $p(\mathbf{a}_k|\mathbf{b}_k) = p(\mathbf{u}_k)/p(\mathbf{b}_k)$ and the definition of a Gaussian PDF in (5.1). It is easy to see that $\hat{p}(\mathbf{u})$ given by $\mathcal{N}(\mathbf{0}_M, \hat{\mathbf{R}})$ is a zero-mean Gaussian PDF and that the elements of $\hat{\mathbf{R}}^{-1}$ are zero at the zero-positionsof $\mathbf{P}$, which proves parts 1) and 2) of the lemma.

The proof of part 3) follows trivially from [54, Corollary 1.2]. For details see [60].

To prove part 4), assume that the factor $p(\mathbf{a}_k|\mathbf{b}_k)$ in $\hat{p}(\mathbf{u})$ can be replaced by $p(\mathbf{u}_k)/p(\mathbf{b}_k)$ while $1/p(\mathbf{b}_k)$ cancels with the argument $\mathbf{u}_l$ of another factor $p(\mathbf{a}_l|\mathbf{b}_l)$, $l \neq k$, of $\hat{p}(\mathbf{u})$. This is possible for symmetric CCREs, since $\mathbf{b}_k$ is contained in $(\mathbf{a}_l, \mathbf{b}_l)$ for some $l < k$, which yields

$$p(\mathbf{a}_l|\mathbf{b}_l)/p(\mathbf{b}_k) = p(\mathbf{u}_l)/p(\mathbf{b}_k)/p(\mathbf{b}_l) = p(\mathbf{u}_l'|\mathbf{b}_k)/p(\mathbf{b}_l),$$

where $\mathbf{u}_l'$ contains the remaining elements of $\mathbf{u}_l$ after taking out all those in $\mathbf{b}_k$. This replacement can be repeated recursively to cancel $p(\mathbf{b}_l)$ with $p(\mathbf{a}_m|\mathbf{b}_m)$ for some $m < l$ and so forth until the empty set $\mathbf{b}_1$ is reached. Thus, with a symmetric CCRE it is possible to factor $\hat{p}(\mathbf{u})$ into $p(\mathbf{u}_k)$ times a product of PDFs where all elements in $\mathbf{u}_k$ appear in the conditioning part, only.

The true source distribution $p(\mathbf{u})$ can always be factored into $p(\mathbf{u}_k)$ times a PDF where $\mathbf{u}_k$ is in the conditioning part using a suitable chain rule expansion. It follows that the variables in $\mathbf{u}_k$ are Gaussian distributed with zero mean and covariance matrix $\mathbf{R}_{\mathbf{u}_k}$ according to either $\hat{p}(\mathbf{u})$ or $p(\mathbf{u})$, i.e. $\hat{\mathbf{R}}$ and $\mathbf{R}$ must have identical entries for all variable pairs $(u_l, u_{l'})$ in $\mathbf{u}_k$.

A.10 Example for the Construction of $\hat{\mathrm{R}}$ for a given CCRE

***Example* A.1** *Suppose we are given a Gaussian source distribution $p(\mathbf{u}) = p(u_1 u_2 u_3 u_4)$ with zero mean and the covariance matrix*

$$\mathbf{R} = \begin{bmatrix} 1 & 0.7 & 0.5 & 0.2 \\ 0.7 & 1 & 0.6 & 0.6 \\ 0.5 & 0.6 & 1 & 0.3 \\ 0.2 & 0.6 & 0.3 & 1 \end{bmatrix}, \tag{A.34}$$

and our task is to compute the inverse covariance matrix of the symmetric CCRE $\hat{p}(\mathbf{u}) = p(u_1 u_2)p(u_3|u_1)p(u_4|u_3)$ from Example 5.3. For each factor, we can write the inverse covariance matrix in a straightforward manner. For instance, taking the factor $p(\mathbf{a}_2|\mathbf{b}_2) = p(u_3|u_1)$ we can compute the latter according to

$$\underbrace{\begin{bmatrix} * & 0 & * & 0 \\ 0 & 0 & 0 & 0 \\ * & 0 & * & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{[\mathbf{R}_{\mathbf{u}_2}^{-1}]-[\mathbf{R}_{\mathbf{b}_2}^{-1}]} = \underbrace{\begin{bmatrix} * & 0 & * & 0 \\ 0 & 0 & 0 & 0 \\ * & 0 & * & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{[\mathbf{R}_{\mathbf{u}_2}^{-1}]} - \underbrace{\begin{bmatrix} * & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{[\mathbf{R}_{\mathbf{b}_2}^{-1}]}.$$

Here, $*$ represents an element that is non-zero in general. Notice that all matrix entries corresponding to arguments u_i that are not present in the factor are set to zero. Now, from the properties of the Gaussian distribution it follows that $\hat{\mathbf{R}}^{-1}$ is equal to the sum of the individual inverse covariance matrices obtained for each factor. Thus, we can write

$$\underbrace{\begin{bmatrix} * & * & * & 0 \\ * & * & 0 & 0 \\ * & 0 & * & * \\ 0 & 0 & * & * \end{bmatrix}}_{\hat{\mathbf{R}}^{-1}} = \underbrace{\begin{bmatrix} * & * & 0 & 0 \\ * & * & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{[\mathbf{R}_{\mathbf{u}_1}^{-1}]-[\mathbf{R}_{\mathbf{b}_1}^{-1}]} + \underbrace{\begin{bmatrix} * & 0 & * & 0 \\ 0 & 0 & 0 & 0 \\ * & 0 & * & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{[\mathbf{R}_{\mathbf{u}_2}^{-1}]-[\mathbf{R}_{\mathbf{b}_2}^{-1}]} + \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & * & * \\ 0 & 0 & * & * \end{bmatrix}}_{[\mathbf{R}_{\mathbf{u}_3}^{-1}]-[\mathbf{R}_{\mathbf{b}_3}^{-1}]}.$$

In numerical terms, this expression translates to

$$\underbrace{\begin{bmatrix} 2.29 & -1.37 & -0.67 & 0 \\ -1.37 & 1.96 & 0 & 0 \\ -0.67 & 0 & 1.43 & -0.33 \\ 0 & 0 & -0.33 & 0.10 \end{bmatrix}}_{\hat{\mathbf{R}}^{-1}} = \underbrace{\begin{bmatrix} 1.96 & -1.37 & 0 & 0 \\ -1.37 & 1.96 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{[\mathbf{R}_{\mathbf{u}_1}^{-1}]-[\mathbf{R}_{\mathbf{b}_1}^{-1}]} + \underbrace{\begin{bmatrix} 0.33 & 0 & -0.67 & 0 \\ 0 & 0 & 0 & 0 \\ -0.67 & 0 & 1.33 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{[\mathbf{R}_{\mathbf{u}_2}^{-1}]-[\mathbf{R}_{\mathbf{b}_2}^{-1}]} + \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.10 & -0.33 \\ 0 & 0 & -0.33 & 0.10 \end{bmatrix}}_{[\mathbf{R}_{\mathbf{u}_3}^{-1}]-[\mathbf{R}_{\mathbf{b}_3}^{-1}]}$$

Thus, for the approximate covariance matrix and its difference to the given covariance matrix $\mathbf{R}$ *in (A.34) we get*

$$\hat{\mathbf{R}} = \begin{bmatrix} 1 & 0.7 & 0.5 & 0.15 \\ 0.7 & 1 & 0.35 & 0.11 \\ 0.5 & 0.35 & 1 & 0.3 \\ 0.15 & 0.11 & 0.3 & 1 \end{bmatrix} \qquad \mathbf{R} - \hat{\mathbf{R}} = \begin{bmatrix} 0 & 0 & 0 & 0.05 \\ 0 & 0 & 0.25 & 0.49 \\ 0 & 0.25 & 0 & 0 \\ 0.05 & 0.49 & 0 & 0 \end{bmatrix}.$$

Now, the covariance matrix $\hat{\mathbf{R}}$ *of the approximation* $\hat{p}(\mathbf{u})$ *is different from the real covariance matrix* $\mathbf{R}$.

Next, we verify the validity of Lemma 5.2. The entries of $\hat{\mathbf{R}}^{-1}$ *are zero for all index pairs* (l, l') *not in* $\mathcal{I}$ *as stated in part (2) of Lemma 5.2. The entries of* $\hat{\mathbf{R}}$ *are equal to those in* $\mathbf{R}$ *for all index pairs* (l, l') *in* $\mathcal{I}$ *because the expansion is symmetric(cf. part (3) of Lemma 5.2). Note also that the index set* $\mathcal{I}$ *belonging to the extension consists of all pairs* (l, l') *which can be selected from the sets* $\{1, 2\}$, $\{1, 3\}$, *or* $\{3, 4\}$ *corresponding to the different factors. The validation of part (4) is obvious.*

A.11 Proof of Theorem 5.3

The first step of the proof is to show that the KLD-optimal functions $f_k(\mathbf{u}_k)$ and, thus, the PDF $\hat{p}(\mathbf{u})$ must be Gaussian given that $p(\mathbf{u})$ is zero-mean Gaussian. This is shown in [60]. The second step is to show that the factors $f_k(\mathbf{u}_k) = p(\mathbf{a}_k|\mathbf{b}_k)$ are the KLD-optimal functions: Let $\mathcal{S}$ be the set of all positive definite $M \times M$ matrices, whose entries are equal to those in $\mathbf{R}$ for all one-positions in $\mathbf{P}$ whereas the other entries are arbitrary. Let $\mathcal{S}'$ be the set of all positive definite $M \times M$ matrices, whose inverse has zero entries for all zero-positions in $\mathbf{P}$ whereas the other entries are arbitrary. From Theorem 2 in [61] follows that for any $\mathbf{A} \in \mathcal{S}$ and any $\mathbf{B} \in \mathcal{S}'$ the following inequality holds

$$D(\mathcal{N}(\mathbf{0}_M, \mathbf{A})||\mathcal{N}(\mathbf{0}_M, \tilde{\mathbf{B}})) \leq D(\mathcal{N}(\mathbf{0}_M, \mathbf{A})||\mathcal{N}(\mathbf{0}_M, \mathbf{B})),$$

where $\tilde{\mathbf{B}}$ is that unique matrix from $\mathcal{S}'$, whose entries are equal to those in $\mathbf{R}$ for all one-positions in $\mathbf{P}$, i.e., $\tilde{\mathbf{B}} \in \mathcal{S}$. A covariance matrix $\hat{\mathbf{R}}$ of the PDF $\hat{p}(\mathbf{u})$ constructed from a symmetric CCRE is an element of both $\mathcal{S}$ (part 2 of Lemma 5.2) and $\mathcal{S}'$ (part 4 of Lemma 5.2), i.e., $\hat{\mathbf{R}}$ is equal to $\tilde{\mathbf{B}}$. Since the true source distribution $\mathbf{R}$ is an element from $\mathcal{S}$, it follows that $D(p(\mathbf{u})||\hat{p}(\mathbf{u}))$ given by $D(\mathcal{N}(\mathbf{0}_M, \mathbf{R})||\mathcal{N}(\mathbf{0}_M, \hat{\mathbf{R}}))$ is the smallest KLD among all Gaussian PDFs, whose covariance matrix is an element from $\mathcal{S}'$. Finally, the elements in $\mathcal{S}'$ represent the admissible factorizations $\prod_{k=1}^{N_f} f_k(\mathbf{u}_k)$ of $\hat{p}(\mathbf{u})$, i.e., a Gaussian PDF $\hat{p}(\mathbf{u})$ constructed from a symmetric CCRE yields the KLD-optimal factors $f_k(\mathbf{u}_k)$ given by $p(\mathbf{a}_k|\mathbf{b}_k)$.

Since $p(\mathbf{u})$ and $\hat{p}(\mathbf{u})$ are Gaussian, computing the KLD $D(p(\mathbf{u})||\hat{p}(\mathbf{u}))$ simplifies to

$$D(p(\mathbf{u})||\hat{p}(\mathbf{u})) = \frac{1}{2}(-\log_2(|\mathbf{R}|\,|\hat{\mathbf{R}}^{-1}|) + \mathrm{tr}(\mathbf{R}\hat{\mathbf{R}}^{-1}) - M) = -\frac{1}{2}\log_2(|\mathbf{R}|\,|\hat{\mathbf{R}}^{-1}|),$$

as shown in [54, 60], where the last line follows from part 3) of Lemma 5.2. Applying (A.33) to $\hat{\mathbf{R}}^{-1}$ yields the formula (5.11) in the theorem.

Nomenclature

$\mathcal{I}^K$	The set of possible blocks of K indices i,	**68**
$\mathcal{J}^K$	The set of possible blocks of K indices j,	**68**
$(\cdot)^T$	The transposition operator,	**94**
β	A positive constant for the exponential decay of the correlation,	**94**
$\mathbf{i}$	A vector of quantization indices,	**95**
$\mathbf{R}_{\mathbf{u}_k}$	A covariance matrix with zero entries except for the variables in $\mathbf{u}_k$,	**105**
$\mathcal{U}$	Discrete alphabet for the source U,	**9**
$\mathcal{X}$	Channel input alphabet,	**9**
$\mathcal{Y}$	Channel output alphabet,	**9**
$\mathcal{R}_{\mathcal{BT}}(\mathcal{D})$	The Berger-Tung inner region,	**137**
Δ	The difference between two values,	**42**
δ	A small positive number,	**10**
$\delta(P_e)$	A positive value that goes to zero as P_e goes to zero,	**124**
δ_i	A small positive number indexed by i,	**68**
ϵ	A small positive number,	**12**
η	A small positive number,	**11**
$\exp(\cdot)$	The exponential operator,	**94**
$\mathrm{E}[\cdot]$	The expected value,	**20**
γ	A normalizing constant,	**95**
$\hat{X}$	An estimate of X,	**15**
$\hat{\mathbf{R}}$	An approximation of the covariance matrix $\mathbf{R}$,	**104**
$\hat{p}(\mathbf{i})$	An approximation of the PMF $p(\mathbf{i})$,	**98**
$\hat{p}(\mathbf{u})$	An approximation of the PDF $p(\mathbf{u})$,	**98**
$\hat{u}_k$	An estimate of the sample u_k observed by sensor k,	**95**
∞	Infinity,	**9**
$\int$	The integral operator,	**100**
λ_i	The i-th coefficient of a convex combination,	**74**
λ_N	A sequence of numbers that goes to zero as P_N goes to zero.,	**36**
$\mathcal{L}$	The set of quantization indices,	**94**
$\log$	The logarithm to base two,	**9**
$\hat{\mathcal{U}}$	Reconstruction alphabet,	**9**
$\mathcal{X}^N$	The set of codewords x^N with N symbols,	**9**
$\mathcal{Y}^N$	The set of channel output sequences y^N with N symbols,	**9**
$\mathbf{z}_m$	The position vector of sensor m,	**94**
$\overline{D}_i$	The expected value of the distortion for source U_i,	**62**
$\mathcal{N}(\boldsymbol{\mu}, \mathbf{R})$	A multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{R}$,	**94**
$\rho_{i,j}$	The correlation between the measurements of sensors i and j,	**94**

σ^2	Noise power in an AWGN channel,	**56**
σ^2	The noise variance,	**95**
τ_i	The time fraction allocated to the i-th user,	**56**
Cov$\{\mathbf{x}, \mathbf{y}\}$	The covariance between $\mathbf{x}$ and $\mathbf{y}$,	**93**
$\tilde{u}(i_k)$	The reconstruction value indexed by i_k,	**94**
$\{U_i^N\}^K$	A random block of K blocks of N source symbols U_i,	**68**
a	A symbol in the alphabet $\mathcal{X}$,	**10**
$a * b$	Shorthand notation for the operation $(1-a)b + (1-b)a$,	**80**
a_i	The i-th letter of alphabet $\mathcal{X}$,	**10**
b_i	The bin index produced by encoder i,	**68**
C	The channel capacity,	**16**
c_i	The i-th letter of the alphabet $\mathcal{Y}$,	**10**
C_{ij}	The capacity of the conference link between encoders i and j,	**25**
D	A prescribed distortion value,	**19**
d	A distortion measure,	**19**
d_f	The degree of a factor node in a factor graph,	**98**
d_i	The distortion measure for source U_i,	**62**
d_v	The degree of a variable node in a factor graph,	**97**
f	Encoding function,	**9**
f_i	Encoding function i,	**35**
$f_k(\cdot)$	The k-th factor of a chain rule expansion for the PDF $p(\mathbf{u})$,	**99**
g	Decoding function,	**9**
$g_k(\cdot)$	The k-th factor of a chain rule expansion for the PMF $p(\mathbf{i})$,	**98**
$H(X)$	The entropy of X,	**11**
$H(XY)$	The joint entropy of X and Y,	**11**
$H_b(\cdot)$	The binary entropy function,	**15**
h_{ik}	Conference function used by encoder i at time k,	**44**
$I(X;Y)$	The mutual information between X and Y,	**12**
i_k	The quantization index produced by sensor k,	**94**
K	A generic block length,	**44**
$l_{i,j}$	The distance between sensors i and j,	**94**
M	Number of information sources, number of sensor nodes in a network,	**23**
$m(i_k)$	The marginal probability of i_k,	**96**
M_i	Number of partition cells for alphabet $\mathcal{U}_i$,	**45**
N	Number of symbols in a sequence, also called block length,	**9**
N_s	Number of channel symbols,	**18**
$p(u)$	Probability distribution of the random variable U,	**9**
$p(u_1 u_2 \dots u_M)$	Joint probability distribution of M sources U_i,	**23**
$p(xy)$	Joint probability distribution of X and Y,	**10**
p_0	The crossover probability of a binary symmetric channel,	**80**
P_e	The single-letter error probability,	**15**
P_i	The power allocated to the i-th user,	**56**
p_k	The index of the main variable in a factor,	**99**
P_N	Block Error Probability for N symbols,	**9**
$P_{x^N y^N}$	Joint type of x^N and y^N,	**10**
P_{x^N}	The type or the empirical distribution of a sequence x^N of N symbols,	**10**
Q	The number of quantization bits per source sample,	**94**

q_k	The index of the conditioning variable in a factor,	**99**
R	The rate of a code,	**9**
$R(D)$	The rate-distortion function,	**20**
$R(S)$	The sum of rates R_i such that $i \in S$,	**23**
$r_i(\cdot)$	Real numbers that depend on the distortion values,	**131**
r_i^*	Rate values for the boundaries of the Berger-Tung region,	**80**
R_s	The channel coding rate,	**18**
R_s	The source coding rate,	**18**
R_{ij}	The conferencing rate at which encoder i communicates with encoder j,	**30**
S	A set of indices,	**23**
S^c	The complement of S,	**23**
U	Random variable describing a source of information,	**9**
$U(S)$	The set of information sources U_i such that $i \in S$,	**23**
U_i'	Auxiliary random variable for the source U_i,	**46**
u_k	The source sample observed by sensor k,	**94**
v_i	The cell index for a partition of the alphabet $\mathcal{U}_i$,	**45**
V_i^N	A block of N conference messages sent by encoder i,	**46**
V_{ik}	The message sent by encoder i at the k-th transmission.,	**43**
X	Random variable describing the channel input,	**9**
X^N	A random codeword,	**13**
x^N	A codeword with N symbols,	**9**
Y	Random variable describing the channel output,	**9**
y^N	A block of channel outputs,	**9**
Z	Auxiliary random variable,	**25**
$Z(S)$	A set of auxiliary random variables Z_{ij} such that $i \in S$ or $j \in S$,	**33**
Z_{ij}	Auxiliary random variable for the conference between encoders i and j,	**33**
$\mathbf{0}_N$	The length-N all-zero column vector,	**93**
$\mathcal{R}(\mathcal{D})$	The multiterminal rate-distortion region,	**84**
$\mathcal{R}(D_1, D_2)$	The multiterminal rate-distortion region for distortions D_1 and D_2,	**62**
$\mathbf{A}$	A generic matrix,	**93**
$\mathbf{I}_N$	The $N \times N$ identity matrix,	**93**
$\mathbf{R}$	The covariance of a multivariate Gaussian distribution,	**94**
$\mathbf{x}_k$	The binary codeword corresponding to the quantization index i_k,	**94**
$\mathcal{A}_\epsilon^N(X)$	The weakly typical set of X subject to ϵ and N,	**12**
$\mathcal{B}$	A generic set,	**17**
$\mathcal{C}_i$	The codebook used by encoder i,	**68**
$\mathcal{S}$	A generic set,	**13**
$\mathcal{T}_\delta^N(X)$	The strongly typical set of X subject to δ and N,	**10**
$\mathcal{T}_\delta^N(XY)$	The strongly jointly typical set of X and Y subject to δ and N,	**10**
$\hat{p}_{\text{ns}}$	A non-symmetric constrained chain rule expansion,	**103**
$\hat{p}_{\text{s}}$	A symmetric constrained chain rule expansion,	**103**
$\mathbb{R}$	The set of real numbers,	**93**
$\mathbb{R}^+$	The set of positive real numbers,	**19**
$\boldsymbol{\mu}$	The mean value of a multivariate Gaussian distribution,	**93**
$x_{k,j}$	The j-th element of the binary codeword $\mathbf{x}_k$,	**94**
AWGN	Additive White Gaussian Noise,	**95**
BP	Belief Propagation,	**97**

BPSK	Binary Phase Shift Kkeying,	**94**
CEO	Chief Executive Officer,	**31**
cf	Compare,	**141**
CME	Conditional Mean Estimator,	**95**
dB	Decibel,	**111**
e.g.	For example,	**30**
FDMA	Frequency Division Multiple Access,	**31**
i.e.	That is,	**4**
KLD	Kullback-Leibler Distance,	**100**
MAC	Medium Access Control,	**32**
MMSE	Minimum Mean Square Error,	**117**
MSE	Mean Square Error,	**95**
PDF	Probability Density Function,	**94**
PMF	Probability Mass Function,	**95**
SNR	Signal-to-Noise Ratio,	**111**
TDMA	Time Division Multiple Access,	**31**

Bibliography

[1] A. Aaron and B. Girod. Compression with side information using turbo codes. In *Proc. IEEE Data Compression Conference, DCC*, Snowbird, UT, 2002.

[2] S. Adireddy and L. Tong. Exploiting decentralized channel state information for random access. Submitted to the *IEEE Transactions on Information Theory*. Available from `http://acsp.ece.cornell.edu/`.

[3] R. Ahlswede. Multi-way communication channels. In *Proc. 2nd International Symposium on Information Theory*, pages 23–52, Prague, 1971.

[4] R. Ahlswede and T. S. Han. On source coding with side information via a multiple-access channel and related problems in multi-user information theory. *IEEE Transactions on Information Theory*, 29(3):396–411, 1983.

[5] S. M. Aji and R. J. McEliece. The generalized distributive law. *IEEE Transactions on Information Theory*, 46(2):325–343, March 2000.

[6] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci. A survey on sensor networks. *IEEE Communications Magazine*, 40(8):102–114, August 2002.

[7] A. Amraoui, S. Dusad, and R. Urbanke. Achieving general points in the 2-user Gaussian MAC without time-sharing or rate-splitting by means of iterative coding. In *IEEE International Symposium on Information Theory*, page 334, Lausanne, Switzerland, June-July 2002.

[8] J. Barros, C. Peraki, and S. D. Servetto. Efficient network architectures for sensor reachback. In *Proceedings of the IEEE International Zuerich Seminar on Communications*, Zuerich, Switzerland, February 2004. Invited paper.

[9] J. Barros and S. D. Servetto. Coding theorems for the sensor reachback problem with partially cooperating nodes. *Network Information Theory*, Edts. G. Kramer and P. Gupta, Discrete Mathematics and Theoretical Computer Science (DIMACS) series, American Mathematical Society (AMS), Providence, Rhode Island, 2004.

[10] J. Barros and S. D. Servetto. The sensor reachback problem. Submitted to the *IEEE Transactions on Information Theory*, November 2003.

[11] J. Barros and S. D. Servetto. On the capacity of the reachback channel in wireless sensor networks. In *Proc. of the IEEE International Workshop Multimedia Signal Processing*, US Virgin Islands, 2002. Invited paper to the special session on *Signal Processing for Wireless Networks*.

[12] J .Barros and S. D. Servetto. An inner bound for the rate/distortion region of the multi-terminal source coding problem. In *Proc. of the 37th Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA, March 2003.

[13] J. Barros and S. D. Servetto. On the rate/distortion region for separate encoding of correlated sources. In *Proc. IEEE Symposium on Information Theory (ISIT)*, Yokohama, Japan, 2003.

[14] J. Barros and S. D. Servetto. Reachback capacity with non-interfering nodes. In *Proc. IEEE Symposium on Information Theory (ISIT)*, Yokohama, Japan, 2003.

[15] J. Barros and S. D. Servetto. A note on cooperative multiterminal source coding. In *Proc. of the 37th Annual Conference on Information Sciences and Systems (CISS)*, Princeton, NJ, USA, March 2004.

[16] J. Barros, M. Tüchler, and Seong P. Lee. Scalable source/channel decoding for large-scale sensor networks. In *Proc. of the IEEE International Conference in Communications (ICC2004), Paris*, June 2004.

[17] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, Inc., 1971.

[18] T. Berger. *The Information Theory Approach to Communications (G. Longo, ed.)*, chapter Multiterminal Source Coding. Springer-Verlag, 1978.

[19] T. Berger and R. W. Yeung. Multiterminal source encoding with one distortion criterion. *IEEE Transactions on Information Theory*, 35(2):228–236, 1989.

[20] T. Berger, Z. Zhang, and H. Viswanathan. The CEO problem. *IEEE Transactions on Information Theory*, 42(3):887–902, 1996.

[21] C. Berrou, A. Glavieux, and P. Thitimajshima. Near Shannon limit error–correcting coding and decoding: Turbo-codes. In *Proc. IEEE International Conference on Communications (ICC)*, pages 1064–1070, Geneva, Switzerland, May 1993.

[22] D. Bertsekas and R. Gallager. *Data Networks (2nd ed)*. Prentice Hall, 1992.

[23] M. H. M. Costa. Writing on dirty paper. *IEEE Transactions on Information Theory*, 29(3):439–441, 1983.

[24] T. M. Cover. Broadcast channels. *IEEE Transactions on Information Theory*, 18:2–14, 1972.

[25] T. M. Cover. Comments on broadcast channels. *IEEE Transactions on Information Theory*, 44:2524–2530, October 1998.

[26] T. M. Cover, A. A. El Gamal, and M. Salehi. Multiple access channels with arbitrarily correlated sources. *IEEE Transactions on Information Theory*, 26(6):648–657, 1980.

[27] T. M. Cover and C. S. K. Leung. An achievable rate region for the multiple-access channel with feedback. *IEEE Transactions on Information Theory*, 27:292–298, May 1981.

[28] T. M. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., 1991.

[29] Imre Csiszár. On the error exponent of source-channel transmission with a distortion threshold. *IEEE Transactions on Information Theory*, 28:823–828, November 1982.

[30] Imre Csiszár and Janos Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Académiai Kiadó, Budapest, 1981.

[31] Z. Dawy and H. Kamoun. The general Gaussian relay channel: Analysis and insights. In 5^{th} *International ITG Conference on Source and Channel Coding (SCC'04)*, Erlangen, Germany, January 2004.

[32] C. R. Dietrich and G. N. Newsam. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107, 1997.

[33] S. Draper and G. Wornell. Side information aware coding strategies for sensor networks. To appear in the IEEE Journal on Selected Areas in Communications, 2004.

[34] G. Dueck. A note on the multiple access channel with correlated sources. *IEEE Transactions on Information Theory*, 27(2):232–235, 1981.

[35] M. Effros, M. Mdard, T. Ho, S. Ray, D. Karger, and R. Koetter. Linear network codes: A unified framework for source channel, and network coding. *Network Information Theory*, Edts. G. Kramer and P. Gupta, Discrete Mathematics and Theoretical Computer Science (DIMACS) series, American Mathematical Society (AMS), Providence, Rhode Island, 2004.

[36] R. G. Gallager. *Information Theory and Reliable Communication*. Wiley, 1968.

[37] H. El Gamal. On the scaling laws of dense wireless sensor networks. Submitted to the IEEE Transactions on Information Theory, April 2003.

[38] J. Garcia-Frias and Y. Zhao. Compression of correlated binary sources using turbo codes. *IEEE Communications Letters*, pages 417–419, 2001.

[39] M. Gastpar and M. Vetterli. Source-channel communication in sensor networks. In *Proceedings of the 2nd International Workshop on Information Processing in Sensor Networks (IPSN'03), Palo Alto, CA. Lecture Notes in Computer Science vol. 2634,* pages 162–177, New York, April 2003.

[40] M. Gondran and M. Minoux. *Graphs and Algorithms*. Wiley, 1984.

[41] R. M. Gray and T. Linder. Mismatch in high rate entropy constrained vector quantization. *IEEE Transactions on Information Theory*, 49:1204–1217, May 2003.

[42] Robert Grisso, Mark Alley, Phil McClellan, Dan Brann, and Steve Donohue. Precision farming: A comprehensive approach. Technical report, Virginia Tech, 2002. Available from `http://www.ext.vt.edu/pubs/bse/442-500/442-500.html`.

[43] J. Hagenauer. Source-controlled channel decoding. *IEEE Transactions on Communications*, 43(9):2449–2457, September 1995.

[44] J. Hagenauer, E. Offer, and L. Papke. Iterative decoding of binary block and convolutional codes. *IEEE Transactions on Information Theory*, 42(2):429–445, March 1996.

[45] Christoph Hausl. Scalable decoding for large-scale sensor networks. Diploma Thesis, Lehrstuhl für Nachrichtentechnik, Technische Universität München, München, Germany, April 2004.

[46] A. Hu and S. D. Servetto. Optimal detection for a distributed transmission array. In *Proc. of the IEEE International Symposium on Information Theory (ISIT),* Yokohama, Japan, June-July 2003.

[47] Zhihua Hu and Baochun Li. Fundamental performance limits of wireless sensor networks. to appear in Ad Hoc and Sensor Networks, Yang Xiao and Yi Pan, Editors, Nova Science Publishers, 2004.

[48] Zhihua Hu and Baochun Li. On the fundamental capacity and lifetime limits of energy-constrained wireless sensor networks. In *Proceedings of the 10th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS 2004),* pages 38–47, Toronto, Canada, May 2004.

[49] J. L. Massey (guest ed.). Special issue on random access communications. *IEEE Transactions on Information Theory*, 31(2), 1985.

[50] S. Jaggi and M. Effros. Universal multiple access source coding. Submitted to the *IEEE Transactions on Information Theory*. Manuscript available at `http://www.its.caltech.edu/`.

[51] S. Jaggi and M. Effros. Universal linked multiple access source codes. In *Proc. of the IEEE International Symposium on Information Theory*, Lausanne, Switzerland, June-July 2002.

[52] N. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice Hall, 1984.

[53] A. H. Kaspi and T. Berger. Rate-distortion for correlated sources with partially separated encoders. *IEEE Transactions on Information Theory*, 28(6):828–840, 1982.

[54] A. Kavcic and J. Moura. Matrices with banded inverses: Inversion algorithms and factorization of Gauss-Markov processes. *IEEE Transactions on Information Theory*, 46:1495–1509, July 2000.

[55] M. A. Khojastepour, A. Sabharwal, and B. Aazhang. Bounds on achievable rates for general multi-terminal networks with practical constraints. In *Proceedings of the 2nd International Workshop on Information Processing in Sensor Networks (IPSN)*, pages 146–161, April 2003.

[56] I. Kozintsev, R. Koetter, and K. Ramchandran. A framework for joint source-channel coding using factor graphs. In *Proc. 33rd Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, October 1999.

[57] G. Kramer, M. Gastpar, and P. Gupta. Capacity theorems for wireless relay channels. In *41st Annual Allerton Conf. on Commun., Control and Comp.*, Allerton, IL, USA, October 2003.

[58] F. R. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[59] J. N. Laneman, D. N. C. Tse, and G. W. Wornell. Cooperative diversity in wireless networks: Efficient protocols and outage behavior. *IEEE Transactions on Information Theory*, April 2003. Accepted for publication.

[60] Seong Per Lee. Iterative decoding of correlated sensor data. Diploma Thesis, Lehrstuhl für Nachrichtentechnik, Technische Universität München, München, Germany, October 2003.

[61] H. Lev-Ari, S. Parker, and T. Kailath. Multidimensional maximum-entropy covariance extension. *IEEE Transactions on Information Theory*, 35:497–508, May 1989.

[62] J. Li, N. Chaddha, and R. Gray. Asymptotic performance of vector quantizers with the perceptual distortion measure. *IEEE Transactions on Information Theory*, 45(4):1082–91, May 1999.

[63] J. Li, Z. Tu, and R. S. Blum. Slepian-Wolf coding for nonuniform sources using turbo codes. In *Proceeding of IEEE Data Compression Conference (DCC)*, Snowbird, UT, USA, March 2004.

[64] H. Liao. *Multiple Access Channels*. PhD thesis, Department of Electrical Engineering, University of Hawaii, Honolulu, Hawaii, 1972.

[65] H. Loeliger. *Least Squares and Kalman Filtering on Forney Graphs*. Codes, Graphs, and Systems, R.E. Blahut and R. Koetter, eds., Kluwer, 2002.

[66] M. Mecking. Multiple-Access With Stripping Receivers. In *Proc. 4th European Mobile Communication Conference (EPMCC)*, Vienna, Austria, 2001.

[67] Neri Merhav and Shlomo Shamai. On joint source-channel coding for the Wyner-Ziv source and the Gel'fand-Pinsker channel. *IEEE Transactions on Information Theory*, 49(11):2844–2855, November 2003.

[68] U. Mitra and A. Sabharwal. On achievable rates of complexity constrained relay channels. In *Proceedings of the 41st Allerton Conference on Communication, Control and Computing*, July 2003.

[69] P. Mitran and J. Bajcsy. Coding for the Wyner-Ziv problem with turbo-like codes. In *Proc. IEEE Int. Symp. Inform. Theory*, Lausanne, Switzerland, 2002.

[70] Murugan, P. Gopala, and H. El Gamal. Correlated sources and wireless channels: Cooperative source-channel coding. To appear in the IEEE Journal on Selected Areas in Communications, 2004.

[71] A. Narula, M. D. Trott, , and G. W. Wornell. Performance limits of coded diversity methods for transmitter antenna arrays. *IEEE Transactions on Information Theory*, 45(7):2418–2433, 1999.

[72] Y. Oohama. Universal coding for correlated sources with linked encoders. *IEEE Transactions on Information Theory*, 42(3):837–847, May 1996.

[73] Y. Oohama. Gaussian multiterminal source coding. *IEEE Transactions on Information Theory*, 43(6):1912–1923, November 1997.

[74] H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer-Verlag, 1994.

[75] S. S. Pradhan and K. Ramchandran. Distributed source coding using syndromes (DISCUS): Design and construction. In *Proc. IEEE Data Compression Conf. (DCC)*, Snowbird, UT, 1999.

[76] S. S. Pradhan and K. Ramchandran. On functional duality between MIMO source and channel coding with one-sided collaboration. In *Proc. IEEE Information Theory Workshop*, Bangalore, India, October 2002.

[77] John G. Proakis. *Digital Communications*. McGraw-Hill, New York, 4th edition, 2001.

[78] A. Scaglione and S. D. Servetto. On the interdependence of routing and data compression in multi-hop sensor networks. In *Proc. ACM MobiCom*, Atlanta, GA, 2002.

[79] A. Sendonaris, E. Erkip, and B. Aazhang. User cooperation diversity–part 1: System description. *IEEE Transactions on Communications*, 51(11), November 2003.

[80] A. Sendonaris, E. Erkip, and B. Aazhang. User cooperation diversity–part 2: Implementation aspects and performance analysis. *IEEE Transactions on Communications*, 51(11), November 2003.

[81] S. D. Servetto. Quantization with side information: Lattice codes, asymptotics, and applications in wireless networks. Accepted for the *IEEE Transactions on Information Theory*. Available from `http://people.ece.cornell.edu/servetto/`.

[82] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423 and 623–656, 1948.

[83] C.E. Shannon. Two-way communication channels. In *Proc. 4th Berkeley Symp. Math. Stat. Prob.*, volume 1, pages 611–644. Univ. California Press, 1961.

[84] D. Slepian and J. K. Wolf. A coding theorem for multiple access channels with correlated sources. *Bell Syst. Tech. J.*, 52(7):1037–1076, 1973.

[85] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, 1973.

[86] N. J. A. Sloane and A. D. Wyner(Edts.). *Claude Elwood Shannon: Collected Papers*. IEEE Press, Piscataway, NJ, 1993.

[87] V. Stankovic, A. Liveris, Z. Xiong, and C. Georghiades. Design of Slepian-Wolf codes by channel code partitioning. In *Proceeding of IEEE Data Compression Conference (DCC)*, Snowbird, UT, USA, March 2004.

[88] R. L. Rivest T. H. Cormen, C. E. Leiserson and C. Stein. *Introduction to Algorithms (2nd ed)*. MIT Press, 2001.

[89] K. Thulasiraman and M. N. S. Swamy. *Graphs: Theory and Algorithms*. John Wiley and Sons, Inc., 1992.

[90] S.Y. Tung. *Multiterminal source coding*. PhD thesis, Cornell University, 1978.

[91] H. Wang and P. Viswanath. Fixed binning schemes for channel and source coding problems: An operational duality. Submitted to the *IEEE Transactions on Information Theory*, September 2003. Available from `http://www.ifp.uiuc.edu/~pramodv/`.

[92] Hanan Weingarten, Yossef Steinberg, and Shlomo Shamai. The capacity region of the Gaussian MIMO broadcast channel. In *Proceedings of the 2004 IEEE International Symposium on Information Theory (ISIT 2004)*, Chicago, Illinois, USA, June/July 2004.

[93] F. M. J. Willems. The discrete memoryless multiple access channel with partially cooperating encoders. *IEEE Transactions on Information Theory*, 29(3):441–445, 1983.

[94] A. D. Wyner and J. Ziv. A theorem on the entropy of certain binary sequences and applications — part 1. *IEEE Transactions on Information Theory*, pages 769–772, November 1973.

[95] A. D. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, 1976.

[96] Y. Yang, V. Stankovic, Z. Xiong, and W. Zhao. Asymmetric code design for remote multiterminal source coding. In *Proc. IEEE Data Compression Conference (DCC'04)*, Snowbird, UT, March 2004.

[97] R. W. Yeung. *A First Course in Information Theory*. Kluwer Academic, 2002.

[98] Raymond Yeung. Separation principles for multi-user communication systems. Unpublished manuscript, available from the author by request.

[99] W. Yu. Duality and the value of cooperation in distributive source and channel coding problems. In *Proceedings of the Allerton Conference on Communications, Control and Computing*, October 2003.

[100] R. Zamir and T. Berger. Multiterminal source coding with high resolution. *IEEE Transactions on Information Theory*, 45(1):106–117, 1999.

[101] R. Zamir, S. Shamai, and U. Erez. Nested linear/lattice codes for structured multiterminal binning. *IEEE Transactions on Information Theory*, 48(6):1250–1276, 2002.

[102] W. Zhong, H. Lou, and J. Garcia-Frias. LDGM codes for joint source-channel coding of correlated sources. In *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.