ABSTRACT

ADAPTIVE DATA ACQUISITION FOR COMMUNICATION NETWORKS

by
Behzad Ahmadi

In an increasing number of communication systems, such as sensor networks or local area networks within medical, financial or military institutions, nodes communicate information sources (e.g., video, audio) over multiple hops. Moreover, nodes have, or can acquire, correlated information sources from the environment, e.g., from data bases or from measurements. Among the new design problems raised by the outlined scenarios, two key issues are addressed in this dissertation: 1) How to preserve the consistency of sensitive information across multiple hops; 2) How to incorporate the design of actuation in the form of data acquisition and network probing in the optimization of the communication network. These aspects are investigated by using information-theoretic (source and channel coding) models, obtaining fundamental insights that have been corroborated by various illustrative examples. To address point 1), the problem of cascade source coding with side information is investigated. The motivating observation is that, in this class of problems, the estimate of the source obtained at the decoder cannot be generally reproduced at the encoder if it depends directly on the side information. In some applications, such as the one mentioned above, this lack of consistency may be undesirable, and a so called Common Reconstruction (CR) requirement, whereby one imposes that the encoder be able to agree on the decoder's estimate, may be instead in order. The rate-distortion region is here derived for some special cases of the cascade source coding problem and of the related Heegard-Berger (HB) problem under the CR constraint. As for point 2), the work is motivated by the fact that, in order to enable, or to facilitate, the exchange of information, nodes of a communication network routinely take various types of

actions, such as data acquisition or network probing. For instance, sensor nodes schedule the operation of their sensing devices to measure given physical quantities of interest, and wireless nodes probe the state of the channel via training. The problem of optimal data acquisition is studied for a cascade source coding problem, a distributed source coding problem and a two-way source coding problem assuming that the side information sequences can be controlled via the selection of cost-constrained actions. It is shown that a joint design of the description of the source and of the control signals used to guide the selection of the actions at downstream nodes is generally necessary for an efficient use of the available communication links. Instead, the problem of optimal channel probing is studied for a broadcast channel and a point-to-point link in which the decoder is interested in estimating not only the message, but also the state sequence. Finally, the problem of embedding information on the actions is studied for both the source and the channel coding set-ups described above.

ADAPTIVE DATA ACQUISITION FOR COMMUNICATION NETWORKS

by
Behzad Ahmadi

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Electrical Engineering

Department of Electrical and Computer Engineering, NJIT

May 2013

APPROVAL PAGE

ADAPTIVE DATA ACQUISITION FOR COMMUNICATION NETWORKS

Behzad Ahmadi

Dr. Osvaldo Simeone, Dissertation Advisor — Date
Associate Professor, New Jersey Institute of Technology

Dr. Yeheskel Bar-Ness, Committee Member — Date
Distinguished Professor, New Jersey Institute of Technology

Dr. Ali Abdi, Committee Member — Date
Associate Professor, New Jersey Institute of Technology

Dr. Alexander Haimovich, Committee Member — Date
Professor, New Jersey Institute of Technology

Dr. Elza Erkip, Committee Member — Date
Professor, Polytechnic Institute of NYU

BIOGRAPHICAL SKETCH

Author: Behzad Ahmadi

Degree: Doctor of Philosophy

Date: May 2013

Undergraduate and Graduate Education:

- Doctor of Philosophy in Electrical Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2013
- Master of Science in Electrical Engineering,
  Isfahan University of Technology, Isfahan, Iran, 2008
- Bachelor of Science in Electrical Engineering,
  Isfahan University of Technology, Isfahan, Iran, 2005

Major: Electrical Engineering

Presentations and Publications:

B. Ahmadi, R. Tandon, O. Simeone and H. V. Poor, "Heegard-Berger and Cascade Source Coding Problems with Common Reconstruction Constraints," *IEEE Trans. Inform. Theory*, vol. 59, no. 3, pp. 1458-1474, Mar. 2013.

B. Ahmadi and O. Simeone, "Distributed and Cascade Lossy Source Coding with a Side Information "Vending Machine"," to appear in *IEEE Trans. Inform. Theory*, arXiv:1109.6665.

B. Ahmadi and O. Simeone, "Two-Way Communication with Adaptive Data Acquisition," to appear in *Transactions on Emerging Telecommunications Technologies*, arXiv:1209.5978.

B. Ahmadi and O. Simeone, "Two-Way Communication with Adaptive Data Acquisition," in *Proc. IEEE International Symposium on Information Theory* (ISIT 2013), Istanbul, Turkey, July 7-12, 2013.

B. Ahmadi, H. Asnani, O. Simeone and H. Permuter, "Information Embedding on Actions," in *Proc. IEEE International Symposium on Information Theory* (ISIT 2013), Istanbul, Turkey, July 7-12, 2013.

B. Ahmadi, O. Simeone, C. Choudhuri and U. Mitra, “On Cascade Source Coding with A Side Information “Vending Machine”,” in *Proc. IEEE Information Theory Workshop* (ITW 2012), Lausanne, Switzerland, Sept. 3-7, 2012.

B. Ahmadi and O. Simeone, “On Channels with Action-Dependent States,” in *Proc. IEEE Information Theory Workshop* (ITW 2012), Lausanne, Switzerland, Sept. 3-7, 2012.

B. Ahmadi and O. Simeone, “Distributed and Cascade Lossy Source Coding with a Side Information “Vending Machine”,” in *Proc. IEEE International Symposium on Information Theory* (ISIT 2012), Cambridge, MA, USA, July 1-6, 2012.

B. Ahmadi, R. Tandon, O. Simeone and H. V. Poor, “On the Heegard-Berger Problem with Common Reconstruction Constraints,” in *Proc. IEEE International Symposium on Information Theory* (ISIT 2012), Cambridge, MA, USA, July 1-6, 2012.

B. Ahmadi and O. Simeone, “Robust Coding for Lossy Computing with Receiver-Side Observation Costs,” in *Proc. IEEE International Symposium on Information Theory* (ISIT 2011), July 31-Aug. 5, Saint Petersburg, Russia, 2011.

B. Ahmadi, C. Choudhuri, O. Simeone and U. Mitra, “Cascade source coding with a side information “vending machine”,” submitted to *IEEE Trans. Inform. Theory*, arXiv:1207.2793.

B. Ahmadi, H. Asnani, O. Simeone and H. Permuter, “Information Embedding on Actions,” submitted to *IEEE Trans. Inform. Theory*, arXiv:1207.6084.

Dedicated to my wife and my parents.

ACKNOWLEDGMENT

I would like to express my heartfelt gratitude and deepest appreciation to my adviser Dr. Osvaldo Simeone for his wisdom, commitment, guidance and inspiration toward my research. His constructive criticism and trusting support made this work a great learning experience.

I would like to express my sincere gratitude to Dr. Yeheskel Bar-Ness, Dr. Ali Abdi, Dr. Alexander Haimovich, and Dr. Elza Erkip for serving as committee members. I appreciate their time as well as their encouraging and constructive comments, feedback and guides on the dissertation. Those are not only my teachers and mentors but dear friends.

Special thanks to Dr. Vincent H. Poor, Dr. Urbashi Mitra, Dr. Haim H. Permuter, Dr. Ravi Tandon, Dr. Chiranjib Choudhuri and Mr. Himanshu Asnani for collaboration with Dr. Simeone and me in this research, for their valuable suggestions and encouragement towards the realization of this project.

Ms. Marlene Toeroek and Ms. Angela Retino deserve a very special acknowledgment from all of us from CWCSPR. They were always ready to help us and they made everything easy.

Further thanks go to Dr. Marino Xanthos, Ms. Clarisa Gonzalez-Lenahan and the staff of the Graduate Studies office of NJIT, Mr. Jeffrey Grundy, Mr. Scott Kline and the staff of the Office for International Students and faculty for their advice, help and support with administrative matters during my PhD studies.

Special thanks go to Hashimoto Fellowship fund and for the financial support during my doctoral studies.

It is my honor to thank my family for being supportive and encouraging. Specifically, I would like to thank my wife, Mina, who has been with me, side by

side and through all ups and downs. Without her, I could not have achieved any of the serious goals.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

MOTIVATION AND OVERVIEW

In communication systems such as data networks, sensor networks and local area networks, sensitive data is transferred through multiple hops. Examples are financial, military and medical. This data often needs to be acquired from the "environment", such as from databases or via sensor measurements. Beside data acquisition, nodes also need to take various type of actions, e.g., to probe the state of the network prior to transmission.

The aim of this thesis is to address two important questions that arise in the design of the outlined communication networks: 1) How to preserve the consistency of sensitive information to be sent over multi-hop networks? 2) How to integrate the optimization of actuation tasks, such as for data acquisition or network probing, in the design of communication networks? These issues are addressed from an information-theoretic point of view. To this end, several source and channel coding systems are investigated that exemplify various key scenarios of interest.

To address the first issue, one has to guarantee that downstream nodes use the locally available information in such as way that no inconsistency is created with the state of knowledge of upstream nodes. This requirement is known in information theory as the Common Reconstruction (CR) constraint. Under this constraint, this thesis derives the rate-distortion performance for the so called Heegard-Berger (HB) problem and for the cascade source coding problem.

The second aforementioned issue is tackled by focusing on two specific actuation tasks, namely adaptive data acquisition and channel probing. Adaptive data acquisition addresses the problem of using efficiently the system resources in order to acquire information to the environment. Applications include sensor networks,

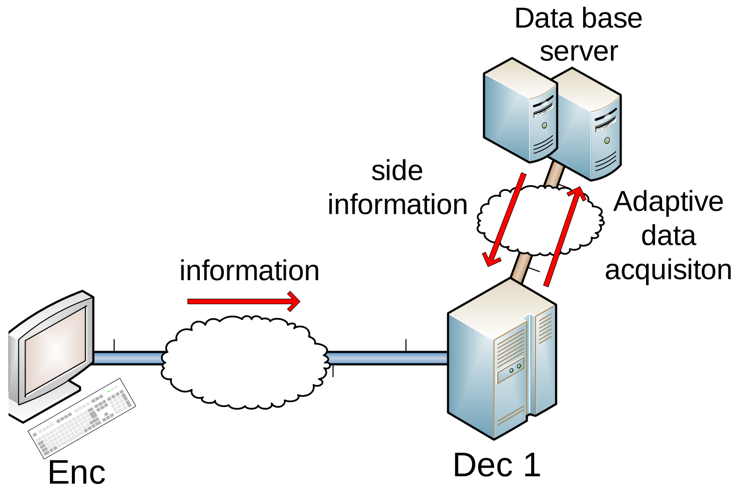

Figure 1.1 Source coding with adaptive data acquisition.

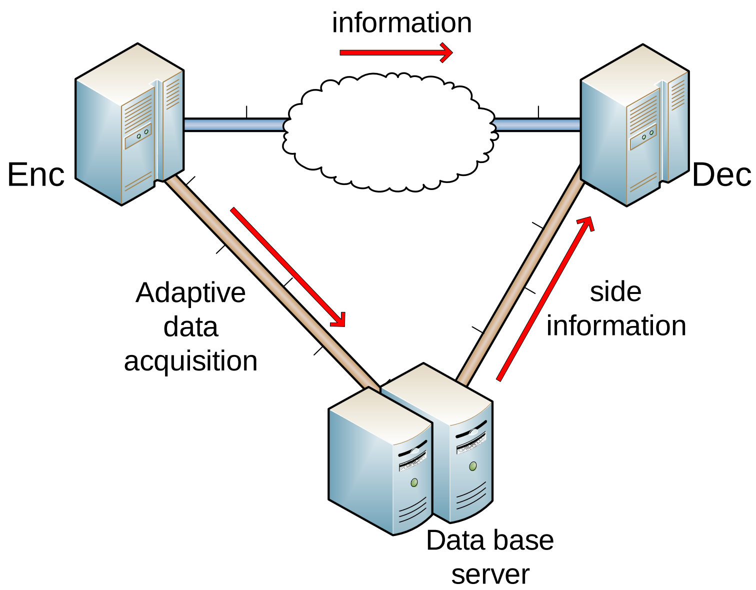

Figure 1.2 Source coding with adaptive data acquisition.

in which acquiring measurements entails an energy cost, and computer networks, in which accessing a database server is bandwidth and time consuming. The recently proposed information-theoretic model of side information "vending machine" (VM) accounts for such scenarios in which the acquisition and/or measurement of information sequences can be controlled via the selection of cost-constrained actions (see Figure 1.1 and Figure 1.2). In thesis, the rate-distortion-cost function is derived for a two-way source coding problem, a distributed source coding problem and a cascade source coding problems in which a side information VM is available and the intermediate and/or at the end node of the cascade (see Figure 1.3). One of

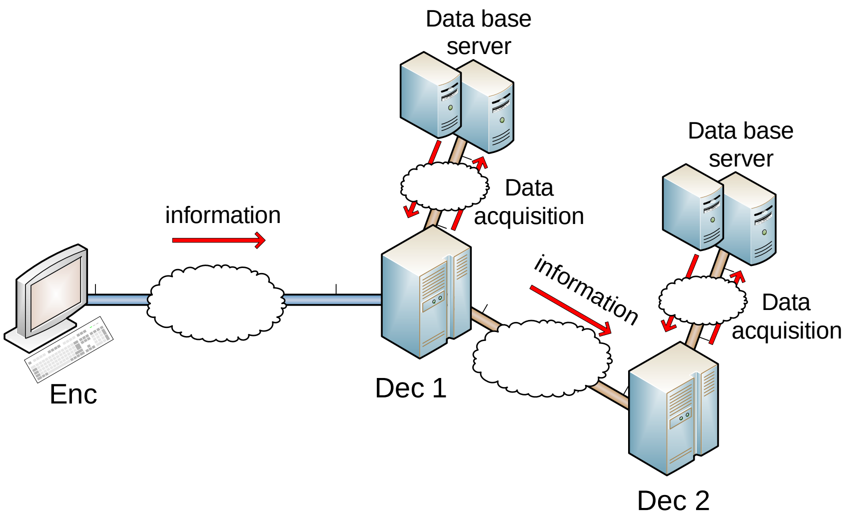

Figure 1.3 Cascade source coding with adaptive data acquisition.

the main conclusions is that a joint design of the description of the source and of the control signals used to guide the selection of the actions at downstream nodes is generally necessary for an efficient use of the available communication links. As for channel probing, the motivating observation is that acquiring information about the current state of the channel or, more generally, of the network, requires the exchange of control information, which uses system resources. For instance, wireless transceivers assess the channel quality via training and/or feedback; routers ascertain network congestion levels via the transmission of probing packets; and radio terminals switch among different operating modes, such as transmit, receive or idle (see Figure 1.4). The recently proposed information-theoretic model of action-dependent channels models such scenarios. In this thesis, the trade-off between capacity and action cost is studied for a broadcast channel with action-dependent state. Finally, the problem of embedding information on the actions is studied for both the source and the channel coding set-ups described above.

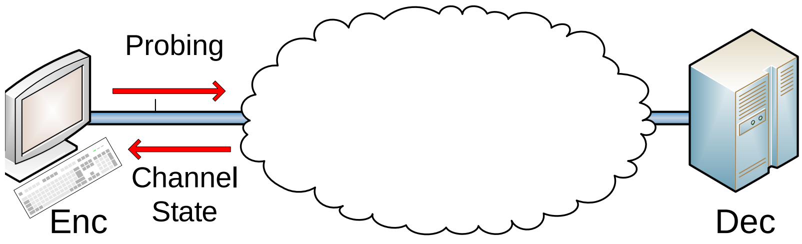

Figure 1.4 Channel probing.

1.1 Organization and Contributions

In this section, the main contributions and organization of the thesis are outlined.

Chapter 2: A very short introduction of the fundamental information theoretic concepts is provided in this Chapter.

Chapter 3: This Chapter investigates the problems Heegard-Berger(HB) and cascade source coding with common reconstruction constraint. The HB problem consists of an encoder broadcasting to two decoders with respective side information. The cascade source coding problem is characterized by a two-hop system with side information available at the intermediate and nal nodes. For the HB problem with the CR constraint, the rate-distortion function is derived under the assumption that the side information sequences are (stochastically) degraded. The rate-distortion function is also calculated explicitly for three examples, namely Gaussian source and side information with quadratic distortion metric, and binary source and side information with erasure and Hamming distortion metrics. The rate-distortion function is then characterized for the HB problem with cooperating decoders and (physically) degraded side information. For the cascade problem with the CR constraint, the rate-distortion region is obtained under the assumption that side information at the nal node is physically degraded with respect to that at the intermediate node. For the latter two cases, it is worth emphasizing that the corresponding problem without the

CR constraint is still open. Outer and inner bounds on the rate-distortion region are also obtained for the cascade problem under the assumption that the side information at the intermediate node is physically degraded with respect to that at the nal node. For the three examples mentioned above, the bounds are shown to coincide. Finally, for the HB problem, the rate-distortion function is obtained under the more general requirement of constrained reconstruction, whereby the decoders estimate must be recovered at the encoder only within some distortion.

The material in this chapter has been reported in the documents:

- B. Ahmadi, R. Tandon, O. Simeone and H. V. Poor, "Heegard-Berger and Cascade Source Coding Problems with Common Reconstruction Constraints," *IEEE Trans. Inform.* Theory, vol. 59, no. 3, pp. 1458,1474, Mar. 2013.

- B. Ahmadi, R. Tandon, O. Simeone and H. V. Poor, "On the Heegard-Berger Problem with Common Reconstruction Constraints," in *Proc. IEEE International Symposium on Information Theory* (ISIT 2012), Cambridge, MA, USA, July 1-6, 2012.

Chapter 4: In this Chapter, the analysis of the trade-offs between rate, distortion and cost associated with the control actions is extended from the previously studied point-to-point set-up to two basic multiterminal models. First, a distributed source coding model is studied, in which two encoders communicate over rate-limited links to a decoder, whose side information can be controlled. The control actions are selected by the decoder based on the messages encoded by both source nodes. For this set-up, inner bounds are derived on the rate-distortion-cost region for both cases in which the side information is available causally and non-causally at the decoder. These bounds are shown to be tight under specic assumptions, including the scenario in which the sequence observed by one of the nodes is a function of the source observed by the other and the side information is available causally at the decoder. Then, a cascade scenario in which three nodes are connected in a cascade and the last node has controllable side information, is also investigated. For this model, the rate-distortion-cost region is derived for general distortion requirements and under the assumption of causal availability of side information at the last node.

The material in this chapter has been reported in the documents:

- B. Ahmadi and O. Simeone, "Distributed and Cascade Lossy Source Coding with a Side Information "Vending Machine"," to appear in *IEEE Trans. Inform. Theory*, arXiv:1109.6665.
- B. Ahmadi and O. Simeone, "Distributed and Cascade Lossy Source Coding with a Side Information "Vending Machine"," in *Proc. IEEE International Symposium on Information Theory* (ISIT 2012), Cambridge, MA, USA, July 1-6, 2012.

Chapter 5:: In this Chapter, a three-node cascade source coding problem is studied under the assumption that a side information VM is available and the intermediate and/or at the end node of the cascade. A single-letter characterization of the achievable trade-off among the transmission rates, the distortions in the reconstructions at the intermediate and at the end node, and the cost for acquiring the side information is derived for a number of relevant special cases. It is shown that a joint design of the description of the source and of the control signals used to guide the selection of the actions at downstream nodes is generally necessary for an efcient use of the available communication links. In particular, for all the considered models, layered coding strategies prove to be optimal, whereby the base layer fullls two network objectives: determining the actions of downstream nodes and simultaneously providing a coarse description of the source. Design of the optimal coding strategy is shown via examples to depend on both the network topology and the action costs. Examples also illustrate the involved performance trade-offs across the network.

The material in this chapter has been reported in the documents:

- B. Ahmadi, C. Choudhuri, O. Simeone and U. Mitra, "Cascade source coding with a side information "vending machine"," submitted to *IEEE Trans. Inform. Theory*, arXiv:1207.2793.
- B. Ahmadi, O. Simeone, C. Choudhuri and U. Mitra, "On Cascade Source Coding with A Side Information "Vending Machine"," in *Proc. IEEE Information Theory Workshop* (ITW 2012), Lausanne, Switzerland, Sept. 3-7, 2012.

Chapter 6: In this Chapter, a bidirectional link is studied in which two nodes, Node 1 and Node 2, communicate to fulfill generally conflicting informational requirements. Node 2 is able to acquire information from the environment, e.g., via access to a remote database or via sensing. Information acquisition is expensive in terms of system resources, e.g., time, bandwidth and energy, and thus should be done efficiently by adapting the acquisition process to the needs of the application. As a result of the forward communication from Node 1 to Node 2, the latter wishes to compute some function, such as a suitable average, of the data available at Node 1 and of the data obtained from the environment. The forward link is also used by Node 1 to query Node 2 with the aim of retrieving suitable information from the environment on the backward link. The problem is formulated in the context of multi-terminal rate-distortion theory and the optimal trade-off between communication rates, distortions of the information produced at the two nodes and costs for information acquisition at Node 2 is derived. The issue of robustness to possible malfunctioning of the data acquisition process at Node 2 is also investigated. The results are illustrated via an example that demonstrates the different roles played by the forward communication, namely data exchange, query and control.

The material in this chapter has been reported in the documents:

- B. Ahmadi and O. Simeone, “Two-way Communication with Adaptive Data Acquisition,” to appear in *Transactions on Emerging Telecommunications Technologies*, arXiv:1209.5978.
- B. Ahmadi and O. Simeone, “Two-Way Communication with Adaptive Data Acquisition,” in *Proc. IEEE International Symposium on Information Theory* (ISIT 2013), Istanbul, Turkey, July 7-12, 2013.

Chapter 7: In this Chapter, two extensions of the original action-dependent channel are studied. In the first, the decoder is interested in estimating not only the message, but also the state sequence within an average per-letter distortion. Under the constraint of common reconstruction (i.e., the decoders estimate of the state must be recoverable also at the encoder) and assuming non-causal state knowledge at the encoder in the second phase, a single-letter characterization of the achievable rate-distortion-cost trade-off is obtained. In the second extension, an action-dependent degraded broadcast channel is studied. Under the assumption that the encoder knows the state sequence causally in the second phase, the capacity-cost region is identified. Various examples, including Gaussian channels and a model with a ”probing” encoder, are also provided to show the advantage of a proper joint design of the two communication phases.

The material in this chapter has been reported in the documents:

- B. Ahmadi and O. Simeone, “On Channels with Action-Dependent States,” in *Proc. IEEE Information Theory Workshop* (ITW 2012), Lausanne, Switzerland, Sept. 3-7, 2012.
- B. Ahmadi and O. Simeone, “On channels with action-dependent states,” arXiv:1202.4438.

Chapter 8: In this Chapter, the problem of embedding information on the actions is studied for both the source and the channel coding set-ups. In both cases, a decoder is present that observes only a function of the actions taken by an encoder or a decoder of an action-dependent point-to-point link. For the source coding model, this decoder wishes to reconstruct a lossy version of the source being transmitted over the point-to-point link, while for the channel coding problem the decoder wishes to retrieve a portion of the message conveyed over the link. For the problem of source coding with actions taken at the decoder, a single letter characterization of the set of all achievable tuples of rate, distortions at the two decoders and action cost is derived, under the assumption that the mentioned decoder observes a function of the actions non-causally, strictly causally or causally. A special case of the problem in which the actions are taken by the encoder is also solved. A single-letter characterization of the achievable capacity-cost region is then obtained for the channel coding set-up with actions. Examples are provided that shed light into the effect of information embedding on the actions for the action-dependent source and channel coding problems.

The material in this chapter has been reported in the documents:

- B. Ahmadi, H. Asnani, O. Simeone and H. Permuter, “Information Embedding on Actions,” in *Proc. IEEE International Symposium on Information Theory* (ISIT 2013), Istanbul, Turkey, July 7-12, 2013.

- B. Ahmadi, H. Asnani, O. Simeone and H. Permuter, “Information Embedding on Actions,” submitted to *IEEE Trans. Inform. Theory*, arXiv:1207.6084.

CHAPTER 2

PRELIMINARIES

In this chapter, first the notation used is discussed and then, some preliminary information-theoretic results on related problems in the field of source coding with side information are studied.

2.1 Notation

For a and b integer with $a \leq b$, $[a, b]$ is defined as the interval $[a, a+1, ..., b]$ and x_a^b is used to denote the sequence $(x_a, \ldots, x_b)$. Also x^b is written for x_1^b for simplicity. Upper case, lower case and calligraphic letters denote random variables, specific values of random variables and their alphabets, respectively. Given discrete random variables, or more generally vectors, X and Y, the notation $p_X(x)$ or $p(x)$ is used for $\Pr[X = x]$, and $p_{X|Y}(x|y)$ or $p(x|y)$ is used for $\Pr[X = x|Y = y]$, where the latter notations are used when the meaning is clear from the context. Given a set $\mathcal{X}$, an n-fold Cartesian product of $\mathcal{X}$ is denoted by $\mathcal{X}^n$. For random variables X and Y, $\sigma^2_{X|Y}$ is the (average) conditional variance of X given Y, i.e., $\mathrm{E}\left[\mathrm{E}[(X - \mathrm{E}[X|Y])^2|Y]\right]$. Function $\delta(x)$ represents the Kronecker delta function, i.e., $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ otherwise. The notation convention in [1] is adopted, in which $\delta(\epsilon)$ represents any function such that $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. Moreover, X—Y—Z form a Markov chain if $p(x, y, z) = p(x)p(y|x)p(z|y)$, that is, X and Z are conditionally independent of each other given Y. The binary entropy function is defined as $H(p) = -p\mathrm{log}_2 p - (1 - p)\mathrm{log}_2(1 - p)$. Finally, $\alpha * \beta = \alpha(1 - \beta) + \beta(1 - \alpha)$.

2.2 Background

From an information theoretic perspective, the baseline setting for the class of source coding problems with side information is one in which a memoryless source $X^n =$

$(X_1, ..., X_n)$, where $X_i \in \mathcal{X}$ has pmf $p(x)$ for $i = 1, ..., n$, is to be communicated by an encoder via a message M of rate R bits per source symbol to a decoder. The decoder has available a correlated sequence Y^n, with $Y_i \in \mathcal{Y}$, that is related to X^n via a memoryless channel $p(y|x)$ (see Figure 2.1). The optimal trade-off between rate R and the average distortion D between the source X^n and reconstruction $\hat{X}^n$ was obtained by Wyner and Ziv in [2] for any given distortion metric $d(x, \hat{x}) : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow \mathcal{R}_+ \cup \{\infty\}$. It was shown that the rate-distortion function is given by

$$R(D) \quad = \min \quad I(X; U|Y), \tag{2.1}$$

where the minimum is taken over all pmfs $p(u|x)$, with $u \in \mathcal{U}$, and deterministic function $\hat{\mathrm{x}}(u, y)$ such that $\mathrm{E}[d(X, \hat{\mathrm{x}}(U, Y))] \leq D$.

The optimal performance can be achieved as follows. The sequence X^n is quantized using a randomly generated codebook of codewords U^n using the standard joint typicality criterion. In order for quantization to be successful $2^{nI(X;U)}$ codewords U^n are sufficient. Thanks to the side information Y^n available at Node 2, the resulting rate $I(X;U)$ (bits per source symbol) can be further decreased to $I(X;U|Y)$ using the technique of binning. The idea is that all the $2^{nI(X;U)}$ codewords U^n are divided into $2^{nI(X;U|Y)}$ bins. An example of a bin is shown in gray in Figure 2.2. The codewords U^n in the same bin are mapped to an identical message M. The bins contain $2^{nI(X;U)}/2^{nI(X;U|Y)} = 2^{nI(U;Y)}$ codewords each, and thus, by the channel coding theorem, the decoder can distinguish among the codewords U^n in the bin based on Y^n. This scheme is known as "Wyner-Ziv" coding in information theory.

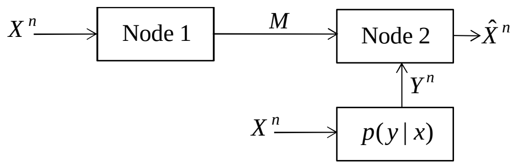

Figure 2.1 Source coding with side information.

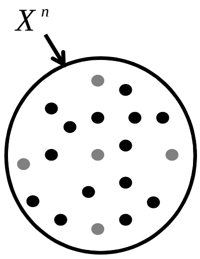

Figure 2.2 Illustration of binning. The set of all dots represent the codebook of codewords U^n and the subset in gray is a bin. The picture assumes for simplicity $\mathcal{X} = \mathcal{U}$.

In sensor networks and cloud computing, reliability of all the computing devices (e.g., sensors or servers) cannot be guaranteed all the time. Therefore, it is appropriate to design the system so as to be robust to system failures. As shown in Figure 2.3, this aspect can be modeled by assuming that the decoder, unbeknownst to the encoder, may not be able to acquire information sequence Y^n. This setting is equivalent to assuming the presence of two decoders, one with the capability to acquire information about Y^n (Node 2) and one without this capability (Node 3). This model is referred to as the Heegard-Berger problem, where Node 2 and Node 3 of Figure 2.3 are interested in estimating $\hat{X}_2^n$ and $\hat{X}_3^n$, respectively. It is emphasized that $\hat{X}_2^n$ and $\hat{X}_3^n$ are two different description of the source sequence X^n to be reconstructed at Node 2 and Node 3 with distortion levels D_2 and D_3, respectively.

It was shown in [3] that the rate-distortion function is given by

$$R(D_2, D_3) \quad = \min \ I(X; \hat{X}_3) + I(X; U|\hat{X}_3, Y), \tag{2.2}$$

where the minimum is taken over all pmfs $p(u, \hat{x}_3|x)$ and deterministic function $\hat{\mathrm{x}}_2(u, y)$ such that $\mathrm{E}[d(X,\ \hat{\mathrm{X}}_3))] \leq D_3$ and $\mathrm{E}[d(X,\ \hat{\mathrm{x}}_2(U, Y))] \leq D_2$. The optimal strategy achieving (2.2) is based on successive refinement and binning. The strategy is identical to that used below in the context of the model of Figure 2.4 (see Figure 2.5 and discussion below for further details).

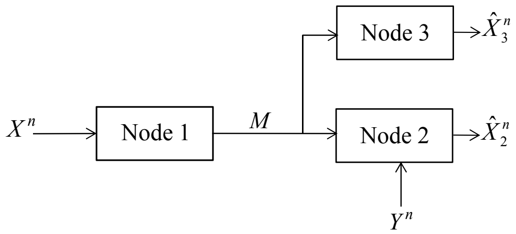

Figure 2.3 Source coding when side information may be absent.

The concept of a side information "vending machine" was introduced in [4] in order to account for source coding scenarios in which acquiring the side information at the receiver entails some cost and thus should be done efficiently. In this class of models, the quality of the side information Y^n can be controlled at the decoder by selecting an action sequence A^n, with $A_i \in \mathcal{A}$, that affects the effective channel between the source X^n and the side information Y^n through a conditional memoryless distribution $p_{Y|X,A}(y|x,a)$. Specifically, given A^n and X^n, the sequence Y^n is distributed as $p(y^n|a^n,x^n) = \prod_{i=1}^n p_{Y|A,X}(y_i|a_i,x_i)$. The cost of the action sequence is defined by a cost function Λ: $\mathcal{A} \rightarrow [0,\Lambda_{\max}]$ with $0 \leq \Lambda_{\max} < \infty$, as $\Lambda(a^n) = \sum_{i=1}^n \Lambda(a_i)$. The estimated sequence $\hat{X}^n$ with $\hat{X}^n \in \hat{\mathcal{X}}^n$ is then obtained as a function of M and Y^n.

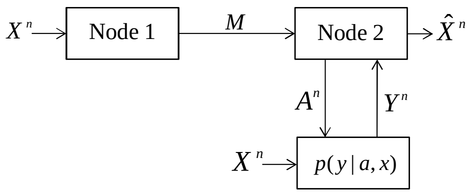

Figure 2.4 Source coding with side information "vending machine".

The optimal trade-off between rate R, the average distortion D and the average action cost Γ was obtained by [4] for any given distortion metric $d(x,\hat{x})$ and action cost function $\Lambda(a)$ is given by

$$R(D,\Gamma) = \min \quad I(X;A) + I(X;U|Y,A), \tag{2.3}$$

where the minimum is taken over all pmfs $p(a, u|x)$ and deterministic function $\hat{\mathrm{x}}(u, y)$ such that $\mathrm{E}[d(X, \hat{\mathrm{x}}(U, Y))] \leq D$ and $\mathrm{E}[\Lambda(A)] \leq \Gamma$.

For later reference, it is useful to describe the optimal strategy as derived in [3]. The basic idea is that of using a successive refinement (or layered) compression strategy, in which the coarse layer is used to inform the decoder about the actions A^n that are better adapted to the source X^n, and the fine layer provides further information that enables the decoder to produce the estimate $\hat{X}^n$. Specifically, as illustrated in Figure 2.5, the sequence X^n is quantized using a codebook of action sequences A^n. As discussed, in order for quantization to be successful, this quantization step requires a rate of $I(X; A)$. Next, a refined description U^n of X^n is also obtained and sent to Node 2. From (3.1), this requires a rate of $I(X; U|A, Y)$ thanks to binning (a bin is shown in gray in Figure 2.5).

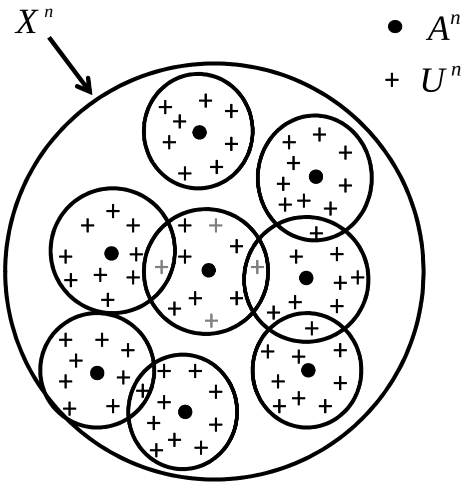

Figure 2.5 Illustration of the optimal strategy for the set-up of Figure 2.4. The picture assumes for simplicity that $\mathcal{X} = \mathcal{A} = \mathcal{U}$.

CHAPTER 3

HEEGARD-BERGER AND CASCADE SOURCE CODING WITH COMMON RECONSTRUCTION CONSTRAINT

3.1 Introduction

Source coding problems with side information at the decoder(s) model a large number of scenarios of practical interest, including video streaming [5] and wireless sensor networks [6]. From an information theoretic perspective, the baseline setting for this class of problems is one in which a memoryless source $X^n = (X_1, ..., X_n)$ is to be communicated by an encoder at a rate R bits per source symbol to a decoder that has available a correlated sequence Y^n that is related to X^n via a memoryless channel $p(y|x)$ (see Figure 3.1[1]). Under the requirement of asymptotically lossless reconstruction $\hat{X}^n$ of the source X^n at the decoder, the minimum required rate was obtained by Slepian and Wolf in [7]. Later, the more general optimal trade-off between rate R and the distortion D between the source X^n and reconstruction $\hat{X}^n$ was obtained by Wyner and Ziv in [2] for any given distortion metric $d(x, \hat{x})$. As shown in Chapter 2, the rate-distortion function is given by

$$R^{WZ}_{X|Y}(D) \quad = \min \quad I(X; U|Y), \tag{3.1}$$

where the minimum is taken over all probability mass functions (pmfs) $p(u|x)$ and deterministic function $\hat{\mathrm{x}}(u, y)$ such that $\mathrm{E}[d(X, \hat{\mathrm{x}}(U, Y))] \leq D$.

3.1.1 Heegard-Berger and Cascade Source Coding Problems

In applications such as the ones discussed above, the point-to-point setting of Figure 3.1 does not fully capture the main features of the source coding problem. For

[1]The presence of the function ψ at the encoder will be explained later.

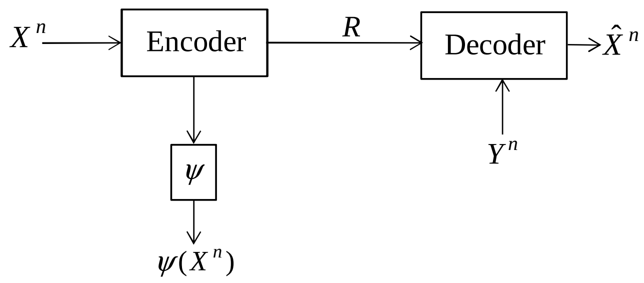

Figure 3.1 Point-to-point source coding with common reconstruction [8].

instance, in video streaming, a transmitter typically *broadcasts* information to a number of decoders. As another example, in sensor networks, data is typically routed over *multiple hops* towards the destination. A model that accounts for the aspect of *broadcasting* to multiple decoders is the *Heegard-Berger (HB)* set-up shown in Figure 3.2. In this model, the link of rate R bits per source symbol is used to communicate to two receivers having different side information sequences, Y_1^n and Y_2^n, which are related to source X^n via a memoryless channel $p(y_1, y_2|x)$. The set of all achievable triples (R, D_1, D_2) for this model, where D_1 and D_2 are the distortion levels at Decoders 1 and 2, respectively was derived in [3] and [9] under the assumption that the side information sequences are (stochastically) degraded versions of the source X^n. In a variation of this model shown in Figure 3.3, decoder cooperation is enabled by a limited capacity link from one decoder (Decoder 1) to the other (Decoder 2). Inner and outer bounds to the rate distortion region for this problem are obtained in [10] under the assumption that the side information of Decoder 2 is (physically) degraded with respect to that of Decoder 1.

As for *multihopping*, a basic model that captures some of the key design issues is shown in Figure 3.4. In this *cascade* set-up, an encoder (Node 1) communicates with rate R_1 to a intermediate node (Node 2), which has side information Y_1^n, and in turns communicates with rate R_2 to a final node (Node 3) with side information Y_2^n. Both Node 2 and Node 3 act as decoders, similar to the HB problem of Figure 3.2, in the sense that they reconstruct a local estimate of the source X^n. The rate-distortion

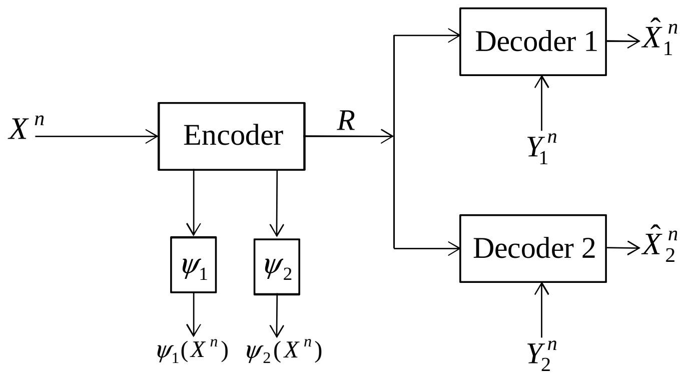

Figure 3.2 Heegard-Berger source coding problem with common reconstruction.

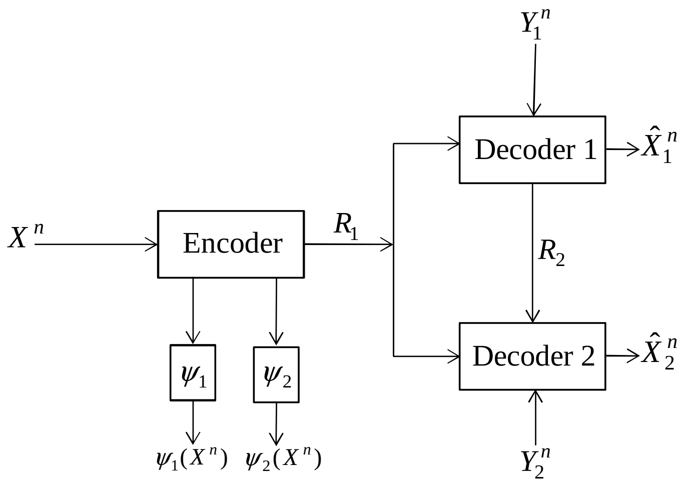

Figure 3.3 Heegard-Berger source coding problem with common reconstruction and decoder cooperation.

function for this problem has been derived for various special cases in [11, 12, 13] and [14] (see Table I in [14] for an overview). Reference [13] derives the set of all achievable quadruples (R_1, R_2, D_1, D_2), i.e., the rate-distortion region, for the case in which Y_1^n is also available at the encoder and Y_2^n is a physically degraded version of X^n with respect to Y_1^n. Instead, [12] derives the rate-distortion region under the assumptions that the source and the side information sequences are jointly Gaussian, that the distortion metric is quadratic, and that the sequence Y_1^n is a physically degraded version of X^n with respect to Y_2^n. The corresponding result for binary source and side information and Hamming distortion metric was derived in [14].

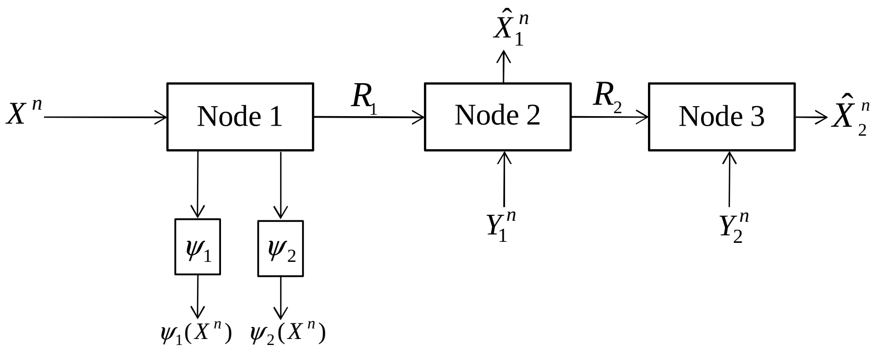

Figure 3.4 Cascade source coding problem with common reconstruction.

3.1.2 Common Reconstruction Constraint

A key aspect of the optimal strategies identified in [2, 3, 9, 12] and [13] is that the side information sequences are, in general, used in two different ways: (i) as a means to reduce the rate required for communication between encoder and decoders via binning; and (ii) as an additional observation that the decoder can leverage, along with the bits received from the encoder, in order to improve its local estimate. For instance, for the point-to-point system of Figure 3.1, the Wyner-Ziv result (3.1) reflects point (i) of the discussion above in the conditioning on side information Y, which reduces the rate, and point (ii) in the fact that the reconstruction $\hat{X}$ is a function $\hat{\mathrm{x}}(U, Y)$ of the signal U received from the encoder and the side information Y.

Leveraging the side information as per point (ii), while advantageous in terms of rate-distortion trade-off, may have unacceptable consequences for some applications. In fact, this use of side information entails that the reconstruction $\hat{X}$ of the decoder cannot be reproduced at the encoder. In other words, encoder and decoder cannot agree on the specific reconstruction $\hat{X}$ obtained at the receiver side, but only on the average distortion level D. In applications such as transmission of sensitive medical, military or financial data, this may not be desirable. Instead, one may want to add the constraint that the reconstruction at the decoder be reproducible by the encoder [8]. This idea, referred to as the Common Reconstruction (CR) constraint, was first proposed in [8], where it is shown for the point-to-point setting of Figure 3.1[2] that the rate-distortion function under the CR constraint is given by

$$R_{X|Y}^{CR}(D) = \min \quad I(X;\hat{X}|Y), \tag{3.2}$$

where the minimum is taken over all pmfs $p(\hat{x}|x)$ such that $\mathrm{E}[d(X,\hat{X})] \leq D$. Comparing (3.2) with the Wyner-Ziv rate-distortion (3.1), it can be seen that the additional CR constraint prevents the decoder from using the side information as a means to improve its estimate $\hat{X}$ (see point (*ii*) above).

The original work of [8] has been recently extended in [15], where a relaxed CR constraint is imposed in which only a distortion constraint is imposed between the decoder's reconstruction and its reproduction at the encoder. This setting is referred to as imposing a *Constrained Reconstruction* (ConR) requirement.

3.1.3 Main Contributions

In this Chapter, the HB source coding problem (Figure 3.2) and the cascade source coding problem (Figure 3.4) are studied under the CR requirement. The considered models are thus relevant for the transmission of sensitive information, which is constrained by CR, via broadcast or multi-hop links – a common occurrence in, e.g., medical, military or financial applications (e.g., for intranets of hospitals or financial institutions). Specifically, our main contributions are:

- For the HB problem with the CR constraint (Figure 3.2), the rate-distortion function is derived under the assumption that the side information sequences are (stochastically) degraded. Also this function is calculated explicitly for three examples, namely Gaussian source and side information with quadratic distortion metric, and binary source and erasure side information with erasure and Hamming distortion metrics (Section 3.2);

[2]The function ψ at the encoder calculates the estimate of the encoder regarding the decoder's reconstruction.

- For the HB problem with the CR constraint and decoder cooperation (Figure 3.3), the rate-distortion region is derived under the assumption that the side information sequences are (physically) degraded in either direction (Section 3.3.1 and Section 3.3.2). It is emphasized that the corresponding problem without the CR constraint is still open as per the discussion above;

- For the cascade problem with the CR constraint (Figure 3.4), the rate-distortion region is obtained under the assumption that side information Y_2 is physically degraded with respect to Y_1 (Section 3.4.2). It is emphasized that the corresponding problem without the CR constraint is still open as per the discussion above;

- For the cascade problem with CR constraint (Figure 3.4), outer and inner bounds on the rate-distortion region are obtained under the assumption that the side information Y_1 is physically degraded with respect to Y_2. Moreover, for the three examples mentioned above in the context of the HB problem, it is shown that the bounds coincide and the corresponding rate-distortion region is explicitly evaluated (Section 3.4.3);

- For the HB problem, the rate-distortion function is finally derived under the more general requirement of ConR (Section 3.5).

3.2 Heegard-Berger Problem with Common Reconstruction

In this section, first the system model for the HB source coding problem in Figure 3.2 with CR is detailed in Section 3.2.1. Next, the characterization of the corresponding rate-distortion performance is derived under the assumption that one of the two side information sequences is a stochastically degraded version of the other in the sense of [3] (see (3.10)). Finally, three specific examples are worked out, namely Gaussian sources under quadratic distortion (Section 3.2.3), and binary sources with side information sequences subject to erasures under Hamming or erasure distortion (Section 3.2.4).

3.2.1 System Model

In this section, the system model for the HB problem with CR is detailed. The system is defined by the pmf $p_{XY_1Y_2}(x, y_1, y_2)$ and discrete alphabets $\mathcal{X}, \mathcal{Y}_1, \mathcal{Y}_2, \hat{\mathcal{X}}_1$, and $\hat{\mathcal{X}}_2$ as follows. The source sequence X^n and side information sequences Y_1^n and Y_2^n, with $X^n \in \mathcal{X}^n$, $Y_1^n \in \mathcal{Y}_1^n$, and $Y_2^n \in \mathcal{Y}_2^n$ are such that the tuples (X_i, Y_{1i}, Y_{2i}) for $i \in [1, n]$ are independent and identically distributed (i.i.d.) with joint pmf $p_{XY_1Y_2}(x, y_1, y_2)$. The encoder measures a sequence X^n and encodes it into a message J of nR bits, which is delivered to the decoders. Decoders 1 and 2 wish to reconstruct the source sequence X^n within given distortion requirements, to be discussed below, as $\hat{X}_1^n \in \hat{\mathcal{X}}_1^n$ and $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$, respectively. The estimated sequence $\hat{X}_j^n$ is obtained as a function of the message J and the side information sequence Y_j^n for $j = 1, 2$. The estimates are constrained to satisfy distortion constraints defined by per-symbol distortion metrics $d_j(x, \hat{x}_j) : \mathcal{X} \times \hat{\mathcal{X}}_j \rightarrow [0, D_{max}]$ with $0 < D_{max} < \infty$. Based on the given distortion metrics, the overall distortion for the estimated sequences $\hat{x}_1^n$ and $\hat{x}_2^n$ is defined as

$$d_j^n(x^n, \hat{x}_j^n) = \frac{1}{n} \sum_{i=1}^{n} d_j(x_i, \hat{x}_{ji}) \text{ for } j = 1, 2. \tag{3.3}$$

The reconstructions $\hat{X}_2^n$ and $\hat{X}_2^n$ are also required to satisfy the CR constraints, as formalized below.

Definition 3.1. An $(n, R, D_1, D_2, \epsilon)$ code for the HB problem with CR consists of an encoding function

$$g\colon \mathcal{X}^n \rightarrow [1, 2^{nR}], \tag{3.4}$$

which maps the source sequence X^n into a message J; a decoding function for Decoder 1,

$$h_1\colon\ [1, 2^{nR}] \times \mathcal{Y}_1^n \rightarrow \hat{\mathcal{X}}_1^n, \tag{3.5}$$

which maps the message J and the side information Y_1^n into the estimated sequence $\hat{X}_1^n$; a decoding function for Decoder 2

$$h_2\colon\ [1, 2^{nR}] \times \mathcal{Y}_2^n \rightarrow \hat{\mathcal{X}}_2^n \tag{3.6}$$

which maps message J and the side information Y_2^n into the estimated sequence $\hat{X}_2^n$; and two reconstruction functions

$$\psi_1\colon\ \mathcal{X}^n \rightarrow \hat{\mathcal{X}}_1^n \tag{3.7a}$$

$$\text{and } \psi_2\colon\ \mathcal{X}^n \rightarrow \hat{\mathcal{X}}_2^n, \tag{3.7b}$$

which map the source sequence into the estimated sequences at the encoder, namely $\psi_1(X^n)$ and $\psi_2(X^n)$, respectively; such that the distortion constraints are satisfied, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[d_j(X_i, \hat{X}_{ji})\right] \ \leq\ D_j \text{ for } j = 1, 2, \tag{3.8}$$

and the CR requirements hold, namely,

$$\Pr\left[\psi_j(X^n) \neq \hat{X}_j^n\right] \ \leq\ \epsilon,\ j = 1, 2. \tag{3.9}$$

Given distortion pairs (D_1, D_2), a rate pair R is said to be achievable if, for any $\epsilon > 0$ and sufficiently large n, there exists an $(n, R, D_1 + \epsilon, D_2 + \epsilon, \epsilon)$ code. The rate-distortion function $R(D_1, D_2)$ is defined as $R(D_1, D_2)$ =inf$\{R :$ the triple (R, D_1, D_2) is achievable$\}$.

3.2.2 Rate-Distortion Function

In this section, a single-letter characterization of the rate-distortion function for the HB problem with CR is derived, under the assumption that the joint pmf $p(x, y_1, y_2)$ is such that there exists a conditional pmf $\tilde{p}(y_1|y_2)$ for which

$$p(x, y_1) = \sum_{y_2 \in \mathcal{Y}_2} p(x, y_2)\tilde{p}(y_1|y_2). \tag{3.10}$$

In other words, the side information Y_1 is a stochastically degraded version of Y_2.

Proposition 3.1. *If the side information Y_1 is stochastically degraded with respect to Y_2, the rate-distortion function for the HB problem with CR is given by*

$$R_{HB}^{CR}(D_1, D_2) \quad = \quad \min I(X; \hat{X}_1|Y_1) + I(X; \hat{X}_2|Y_2\hat{X}_1) \tag{3.11}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x, y_1, y_2, \hat{x}_1, \hat{x}_2) = p(x, y_1, y_2)p(\hat{x}_1, \hat{x}_2|x), \tag{3.12}$$

and minimization is performed with respect to the conditional pmf $p(\hat{x}_1, \hat{x}_2|x)$ under the constraints

$$\mathrm{E}[d_j(X, \hat{X}_j)] \leq D_j, \text{ for } j = 1, 2. \tag{3.13}$$

The proof of the converse can be found in Appendix A. Achievability follows as a special case of Theorem 3 of [3] and can be easily shown using standard arguments. In particular, the encoder randomly generates a standard lossy source code $\hat{X}_1^n$ for the source X^n with rate $I(X; \hat{X}_1)$ bits per source symbol. Random binning is used to reduce the rate to $I(X; \hat{X}_1|Y_1)$. By the Wyner-Ziv theorem [1, p. 280], this guarantees that both Decoder 1 and Decoder 2 are able to recover $\hat{X}_1^n$ (since Y_1 is a degraded version of Y_2). The encoder then maps the source X^n into the reconstruction sequence $\hat{X}_2^n$ using a codebook that is generated conditional on $\hat{X}_1^n$ with rate $I(X; \hat{X}_2|\hat{X}_1)$ bits per source symbol. Random binning is again used to reduce the rate to $I(X;\hat{X}_2|Y_2\hat{X}_1)$. From the Wyner-Ziv theorem, and the fact that Decoder 2 knows the sequence $\hat{X}_1^n$, it follows that Decoder 2 can recover the reconstruction $\hat{X}_2^n$ as well. Note that, since the reconstruction sequences $\hat{X}_1^n$ and $\hat{X}_2^n$ are generated by the encoder, functions ψ_1 and ψ_2 that guarantees the CR constraints (3.9) exist by construction.

Remark 3.1. Under the physical degradedness assumption that the Markov chain condition X—Y_2—Y_1 holds, equation (3.11) can be rewritten as

$$R = \min I(X;\hat{X}_1\hat{X}_2|Y_2) + I(\hat{X}_1;Y_2|Y_1), \tag{3.14}$$

with the minimization defined as in (3.11). This expression quantifies by $I(\hat{X}_1;Y_2|Y_1)$ the additional rate that is required with respect to the ideal case in which both decoders have the better side information Y_2.

Remark 3.2. Without the CR constraint, the rate-distortion function under the assumption of Proposition 3.1 is given by [3] (see also eq. (2.2))

$$R_{HB}(D_1,D_2) \quad = \quad \min I(X;U_1|Y_1) + I(X;U_2|Y_2U_1), \tag{3.15}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x,y_1,y_2,u_1,u_2,\hat{x}_1,\hat{x}_2) = p(x,y_1,y_2)p(u_1,u_2|x)\delta(\hat{x}_1 - \hat{\mathrm{x}}_1(u_1,y_1))\delta(\hat{x}_2 - \hat{\mathrm{x}}_2(u_2,y_2)), \tag{3.16}$$

and minimization is performed with respect to the conditional pmf $p(u_1,u_2|x)$ and the deterministic functions $\hat{\mathrm{x}}_\mathrm{j}(u_j,y_j)$, for $j = 1,2$, such that distortion constraints (3.13) are satisfied. Comparison of (3.11) with (3.15) reveals that, similar to the discussion around (3.1) and (3.2), the CR constraint permits the use of side information only to reduce the rate via binning, but not to improve the decoder's estimates via the use of the auxiliary codebooks represented by variables U_1 and U_2, and functions $\hat{\mathrm{x}}_\mathrm{j}(u_j,y_j)$, for $j = 1,2$, in (3.16).

Remark 3.3. Consider the case in which the side information sequences are available in a causal fashion in the sense of [16], that is, the decoding functions (3.5)-(3.6) are modified as h_{ji}: $[1, 2^{nR}] \times \mathcal{Y}_j^i \to \hat{\mathcal{X}}_{ji}$, for $i \in [1, n]$ and $j = 1, 2$, respectively. Following similar steps as in the proof of Proposition 2 and in [16], it can be concluded that, under the CR constraint, the rate-distortion function in this case is the same as if the two side information sequences were not available at the decoders, and is thus given by (3.11) upon removing the conditioning on the side information. Note that this is true irrespective of the joint pmf $p(x, y_1, y_2)$ and hence it holds also for non-degraded side information. This result can be explained by noting that, as explained in [16], causal side information prevents the possibility of reducing the rate via binning. Since the CR constraint also prevents the side information from being used to improve the decoders' estimates, it follows that the side information is useless in terms of rate-distortion performance, if used causally under the CR constraint.

On a similar note, if only side information Y_1 is causally available, while Y_2 can still be used in the conventional non-causal fashion, then it can be proved that Y_1 can be neglected without loss of optimality. Therefore, the rate-distortion function follows from (3.11) by removing the conditioning on Y_1.

Remark 3.4. In [17], a related model is studied in which the source is given as $X = (Y_1, Y_2)$ and each decoder is interested in reconstructing a lossy version of the side information available at the other decoder. The CR constraint is imposed in a different way by requiring that each decoder be able to reproduce the estimate reconstructed at the other decoder.

3.2.3 Gaussian Sources and Quadratic Distortion

In this section, the result of Proposition 3.1 is highlighted by considering a zero-mean Gaussian source $X \sim \mathcal{N}(0, \sigma_x^2)$, with side information variables

$$\begin{aligned} Y_1 &= X + Z_1 && \text{(3.17a)} \\ \text{and } Y_2 &= X + Z_2, && \text{(3.17b)} \end{aligned}$$

where $Z_1 \sim \mathcal{N}(0, N_1 + N_2)$ and $Z_2 \sim \mathcal{N}(0, N_2)$ are independent of each other and of Y_2 and X. Note that the joint distribution of (X, Y_1, Y_2) satisfies the stochastic degradedness condition. Quadratic distortion $d_j(x, \hat{x}_j) = (x - \hat{x}_j)^2$ for $j = 1, 2$ is considered. By leveraging standard arguments that allow us to apply Proposition 3.1 to Gaussian sources under mean-square-error constraint (see [1, pp. 50-51] and [18]), a characterization of the rate-distortion function is obtained for the given distortion and metrics.

It is first recalled that for the point-to-point set-up in Figure 3.1 with $X \sim \mathcal{N}(0, \sigma_x^2)$ and side information $Y = X + Z$, with $Z \sim \mathcal{N}(0, N)$ independent of X, the rate-distortion function with CR under quadratic distortion is given by [8]

$$R_{X|Y}^{CR}(D) = \begin{cases} R_{\mathcal{G}}^{CR}(D, N) \triangleq \frac{1}{2}\log_2\left(\frac{\sigma_x^2}{\sigma_x^2 + N} \cdot \frac{D+N}{D}\right) & \text{for } D \leq \sigma_x^2 \\ 0 & \text{for } D > \sigma_x^2, \end{cases} \tag{3.18}$$

where an explicit dependence on N of function $R_{\mathcal{G}}^{CR}(D, N)$ is made for convenience. The rate-distortion function (3.18) for $D \leq \sigma_x^2$ is obtained from (3.2) by choosing the distribution $p(\hat{x}|x)$ such that $X = \hat{X} + Q$ where $Q \sim \mathcal{N}(0, D)$ is independent of $\hat{X}$.

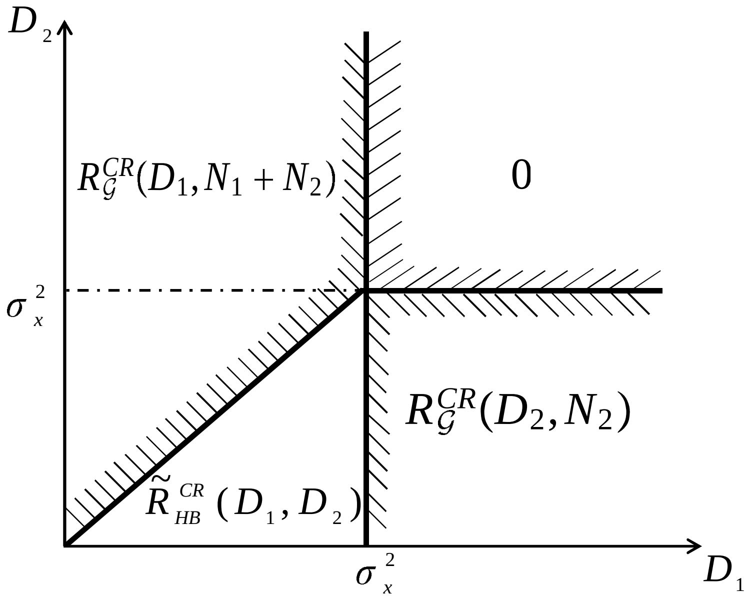

Figure 3.5 Illustration of the distortion regions in the rate-distortion function (3.19) for Gaussian sources and quadratic distortion.

Proposition 3.2. *The rate-distortion function for the HB problem with CR for Gaussian sources (3.17) and quadratic distortion is given by*

$$R_{HB}^{CR}(D_1, D_2) \;=\; \begin{cases} 0 & \text{if } D_1 \geq \sigma_x^2 \text{ and } D_2 \geq \sigma_x^2, \\ R_{\mathcal{G}}^{CR}(D_1, N_1+N_2) & \text{if } D_1 \leq \sigma_x^2 \text{ and } D_2 \geq \min(D_1, \sigma_x^2) \\ R_{\mathcal{G}}^{CR}(D_2, N_2) & \text{if } D_1 \geq \sigma_x^2 \text{ and } D_2 \leq \sigma_x^2 \\ \tilde{R}_{HB}^{CR}(D_1, D_2) & \text{if } D_2 \leq D_1 \leq \sigma_x^2 \end{cases} \tag{3.19}$$

where $R_{\mathcal{G}}^{CR}(D, N)$ *is defined in (3.18) and*

$$\tilde{R}_{HB}^{CR}(D_1, D_2) \triangleq \frac{1}{2}\log_2\left(\frac{\sigma_x^2}{(\sigma_x^2+N_1+N_2)} \cdot \frac{(D_1+N_1+N_2)(D_2+N_2)}{(D_1+N_2)D_2}\right). \tag{3.20}$$

Remark 3.5. The rate-distortion function for the HB problem for Gaussian sources (3.17) without the CR constraint can be found in [3]. Comparison with (3.19) confirms the performance loss discussed in Remark 3.2.

Definition of the rate distortion function (3.19) requires different consideration for the four subregions of the (D_1, D_2) plane sketched in Figure 3.5. In fact, for

$D_1 \geq \sigma_x^2$ and $D_2 \geq \sigma_x^2$, the required rate is zero, since the distortion constraints are trivially met by setting $\hat{X}_1 = \hat{X}_2 = 0$ in the achievable rate (3.11). For the case $D_1 \geq \sigma_x^2$ and $D_2 \leq \sigma_x^2$, it is sufficient to cater only to Decoder 2 by setting $\hat{X}_1 = 0$ and $X = \hat{X}_2 + Q_2$, with $Q_2 \sim \mathcal{N}(0, D_2)$ independent of $\hat{X}_2$, in the achievable rate (3.11). That this rate cannot be improved upon follows from the trivial converse

$$R_{HB}^{CR}(D_1, D_2) \geq \max\{R_{\mathcal{G}}^{CR}(D_1, N_1 + N_2), R_{\mathcal{G}}^{CR}(D_2, N_2)\}, \tag{3.21}$$

which follows by cut-set arguments. The same converse suffices also for the regime $D_1 \leq \sigma_x^2$ and $D_2 \geq \min(D_1, \sigma_x^2)$. For this case, achievability follows by setting $X = \hat{X}_1 + Q_1$ and $\hat{X}_1 = \hat{X}_2$ in (3.11), where $Q_1 \sim \mathcal{N}(0, D_1)$ is independent of $\hat{X}_1$. In the remaining case, namely $D_2 \leq D_1 \leq \sigma_x^2$, the rate-distortion function does not follow from the point-to-point result (3.18) as for the regimes discussed thus far. The analysis of this case requires use of entropy-power inequality (EPI) and can be found in Appendix B.

Figure 3.6 depicts the rate $R_{HB}^{CR}(D_1, D_2)$ in (3.19) versus D_1 for different values of D_2 with $\sigma_x^2 = 4$, $N_1 = 2$, and $N_2 = 3$. As discussed above, for $D_2 = 5$, which is larger than σ_x^2, $R_{HB}^{CR}(D_1, D_2)$ becomes zero for values of D_1 larger than $\sigma_x^2 = 4$, while this is not the case for values $D_2 < \sigma_x^2 = 4$.

3.2.4 Binary Source with Erased Side Information and Hamming or Erasure Distortion

In this section, a binary source $X \sim \text{Ber}(\frac{1}{2})$ with erased side information sequences Y_1 and Y_2 are considered. The source Y_2 is an erased version of the source X with erasure probability p_2 and Y_1 is an erased version of X with erasure probability $p_1 > p_2$. This means that $Y_j = e$, where e represents an erasure, with probability p_j and $Y_j = X$ with probability $1 - p_j$. Note that, with these assumptions, the side information Y_1 is stochastically degraded with respect to Y_2. In fact, consider the factorization (3.10),

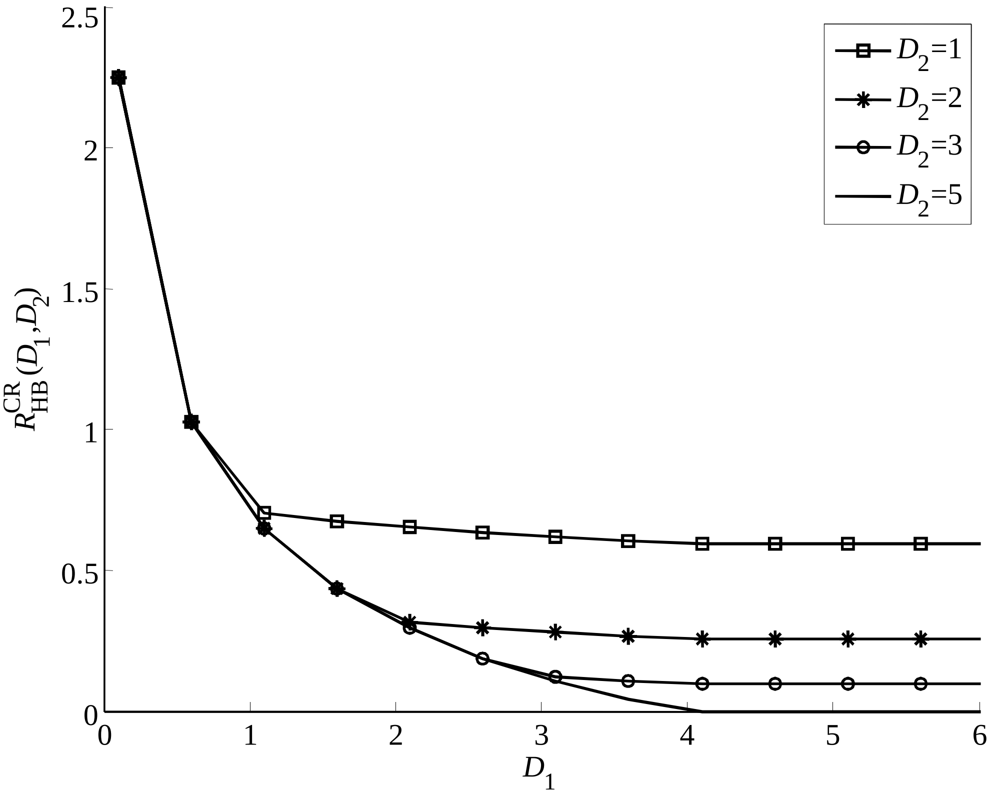

Figure 3.6 The rate-distortion function $R_{HB}^{CR}(D_1, D_2)$ in (3.19) versus distortion D_1 for different values of distortion D_2 and for $\sigma_x^2 = 4$, $N_1 = 2$, and $N_2 = 3$.

where additional distributions $p(y_2|x)$ and $\tilde{p}(y_1|y_2)$ are illustrated in Figure 3.7. As seen in Figure 3.7, the pmf $\tilde{p}(y_1|y_2)$ is characterized by the probability $\tilde{p}_1$ that satisfies the equality $p_1 = p_2+\tilde{p}_1(1-p_2)$. Hamming and erasure distortions are considered. For the Hamming distortion, the reconstruction alphabets are binary, $\hat{\mathcal{X}}_1 = \hat{\mathcal{X}}_2 = \{0, 1\}$, and $d_j(x, \hat{x}_j) = 0$ if $x = \hat{x}_j$ and $d_j(x, \hat{x}_j) = 1$ otherwise for $j = 1, 2$. Instead, for the erasure distortion the reconstruction alphabets are $\hat{\mathcal{X}}_1 = \hat{\mathcal{X}}_2 = \{0, 1, e\}$, and for $j = 1, 2$:

$$d_j(x, \hat{x}_j) = \begin{cases} 0 & \text{for } \hat{x}_j = x \\ 1 & \text{for } \hat{x}_j = e \\ \infty & \text{otherwise} \end{cases} \tag{3.22}$$

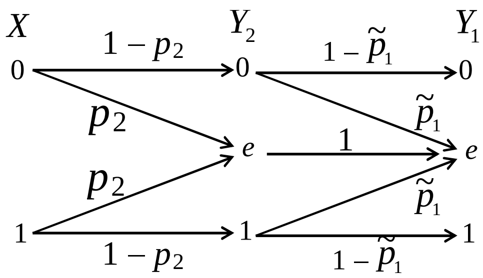

Figure 3.7 Illustration of the pmfs in the factorization (3.10) of the joint distribution $p(x, y_1, y_2)$ for a binary source X and erased side information sequences (Y_1, Y_2).

In Appendix C, it is proved that for the point-to-point set-up in Figure 3.1 with $X \sim \text{Ber}(\frac{1}{2})$ and erased side information Y, with erasure probability p, the rate-distortion function with CR under Hamming distortion is given by

$$R^{CR}_{X|Y}(D) = \begin{cases} R^{CR}_{\mathcal{B}}(D, p) \triangleq p(1 - H(D)) & \text{for } D \leq 1/2 \\ 0 & \text{for } D > 1/2, \end{cases} \tag{3.23}$$

where an explicit the dependence on p of function $R^{CR}_{\mathcal{B}}(D, p)$ is made for convenience. The rate-distortion function (3.23) for $D \leq 1/2$ is obtained from (3.2) by choosing the distribution $p(\hat{x}|x)$ such that $X = \hat{X} \oplus Q$ where $Q \sim \text{Ber}(D)$ is independent of $\hat{X}$. Following the same steps as in Appendix C, it can be also proved that for the point-to-point set-up in Figure 3.1 with $X \sim \text{Ber}(\frac{1}{2})$ and erased side information Y, with erasure probability p, the rate-distortion function with CR under erasure distortion is given by

$$R^{CR}_{X|Y}(D) \;=\; R^{CR}_{\mathcal{BE}}(D, p) \triangleq p(1 - D). \tag{3.24}$$

The rate-distortion function (3.24) is obtained from (3.2) by choosing the distribution $p(\hat{x}|x)$ such that $\hat{X} = X$ with probability $1 - D$ and $\hat{X} = e$ with probability D.

Remark 3.6. The rate-distortion function with erased side information and Hamming distortion without the CR constraint is derived in [19] (see also [20]). Comparison with (3.23) shows again the limitation imposed by the CR constraint on the use of side information (see Remark 3.2).

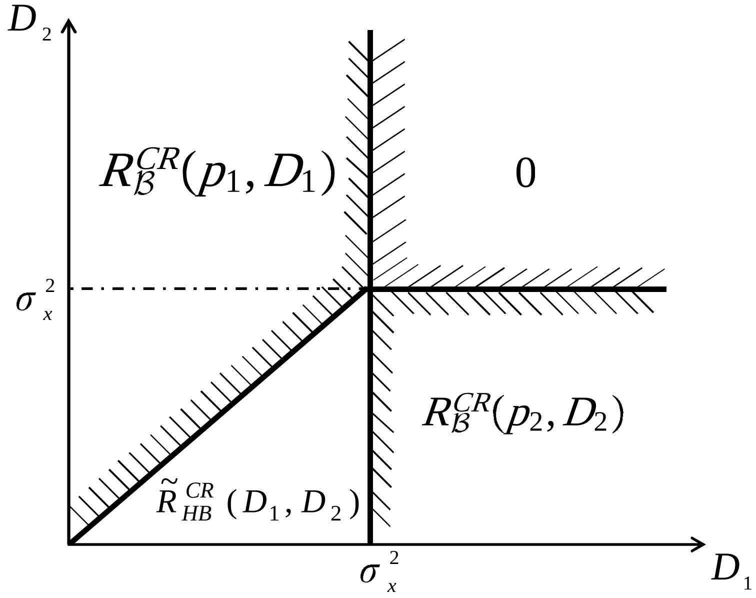

Figure 3.8 Illustration of the distortion regions in the rate-distortion function (3.25) for a binary source with degraded erased side information and Hamming distortion.

Proposition 3.3. *The rate-distortion function for the HB problem with CR for the binary source with the stochastically degraded erased side information sequences illustrated in Figure 3.7 under Hamming distortion is given by*

$$R_{HB}^{CR}(D_1, D_2) = \begin{cases} 0 & \text{if } D_1 \geq 1/2 \text{ and } D_2 \geq 1/2, \\ R_{\mathcal{B}}^{CR}(D_1, p_1) & \text{if } D_1 \leq 1/2 \text{ and } D_2 \geq \min(D_1, 1/2) \\ R_{\mathcal{B}}^{CR}(D_2, p_2) & \text{if } D_1 \geq 1/2 \text{ and } D_2 \leq 1/2 \\ \tilde{R}_{HB}^{CR}(D_1, D_2) & \text{if } D_2 \leq D_1 \leq 1/2 \end{cases} \tag{3.25}$$

where $R_{\mathcal{B}}^{CR}(D, N)$ *is defined in (3.23) and*

$$\tilde{R}_{HB}^{CR}(D_1, D_2) \triangleq p_1(1 - H(D_1)) + p_2(H(D_1) - H(D_2)). \tag{3.26}$$

Moreover, for the same source under erasure distortion the rate-distortion function is given by (3.25) by substituting $R_{\mathcal{B}}^{CR}(D_j, p_j)$ *with* $R_{\mathcal{BE}}^{CR}(D_j, p_j)$ *as defined in (3.24) for*

$j = 1, 2$ *and by substituting (3.26) with*

$$\tilde{R}^{CR}_{HB,E}(D_1, D_2) \triangleq p_1(1 - D_1) + p_2(D_1 - D_2). \tag{3.27}$$

Similar to the Gaussian example, the characterization of the rate distortion function (3.25) requires different considerations for the four subregions of the (D_1, D_2) plane sketched in Figure 3.8. In fact, for $D_1 \geq 1/2$ and $D_2 \geq 1/2$, the required rate is zero, since the distortion constraints are trivially met by setting $\hat{X}_1 = \hat{X}_2 = 0$ in the achievable rate (3.11). For the case $D_1 \geq 1/2$ and $D_2 \leq 1/2$, it is sufficient to cater only to Decoder 2 by setting $\hat{X}_1 = 0$ and $X = \hat{X}_2 \oplus Q_2$, with $Q_2 \sim \text{Ber}(D_2)$ independent of X, in the achievable rate (3.11). That this rate cannot be improved upon is a consequence from the trivial converse

$$R^{CR}_{HB}(D_1, D_2) \geq \max\{R^{CR}_{\mathcal{B}}(D_1, p_1), R^{CR}_{\mathcal{B}}(D_2, p_2)\}, \tag{3.28}$$

which follows by cut-set arguments. The same converse suffices also for the regime $D_1 \leq 1/2$ and $D_2 \geq \min(D_1, 1/2)$. For this case, achievability follows by setting $X = \hat{X}_1 \oplus Q_1$ and $\hat{X}_1 = \hat{X}_2$ in (3.11), where $Q_1 \sim \text{Ber}(D_1)$ is independent of $\hat{X}_1$. In the remaining case, namely $D_2 \leq D_1 \leq 1/2$, the rate-distortion function does not follow from the point-to-point result (3.23) as for the regimes discussed thus far. The analysis of this case can be found in Appendix D. Similar arguments apply also for the erasure distortion metric.

The rate-distortion function for the binary source $X \sim \text{Ber}(\frac{1}{2})$ is now compared with erased side information under Hamming distortion for three settings. In the first setting, known as the Kaspi model [9], the encoder knows the side information, and thus the position of the erasures. For this case, the rate-distortion function $R_{Kaspi}(D_1, D_2)$ for the example at hand was calculated in [19]. Note that in the Kaspi model, the CR constraint does not affect the rate-distortion performance since the encoder has all the information available at the decoders. The second model of interest is the standard HB setting with no CR constraint, whose rate-distortion function $R_{HB}(D_1, D_2)$ for the example at hand was derived [14]. The third model is the HB setup with CR studied here. The inequalities

$$R_{Kaspi}(D_1, D_2) \leq R_{HB}(D_1, D_2) \leq R_{HB}^{CR}(D_1, D_2), \tag{3.29}$$

hold, where the first inequality in (3.29) accounts for the impact of the availability of the side information at the encoder, while the second reflects the potential performance loss due to the CR constraint.

Figure 3.9 shows the aforementioned rate-distortion functions with $p_1 = 1$ and $p_2 = 0.35$, which corresponds to the case where Decoder 1 has no side information, for two values of the distortion D_2 versus the distortion D_1. For $D_2 \geq \frac{p_2}{2} = 0.175$, the given settings reduce to one in which the encoder needs to communicate information only to Decoder 1. Since Decoder 1 has no side information, the Kaspi and HB settings yield equal performance i.e., $R_{Kaspi}(D_1, D_2) = R_{HB}(D_1, D_2)$. Moreover, if D_1 is sufficiently smaller than D_2, the operation of the encoder is limited by the distortion requirements of Decoder 1. In this case, Decoder 2 can in fact reconstruct as $\hat{X}_1 = \hat{X}_2$ while still satisfying its distortion constraints. Therefore, the same performance is obtained in all of the three settings, i.e., $R_{Kaspi}(D_1, D_2) = R_{HB}(D_1, D_2) = R_{HB}^{CR}(D_1, D_2)$. It is also noted that the general performance loss due to the CR constraint, unless, as discussed above, distortion D_1 is sufficiently smaller than D_2.

3.3 Heegard-Berger Problem with Cooperative Decoders

The system model for the HB problem with CR and decoder cooperation is similar to the one provided in Section 3.2.1 with the following differences. Here, in addition to encoding function given in (8.33) which maps the source sequence X^n into a message

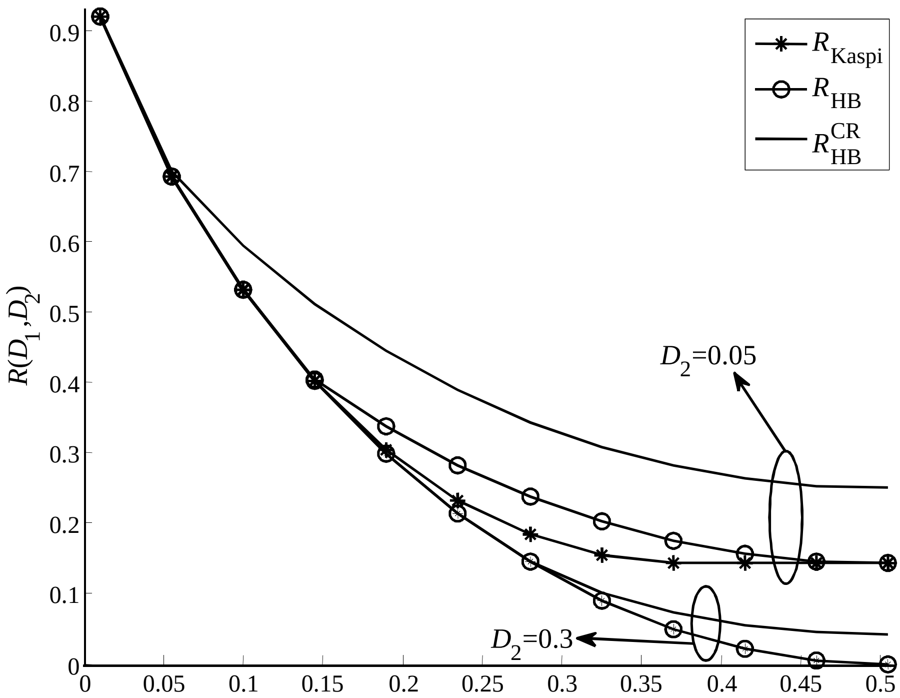

Figure 3.9 Rate-distortion functions $R_{Kaspi}(D_1, D_2)$ [19], $R_{HB}(D_1, D_2)$ [14] and $R_{HB}^{CR}(D_1, D_2)$ (3.25) for a binary source under erased side information versus distortion D_1 ($p_1 = 1$, $p_2 = 0.35$, $D_2 = 0.05$ and $D_2 = 0.3$).

J_1 of nR_1 bits, there is an encoder at Decoder 1 given by

$$g_1\colon\ [1, 2^{nR_1}] \times \mathcal{Y}_1^n \to [1, 2^{nR_2}], \tag{3.30}$$

which maps message J_1 and the source sequence Y_1^n into a message J_2. Moreover, instead of the decoding function given in (3.5), the decoding function for Decoder 2 is

$$h_2\colon\ [1, 2^{nR_1}] \times [1, 2^{nR_2}] \times \mathcal{Y}_2^n \to \hat{\mathcal{X}}_2^n, \tag{3.31}$$

which maps the messages J_1 and J_2 and the side information Y_2^n into the estimated sequence $\hat{X}_2^n$.

3.3.1 Rate-Distortion Region for $X - Y_1 - Y_2$

In this section, a single-letter characterization of the rate-distortion region is derived under the assumption that the joint pmf $p(x, y_1, y_2)$ is such that the Markov chain $X - Y_1 - Y_2$ holds[3].

Proposition 3.4. *The rate-distortion region $\mathcal{R}^{CR}(D_1, D_2)$ for the HB source coding problem with CR and cooperative decoders under the assumption $X - Y_1 - Y_2$ is given by the union of all rate pairs (R_1, R_2) that satisfy the conditions*

$$\begin{aligned} R_1 &\geq I(X; \hat{X}_1 \hat{X}_2 | Y_1) && \text{(3.32a)} \\ \textit{and } R_1 + R_2 &\geq I(X; \hat{X}_2 | Y_2) + I(X; \hat{X}_1 | Y_1, \hat{X}_2), && \text{(3.32b)} \end{aligned}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x, y_1, y_2, \hat{x}_1, \hat{x}_2) = p(x, y_1) p(y_2 | y_1) p(\hat{x}_1, \hat{x}_2 | x), \tag{3.33}$$

for some pmf $p(\hat{x}_1, \hat{x}_2 | x)$ such that the constraints (3.13) are satisfied.

The proof of the converse can be easily established following cut-set arguments for bound (3.32a), while the bound (3.32b) on the sum-rate $R_1 + R_2$ can be proved following the same step as in Appendix A and substituting J with (J_1, J_2). As for the achievability, it follows as a straightforward extension of [10, Section III] to the setup at hand where Decoder 2 has side information as well. It is worth emphasizing that the reconstruction $\hat{X}_2$ for the Decoder 2, which has degraded side information, is conveyed by using both the direct link from the Encoder, of rate R_1, and the path Encoder-Decoder 1-Decoder 2. The latter path leverages the the better side information at Decoder 1 and the cooperative link of rate R_2.

[3] Note that, unlike the conventional HB problem studied in Section 3.2, the rate-distortion region with cooperative decoders depends on the joint distribution of the variables (Y_1, Y_2), and thus stochastic and physical degradedness of the side information sequences lead to different results.

Remark 3.7. Without the CR constraint, the problem of determining the rate-distortion region for the setting of Figure 3.3 under the assumption $X - Y_1 - Y_2$ is still open. In [10], inner and outer bounds are obtained to the rate distortion region, for the case which the side information Y_2 is absent. The bounds were shown to coincide for the case where Decoder 1 wishes to recover X losslessly (i.e., $D_1 = 0$) and also for certain distortion regimes in the quadratic Gaussian case. Moreover, the rate distortion tradeoff is completely characterized in [10] for the case in which the encoder also has access to the side information. It is noted that, as per the discussion in Section 3.2.4, these latter result immediately carry over to the case with CR constraint since the encoder is informed about the side information.

Remark 3.8. To understand why imposing the CR constraint simplifies the problem of obtaining a single-letter characterization of the rate-distortion function, let us consider the degrees of freedom available at Decoder 1 in Figure 3.3 for the use of the link of rate R_2. In general, Decoder 1 can follow two possible strategies: the first is forwarding, whereby Decoder 1 simply forwards some of the bits received from the encoder to Decoder 2; while the second is recompression, whereby the data received from the encoder is combined with the available side information Y_1^n, compressed to at most R_2 bits per symbol, and then sent to Decoder 2. It is the interplay and contrast between these two strategies that makes the general problem hard to solve. In particular, while the strategies of forwarding/recompression and combinations thereof appear to be natural candidates for the problem, finding a matching converse when both such degrees of freedom are permissible at the decoder is difficult (see, e.g., [21]). However, under the CR constraint, the strategy of recompression becomes irrelevant, since any information about the side information Y_1^n that is not also available at the encoder cannot be leveraged by Decoder 2 without violating the CR constraint. This restriction in the set of available strategies for Decoder 1 makes the problem easier to address under the CR constraint."

3.3.2 Rate-Distortion Region for $X - Y_2 - Y_1$

In this section, a single-letter characterization of the rate-distortion region is derived under the assumption that the joint pmf $p(x, y_1, y_2)$ is such that the Markov chain relationship $X - Y_2 - Y_1$ holds.

Proposition 3.5. *The rate-distortion region $\mathcal{R}^{CR}(D_1, D_2)$ for the HB source coding problem with CR and cooperative decoders under the assumption the Markov chain relationship $X - Y_2 - Y_1$ is given by the union of all rate pairs (R_1, R_2) that satisfy the conditions*

$$R_1 \geq I(X; \hat{X}_1|Y_1) + I(X; \hat{X}_2|Y_2, \hat{X}_1) \tag{3.34a}$$

$$\text{and } R_2 \geq 0, \tag{3.34b}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x, y_1, y_2, \hat{x}_1, \hat{x}_2) = p(x, y_2)p(y_1|y_2)p(\hat{x}_1, \hat{x}_2|x), \tag{3.35}$$

for some pmf $p(\hat{x}_1, \hat{x}_2|x)$ such that the constraints (3.13) are satisfied.

The proof of achievability follows immediately by neglecting the link of rate R_2 and using rate R_1 as per the HB scheme of Proposition 3.1. The converse follows by considering an enhanced system in which Decoder 2 is provided with the side information of Decoder 1. In this system, link R_2 becomes useless since Decoder 2 possesses all the information available at Decoder 1. It follows that the system reduces to the HB problem with degraded sources studied in the previous section and the bound (3.34a) follows immediately from Proposition 3.1.

Remark 3.9. In the case without CR, the rate-distortion function is given similarly to (3.34), but with the HB rate-distortion function (3.15) in lieu of the rate-distortion function of the HB problem with CR in (3.34a).

3.4 Cascade Source Coding with Common Reconstruction

In this section, the system model in Figure 3.4 of cascade source coding with CR is first detailed. As mentioned in Section 3.1, the motivation for studying this class of models comes from multi-hop applications. Next, the characterization of the corresponding rate-distortion performance is presented under the assumption that one of the two side information sequences is a degraded version of the other. Finally, following the previous section, three specific examples are worked out, namely Gaussian sources under quadratic distortion (Section 3.4.3), and binary sources with side information subject to erasures under Hamming or erasure distortion (Section 3.4.3).

3.4.1 System Model

In this section, the system model for the cascade source coding problem with CR is detailed similar to Section 3.2.1. The problem is defined by the pmf $p_{XY_1Y_2}(x, y_1, y_2)$ and discrete alphabets $\mathcal{X}, \mathcal{Y}_1, \mathcal{Y}_2, \hat{\mathcal{X}}_1$, and $\hat{\mathcal{X}}_2$ as follows. The source sequence X^n and side information sequences Y_1^n and Y_2^n, with $X^n \in \mathcal{X}^n$, $Y_1^n \in \mathcal{Y}_1^n$, and $Y_2^n \in \mathcal{Y}_2^n$ are such that the tuples (X_i, Y_{1i}, Y_{2i}) for $i \in [1, n]$ are i.i.d. with joint pmf $p_{XY_1Y_2}(x, y_1, y_2)$. Node 1 measures a sequence X^n and encodes it into a message J_1 of nR_1 bits, which is delivered to Node 2. Node 2 estimates a sequence $\hat{X}_1^n \in \hat{\mathcal{X}}_1^n$ within given distortion requirements. Node 2 also encodes the message J_1 received from Node 1 and the sequence Y_1^n into a message J_2 of nR_2 bits, which is delivered to Node 3. Node 3 estimates a sequence $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ within given distortion requirements. Distortion and CR requirements are defined as in Section 3.2.1, leading to the following definition.

Definition 3.2. An $(n, R_1, R_2, D_1, D_2, \epsilon)$ code for the cascade source coding problem with CR consists an encoding function for Node 1,

$$g_1\colon \mathcal{X}^n \to [1, 2^{nR_1}], \tag{3.36}$$

which maps the source sequence X^n into a message J_1; an encoding function for Node 2,

$$g_2\colon\ [1, 2^{nR_1}] \times \mathcal{Y}_1^n \to [1, 2^{nR_2}], \tag{3.37}$$

which maps the source sequence Y_1^n and message J_1 into a message J_2; a decoding function for Node 2

$$h_1\colon\ [1, 2^{nR_1}] \times \mathcal{Y}_1^n \to \hat{\mathcal{X}}_1^n, \tag{3.38}$$

which maps message J_1 and the side information Y_1^n into the estimated sequence $\hat{X}_1^n$; a decoding function for Node 3

$$h_2\colon\ [1, 2^{nR_2}] \times \mathcal{Y}_2^n \to \hat{\mathcal{X}}_2^n, \tag{3.39}$$

which maps message J_2 and the side information Y_2^n into the estimated sequence $\hat{X}_2^n$; two encoder reconstruction functions as in (3.7), which map the source sequence into estimated sequences $\psi_1(X^n)$ and $\psi_2(X^n)$ at Node 1; such that the distortion constraints (3.8) and (3.9) are satisfied.

Given a distortion pair (D_1, D_2), a rate pair (R_1, R_2) is said to be achievable if, for any $\epsilon > 0$ and sufficiently large n, there a exists an $(n, R_1, R_2, D_1 + \epsilon, D_2 + \epsilon, \epsilon)$ code. The rate-distortion region $\mathcal{R}(D_1, D_2)$ is defined as the closure of all rate pairs (R_1, R_2) that are achievable given the distortion pair (D_1, D_2).

3.4.2 Rate-Distortion Region for $X - Y_1 - Y_2$

In this section, a single-letter characterization of the rate-distortion region is derived under the assumption that the joint pmf $p(x, y_1, y_2)$ is such that the Markov chain relationship $X - Y_1 - Y_2$ holds [4].

[4]As for the HB problem with cooperative decoders studied in Section 3.3, the rate-distortion region of the cascade source coding problem depends on the joint distribution of the variables

Proposition 3.6. *The rate-distortion region $\mathcal{R}^{CR}(D_1, D_2)$ for the cascade source coding problem with CR is given by the union of all rate pairs (R_1, R_2) that satisfy the conditions*

$$R_1 \geq I(X; \hat{X}_1\hat{X}_2|Y_1) \tag{3.40a}$$

$$\text{and } R_2 \geq I(X; \hat{X}_2|Y_2), \tag{3.40b}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x, y_1, y_2, \hat{x}_1, \hat{x}_2) = p(x, y_1)p(y_2|y_1)p(\hat{x}_1, \hat{x}_2|x), \tag{3.41}$$

for some pmf $p(\hat{x}_1, \hat{x}_2|x)$ such that the constraints (3.13) are satisfied.

The proof of the converse is easily established following cut-set arguments. To prove achievability, it is sufficient to consider a scheme based on binning at Node 1 and decode and rebin at Node 2 (see [13]). Specifically, Node 1 randomly generates a standard lossy source code $\hat{X}_1^n$ for the source X^n with rate $I(X; \hat{X}_1)$ bits per source symbol. Random binning is used to reduce the rate to $I(X; \hat{X}_1|Y_1)$. Node 1 then maps the source X^n into the reconstruction sequence $\hat{X}_2^n$ using a codebook that is generated conditional on $\hat{X}_1^n$ with rate $I(X; \hat{X}_2|\hat{X}_1)$ bits per source symbol. Using the side information Y_1^n available at Node 2, random binning is again used to reduce the rate to $I(X; \hat{X}_2|Y_1\hat{X}_1)$. The codebook of $\hat{X}_2^n$ is also randomly binned to the rate $I(X; \hat{X}_2|Y_2)$. Node 2, having recovered $\hat{X}_2^n$, forwards the corresponding bin index to Node 3. The latter, by choice of the binning rate, is able to obtain $\hat{X}_2^n$. Note that, since the reconstruction sequences $\hat{X}_1^n$ and $\hat{X}_2^n$ are generated by the encoder, functions ψ_1 and ψ_2 that guarantees the CR constraints (3.9) exist by construction.

Remark 3.10. Without the CR constraint, the problem of determining the rate-distortion region for the setting of Figure 3.4 under the Markov condition $X - Y_1 - Y_2$

(Y_1, Y_2), and thus stochastic and physical degradedness of the side information sequences lead to different results.

is still open. In the special case in which $Y_1 = Y_2$ the problem has been solved in [12] for Gaussian sources under quadratic distortion and in [14] for binary sources with erased side information under Hamming distortion.

Remark 3.11. Following Remark 3.3, if both side information sequences are causal, it can be shown that they have no impact on the rate-distortion function (3.40). Therefore, the rate-distortion region follows immediately from the results in (3.40) by removing both of the side information terms. Note that with causal side information sequences the rate-distortion function holds for any joint pmf $p(x, y_1, y_2)$ with no degradedness requirements. Moreover, if only the side information Y_2 is causal, while Y_1 is still observed non-causally, then the side information Y_2 can be neglected without loss of optimality, and the rate-distortion region follows from (3.40) by removing the conditioning on Y_2.

3.4.3 Bounds on the Rate-Distortion Region for $X - Y_2 - Y_1$

In this section, outer and inner bounds are derived for the rate-distortion region under the assumption that the joint pmf $p(x, y_1, y_2)$ is such that the Markov chain relationship $X - Y_2 - Y_1$ holds. The bounds are then shown to coincide in Section 3.4.3 for Gaussian sources and in Section 3.4.3 for binary sources with erased side information.

Proposition 3.7. *(Outer bound) The rate-distortion region $\mathcal{R}^{CR}(D_1, D_2)$ for the cascade source coding problem with CR is contained in the region $\mathcal{R}^{CR}_{out}(D_1, D_2)$, which is given by the set of all rate pairs (R_1, R_2) that satisfy the conditions*

$$R_1 \geq R^{CR}_{HB}(D_1, D_2) \tag{3.42a}$$

$$\text{and } R_2 \geq R^{CR}_{X|Y_2}(D_2), \tag{3.42b}$$

where $R_{HB}^{CR}(D_1, D_2)$ *is defined in (3.11) and* $R_{X|Y_2}^{CR}(D_2) = \min I(X; \hat{X}_2|Y_2)$*, where the minimization is performed with respect to the conditional pmf* $p(\hat{x}_2|x)$ *under the distortion constraints (3.13) for* $j = 2$*.*

Proposition 3.8. *(Inner bound) The rate-distortion region* $\mathcal{R}^{CR}(D_1, D_2)$ *for the cascade source coding problem with CR contains the region* $\mathcal{R}_{in}^{CR}(D_1, D_2)$*, which is given by the union of all rate pairs* (R_1, R_2) *that satisfy the conditions*

$$
\begin{aligned}
R_1 &\geq I(X; \hat{X}_1|Y_1) + I(X; \hat{X}_2|Y_2\hat{X}_1) && \text{(3.43a)}\\
\textit{and } R_2 &\geq I(X; \hat{X}_1|Y_2) + I(X; \hat{X}_2|\hat{X}_1 Y_2) && \text{(3.43b)}\\
&= I(X; \hat{X}_1\hat{X}_2|Y_2) && \text{(3.43c)}
\end{aligned}
$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x, y_1, y_2, \hat{x}_1, \hat{x}_2) = p(x, y_2)p(y_1|y_2)p(\hat{x}_1, \hat{x}_2|x), \tag{3.44}$$

for some pmf $p(\hat{x}_1, \hat{x}_2|x_1)$ *such that the distortion constraints (3.13) are satisfied.*

The outer bound in Proposition 3.7 follows immediately from cut-set arguments similar to those in [12] and [14]. As for the inner bound of Proposition 19, the strategy works as follows. Node 1 sends the description $\hat{X}_1^n$ to Node 2 using binning with rate $I(X; \hat{X}_1|Y_1)$. It also maps the sequence X^n into the sequence $\hat{X}_2^n$ using a conditional codebook with respect to $\hat{X}_1^n$, which is binned in order to leverage the side information Y_2^n at Node 3 with rate $I(X; \hat{X}_2|\hat{X}_1, Y_2)$. Node 2 recovers $\hat{X}_1^n$, whose codebook is then binned to rate $I(X; \hat{X}_1|Y_2)$. Then, it forwards the so obtained bin index for $\hat{X}_1^n$ and the bin index for the codebook of $\hat{X}_2^n$ produced by Node 1 to Node 3. By the choice of the rates, the latter can recover both $\hat{X}_1^n$ and $\hat{X}_2^n$. Since both descriptions are produced by Node 1, the CR constraint is automatically satisfied.

The inner and outer bounds defined above do not coincide in general. However, in the next sections, two examples are provided in which they coincide and thus characterize the rate-distortion region of the corresponding settings.

Remark 3.12. Without the CR constraint, the problem of deriving the rate-distortion region for the setting at hand under the Markov chain condition $X - Y_2 - Y_1$ is open. The problem has been solved in [12] for Gaussian sources under quadratic distortion and in [14] for binary sources with erased side information under Hamming distortion for $Y_1 = Y_2$.

Gaussian Sources and Quadratic Distortion In this section, Gaussian sources in (3.17) and the quadratic distortion are assumed as in Sec 3.2.3, and derive the rate-distortion region for the cascade source coding problem with CR.

Proposition 3.9. *The rate-distortion region* $\mathcal{R}^{CR}(D_1, D_2)$ *for the cascade source coding problem with CR for the Gaussian sources in (3.17) and quadratic distortion is given by (3.42) with* $R_{HB}^{CR}(D_1, D_2)$ *in* (3.19) *and* $R_{X|Y_2}^{CR}(D_2) = R_{\mathcal{G}}^{CR}(D_2, N_2)$ *(see* (3.18)).

The proof is given in Appendix E.

Binary Sources with Erased Side Information and Hamming Distortion In this section, the binary sources in Figure 3.7 and the Hamming distortion are assumed as in Sec 3.2.4, and derive the rate-distortion region for the cascade source coding problem with CR.

Proposition 3.10. *The rate-distortion region* $\mathcal{R}^{CR}(D_1, D_2)$ *for the cascade source coding problem with CR for the binary sources in Figure 3.7 and Hamming distortion is given by (3.42) with* $R_{HB}^{CR}(D_1, D_2)$ in (3.25) and $R_{X|Y_2}^{CR}(D_2) = R_{\mathcal{B}}^{CR}(D_2, p_2)$ (see (3.23)).

The proof is given in Appendix F.

3.5 Heegard-Berger Problem with Constrained Reconstruction

In this section, the HB problem is revisited and relax the CR constraint to the ConR constraint of [13]. This implies that the code as per Definition 1 is still adopted, but (3.9) is substituted with the less stringent constraint

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d_{e,j}(\hat{X}_{ji},\psi_{ji}(X^n))\right] \leq D_{e,j} \text{ for } j=1,2, \tag{3.45}$$

where $d_{e,j}(\hat{x}_j, \hat{x}_{e,j})$: $\hat{\mathcal{X}}_j \times \hat{\mathcal{X}}_j \rightarrow [0, D_{e,max}]$ is a per-symbol distortion metric and $\psi_{ji}(X^n)$, for $j = 1, 2$ is used to denote the ith letter of the vector $\psi_j(X^n) = (\psi_{j1}(X^n), ..., \psi_{jn}(X^n))$.

Definition 3.3. Given a distortion tuple $(D_{e,1}, D_{e,2}, D_1, D_2)$, a rate R is said to be achievable if, for any $\epsilon > 0$ and sufficiently large n, there a exists an $(n, R, D_{e,1} + \epsilon, D_{e,2} + \epsilon, D_1 + \epsilon, D_2 + \epsilon, \epsilon)$ code. The rate-distortion function $R(D_{e,1}, D_{e,2}, D_1, D_2)$ is defined as $R(D_{e,1}, D_{e,2}, D_1, D_2) = \inf\{R$: the tuple $(D_{e,1}, D_{e,2}, D_1, D_2)$ is achievable$\}$.

Note that, by setting $D_{e,j} = 0$ for $j = 1, 2$, and letting $d_{e,j}(\hat{x}_j, \hat{x}_{e,j})$ be the Hamming distortion metric (i.e., $d_{e,j}(\hat{x}_j, \hat{x}_{e,j}) = 1$ if $x \neq \hat{x}_j$ and $d_{e,j}(\hat{x}_j, \hat{x}_{e,j}) = 0$ if $x = \hat{x}_j$), a relaxed CR constraint is obtained in which the average per-symbol, rather than per-block, error probability criterion is adopted.

Remark 3.13. The problem at hand reduces to the one studied in [15] by setting $D_1 = D_{max}$ and $D_{e,1} = D_{e,max}$.

Proposition 3.11. *If the side information Y_1 is stochastically degraded with respect to Y_2, the rate-distortion function for the HB problem with ConR is given by*

$$\begin{aligned} R_{HB}^{ConR}(D_{e,1}, D_{e,2}, D_1, D_2) &= \min I(X;U_1|Y_1) + I(X;U_2|Y_2U_1) && \text{(3.46a)}\\ &= \min I(X;U_1U_2|Y_2) + I(U_1;Y_2|Y_1), && \text{(3.46b)}\end{aligned}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x, y_1, y_2, u_1, u_2) = p(x, y_1, y_2)p(u_1, u_2|x), \tag{3.47}$$

and minimization is performed with respect to the conditional pmf $p(u_1, u_2|x)$ *and the deterministic functions* $\hat{\mathrm{x}}_{\mathrm{j}}(u_j, y_j)$: $\mathcal{U}_j \times \mathcal{Y}_j \to \hat{\mathcal{X}}_j$ and $\hat{\mathrm{x}}_{\mathrm{e,j}}(u_j, x)$: $\mathcal{U}_j \times \mathcal{X} \to \hat{\mathcal{X}}_{e,j}$ *for* $j = 1, 2$*, such that the distortion constraints* $\mathrm{E}[d_j(X, \hat{\mathrm{x}}_j(U_j, Y_j))] \leq D_j$ for $j = 1, 2$, *and the ConR requirements*

$$\mathrm{E}[d_{e,j}(\hat{\mathrm{x}}_j(U_j, Y_j), \hat{\mathrm{x}}_{e,j}(U_j, X))] \leq D_{e,j}, \text{ for } j = 1, 2, \tag{3.48}$$

are satisfied. Finally, (U_1, U_2) *are auxiliary random variables whose alphabet cardinalities can be constrained as* $|\mathcal{U}_1| \leq |\mathcal{X}| + 4$ *and* $|\mathcal{U}_2| \leq (|\mathcal{X}| + 2)^2$.

The proof is given in Appendix G.

Remark 3.14. Proposition 3.11 reduces to [15, Theorem 2] when setting $D_1 = D_{max}$ and $D_{e,1} = D_{e,max}$.

Remark 3.15. Similar to [15, Theorem 2], it can be proved that, by setting $D_{e,1} = D_{e,2} = 0$ and letting $d_{e,j}$ be the Hamming distortion for $j = 1, 2$, the rate-distortion function (3.46), $R_{HB}^{ConR}(0, 0, D_1, D_2)$, reduces to the rate-distortion function with CR (3.11).

Remark 3.16. Similar to Remark 3.15, if $D_{e,1} = 0$ and $D_{e,2} = D_{e,max}$, the rate-distortion function (3.46) is given by

$$R_{HB}^{CR}(0, D_{e,max}, D_1, D_2) \quad = \quad \min I(X; \hat{X}_1|Y_1) + I(X; U_2|Y_2\hat{X}_1), \tag{3.49}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x, y_1, y_2, u_2, \hat{x}_1) = p(x, y_1, y_2)p(\hat{x}_1, u_2|x), \tag{3.50}$$

and minimization is performed with respect to the conditional pmf $p(\hat{x}_1, u_2|x)$ and the deterministic functions $\hat{\mathrm{x}}_2(u_2, y_2)$: $\mathcal{U}_2 \times \mathcal{Y}_2 \to \hat{\mathcal{X}}_2$ and $\hat{\mathrm{x}}_{\mathrm{e},2}(u_2, x)$: $\mathcal{U}_2 \times \mathcal{X} \to \hat{\mathcal{X}}_{e,2}$, such that the distortion constraints $\mathrm{E}[d_1(X, \hat{X}_1)] \leq D_1$ and $\mathrm{E}[d_2(X, \hat{\mathrm{x}}_2(U_2, Y_2))] \leq D_2$ and the ConR requirement $\mathrm{E}[d_{e,2}(\hat{\mathrm{x}}_2(U_2, Y_2), \hat{x}_{e,2}(U_2, X))] \leq D_{e,2}$ are satisfied. It can be proved that this is also the rate-distortion function under the *partial CR* requirement that there exists a function $\psi_1(X^n)$ such that (3.9) holds for $j = 1$ only. Similar conclusions apply symmetrically to the case where CR and ConR requirements are imposed only on the reconstruction of Decoder 2.

Remark 3.17. If both side information sequences are causally available at the decoders, it can be proved that they have no impact on the rate-distortion function (3.46). In this case, the rate-distortion function follows immediately from the results in (3.46) by removing conditioning on both side information sequences. Moreover, the result can be simplified by introducing a single auxiliary random variable. Similarly, if only side information Y_1 is causal, then it can be neglected with no loss of optimality, and the results follow from (3.46) by removing the conditioning on Y_1.

Remark 3.18. It is noted that the ConR formulation studied in this section is more general than the conventional formulation with distortion constraints for the decoders only. Therefore, problems that are open with the conventional formulation, such as HB with cooperative decoders (Section 3.3) and cascade source coding (Section 3.4), are *a fortiori* also open in the ConR set-up.

3.6 Concluding Remarks

The Common Reconstruction requirement [8], and its generalization in [15], substantially modify the problem of source coding in the presence of side information at the decoders. From a practical standpoint, in various applications, such as transmission of medical records, CR is a design constraint. In these cases, evaluation of the rate-distortion performance under CR thus reveals the cost, in terms of transmission resources, associated with this additional requirement. From a theoretical perspective, adding the CR constraint to standard source coding problems with decoder side information proves instrumental in concluding about the optimality of various known strategies in settings in which the more general problem, without the CR constraint, is open [8]. This chapter has extended these considerations from a point-to-point setting to three baseline multiterminal settings, namely the Heegard-Berger problem, the HB problem with cooperating decoders and the cascade problems. The optimal rate-distortion trade-off has been derived in a number of cases and explicitly evaluated in various examples.

A general subject of theoretical interest is identifying those models for which the CR requirements enables a solution of problems that have otherwise resisted solutions for decades. Examples include the Heegard-Berger and cascade source coding problems with no assumptions on side information degradedness and the one-helper lossy source coding problem.

CHAPTER 4

DISTRIBUTED AND CASCADE SOURCE CODING WITH SIDE INFORMATION "VENDING MACHINE"

4.1 Introduction

As reviewed in Chapter 2, reference [4] introduced the notion of a side information "vending machine". To illustrate the idea, consider the setting in Figure 4.1, as studied in [4]. Here, as explained in Chapter 2, unlike the conventional Wyner-Ziv set-up (see, e.g., [1, Chapter 12]), the joint distribution of the side information Y^n available at the decoder (Node 2) and of the source X^n observed at the encoder (Node 1) is not given. Instead, it can be controlled through the selection of an "action" A^n, so that, for a given action A^n and source symbol X^n, the side information Y^n is distributed according to a given conditional distribution $p(y|a,x)$. Action A^n is selected by the decoder based on the message M, of R bits per source symbol, received from the encoder, and is subject to a cost constraint. The latter limits the "quality" of the side information that can be collected by the decoder.

The source coding problem with a vending machine provides a useful model for scenarios in which acquiring data as side information is costly and thus should be done effectively. Examples include computer networks, in which data must be obtained from remote data bases, and sensor networks, where data is acquired via measurements.

The key aspect of this model is that the message M produced by the encoder plays a double role. In fact, on the one hand, it needs to carry the description of the source X^n itself, as in, e.g., the standard Wyner-Ziv model. On the other hand, it can also carry *control* information aimed at enabling the decoder to make an appropriate selection of action A^n. The goal of such a selection is to obtain a side information

Y^n that is better suited to provide partial information about the source X^n to the decoder. This in turn can potentially reduce the rate R necessary for the decoder to reconstruct source X^n at a given distortion level (or, vice versa, to reduce the distortion level for a given rate R).

The performance of the system in Figure 4.1 is expressed in terms of the interplay among three metrics, namely the rate R, the cost budget Γ on the action A^n, and the distortion D of the reconstruction $\hat{X}$ at the decoder. This trade-off is summarized by the *rate-distortion-cost* function $R(D,\Gamma)$. This function characterizes the infimum of all rates R for which a distortion level D can be achieved under an action cost budget Γ, by allowing encoding of an arbitrary number n of source symbols $X^n = (X_1, ..., X_n)$. This function is derived in [4] for both cases in which the side information Y^n is available "non-causally" to the decoder, as in the standard Wyner-Ziv model, or "causally", as introduced in [16]. In the former case (Figure 4.1-(a)), the estimated sequence $\hat{X}^n = (\hat{X}_1, ..., \hat{X}_n)$ is a function of message M and of the entire side information sequence $Y^n = (Y_1, ..., Y_n)$, while, in the latter (Figure 4.1-(b)), each estimated sample $\hat{X}_i$ is a function of message M and the side information as received up to time i, i.e., $Y^i = (Y_1, ..., Y_i)$ for $i = 1, ..., n$. It is noted that the model with causal side information is appropriate, for instance, when there are delay constraints on the reproduction at the decoder or when the decoder operates by filtering the side information sequence. The reader is referred to [16, Sec I] for an extensive discussion on these points.

Following reference [4], recent works [22] and [23] generalized the characterization of the rate-distortion-cost function for the models in Figure 4.1 to a set-up analogous to the so called Kaspi-Heegard-Berger problem [3][9], in which the side information vending machine may or may not be available at the decoder. This entails the presence of *two decoders*, rather than only one as in Figure 4.1, one with access to the vending machine and one without any side information. Reference [23]

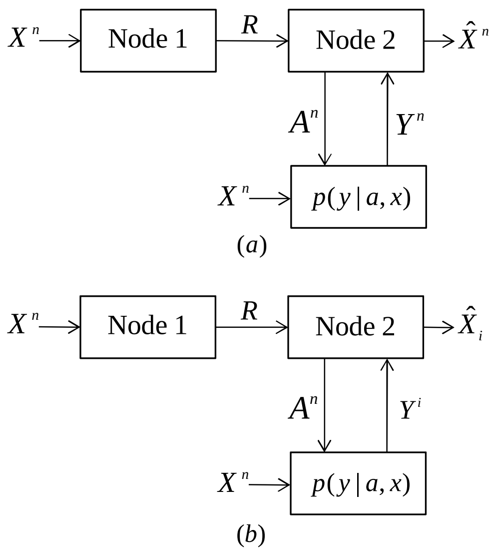

Figure 4.1 Source coding with a vending machine at the decoder [4] with: (a) "non-causal" side information; (b) "causal" side information.

also solved the more general case in which both decoders have access to the same vending machine, and either the side informations produced by the vending machine at the two decoders satisfy a degradedness condition, or lossless source reconstructions are required at the decoders. The papers [24][25] studied the setting of Figure 4.1 but under the additional constraints of common reconstruction, in the sense of [8], in [24], and of secrecy with respect to an "eavesdropping" node in [25], providing characterizations of the corresponding achievable performance. The impact of actions that adapt to the previously measured samples of the side information is studied in [26]. In [27], a related problem is considered in which the sequence to be compressed is dependent on the actions taken by a separate encoder. Finally, real-time constraints are investigated in [28].

4.1.1 Contributions and Overview

In this chapter, two multi-terminal extensions of the set-up in Figure 4.1 are studied, namely the *distributed source coding* setting of Figure 4.2, and the *cascade* model of Figure 4.3. The analysis of these scenarios is motivated by the observation that they constitute key components of computer and sensor networks. In fact, as discussed above, an important aspect of these networks is the need to effectively acquire side information data, which can be modeled by including a side information vending machine. The two extensions and the corresponding main results are evaluated below.

1) *Distributed source coding with a side information vending machine* (Section 4.2): In the distributed source coding setting of Figure 4.2, *two encoders* (Node 1 and Node 2), which measure correlated sources X_1^n and X_2^n, respectively, communicate over rate-limited links, of rates R_1 and R_2, respectively, to a single decoder (Node 3). The decoder has side information Y^n on sources X_1^n and X_2^n, which can be controlled through an action A^n. The action sequence is selected by the decoder based on the messages M_1 and M_2 received from Node 1 and Node 2, respectively, and needs to satisfy a cost constraint of Γ. Inner bounds are derived to the rate-distortion-

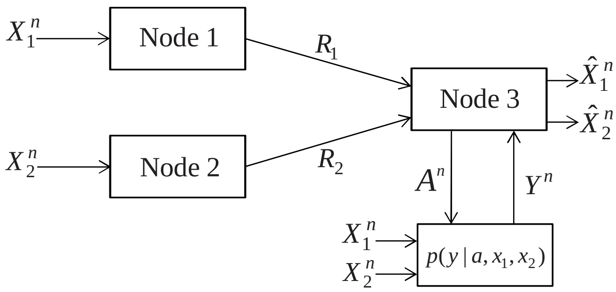

Figure 4.2 Distributed source coding with a side information vending machine at the decoder.

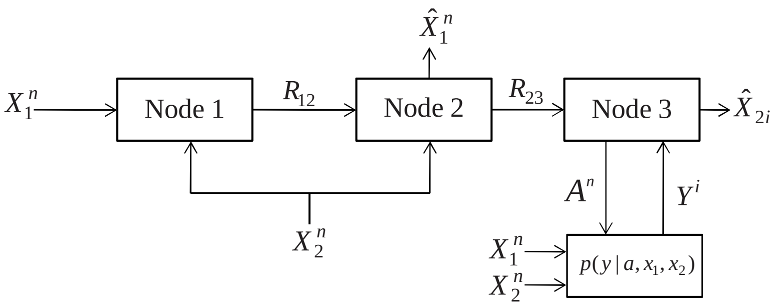

Figure 4.3 Cascade source coding with a side information vending machine. Side information is assumed to be available "causally" to the decoder.

cost region $\mathcal{R}(D_1, D_2, \Gamma)$ under non-causal and causal side information by combining the strategies proposed in [4] with the Berger-Tung strategy [29] and its extension to the Wyner-Ziv set-up [30]. These bounds are shown to be tight under specific assumptions, including the scenario where the sequence observed by one of the nodes is a function of the source observed by the other and the side information is available causally at the decoder.

2) *Cascade source coding with a side information vending machine* (Section 4.3): In the cascade model of Figure 4.3, Node 1 is connected via a rate-limited link, of rate R_{12}, to Node 2, which is in turn communicates with Node 3 with rate R_{23}. Source X_1^n is measured by Node 1 and the correlated source X_2^n by both Node 1 and Node 2. Similarly to the distributed coding setting described above, Node 3 has side information Y^n on sources X_1^n and X_2^n, which can be controlled via an action A^n. Action A^n is selected by Node 3 based on the message received from Node 2 and needs to satisfy a cost constraint of Γ. The set $\mathcal{R}(D_1, D_2, \Gamma)$ of all achievable rates (R_{12}, R_{23}) are derived for given distortion constraints (D_1, D_2) on the reconstructions $\hat{X}_1^n$ and $\hat{X}_2^n$ at Node 2 and Node 3, respectively, and for cost constraint Γ. This characterization is obtained under the assumption that the side information Y be available causally at Node 3. It is mentioned that, following the submission of this work, the analysis of the case with non-causal side information at Node 3 was carried out in [31].

4.2 Distributed Source Coding with a Side Information Vending Machine

In this section, the system model for the problem of distributed source coding with a side information vending machine is detailed in Section 4.2.1. Then, an achievable strategy is proposed in Section 4.2.2 for both the cases with non-causal and causal side information at the decoder. In Section 4.2.3 and Section 4.2.4 scenarios are discussed in which the achievable strategies match given outer bounds. A numerical example is then developed in Section 4.2.5.

4.2.1 System Model

The problem of distributed lossy source coding with a vending machine and non-causal side information is illustrated in Figure 4.2. It is defined by the probability mass functions (pmfs) $p_{X_1X_2}(x_1,x_2)$ and $p_{Y|AX_1X_2}(y|a,x_1,x_2)$ and discrete alphabets $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}, \mathcal{A}, \hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2$ as follows. The source sequences X_1^n and X_2^n with $X_1^n \in \mathcal{X}_1^n$ and $X_2^n \in \mathcal{X}_2^n$, respectively, are such that the tuples (X_{1i}, X_{2i}) for $i \in [1,n]$ are independent identically distributed (i.i.d.) with joint pmf $p_{X_1X_2}(x_1,x_2)$. Node 1 measures sequences X_1^n and encodes it into message M_1 of nR_1 bits, while Node 2 measures sequences X_2^n and encodes it into message M_2 of nR_2 bits. Node 3 wishes to reconstruct the two sources within given distortion requirements, to be discussed below, as $\hat{X}_1^n \in \hat{\mathcal{X}}_1^n$ and $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$.

To this end, Node 3 selects an action sequence A^n, where $A^n \in \mathcal{A}^n$, based on the messages M_1 and M_2 received from Node 1 and Node 2, respectively. The side information sequence Y^n is then realized as the output of a memoryless channel with inputs (A^n, X_1^n, X_2^n). Specifically, given A^n, X_1^n and X_2^n, the sequence Y^n is distributed as

$$p(y^n|a^n, x_1^n, x_2^n) = \prod_{i=1}^{n} p_{Y|AX_1X_2}(y_i|a_i, x_{1i}, x_{2i}). \tag{4.1}$$

The overall cost of an action sequence a^n is defined by a per-symbol cost function Λ: $\mathcal{A} \rightarrow [0, \Lambda_{\max}]$ with $0 \leq \Lambda_{\max} < \infty$, as

$$\Lambda^n(a^n) = \frac{1}{n}\sum_{i=1}^{n} \Lambda(a_i). \tag{4.2}$$

The estimated sequences $\hat{X}_1^n$ and $\hat{X}_2^n$ are obtained as a function of both messages M_1 and M_2 and of the side information Y^n. The estimates $\hat{X}_1^n$ and $\hat{X}_2^n$ are constrained to satisfy distortion constraints defined by two per-symbol distortion measures, namely $d_j(x_1, x_2, y, \hat{x}_j)$: $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y} \times \hat{\mathcal{X}}_j \rightarrow [0, D_{\max}]$ for $j = 1, 2$ with $0 \leq D_{\max} < \infty$. Based on such scalar measures, the overall distortion for the estimated sequences $\hat{x}_1^n$ and $\hat{x}_2^n$ is defined as

$$d_j^n(x_1^n, x_2^n, y^n, \hat{x}_j^n) = \frac{1}{n} \sum_{i=1}^{n} d_j(x_{1i}, x_{2i}, y_i, \hat{x}_{ji}) \text{ for } j = 1, 2. \tag{4.3}$$

Note that, based on (4.3), the estimate $\hat{X}_j^n$ for $j = 1, 2$ can be required to be a lossy version of an arbitrary (per-letter) function of both sources X_1^n and X_2^n and of the side information sequence Y^n. A formal description of the operations at encoders and decoder, and of cost and distortion constraints, is presented below for both the cases in which the side information is available causally or non-causally at the decoder.

Definition 4.1. An $(n, R_1, R_2, D_1, D_2, \Gamma)$ code for the case of *non-casual* side information at Node 3 consists of two source encoders

$$\begin{aligned} \mathrm{g}_1&: \mathcal{X}_1^n \rightarrow [1, 2^{nR_1}], \\ \text{and } \mathrm{g}_2&: \mathcal{X}_2^n \rightarrow [1, 2^{nR_2}], \end{aligned} \tag{4.4}$$

which map the sequences X_1^n and X_2^n into messages M_1 and M_2 at Node 1 and Node 2, respectively; an "action" function

$$\ell\colon\ [1, 2^{nR_1}] \times [1, 2^{nR_2}] \rightarrow \mathcal{A}^n, \tag{4.5}$$

which maps the message (M_1, M_2) into an action sequence A^n at Node 3; and two decoding functions

$$\mathrm{h}_1\colon [1, 2^{nR_1}] \times [1, 2^{nR_2}] \times \mathcal{Y}^n \rightarrow \hat{\mathcal{X}}_1^n, \tag{4.6}$$

$$\text{and } \mathrm{h}_2\colon [1, 2^{nR_1}] \times [1, 2^{nR_2}] \times \mathcal{Y}^n \rightarrow \mathcal{X}_2^n, \tag{4.7}$$

which map the messages M_1 and M_2, and the side information sequence Y^n into the estimated sequences $\hat{X}_1^n$ and $\hat{X}_2^n$ at Node 3; such that the action cost constraint Γ is satisfied as

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\Lambda(A_i)\right] \leq \Gamma, \tag{4.8}$$

and the distortion constraints D_1 and D_2 hold, namely

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d_j(X_{1i}, X_{2i}, Y_i, \hat{X}_{ji})\right] \leq D_j, \text{ for } j = 1, 2. \tag{4.9}$$

Definition 4.2. A $(n, R_1, R_2, D_1, D_2, \Gamma)$ code for the case of *causal* side information at Node 3 is as in Definition 4.1 with the only difference that, in lieu of (6)-(7), the sequence of decoding functions are

$$\mathrm{h}_{1i}\colon [1, 2^{nR_1}] \times [1, 2^{nR_2}] \times \mathcal{Y}^i \rightarrow \hat{\mathcal{X}}_{1i}, \tag{4.10}$$

$$\text{and } \mathrm{h}_{2i}\colon [1, 2^{nR_1}] \times [1, 2^{nR_2}] \times \mathcal{Y}^i \rightarrow \mathcal{X}_{2i}, \tag{4.11}$$

for $i \in [1, n]$, which map the message (M_1, M_2) and the measured sequence Y^i into the ith estimated symbol $\hat{X}_{ji} = \mathrm{h}_{ji}(M_1, M_2, Y^i)$ for $j = 1, 2$ at Node 3.

Definition 4.3. Given a distortion-cost tuple (D_1, D_2, Γ), a rate pair (R_1, R_2) is said to be achievable for the case with non-causal or causal side information if, for any $\epsilon > 0$ and sufficiently large n, there exists a corresponding $(n, R_1, R_2, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)$ code.

Definition 4.4. The *rate-distortion-cost region* $\mathcal{R}_{NC}(D_1, D_2, \Gamma)$ is defined as the closure of all rate pairs (R_1, R_2) that are achievable with non-causal side information given the distortion-cost tuple (D_1, D_2, Γ). The rate-distortion-cost region $\mathcal{R}_C(D_1, D_2, \Gamma)$ is similarly defined for the case of casual side information.

4.2.2 Achievable Strategies

In this section, inner bounds to the rate-distortion-cost regions are obtained for the cases with non-causal and causal side information.

Proposition 4.1. *The rate-distortion-cost region with non-causal side information at Node 3 satisfies the inclusion* $\mathcal{R}_{NC}(D_1, D_2, \Gamma) \supseteq \mathcal{R}^a_{NC}(D_1, D_2, \Gamma)$*, where the region* $\mathcal{R}^a_{NC}(D_1, D_2, \Gamma)$ *is given by the union of the set of all of rate tuples* (R_1, R_2) *that satisfy the inequalities*

$$R_1 \geq I(X_1; V_1|V_2, Q) + I(X_1; U_1|V_1, V_2, U_2, Y, Q) \tag{4.12a}$$

$$R_2 \geq I(X_2; V_2|V_1, Q) + I(X_2; U_2|V_1, V_2, U_1, Y, Q) \tag{4.12b}$$

$$\text{and } R_1 + R_2 \geq I(X_1, X_2; V_1, V_2|Q) + I(X_1, X_2; U_1, U_2|V_1, V_2, Y, Q), \tag{4.12c}$$

for some joint pmfs that factorizes as

$$\begin{aligned} p(q, x_1, x_2, y, v_1, v_2, u_1, u_2, a, \hat{x}_1, \hat{x}_2) = \\ p(q)p(x_1, x_2)p(v_1, u_1|x_1, q)p(v_2, u_2|x_2, q)\delta(a - \mathrm{a}(v_1, v_2, q)) \\ p(y|a, x_1, x_2)\delta(\hat{x}_1 - \hat{\mathrm{x}}_1(u_1, u_2, y, q))\delta(\hat{x}_2 - \hat{\mathrm{x}}_2(u_1, u_2, y, q)), \end{aligned} \tag{4.13}$$

with pmfs $p(q)$ *and* $p(v_1, u_1|x_1, q)$ *and* $p(v_2, u_2|x_2, q)$ *and deterministic functions* $\mathrm{a}\colon \mathcal{V}_1 \times \mathcal{V}_2 \times \mathcal{Q} \to \mathcal{A}$, $\hat{\mathrm{x}}_j\colon \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{Y} \times Q \to \hat{\mathcal{X}}_j$ *for* $j = 1, 2$*, such that the action and the distortion constraints*

$$\mathrm{E}\left[\Lambda(A)\right] \leq \Gamma \tag{4.14a}$$

$$\text{and } \mathrm{E}\left[d_j(X_1, X_2, Y, \hat{X}_j)\right] \leq D_j, \text{ for } j = 1, 2, \tag{4.14b}$$

hold. Finally, any extreme point of the region $\mathcal{R}^a_{NC}(D_1, D_2, \Gamma)$ *can be obtained by limiting the cardinalities of the random variables* (V_1, V_2, U_1, U_2) *as* $|\mathcal{V}_j| \leq |\mathcal{X}_j| + 6$ *and* $|\mathcal{U}_j| \leq |\mathcal{X}_j|\,|\mathcal{V}_j| + 5$*, for* $j = 1, 2$.

Remark 4.1. If $p(y|a,x_1,x_2) = p(y|x_1,x_2)$, so that the side information is action-independent, Proposition 4.1 reduces to the extension of the Berger-Tung scheme [29] to the Wyner-Ziv set-up studied in [30, Theorem 2]. Moreover, in the special case in which there is only one encoder, the achievable rate coincides with that derived in [4, Theorem 1].

The proof of Proposition 4.1 follows easily from standard arguments, and thus it is only briefly discussed here. The proposed scheme combines the Berger-Tung distributed source coding strategy [29] and the distributed Wyner-Ziv approach proposed in [30, Theorem II] with the layered two-stage coding scheme that is proved to be optimal in [4] for the special case of a single encoder. Throughout the discussion, the time-sharing variable is neglected Q for simplicity. This can be handled in the standard way (see, e.g., [1, Section 4.5.3]). The encoding scheme at Node 1 and Node 2 multiplexes two descriptions, which are obtained in two encoding stages. In the first encoding stage, the distributed source coding strategy of [29], conventionally referred to as the Berger-Tung scheme, is adopted by Node 1 and Node 2 to convey descriptions V_1^n and V_2^n, respectively, to Node 3. In order for the decoder to be able to recover these descriptions the rates $R_1^{'}$ and $R_2^{'}$ allocated by Node 1 and Node 2 have to satisfy the conditions [29][1, Chapter 13]

$$R_1^{'} \geq I(X_1;V_1|V_2) \tag{4.15a}$$

$$R_2^{'} \geq I(X_2;V_2|V_1) \tag{4.15b}$$

$$\text{and } R_1^{'} + R_2^{'} \geq I(X_1,X_2;V_1,V_2). \tag{4.15c}$$

Having decoded the descriptions (V_1^n, V_2^n), Node 3 selects the action sequence A^n as the per-symbol function $A_i = \mathrm{a}(V_{1i}, V_{2i})$ for $i \in [1,n]$. Node 3 thus measures the side information sequence Y^n. The sequences (Y^n, V_1^n, V_2^n) can then be regarded as side information available at the decoder. Therefore, in the second encoding stage, the distributed Wyner-Ziv scheme proposed in [30, Theorem 2] is used to convey the descriptions U_1^n and U_2^n by Node 1 and Node 2, respectively, to Node 3. Note that the fact that sequences (Y^n, V_1^n, V_2^n) are not i.i.d. does not affect the achievability of the rate region derived in [30]. This is because, as shown in [1, Lemma 3.1], the packing lemma leveraged to ensure the correctness of the decoding process applies for an arbitrary distribution of the sequences (Y^n, V_1^n, V_2^n). In order for the decoder to correctly retrieve the descriptions U_1^n and U_2^n, the rates $R_1^{''}$ and $R_2^{''}$ allocated by Node 1 and Node 2 must satisfy the inequalities [30]

$$R_1^{''} \geq I(X_1; U_1|V_1, V_2, U_2, Y) \tag{4.16a}$$

$$R_2^{''} \geq I(X_2; U_2|V_1, V_2, U_1, Y) \tag{4.16b}$$

$$\text{and } R_1^{''} + R_2^{''} \geq I(X_1, X_2; U_1, U_2|V_1, V_2, Y). \tag{4.16c}$$

Node 1 and Node 2 multiplex the source indices obtained in the two phases and hence the overall rates are $R_1 = R_1^{'} + R_1^{''}$ and $R_2 = R_2^{'} + R_2^{''}$. Using these equalities, along with (4.15) and (4.16), leads to (4.12). Finally, the decoder j estimates $\hat{X}_j^n$ with $j = 1, 2$ sample by sample as a function of U_{1i}, U_{2i} and Y_i. The proof of the cardinality bounds follows from standard arguments and is sketched in Appendix H[1]. Let's consider a similar achievable strategy for the case with causal side information.

Proposition 4.2. *The rate-distortion-cost region with causal side information at Node 3 satisfies the inclusion $\mathcal{R}_C(D_1, D_2, \Gamma) \supseteq \mathcal{R}_C^a(D_1, D_2, \Gamma)$, where the region $\mathcal{R}_C^a(D_1, D_2, \Gamma)$ is given by the union of the set of all of rate tuples (R_1, R_2) that satisfy the inequalities*

$$R_1 \geq I(X_1; U_1|U_2, Q) \tag{4.17a}$$

$$R_2 \geq I(X_2; U_2|U_1, Q) \tag{4.17b}$$

$$\textit{and } R_1 + R_2 \geq I(X_1, X_2; U_1, U_2|Q), \tag{4.17c}$$

[1]It is noted that, using the approach of [32], it may be possible to improve the cardinality bounds. This aspect is not further explored here.

for some joint pmfs that factorizes as

$$\begin{aligned} p(q,x_1,x_2,y,u_1,u_2,a,\hat{x}_1,\hat{x}_2) &= p(q)p(x_1,x_2)p(u_1|x_1,q)p(u_2|x_2,q)\delta(a-\mathrm{a}(u_1,u_2,q)) \\ &\quad p(y|a,x_1,x_2)\delta(\hat{x}_1-\hat{\mathrm{x}}_1(u_1,u_2,y,q)) \\ &\quad \delta(\hat{x}_2-\hat{\mathrm{x}}_2(u_1,u_2,y,q)), \end{aligned} \tag{4.18}$$

with pmfs $p(q)$, $p(u_1|x_1,q)$ *and* $p(u_2|x_2,q)$ *and deterministic functions* a: $\mathcal{U}_1\times\mathcal{U}_2\times\mathcal{Q}\rightarrow\mathcal{A}$ *and* $\hat{\mathrm{x}}_j$: $\mathcal{U}_1\times\mathcal{U}_2\times\mathcal{Y}\times Q\rightarrow\hat{\mathcal{X}}_j$ *for* $j=1,2$*, such that the action and the distortion constraints (4.14a)-(4.14b) hold, respectively. Finally, any extreme point in the region* $\mathcal{R}_C^a(D_1,D_2,\Gamma)$ *can be obtained by constraining the cardinalities of random variables* (U_1,U_2) *as* $|\mathcal{U}_1|\leq|\mathcal{X}_1|+5$ *and* $|\mathcal{U}_2|\leq|\mathcal{X}_2|+5$*.*

The proof follows by similar arguments as the ones in the proof of Proposition 4.1 with the only difference that only one stage of encoding is sufficient. Specifically, as in Proposition 4.1, Berger-Tung coding is adopted to convey the descriptions U_1^n and U_2^n to Node 3. Note that, with causal side information, there is no advantage in having a second encoding stage, since the side information sequence cannot be leveraged for binning in contrast to the case with non-causal side information [16][1, Chapter 12]. The cardinality bounds follow from arguments similar to Appendix H.

4.2.3 Degraded Source Sets and Causal Side Information

In this section, a special case in which the sequence observed by Node 2 is a symbol-by-symbol function of the source observed at Node 1 is considered [33, Section V.] (see also [34]). In other words, $X_{1i}=(X'_{1i},X_{2i})$ for $i\in[1,n]$, where $X_1'^n$ is an i.i.d. sequence independent of X_2^n. This set-up is referred to as having *degraded source sets*. Moreover, it is assumed that the side information Y is available causally at Node 3. The next proposition proves that the achievable strategy of Proposition 4.2 is optimal in this case.

Proposition 4.3. *The rate-distortion-cost region* $\mathcal{R}_C(D_1, D_2, \Gamma)$ *for the set-up with degraded source sets and with causal side information at Node 3 satisfies* $\mathcal{R}_C(D_1, D_2, \Gamma) = \mathcal{R}_C^a(D_1, D_2, \Gamma)$.

For the proof of converse, the reader is referred to Appendix I.

Remark 4.2. Proposition 3 generalizes to the case with action-dependent side information the result in [33, Section V] for the case with no side information.

4.2.4 One-Distortion Criterion and Non-Causal Side Information

In this section, a variation on the set-up of source coding with action-dependent non-causal side information described in Definition 4.1 is considered. Specifically, Node 3 selects the action sequence A^n based only on the message M_1 received from Node 1. In other words, the action function (4.5) is modified to

$$\ell\colon\ [1, 2^{nR_1}] \to \mathcal{A}^n, \tag{4.19}$$

which maps the message M_1 into an action sequence A^n at Node 3. This may be the case in scenarios in which there is a hierarchy between Node 1 and Node 2, e.g., in a sensor network, and the functionality of remote control of the side information is assigned solely to Node 1. The next proposition characterizes the rate-distortion-cost function $\mathcal{R}_{NC}(D_1, 0, \Gamma)$ under the mentioned assumption when Hamming distortion is selected for $\hat{X}_2$. That is, distortion measure $d_2(x_2, \hat{x}_2)$ is chosen as $d_H(x_2, \hat{x}_2) = 0$ if $x_2 = \hat{x}_2$ and $d_H(x_2, \hat{x}_2) = 1$ otherwise. This implies that the constraint of vanishingly small per-symbol Hamming distortion between source X_2^n and estimate $\hat{X}_2^n$ is imposed, or equivalently the constraint $\frac{1}{n}\sum_{i=1}^{n} \Pr[\hat{X}_{2i} \neq X_{2i}] \to 0$ for $n \to \infty$. This assumption will be referred to as by saying that source sequence X_2^n must be recovered losslessly at the decoder.

Proposition 4.4. *If the action function is given by (4.19) and* X_2^n *must be recovered losslessly at Node 3, the* rate-distortion-cost region $\mathcal{R}_{NC}(D_1, 0, \Gamma)$ *is given by union* *of the set of all of rate tuples* (R_1, R_2) *that satisfy the inequalities*

$$R_1 \geq I(X_1; A|Q) + I(X_1; U_1|A, X_2, Y, Q) \tag{4.20a}$$

$$R_2 \geq H(X_2|A, Y, U_1, Q) \tag{4.20b}$$

$$\text{and } R_1 + R_2 \geq I(X_1; A|Q) + H(X_2|A, Y, Q) + I(X_1; U_1|A, X_2, Y, Q), \tag{4.20c}$$

for some joint pmfs that factorizes as

$$p(q, x_1, x_2, y, u_1, a, \hat{x}_1) = p(q)p(x_1, x_2)p(a, u_1|x_1, q)p(y|a, x_1, x_2)\delta(\hat{x}_1 - \hat{\mathrm{x}}_1(u_1, x_2, y, q)), \tag{4.21}$$

with pmfs $p(q)$ *and* $p(a, u_1|x_1, q)$ *and deterministic function* $\hat{x}_1(u_1, x_2, y, q)$*, such that the action and the distortion constraints*

$$\mathrm{E}[\Lambda(A)] \leq \Gamma \tag{4.22a}$$

$$\text{and } \mathrm{E}\left[d_1(X_1, X_2, Y, \hat{X}_1)\right] \leq D_1 \tag{4.22b}$$

hold. Finally, Q *and* U_1 *are auxiliary random variables whose alphabet cardinality can be constrained as* $|Q| \leq 6$ *and* $|\mathcal{U}_1| \leq 6|\mathcal{X}_1||\mathcal{A}| + 3$ *without loss of optimality.*

Remark 4.3. In the case in which there is no side information, Proposition 4.4 reduces to [35, Theorem 1].

For the proof of converse, the reader is referred to Appendix J. The achievability follows from Proposition 4.1 by setting $V_2 = \emptyset$, $V_1 = A$ and $U_2 = X_2$.

Remark 4.4. Extension of the result in Proposition to an arbitrary number K of encoders can be found in [36].

4.2.5 A Binary Example

In this section, a specific numerical example is considered in order to illustrate the result derived in Proposition 4.1 and Proposition 4.4 and the advantage of selecting actions at Node 3 based on the message received from one of the nodes. Specifically, it is assumed that all alphabets are binary and that (X_1, X_2) is a doubly symmetric binary source (DSBS) characterized by probability p, with $0 \leq p \leq 1/2$, so that $p(x_1) = p(x_2) = 1/2$ for $x_1, x_2 \in \{0, 1\}$ and $\Pr[X_1 \neq X_2] = p$. Moreover, Hamming distortion is adopted for both sources to reconstruct both X_1 and X_2 losslessly in the sense discussed above. Note that, this implies that $d_1(x_1, x_2, y, \hat{x}_1) = d_H(x_1, \hat{x}_1)$ and $D_1 = 0$. The side information Y_i is such that

$$Y_i = \begin{cases} \mathrm{f}(X_{1i}, X_{2i}) & \text{if } A_i = 1 \\ 1 & \text{if } A_i = 0 \end{cases}, \tag{4.23}$$

where $\mathrm{f}(x_1, x_2)$ is a deterministic function to be specified. Therefore, when action $A_i = 1$ is selected, then $Y_i = \mathrm{f}(X_{1i}, X_{2i})$ is measured at the receiver, while with $A_i = 0$ no useful information is collected by the decoder. The action sequence A^n must satisfy the cost constraint (4.8), where the cost function is defined as $\Lambda(A_i) = 1$ if $A_i = 1$ and $\Lambda(A_i) = 0$ if $A_i = 0$. It follows that, given (4.23), a cost Γ implies that the decoder can observe $\mathrm{f}(X_{1i}, X_{2i})$ only for at most $n\Gamma$ symbols. As for the function $\mathrm{f}(x_1, x_2)$, two cases are considered, namely $\mathrm{f}(x_1, x_2) = x_1 \oplus x_2$, where $\oplus$ is the binary sum and $\mathrm{f}(x_1, x_2) = x_1 \odot x_2$, where $\odot$ is the binary product. It is assumed that the side information is available non-causally at the decoder.

To start with, observe that the sum-rate is a non-increasing function of the action cost Γ and hence, the minimum sum-rate is obtained when $\Gamma = 1$. With $\Gamma = 1$, it is clearly optimal to set $A = 1$, irrespective of the value of X_1. In this case, from the Slepian-Wolf theorem, the sum rate equals $R_{sum}(1) = H(X_1, X_2|Y)$. Specifically, with sum side information

$$R^{\oplus}_{sum}(1) = 1, \tag{4.24}$$

since $R^{\oplus}_{sum}(1) = H(X_1, X_2|X_1 \oplus X_2) = H(X_1|X_1 \oplus X_2) = H(X_1)$ hold, where the second equality follows from the chain rule and the third from the crypto-lemma [37, Lemma 2]. Instead, with product side information,

$$R^{\odot}_{sum}(1) = H\left(\frac{1-p}{1+p}, \frac{p}{1+p}, \frac{p}{1+p}\right)\left(\frac{1+p}{2}\right), \tag{4.25}$$

where the definition $H(p_1, p_2, ..., p_k) = -\sum_{i=1}^{k} p_k \log_2 p_k$ is used. Equation (4.25) follows since

$$\begin{aligned} R^{\odot}_{sum}(1) &= H(X_1, X_2|X_1 \odot X_2) \\ &= H(X_1, X_2|X_1 \odot X_2 = 0) \Pr[X_1 \odot X_2 = 0], \end{aligned} \tag{4.26}$$

where the second equality is a consequence of the fact that $X_1 \odot X_2 = 1$ implies that $X_1 = 1$ and $X_2 = 1$. Sum-rate (4.25) is then obtained by evaluating (4.26) for the DSBS at hand. Figure 4.4 shows the sum-rates (4.24) and (4.25), demonstrating

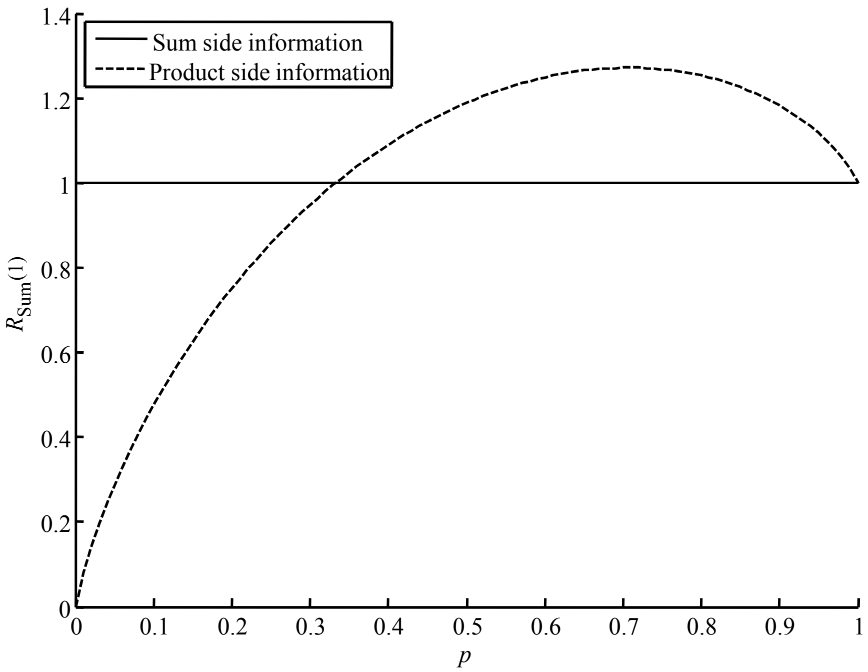

Figure 4.4 Sum-rates versus p for sum and product side informations ($\Gamma = 1$).

that, if p is sufficiently small, namely if $p \lesssim 0.33$, $R^{\odot}_{sum}(1) < R^{\oplus}_{sum}(1)$ is true and thus product side information is more informative than the sum, while for $p \gtrsim 0.33$ the opposite is true (and for $p = 1$, they are equally informative).

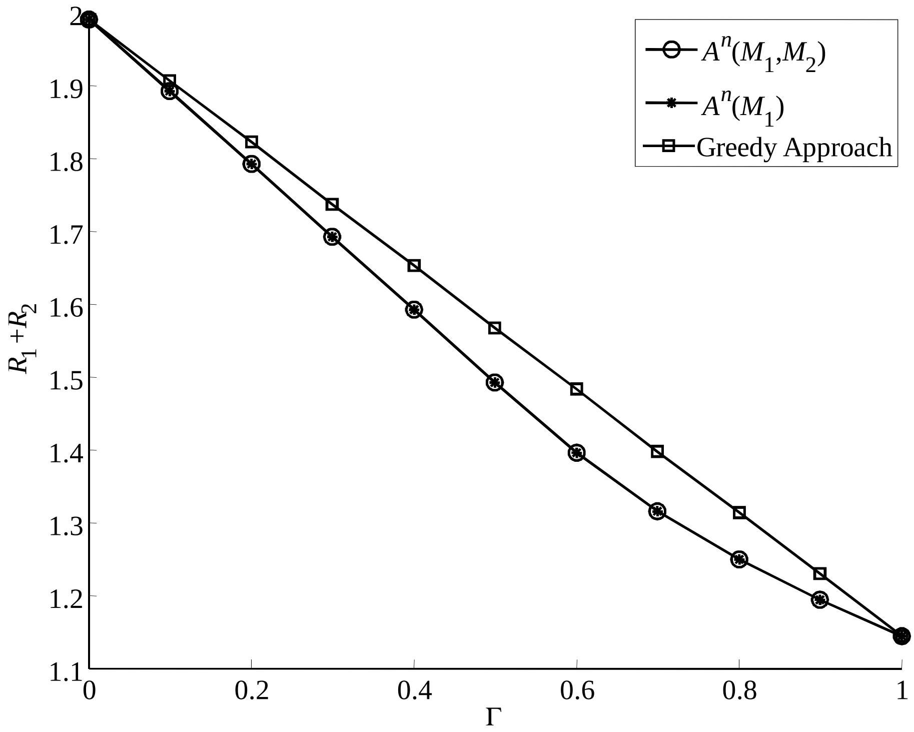

Figure 4.5 Sum-rates versus the action cost Γ for product side information ($p = 0.45$).

Considering a general cost budget $0 \leq \Gamma \leq 1$, in order to emphasize the role of both data and control information for the system performance, the sum-rate attainable by imposing that the action A be selected by Node 3 a priori, that is, without any control from Node 1, is now evaluated. This can be easily seen to be given by [4]

$$\begin{aligned} R_{sum,\ greedy}(\Gamma) &= \Gamma H(X_1, X_2|Y) + (1-\Gamma)H(X_1, X_2) \\ &= \Gamma H(X_1, X_2|Y) + (1-\Gamma)(1 + H(p)). \end{aligned} \tag{4.27}$$

This sum-rate will be compared below with the performance of the scheme in Proposition 4.1, in which the actions are selected based on both messages (M_1, M_2),

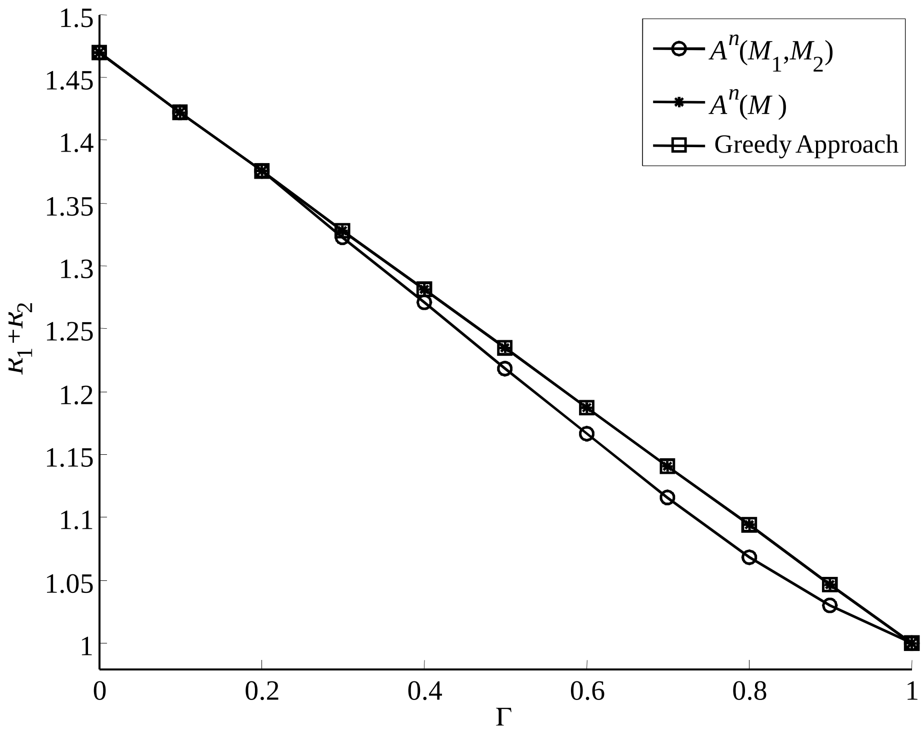

Figure 4.6 Sum-rates versus the action cost Γ for sum side information ($p = 0.1$).

and that of Proposition 4.4, in which the actions are selected based only on message M_1.

Figure 4.5 depicts the mentioned sum-rates[2] versus the action cost Γ for $p = 0.45$ and product side information. It can be seen that the greedy approach suffers from a significant performance loss with respect to the approaches in which actions are selected based on the messages received from one encoder or both encoders. It can be also observed that no gains are obtained by selecting the actions based on both messages. The fact that choosing the action based on the message received from Node 1 provides performance benefits can be explained as follows. If $X_1 = 0$, the value of the side information is always $Y = X_1 \odot X_2 = 0$ irrespective of the value of X_2. Therefore, if $X_1 = 0$, the side information is less informative than if $X_1 = 1$

[2]The sum-rate from Proposition 4.1 is calculated by assuming binary auxiliary variables V_1 and V_2 and performing global optimization.

and hence it may be advantageous to save on the action cost by setting $A = 0$. Consequently, choosing actions based on the message received from Node 1 can result in a lower sum-rate.

The scenario with sum side information is considered in Figure 4.6 for $p = 0.1$. A first observation is that, as proved in Appendix K, choosing the action based only on M_1 cannot improve the sum-rate with respect to the greedy case. This contrasts with the product side information case, and is due to the fact that X_1 is independent of the side information Y. Instead, choosing the actions based on both messages allows to save on the necessary communication sum-rate.

4.3 Cascade Source Coding with a Side Information Vending Machine

In this section, the system model for the setting of Figure 4.3 of cascade source coding with a side information vending machine is first described. It is recalled that side information Y is here assumed to be available causally at the decoder (Node 3). The corresponding model with non-causal side information is studied in [31]. Next, the characterization of the corresponding rate-distortion-cost performance is presented in Section 4.3.2.

4.3.1 System Model

The problem of cascade lossy computing with causal observation costs at second user, illustrated in Figure 4.3, is defined by the pmfs $p_{X_1X_2}(x_1, x_2)$ and $p_{Y|AX_1X_2}(y|a, x_1, x_2)$ and discrete alphabets $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}, \mathcal{A}, \hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2$, as follows. The source sequences X_1^n and X_2^n with $X_1^n \in \mathcal{X}_1^n$ and $X_2^n \in \mathcal{X}_2^n$, respectively, are such that the pairs (X_{1i}, X_{2i}) for $i \in [1, n]$ are i.i.d. with joint pmf $p_{X_1X_2}(x_1, x_2)$. Node 1 measures sequences X_1^n and X_2^n and encodes them in a message M_{12} of nR_{12} bits, which is delivered to Node 2. Node 2 estimates a sequence $\hat{X}_1^n \in \hat{\mathcal{X}}_1^n$ within given distortion requirements to be discussed below. Moreover, Node 2 encodes the message M_{12}, received from Node

1, and the locally available sequence X_2^n in a message M_{23} of nR_{23} bits, which is delivered to node 3. Node 3 wishes to estimate a sequence $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ within given distortion requirements to be discussed. To this end, Node 3 receives message M_{23} and based on this, selects an action sequence A^n, where $A^n \in \mathcal{A}^n$. The action sequence affects the quality of the measurement Y^n of sequence X_1^n and X_2^n obtained at the Node 3. Specifically, given A^n, X_1^n and X_2^n, the sequence Y^n is distributed as in (4.1). The cost of the action sequence is defined by a cost function Λ: $\mathcal{A} \rightarrow [0, \Lambda_{\max}]$ with $0 \leq \Lambda_{\max} < \infty$, as in (4.2). The estimated sequence $\hat{X}_2^n$ with $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ is then obtained as a function of M_{23} and Y^n.

Estimated sequences $\hat{X}_j^n$ for $j = 1, 2$ must satisfy distortion constraints defined by functions $d_j(x_1, x_2, y, \hat{x}_j)$: $\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y} \times \hat{\mathcal{X}}_j \rightarrow [0, D_{\max}]$ with $0 \leq D_{\max} < \infty$ for $j = 1, 2$, respectively. A formal description of the operations at encoder and decoder follows.

Definition 4.5. An $(n, R_{12}, R_{23}, D_1, D_2, \Gamma)$ code for the set-up of Figure 4.3 consists of two source encoders, namely

$$\mathrm{g}_1\colon\ \mathcal{X}_1^n \times \mathcal{X}_2^n \rightarrow [1, 2^{nR_{12}}], \tag{4.28}$$

which maps the sequences X_1^n and X_2^n into a message M_{12};

$$\mathrm{g}_2\colon\ \mathcal{X}_2^n \times [1, 2^{nR_{12}}] \rightarrow [1, 2^{nR_{23}}] \tag{4.29}$$

which maps the sequence X_2^n and message M_{12} into a message M_{23}; an "action" function

$$\ell\colon\ [1, 2^{nR_{23}}] \rightarrow \mathcal{A}^n, \tag{4.30}$$

which maps the message M_{23} into an action sequence A^n; a decoding function

$$\mathrm{h}_1\colon\ [1, 2^{nR_{12}}] \times \mathcal{X}_2^n \rightarrow \hat{\mathcal{X}}_1^n, \tag{4.31}$$

which maps the message M_{12} and the measured sequence X_2^n into the estimated sequence $\hat{X}_1^n$; and a sequence of decoding functions

$$\mathrm{h}_{2i}\colon\ [1, 2^{nR_{23}}] \times \mathcal{Y}^i \to \hat{\mathcal{X}}_2, \tag{4.32}$$

for $i \in [1, n]$ which maps the message M_{23} and the measured sequence Y^i into the ith estimated symbol $\hat{X}_{2i} = \mathrm{h}_{2i}(M_{23}, Y^i)$; such that the action cost constraint Γ and distortion constraints D_j for $j = 1, 2$ are satisfied, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\Lambda(A_i)\right] \leq \Gamma \tag{4.33}$$

$$\text{and } \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[d_j(X_{1i}, X_{2i}, Y_i, \hat{X}_{ji})\right] \leq D_j \text{ for } j = 1, 2, \tag{4.34}$$

respectively.

Definition 4.6. Given a distortion-cost tuple (D_1, D_2, Γ), a rate tuple (R_{12}, R_{23}) is said to be achievable if, for any $\epsilon > 0$, and sufficiently large n, there exists a $(n, R_{12}, R_{23}, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)$ code.

Definition 4.7. The *rate-distortion-cost region* $\mathcal{R}(D_1, D_2, \Gamma)$ is defined as the closure of all rate tuples (R_{12}, R_{23}) that are achievable given the distortion-cost tuple (D_1, D_2, Γ).

Remark 4.5. For side information Y independent of the action A given X_1 and X_2, i.e., for $p(y|a, x_1, x_2) = p(y|x_1, x_2)$, the rate-distortion region $\mathcal{R}(D_1, D_2, \Gamma)$ has been derived in [38].

4.3.2 Rate-Distortion-Cost Region

The following characterize the rate-distortion-cost region.

Proposition 4.5. *The rate-distortion-cost region $\mathcal{R}(D_1, D_2, \Gamma)$ for the set-up of Figure 4.3 is given by the union of all rate pairs (R_{12}, R_{23}) satisfying the inequalities*

$$R_{12} \geq I(X_1; U, A, \hat{X}_1 | X_2) \tag{4.35a}$$

$$\text{and } R_{23} \geq I(X_1, X_2; U, A), \tag{4.35b}$$

for some joint pmf that factorizes as

$$p(x_1, x_2, y, a, u, \hat{x}_1, \hat{x}_2) = p(x_1, x_2)p(a, u, \hat{x}_1 | x_1, x_2)p(y|a, x_1, x_2) \cdot \delta(\hat{x}_2 - \hat{\mathrm{x}}_2(u, y)), \tag{4.36}$$

with pmf $p(a, u, \hat{x}_1 | x_1, x_2)$ and deterministic function $\hat{\mathrm{x}}_2(u, y)$, such that the action and the distortion constraints

$$\mathrm{E}\left[\Lambda(A)\right] \leq \Gamma \tag{4.37}$$

$$\text{and } \mathrm{E}[d_j(X_1, X_2, Y, \hat{X}_j)] \leq D_j, \text{for } j = 1, 2, \tag{4.38}$$

respectively, hold. Finally, U is an auxiliary random variable whose alphabet cardinality can be constrained as $|\mathcal{U}| \leq |\mathcal{X}_1|\,|\mathcal{X}_2| + 4$, without loss of optimality.

Remark 4.6. If $p(y|a, x_1, x_2) = p(y|x_1, x_2)$, Proposition 4.5 reduces to [38, Theorem 1].

The proof of converse is provided in Appendix L. The coding strategy that proves achievability is a combination of the techniques proposed in [4] and [38, Theorem 1]. Here, the main ideas are briefly outlined, since the technical details follow from standard arguments. In the scheme at hand, Node 1 first maps sequences X_1^n and X_2^n into the action sequence A^n and an auxiliary codeword U^n using the standard joint typicality criterion. This mapping operation requires a codebook of rate $I(X_1, X_2; U, A)$ (see, e.g., [1, Chapter 3]). Then, given the so obtained sequences A^n and U^n, source sequences X_1^n and X_2^n are further mapped into the estimate

$\hat{X}_1^n$ for Node 2 so that the sequences $(X_1^n, X_2^n, A^n, U^n, \hat{X}_1^n)$ are jointly typical. This requires rate $I(X_1, X_2; \hat{X}_1|U, A)$ [1, Chapter 3]. Leveraging the side information X_2^n available at Node 2, conveying the codewords A^n, $\hat{X}_1^n$ and U^n to Node 2 requires rate $I(X_1, X_2; U, A) + I(X_1, X_2; \hat{X}_1|U, A) - I(U, A, \hat{X}_1; X_2)$ [1, Chapter 12], which equals the right-hand side of (4.35a). Node 2 conveys U^n and A^n to Node 3 by simply forwarding the index received from Node 1 (of rate $I(X_1, X_2; U, A)$). Finally, Node 3 estimates $\hat{X}_2^n$ through a symbol-by-symbol function as $\hat{X}_{2i} = \hat{\mathrm{x}}_2(U_i, Y_i)$ for $i \in [1, n]$.

4.4 Concluding Remarks

In the setting of source coding with a side information vending machine introduced in [4], the decoder can control the quality of the side information through a control, or action, sequence that is selected based on the message encoded by the source node. Since this message must also carry information directly related to the source to be reproduced at the decoder, a key aspect of the model is the interplay between encoding data and control information.

In this chapter, the original work [4] was generalized to two standard multi-terminal scenarios, namely distributed source coding and cascade source coding. For the former, inner bounds to the rate-distortion-cost regions are obtained for the cases with non-causal and causal side information at the decoder. These bounds have been found to be tight in two special cases. Some numerical examples have also been provided to shed some light on the advantages of an optimized trade-off between data and control transmission. As for the cascade source coding problem, a single-letter characterizations of achievable rate-distortion-cost trade-offs has been derived under the assumption of causal side information at the decoder.

A number of open problems have been left unsolved by this work, including the identification of more general conditions under which the inner bounds of Proposition 1 and Proposition 2 are tight. The technical challenges faced in this task are related to the well-known issues that arise when identifying auxiliary random variables that satisfy the desired Markov chain conditions in distributed source coding problems (see, e.g., [1, Chapter 13]).

CHAPTER 5

CASCADE SOURCE CODING WITH A SIDE INFORMATION "VENDING MACHINE"

5.1 Introduction

As reviewed in Chapter 4, the concept of a side information "vending machine" (VM) accounts for source coding scenarios in which acquiring the side information at the receiver entails some cost and thus should be done efficiently. In this class of models, the quality of the side information Y can be controlled at the decoder by selecting an action A that affects the effective channel between the source X and the side information Y through a conditional distribution $p_{Y|X,A}(y|x,a)$. Each action A is associated with a cost, and the problem is that of characterizing the available trade-offs between rate, distortion and action cost.

Extending the point-to-point set-up, cascade models provide baseline scenarios in which to study fundamental aspects of communication in multi-hop networks, which are central to the operation of, e.g., sensor or computer networks (see Figure 5.1). Standard information-theoretic models for cascade scenarios assume the availability of given side information sequences at the nodes (see e.g., [14]-[13]). In this chapter, instead, it is accounted for the cost of acquiring the side information by introducing side information VMs at an intermediate node and/ or at the final destination of a cascade model. As an example of the applications of interest, consider the computer network of Figure 5.1, where the intermediate and end nodes can obtain side information from remote data bases, but only at the cost of investing system resources such as time or bandwidth. Another example is a sensor network in which acquiring measurements entails an energy cost.

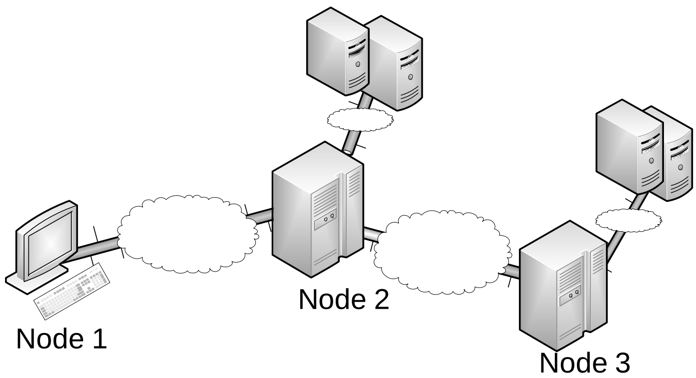

Figure 5.1 A multi-hop computer network in which intermediate and end nodes can access side information by interrogating remote data bases via cost-constrained actions.

As shown in Chapter 4 for a distributed system, the optimal operation of a VM at the decoder requires taking actions that are guided by the message received from the encoder. This implies the exchange of an explicit control signal embedded in the message communicated to the decoder that instructs the latter on how to operate the VM. Generalizing to the cascade models under study, a key issue to be tackled in this work is the design of communication strategies that strike the right balance between control signaling and source compression across the two hops.

The problem of characterizing the rate-distortion region for cascade source coding models, even with conventional side information sequences (i.e., without VMs as in Figure 5.2) at Node 2 and Node 3, is generally open. [14] and references therein review the state of the art on the cascade problem and [12] reviews that for the cascade-broadcast problem.

In this chapter, it is focused on the cascade source coding problem with side information VMs. The basic cascade source coding model consists of three nodes arranged so that Node 1 communicates with Node 2 and Node 2 to Node 3 over finite-rate links, as illustrated for a computer network scenario in Figure 5.1 and schematically in Figure 5.2-(a). Both Node 2 and Node 3 wish to reconstruct a, generally lossy, version of source X and have access to different side information sequences. An extension of the cascade model is the cascade-broadcast model of Figure 5.2-(b), in which an additional "broadcast" link of rate R_b exists that is received by both Node 2 and Node 3.

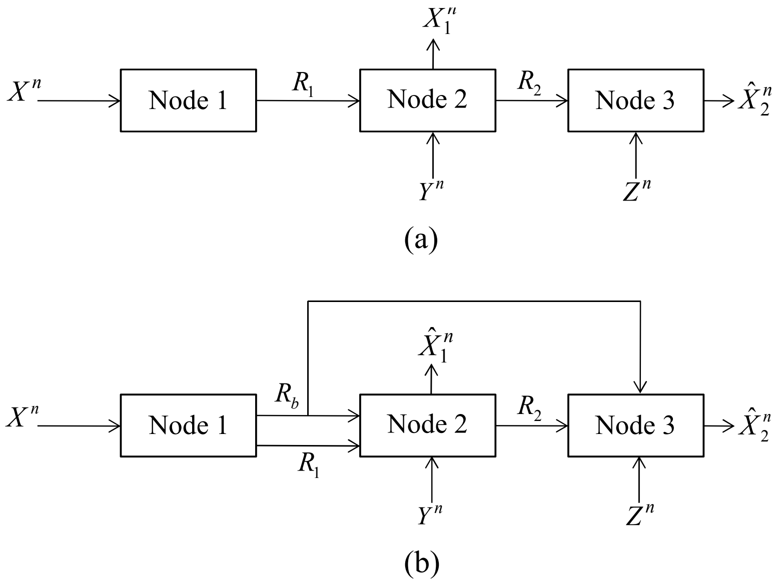

Figure 5.2 (a) Cascade source coding problem and (b) cascade-broadcast source coding problem.

Two specific instances of the models in Figure 5.2 for which a characterization of the rate-distortion performance has been found are the settings considered in [13] and that in [39], which it briefly reviewed here for their relevance to the present work. In [13], the cascade model in Figure 5.2(a) was considered for the special case in which the side information Y measured at Node 2 is also available at Node 1 (i.e., $X = (X, Y)$) and the Markov chain $X - Y - Z$ holds so that the side information at Node 3 is degraded with respect to that of Node 2. Instead, in [39], the cascade-broadcast model in Figure 5.2-(b) was considered for the special case in which either rate R_b or R_1 is zero, and the reconstructions at Node 1 and Node 2 are constrained to be retrievable also at the encoder in the sense of the Common Reconstruction (CR) introduced in [8] (see below for a rigorous definition).

5.1.1 Contributions

In this chapter, the source coding models of Figure 5.2 is investigated by assuming that some of the side information sequences can be affected by the actions taken by the corresponding nodes via VMs. The main contributions are as follows.

- *Cascade source coding problem with VM at Node 3* (Figure 5.3): In Section 5.2.2, the achievable rate-distortion-cost trade-offs are derived for the set-up in Figure 5.3, in which a side information VM exists at Node 3, while the side information Y is known at both Node 1 and Node 2 and satisfies the Markov chain X—Y—Z. This characterization extends the result of [13] discussed above to a model with a VM at Node 3. It should be mentioned that in [36], the rate-distortion-cost characterization for the model in Figure 5.3 was obtained, but under the assumption that the side information at Node 3 be available in a causal fashion in the sense of [16];

- *Cascade-broadcast source coding problem with VM at Node 2 and Node 3, lossless compression* (Figure 5.4): In Section 5.3.2, the cascade-broadcast model in Figure 5.4 is studied in which a VM exists at both Node 2 and Node 3. In order to enable the action to be taken by both Node 2 and Node 3, it is assumed that the information about which action should be taken by Node 2 and Node 3 is sent by Node 1 on the broadcast link of rate R_b. Under the constraint of lossless reconstruction at Node 2 and Node 3, a characterization of the rate-cost performance is obtained. This conclusion generalizes the result in [23] discussed above to the case in which the rate R_1 and/or R_2 are non-zero;

- *Cascade-broadcast source coding problem with VM at Node 2 and Node 3, lossy compression with CR constraint* (Figure 5.4): In Section 5.3.4, the problem in Figure 5.4 is tackled but under the more general requirement of lossy reconstruction. Conclusive results are obtained under the additional constraints that the side information at Node 3 is degraded and that the source reconstructions at Node 2 and Node 3 can be recovered with arbitrarily small error probability at Node 1. This is referred to as the CR constraint following [8], and is of relevance in applications in which the data being sent is of sensitive nature and unknown distortions in the receivers' reconstructions are not acceptable (see [8] for further discussion). This characterization extends the result of [39] mentioned above to the set-up with a side information VM, and also in that both rates R_1 and R_b are allowed to be non-zero;

- *Adaptive actions*: Finally, the results above are revisited by allowing the decoders to select their actions in an adaptive way, based not only on the received messages but also on the previous samples of the side information extending [26]. Note that the effect of adaptive actions on rate–distortion–cost region was open even for simple point-to-point communication channel with decoder side non-causal side information VM until recently, when [26] has shown that adaptive action does not decrease the rate–distortion–cost region of point-to-point system. In this chapter, this result is extended to the multi-terminal framework and it is conclude that, in all of the considered examples, where applicable, adaptive selection of the actions does not improve the achievable rate-distortion-cost trade-offs.

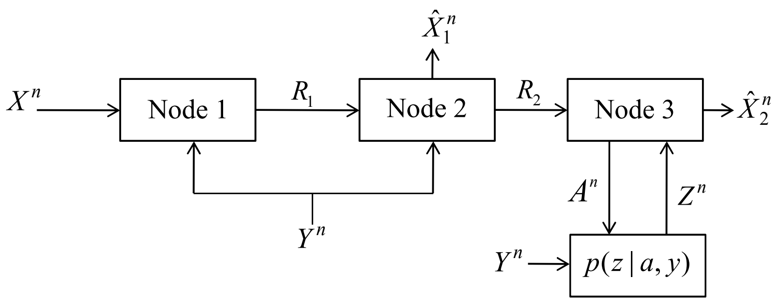

Figure 5.3 Cascade source coding problem with a side information "vending machine" at Node 3.

Our results extends to multi-hop scenarios the conclusion in [4] that a joint representation of data and control messages enables an efficient use of the available communication links. In particular, layered coding strategies prove to be optimal for all the considered models, in which, the base layer fulfills two objectives: determining the actions of downstream nodes and simultaneously providing a coarse description of the source. Moreover, the examples provided in this chapter demonstrate the dependence of the optimal coding design on network topology action costs.

5.2 Cascade Source Coding with A Side information Vending Machine

In this section, the system model for the cascade source coding problem with a side information vending machine of Figure 5.3 is described. Then the characterization of the corresponding rate-distortion-cost performance is presented in Section 5.2.2.

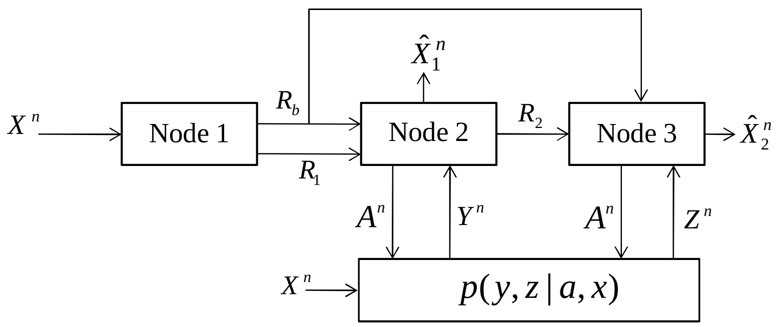

Figure 5.4 Cascade source coding problem with a side information "vending machine" at Node 2 and Node 3.

5.2.1 System Model

The problem of cascade source coding of Figure 5.3, is defined by the probability mass functions (pmfs) $p_{XY}(x,y)$ and $p_{Z|AY}(z|a,y)$ and discrete alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{A}, \hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2$, as follows. The source sequences X^n and Y^n with $X^n \in \mathcal{X}^n$ and $Y^n \in \mathcal{Y}^n$, respectively, are such that the pairs (X_i, Y_i) for $i \in [1, n]$ are independent and identically distributed (i.i.d.) with joint pmf $p_{XY}(x,y)$. Node 1 measures sequences X^n and Y^n and encodes them in a message M_1 of nR_1 bits, which is delivered to Node 2. Node 2 estimates a sequence $\hat{X}_1^n \in \hat{\mathcal{X}}_1^n$ within given distortion requirements to be discussed below. Moreover, Node 2 maps the message M_1 received from Node 1 and the locally available sequence Y^n in a message M_2 of nR_2 bits, which is delivered to Node 3. Node 3 wishes to estimate a sequence $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ within given distortion requirements. To this end, Node 3 receives message M_2 and based on this, it selects an action sequence A^n, where $A^n \in \mathcal{A}^n$. The action sequence affects the quality of the measurement Z^n of sequence Y^n obtained at the Node 3. Specifically, given A^n and Y^n, the sequence Z^n is distributed as $p(z^n|a^n, y^n) = \prod_{i=1}^n p_{Z|A,Y}(z_i|y_i, a_i)$. The cost of the action sequence is defined by a cost function Λ: $\mathcal{A} \to [0, \Lambda_{\max}]$ with $0 \leq \Lambda_{\max} < \infty$, as $\Lambda(a^n) = \sum_{i=1}^n \Lambda(a_i)$. The estimated sequence $\hat{X}_2^n$ with $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ is then obtained as a function of M_2 and Z^n. The estimated sequences $\hat{X}_j^n$ for $j = 1, 2$ must satisfy distortion constraints defined by functions $d_j(x, \hat{x}_j)$: $\mathcal{X} \times \hat{\mathcal{X}}_j \to [0, D_{\max}]$ with $0 \leq D_{\max} < \infty$ for $j = 1, 2$, respectively. A formal description of the operations at the encoder and the decoder follows.

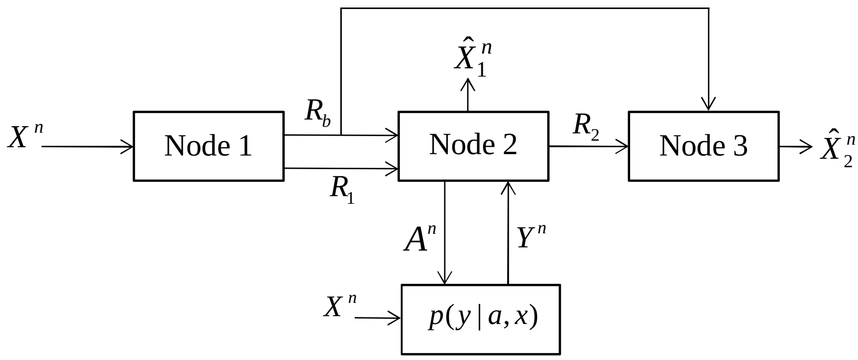

Figure 5.5 Cascade-broadcast source coding problem with a side information "vending machine" at Node 2.

Definition 5.1. An $(n, R_1, R_2, D_1, D_2, \Gamma, \epsilon)$ code for the set-up of Figure 5.3 consists of two source encoders, namely

$$\mathrm{g}_1\colon\ \mathcal{X}^n \times \mathcal{Y}^n \to [1, 2^{nR_1}], \tag{5.1}$$

which maps the sequences X^n and Y^n into a message M_1;

$$\mathrm{g}_2\colon \mathcal{Y}^n \times [1, 2^{nR_1}] \to [1, 2^{nR_2}], \tag{5.2}$$

which maps the sequence Y^n and message M_1 into a message M_2; an "action" function

$$\ell\colon\ [1, 2^{nR_2}] \to \mathcal{A}^n, \tag{5.3}$$

which maps the message M_2 into an action sequence A^n; two decoders, namely

$$\mathrm{h}_1\colon\ [1, 2^{nR_1}] \times \mathcal{Y}^n \to \hat{\mathcal{X}}_1^n, \tag{5.4}$$

which maps the message M_1 and the measured sequence Y^n into the estimated sequence $\hat{X}_1^n$;

$$\mathrm{h}_2\colon\ [1, 2^{nR_2}] \times \mathcal{Z}^n \to \hat{\mathcal{X}}_2^n, \tag{5.5}$$

which maps the message M_2 and the measured sequence Z^n into the the estimated sequence $\hat{X}_2^n$; such that the action cost constraint Γ and distortion constraints D_j for $j = 1, 2$ are satisfied, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\Lambda(A_i)\right] \le \Gamma \tag{5.6}$$

$$\text{and } \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d_j(X_{ji}, \mathrm{h}_{ji})\right] \le D_j \text{ for } j = 1, 2, \tag{5.7}$$

where h_{1i} and h_{2i} are the ith symbol of the function $\mathrm{h}_1(M_1, Y^n)$ and $\mathrm{h}_2(M_2, Z^n)$, respectively.

Definition 5.2. Given a distortion-cost tuple (D_1, D_2, Γ), a rate tuple (R_1, R_2) is said to be achievable if, for any $\epsilon > 0$, and sufficiently large n, there exists a $(n, R_1, R_2, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)$ code.

Definition 5.3. The *rate-distortion-cost region* $\mathcal{R}(D_1, D_2, \Gamma)$ is defined as the closure of all rate tuples (R_1, R_2) that are achievable given the distortion-cost tuple (D_1, D_2, Γ).

Remark 5.1. For side information Z available causally at Node 3, i.e., with decoding function (8.4) at Node 3 modified so that $\hat{X}_i$ is a function of M_2 and Z^i only, the rate-distortion region $\mathcal{R}(D_1, D_2, \Gamma)$ has been derived in [36].

5.2.2 Rate-Distortion-Cost Region

In this section, a single-letter characterization of the rate-distortion-cost region is derived.

Proposition 5.1. *The rate-distortion-cost region $\mathcal{R}(D_1, D_2, \Gamma)$ for the cascade source coding problem illustrated in Figure 5.3 is given by the union of all rate pairs (R_1, R_2) that satisfy the conditions*

$$\begin{aligned} R_1 &\geq I(X; \hat{X}_1, A, U|Y) && (5.8a) \\ \text{and } R_2 &\geq I(X, Y; A) + I(X, Y; U|A, Z), && (5.8b) \end{aligned}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x, y, z, a, \hat{x}_1, u) = p(x, y)p(\hat{x}_1, a, u|x, y)p(z|y, a), \tag{5.9}$$

for some pmf $p(\hat{x}_1, a, u|x, y)$ such that the inequalities

$$\begin{aligned} \mathrm{E}[d_1(X, \hat{X}_1)] &\leq D_1, && (5.10a) \\ \mathrm{E}[d_2(X, \mathrm{f}(U, Z))] &\leq D_2, && (5.10b) \\ \text{and } \mathrm{E}[\Lambda(A)] &\leq \Gamma, && (5.10c) \end{aligned}$$

are satisfied for some function f*:* $\mathcal{U} \times \mathcal{Z} \rightarrow \hat{\mathcal{X}}_2$*. Finally,* U *is an auxiliary random variable whose alphabet cardinality can be constrained as* $|\mathcal{U}| \leq |\mathcal{X}||\mathcal{Y}||\mathcal{A}|+3$*, without loss of optimality.*

Remark 5.2. For side information Z independent of the action A given Y, i.e., for $p(z|a,y) = p(z|y)$, the rate-distortion region $\mathcal{R}(D_1, D_2, \Gamma)$ in Proposition 5.1 reduces to that derived in [13].

The proof of the converse is provided in Appendix M for a more general case of adaptive action to be defined in Sec 5.4. The achievability follows as a combination of the techniques proposed in [4] and [13, Theorem 1]. Here, the main ideas are briefly outlined, since the technical details follow from standard arguments. For the scheme at hand, Node 1 first maps sequences X^n and Y^n into the action sequence A^n using the standard joint typicality criterion. This mapping requires a codebook of rate $I(X,Y;A)$ (see, e.g., [1, pp. 62-63]). Given the sequence A^n, the sequences X^n and Y^n are further mapped into a sequence U^n. This requires a codebook of size $I(X,Y;U|A)$ for each action sequence A^n from standard rate-distortion considerations [1, pp. 62-63]. Similarly, given the sequences A^n and U^n, the sequences X^n and Y^n are further mapped into the estimate $\hat{X}_1^n$ for Node 2 using a codebook of rate $I(X,Y;\hat{X}_1|U,A)$ for each codeword pair (U^n, A^n). The thus obtained codewords are then communicated to Node 2 and Node 3 as follows. By leveraging the side information Y^n available at Node 2, conveying the codewords A^n, U^n and $\hat{X}_1^n$ to Node 2 requires rate $I(X,Y;U,A) + I(X,Y;\hat{X}_1|U,A) - I(U,A,\hat{X}_1;Y)$ by the Wyner-Ziv theorem [1, p. 280], which equals the right-hand side of (8.39a). Then, sequences A^n and U^n are sent by Node 2 to Node 3, which requires a rate equal to the right-hand side of (6.9b). This follows from the rates of the used codebooks and from the Wyner-Ziv theorem, due to the side information Z^n available at Node 3 upon application of the action sequence A^n. Finally, Node 3 produces $\hat{X}_2^n$ that leverages through a symbol-by-symbol function as $\hat{X}_{2i} = \mathrm{f}(U_i, Z_i)$ for $i \in [1, n]$.

5.2.3 Lossless Compression

Suppose that the source sequence X^n needs to be communicated *losslessly* at both Node 2 and Node 3, in the sense that $d_j(x, \hat{x}_j)$ is the Hamming distortion measure for $j = 1, 2$ ($d_j(x, \hat{x}_j) = 0$ if $x = \hat{x}_j$ and $d_j(x, \hat{x}_j) = 1$ if $x \neq \hat{x}_j$) and $D_1 = D_2 = 0$. The following immediate consequence of Proposition 5.1 is established.

Corollary 5.1. *The rate-distortion-cost region $\mathcal{R}(0, 0, \Gamma)$ for the cascade source coding problem illustrated in Figure 5.3 with Hamming distortion metrics is given by the union of all rate pairs (R_1, R_2) that satisfy the conditions*

$$R_1 \geq I(X; A|Y) + H(X|A, Y) \tag{5.11a}$$

$$\text{and } R_2 \geq I(X, Y; A) + H(X|A, Z), \tag{5.11b}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x, y, z, a) = p(x, y)p(a|x, y)p(z|y, a), \tag{5.12}$$

for some pmf $p(a|x, y)$ such that $\mathrm{E}[\Lambda(A)] \leq \Gamma$.

5.3 Cascade-Broadcast Source Coding with A Side Information Vending Machine

In this section, the cascade-broadcast source coding problem with a side information vending machine illustrated in Figure 5.4 is studied. At first, the rate-cost performance is characterized for the special case in which the reproductions at Node 2 and Node 3 are constrained to be lossless. Then, the lossy version of the problem is considered in Section 5.3.4, with an additional common reconstruction requirement in the sense of [8] and assuming degradedness of the side information sequences.

5.3.1 System Model

In this section, the general system model for the cascade-broadcast source coding problem with a side information vending machine is described. It is emphasized that, unlike the setup of Figure 5.3, here, the vending machine is at both Node 2 and Node 3. Moreover, it is assumed that an additional broadcast link of rate R_b is available that is received by Node 2 and 3 to enable both Node 2 and Node 3 so as to take concerted actions in order to affect the side information sequences. It is also assumed the action sequence taken by Node 2 and Node 3 to be a function of only the broadcast message M_b sent over the broadcast link of rate R_b.

The problem is defined by the pmfs $p_X(x)$, $p_{YZ|AX}(y,z|a,x)$ and discrete alphabets $\mathcal{X},\mathcal{Y},\mathcal{Z},\mathcal{A}$, $\hat{\mathcal{X}}_1,\hat{\mathcal{X}}_2$, as follows. The source sequence X^n with $X^n \in \mathcal{X}^n$ is i.i.d. with pmf $p_X(x)$. Node 1 measures sequence X^n and encodes it into messages M_1 and M_b of nR_1 and nR_b bits, respectively, which are delivered to Node 2. Moreover, message M_b is broadcast also to Node 3. Node 2 estimates a sequence $\hat{X}_1^n \in \hat{\mathcal{X}}_1^n$ and Node 3 estimates a sequence $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$. To this end, Node 2 receives messages M_1 and M_b and, based only on the latter message, it selects an action sequence A^n, where $A^n \in \mathcal{A}^n$. Node 2 maps messages M_1 and M_b, received from Node 1, and the locally available sequence Y^n in a message M_2 of nR_2 bits, which is delivered to Node 3. Node 3 receives messages M_2 and M_b and based only on the latter message, it selects an action sequence A^n, where $A^n \in \mathcal{A}^n$. Given A^n and X^n, the sequences Y^n and Z^n are distributed as $p(y^n,z^n|a^n,x^n) = \prod_{i=1}^n p_{YZ|A,X}(y_i,z_i|a_i,x_i)$. The cost of the action sequence is defined as in previous section. A formal description of the operations at encoder and decoder follows.

Definition 5.4. An $(n, R_1, R_2, R_b, D_1, D_2, \Gamma, \epsilon)$ code for the set-up of Figure 5.5 consists of two source encoders, namely

$$\mathrm{g}_1\colon\ \mathcal{X}^n \rightarrow [1, 2^{nR_1}] \times [1, 2^{nR_b}], \tag{5.13}$$

which maps the sequence X^n into messages M_1 and M_b, respectively;

$$\mathrm{g}_2\colon [1, 2^{nR_1}] \times [1, 2^{nR_b}] \times \mathcal{Y}^n \to [1, 2^{nR_2}] \tag{5.14}$$

which maps the sequence Y^n and messages (M_1, M_b) into a message M_2; an "action" function

$$\ell\colon [1, 2^{nR_b}] \to \mathcal{A}^n, \tag{5.15}$$

which maps the message M_b into an action sequence A^n; two decoders, namely

$$\mathrm{h}_1\colon \;\; [1, 2^{nR_1}] \times [1, 2^{nR_b}] \times \mathcal{Y}^n \to \hat{\mathcal{X}}_1^n, \tag{5.16}$$

which maps messages M_1 and M_b and the measured sequence Y^n into the estimated sequence $\hat{X}_1^n$; and

$$\mathrm{h}_2\colon [1, 2^{nR_2}] \times [1, 2^{nR_b}] \times \mathcal{Z}^n \to \hat{\mathcal{X}}_2^n, \tag{5.17}$$

which maps the messages M_2 and M_b into the the estimated sequence $\hat{X}_2^n$; such that the action cost constraint (8.5) and distortion constraint (8.6) are satisfied.

Achievable rates (R_1, R_2, R_b) and rate-distortion-cost region are defined analogously to Definitions 8.2 and 8.3.

The rate–distortion–cost region for the system model described above is open even for the case without VM at Node 2 and Node 3 (see [12]). Hence, in the following subsections, the rate region is characterized for a few special cases. As in the previous section, subscripts are dropped from the pmf for simplicity of notation.

5.3.2 Lossless Compression

In this section, a single-letter characterization of the rate-cost region $\mathcal{R}(0, 0, \Gamma)$ is derived for the special case in which the distortion metrics are assumed to be Hamming and the distortion constraints are $D_1 = 0$ and $D_2 = 0$.

Proposition 5.2. *The rate-cost region* $\mathcal{R}(0,0,\Gamma)$ *for the cascade-broadcast source coding problem illustrated in Figure 5.4 with Hamming distortion metrics is given by the union of all rate triples* (R_1, R_2, R_b) *that satisfy the conditions*

$$R_b \geq I(X;A) \tag{5.18a}$$

$$R_1 + R_b \geq I(X;A) + H(X|A,Y) \tag{5.18b}$$

$$\text{and } R_2 + R_b \geq I(X;A) + H(X|A,Z) \tag{5.18c}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x,y,z,a) = p(x,a)p(y,z|a,x), \tag{5.19}$$

for some pmf $p(a|x)$ *such that* $E[\Lambda(A)] \leq \Gamma$.

Remark 5.3. If $R_1 = 0$ and $R_2 = 0$, the rate-cost region $\mathcal{R}(\Gamma)$ of Proposition 5.2 reduces to the one derived in [23, Theorem 1].

Remark 5.4. The rate region (5.18) also describes the rate-distortion region under the more restrictive requirement of lossless reconstruction in the sense of the probabilities of error $\Pr[X^n \neq \hat{X}_j^n] \leq \epsilon$ for $j = 1, 2$, as it follows from standard arguments (see [1, Section 3.6.4]). A similar conclusion applies for Corollary 5.1.

The converse proof for bound (5.18a) follows immediately since A^n is selected only as a function of message M_b. As for the other two bounds, namely (5.18b)-(5.18c), the proof of the converse can be established following cut-set arguments and using the point-to-point result of [4]. For achievability, the code structure proposed in [4] is used along with rate splitting. Specifically, Node 1 first maps sequence X^n into the action sequence A^n. This mapping requires a codebook of rate $I(X;A)$. This rate has to be conveyed over link R_b by the definition of the problem and is thus received by both Node 2 and Node 3. Given the so obtained sequence A^n, communicating X losslessly to Node 2 requires rate $H(X|A,Y)$. This rate is split into two rates r_{1b} and

r_{1d}, such that the message corresponding to the first rate is carried over the broadcast link of rate R_b and the second on the direct link of rate R_1. Note that Node 2 can thus recover sequence X losslessly. The rate $H(X|A,Z)$ which is required to send X losslessly to Node 3, is then split into two parts, of rates r_{2b} and r_{2d}. The message corresponding to the rate r_{2b} is sent to Node 3 on the broadcast link of the rate R_b by Node 1, while the message of rate r_{2d} is sent by Node 2 to Node 3. This way, Node 1 and Node 2 cooperate to transmit X to Node 3. As per the discussion above, the following inequalities have to be satisfied

$$\begin{aligned} r_{2b}+r_{2d}+r_{1b} &\geq H(X|A,Z), \\ r_{1b}+r_{1d} &\geq H(X|A,Y), \\ R_1 &\geq r_{1d}, \\ R_2 &\geq r_{2d}, \\ \text{and } R_b &\geq r_{1b}+r_{2b}+I(X;A), \end{aligned}$$

Applying Fourier-Motzkin elimination [1, Appendix C] to the inequalities above, the inequalities in (5.18) are obtained.

5.3.3 Example: Switching-Dependent Side Information

In this section, a special case of the model in Figure 5.4 is considered in which the actions $A \in \mathcal{A} = \{0,1,2,3\}$ acts a switch that decides whether Node 2, Node 3 or either node gets to observe a side information W. The side information W is jointly distributed with source X according to the joint pmf $p(x,w)$. Moreover, defining as e an "erasure" symbol, the conditional pmf $p(y,z|x,a)$ is as follows: $Y = Z = \mathrm{e}$ for $A = 0$ (neither Node 2 nor Node 3 observes the side information W); $Y = W$ and $Z = \mathrm{e}$ for $A = 1$ (only Node 2 observes the side information W); $Y = \mathrm{e}$ and $Z = W$ for $A = 2$ (only Node 3 observes the side information W); and $Y = Z = W$ for $A = 3$

(both nodes observe the side information W)[1]. The cost function is selected such that $\Lambda(j) = \lambda_j$ for $j \in \mathcal{A}$. When $R_1 = R_2 = 0$, this model reduces to the ones studied in [23, Section III]. The following is a consequence of Proposition 2.

Corollary 5.2. *For the setting of switching-dependent side information described above, the rate-cost region (5.18) is given by*

$$R_b \geq I(X;A) \tag{5.20a}$$

$$R_1 + R_b \geq H(X) - p_1 I(X;W|A=1) - p_3 I(X;W|A=3) \tag{5.20b}$$

$$\text{and } R_2 + R_b \geq H(X) - p_2 I(X;W|A=2) - p_3 I(X;W|A=3) \tag{5.20c}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x,y,z,a) = p(x,a)p(y,z|a,x), \tag{5.21}$$

for some pmf $p(a|x)$ *such that* $\sum_{j=0}^{3} p_j \lambda_j \leq \Gamma$*, where* $p_j = \Pr[A=j]$ *for* $j \in \mathcal{A}$*.*

Proof. The region (5.20) is obtained from the rate-cost region (5.18) by noting that in (5.18b), $I(X;A)+H(X|A,Y) = H(X)-I(X;Y|A)$ holds and similarly for (5.18c). □

In the following, two specific instances of the switching-dependent side information example are evaluated.

Binary Symmetric Channel (BSC) between X and W: Let (X,W) be binary and symmetric so that $p(x) = p(w) = 1/2$ for $x, w \in \{0,1\}$ and $\Pr[X \neq W] = \delta$ for $\delta \in [0,1]$. Moreover, let $\lambda_j = \infty$ for $j = 0,3$ and $\lambda_j = 1$ otherwise. The action cost constraint is set to $\Gamma = 1$. Note that, given this definition of $\Lambda(a)$, at each time, Node 1 can choose whether to provide the side information W to Node 2 *or* to Node 3 with no further constraints. By symmetry, the pmf $p(a|x)$ with $x \in \{0,1\}$ and $a \in \{1,2\}$ is

[1]This implies that $p(y,z|x,a) = \sum_w p(w|x)\delta(y-w)\delta(z-\mathrm{e})$ for $a = 1$ and similarly for other values of a.

set to be a BSC with transition probability q. This implies that $p_1 = \Pr[A = 1] = q$ and $p_2 = \Pr[A = 2] = 1 - q$. Let's now evaluate the inequality (5.20a) as $R_b \geq 0$; inequality (5.20b) as $R_1 + R_b \geq 1 - p_1 I(X;W|A = 1) = 1 - qH(\delta)$; and similarly inequality (5.18c) as $R_2 + R_b \geq 1 - (1 - q)H(\delta)$. From these inequalities, it can be seen that, in order to trace the boundary of the rate-cost region, in general, one needs to consider all values of q in the interval $[0, 1]$. This corresponds to appropriate

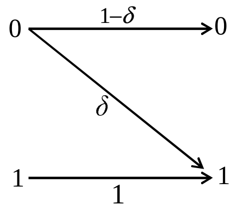

Figure 5.6 The side information S-channel $p(w|x)$ used in the example of Section 5.3.3.

time-sharing between providing side information to Node 2 (for a fraction of time q) and Node 3 (for the remaining fraction of time). Note that, as shown in [23, Section III], if $R_1 = R_2 = 0$, it is optimal to set $q = \frac{1}{2}$, and thus equally share the side information between Node 2 and Node 3, in order to minimize the rate R_b. This difference is due to the fact that in the cascade model at hand, it can be advantageous to provide more side information to one of the two encoders depending on the desired trade-off between the rates R_1 and R_2 in the achievable rate-cost region.

S-*Channel between X and W*: Let's now consider the special case of Corollary 5.2 in which (X, W) are jointly distributed so that $p(x) = 1/2$ and $p(w|x)$ is the S-channel characterized by $p(0|0) = 1 - \delta$ and $p(1|1) = 1$ (see Figure 5.6). Moreover, let $\lambda_1 = 1$, $\lambda_2 = 0$, $\lambda_0 = \lambda_3 = \infty$ as above, while the cost constraint is set to $\Gamma \leq 1$. As discussed in [23, Section III] for this example with $R_1 = R_2 = 0$, providing side information to Node 2 is more costly and thus should be done efficiently. In particular, given Figure 5.6, it is expected that biasing the choice $A = 2$ when $X = 1$ (i.e., providing side information to Node 2) may lead to some gain (see [23]). Here it is shown that in the cascade model, this gain depends on the relative importance of rates R_1 and R_2.

To this end, $p(a|x)$ is set as $p(1|0) = \alpha$ and $p(1|1) = \beta$ for $\alpha, \beta \in [0, 1]$. Now, the inequality (5.20a) is evaluated as $R_b \geq 0$; inequality (5.20b) as

$$R_1 + R_b \geq 1 - \Big(\frac{\alpha+\beta}{2}\Big)\Big(H\Big(\frac{(1-\delta)\alpha}{\alpha+\beta}\Big) - H(1-\delta)\frac{\alpha}{\alpha+\beta}\Big); \tag{5.22}$$

and inequality (5.20c) as

$$R_2 + R_b \geq 1 - \Big(\frac{2-\alpha-\beta}{2}\Big)\Big(H\Big(\frac{(1-\delta)(1-\alpha)}{2-\alpha-\beta}\Big) - H(1-\delta)\frac{1-\alpha}{2-\alpha-\beta}\Big) \tag{5.23}$$

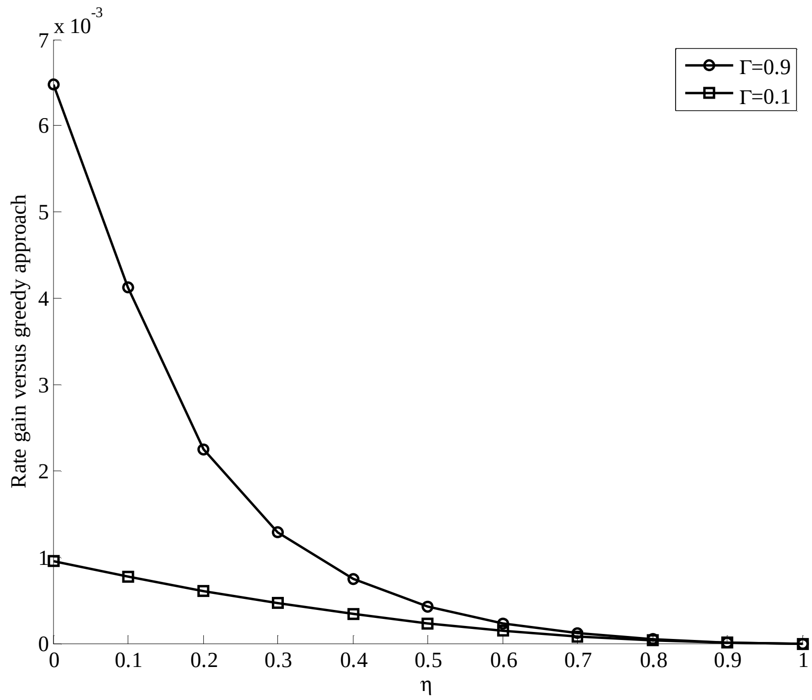

Figure 5.7 Difference between the weighted sum-rate $R_1 + \eta R_2$ obtained with the greedy and with the optimal strategy as per Corollary 5.2 ($R_b = 0.4$, $\delta = 0.6$).

Now the minimum weighted sum-rate $R_1 + \eta R_2$ obtained from (5.22)-(5.23) is evaluated for $R_b = 0.4$, $\delta = 0.6$ and both $\Gamma = 0.1$ and $\Gamma = 0.9$. Parameter $\eta \geq 0$ rules on the relative importance of the two rates. For comparison, the performance attainable by imposing that the action A be selected independent of X, referred to as the greedy approach [4], is computed. Figure 5.7 plots the difference between the two weighted sum-rates $R_1+\eta R_2$. It can be seen that, as η decreases and thus minimizing rate R_1 to Node 2 becomes more important, one can achieve larger gains by choosing the action A to be dependent on X. Moreover, this gain is more significant when the action cost budget Γ allows Node 2 to collect a larger fraction of the side information samples.

5.3.4 Lossy Compression with Common Reconstruction Constraint

In this section, the problem of characterizing the rate-distortion-cost region $\mathcal{R}(D_1, D_2, \Gamma)$ for $D_1, D_2 > 0$ is considered. In order to make the problem tractable [2], the degradedness condition $X - (A, Y) - Z$ (as in [23]) is imposed, which implies the factorization

$$p(y, z|a, x) \quad = \quad p(y|a, x)p(z|y, a); \tag{5.24}$$

and that the reconstructions at Nodes 2 and 3 be reproducible by Node 1. As discussed, this latter condition is referred to as the CR constraint [8]. Note that this constraint is automatically satisfied in the lossless case. To be more specific, an $(n, R_1, R_2, R_b, D_1, D_2, \Gamma, \epsilon)$ code is defined per Definition 5.4 with the difference that there are two additional functions for the encoder, namely

$$\psi_1\colon\ \mathcal{X}^n \to \hat{\mathcal{X}}_1^n \tag{5.25a}$$

$$\text{and } \psi_2\colon\ \mathcal{X}^n \to \hat{\mathcal{X}}_2^n, \tag{5.25b}$$

[2] As noted earlier, the problem is open even in the case with no VM [12].

which map the source sequence into the estimated sequences at the encoder, namely $\psi_1(X^n)$ and $\psi_2(X^n)$, respectively; and the CR requirements are imposed, i.e.,

$$\Pr\left[\psi_1(X^n) \neq \mathrm{h}_1(M_1, M_b, Y^n)\right] \leq \epsilon \tag{5.26a}$$

$$\text{and } \Pr\left[\psi_2(X^n) \neq \mathrm{h}_2(M_2, M_b, Z^n)\right] \leq \epsilon, \tag{5.26b}$$

so that the encoder's estimates $\psi_1(\cdot)$ and $\psi_2(\cdot)$ are equal to the decoders' estimates (cf. (5.16)-(5.17)) with high probability.

Proposition 5.3. *The rate-distortion region $\mathcal{R}(D_1, D_2, \Gamma)$ for the cascade-broadcast source coding problem illustrated in Figure 5.4 under the CR constraint and the degradedness condition (7.24) is given by the union of all rate triples (R_1, R_2, R_b) that satisfy the conditions*

$$R_b \geq I(X; A) \tag{5.27a}$$

$$R_1 + R_b \geq I(X; A) + I(X; \hat{X}_1, \hat{X}_2 | A, Y) \tag{5.27b}$$

$$R_2 + R_b \geq I(X; A) + I(X; \hat{X}_2 | A, Z) \tag{5.27c}$$

$$\textit{and } R_1 + R_2 + R_b \geq I(X; A) + I(X; \hat{X}_2 | A, Z) + I(X; \hat{X}_1 | A, Y, \hat{X}_2), \tag{5.27d}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x, y, z, a, \hat{x}_1, \hat{x}_2) = p(x)p(a|x)p(y|x, a)p(z|a, y)p(\hat{x}_1, \hat{x}_2|x, a), \tag{5.28}$$

such that the inequalities

$$\mathrm{E}[d_j(X, \hat{X}_j)] \leq D_j, \ \textit{for } j = 1, 2, \tag{5.29a}$$

$$\textit{and } \mathrm{E}[\Lambda(A)] \leq \Gamma, \tag{5.29b}$$

are satisfied.

Remark 5.5. If either $R_1 = 0$ or $R_b = 0$ and the side information Y is independent of the action A given X, i.e., $p(y|a,x) = p(y|x)$, the rate-distortion region $\mathcal{R}(D_1, D_2, \Gamma)$ of Proposition 5.3 reduces to the one derived in [39, Proposition 10].

The proof of the converse is provided in Appendix N. The achievability follows similar to Proposition 5.2. Specifically, Node 1 first maps sequence X^n into the action sequence A^n. This mapping requires a codebook of rate $I(X;A)$. This rate has to be conveyed over link R_b by the definition of the problem and is thus received by both Node 2 and Node 3. The source sequence X^n is mapped into the estimate $\hat{X}_2^n$ for Node 3 using a codebook of rate $I(X;\hat{X}_2|A)$ for each sequence A^n. Communicating $\hat{X}_2^n$ to Node 2 requires rate $I(X;\hat{X}_2|A,Y)$ by the Wyner-Ziv theorem. This rate is split into two rates r_{2b} and r_{2d}, such that the message corresponding to the first rate is carried over the broadcast link of rate R_b and the second on the direct link of rate R_1. Note that Node 2 can thus recover sequence $\hat{X}_2^n$. Communicating $\hat{X}_2^n$ to Node 3 requires rate $I(X;\hat{X}_2|A,Z)$ by the Wyner-Ziv theorem. This rate is split into two rates r_{0b} and r_{0d}. The message corresponding to the rate r_{0b} is send to Node 3 on the broadcast link of the rate R_b by Node 1, while the message of rate r_{0d} is sent by Node 2 to Node 3. This way, Node 1 and Node 2 cooperate to transmit $\hat{X}_2$ to Node 3. Finally, the source sequence X^n is mapped by Node 1 into the estimate $\hat{X}_1^n$ for Node 2 using a codebook of rate $I(X;\hat{X}_1|A,\hat{X}_2)$ for each pair of sequences $(A^n, \hat{X}_2^n)$. Using the Wyner-Ziv coding, this rate is reduced to $I(X;\hat{X}_1|A,Y,\hat{X}_2)$ and split into two rates r_{1b} and r_{1d}, which are sent through links R_b and R_1, respectively. As per discussion above, the following inequalities have to be satisfied

$$\begin{aligned}
r_{0b} + r_{0d} + r_{2b} &\geq I(X;\hat{X}_2|A,Z), \\
r_{2b} + r_{2d} &\geq I(X;\hat{X}_2|A,Y),
\end{aligned}$$

$$
\begin{aligned}
r_{1b} + r_{1d} &\geq I(X;\hat{X}_1|A,Y,\hat{X}_2), \\
R_1 &\geq r_{1d} + r_{2d}, \\
R_2 &\geq r_{0d}, \\
\text{and } R_b &\geq r_{1b} + r_{2b} + r_{0b} + I(X;A),
\end{aligned}
$$

Applying Fourier-Motzkin elimination [1, Appendix C] to the inequalities above, the inequalities in (5.27) are obtained.

5.4 Adaptive Actions

In this section, it is assume that actions taken by the nodes are not only a function of the message M_2 for the model of Figure 5.3 or M_b for the models of Figure 5.4 and Figure 5.5, respectively, but also a function of the past observed side information samples. Following [26], this case is referred to as the one with *adaptive actions*. Note that for the cascade-broadcast problem, the model in Figure 5.5 is considered, which differs from the one in Figure 5.4 considered thus far in that the side information Z is not available at Node 3. At this time, it appears to be problematic to define adaptive actions in the presence of two nodes that observe different side information sequences. For the cascade model in Figure 5.3, a $(n, R_1, R_2, D_1, D_2, \Gamma)$ code is defined per Definition 8.1 with the difference that the action encoder (8.2) is modified to be

$$\ell\colon\ [1, 2^{nR_2}] \times \mathcal{Z}^{i-1} \to \mathcal{A}, \tag{5.30}$$

which maps the message M_2 and the past observed decoder side information sequence Z^{i-1} into the ith symbol of the action sequence A_i. Moreover, for the cascade-broadcast model of Figure 5.5, the "action" function (5.15) in Definition 5.4 is modified as

$$\ell\colon\ [1, 2^{nR_b}] \times \mathcal{Y}^{i-1} \to \mathcal{A}, \tag{5.31}$$

which maps the message M_b and the past observed decoder side information sequence Y^{i-1} into the ith symbol of the action sequence A_i.

Proposition 5.4. *The rate-distortion-cost region $\mathcal{R}(D_1, D_2, \Gamma)$ for the cascade source coding problem illustrated in Figure 5.3 with adaptive action-dependent side information is given by the rate region described in Proposition 5.1.*

Proposition 5.5. *The rate-distortion-cost region $\mathcal{R}(D_1, D_2, \Gamma)$ for the cascade-broadcast source coding problem under the CR illustrated in Figure 5.5 with adaptive action-dependent side information is given by the region described in Proposition 5.3 by setting $Z = \emptyset$.*

Remark 5.6. The results above show that enabling adaptive actions does not increase the achievable rate-distortion-cost region. These results generalize the observations in [26] for the point-to-point setting, wherein a similar conclusion is drawn.

To establish the propositions above, it is sufficient to prove the converse. The proofs for Proposition 5.4 and Proposition 5.5 are given in Appendix M and N, respectively.

5.5 Concluding Remarks

In an increasing number of applications, communication networks are expected to be able to convey not only data, but also information about control for actuation over multiple hops. In this chapter, a baseline communication model with three nodes connected in a cascade with the possible presence of an additional broadcast link is evaluated. The optimal trade-off between rate, distortion and cost is characterized for actuation in a number of relevant cases of interest. In general, the results point to the advantages of leveraging a joint representation of data and control information in order to utilize in the most efficient way the available communication links. Specifically, in all the considered models, a layered coding strategy, possibly coupled with rate splitting, has been proved to be optimal. This strategy is such that the base layer has the double role of guiding the actions of the downstream nodes and of providing a coarse description of the source, similar to [4]. Moreover, it is shown that this base compression layer should be designed in a way that depends on the network topology and on the relative cost of activating the different links.

CHAPTER 6

TWO-WAY COMMUNICATION WITH ADAPTIVE DATA ACQUISITION

6.1 Introduction

In computer networks and machine-to-machine links, communication is often interactive and serves a number of integrated functions, such as data exchange, query and control. As an illustrative example, consider the set-up in Figure 6.1 in which the terminals labeled Node 1 and Node 2 communicate on bidirectional links. Node 2 has access to a database or, more generally, is able to acquire information from the environment, e.g., through sensors. As a result of the communication on the forward link (Figure 6.1-(a)), Node 2 wishes to compute some functions $\hat{X}_2^n$, e.g., a suitable average, of the data X^n available at Node 1 and of the information it can retrieve from the environment[1]. The latter is measured as Y^n by Node 2 (Fig 6.1-(b)). The forward link (Fig 6.1-(a)) is also used by Node 1 to query Node 2 with the aim of retrieving some information $\hat{X}_1^n$ about Y^n from the environment through the backward link (Fig 6.1-(c)).

As reviewed in Chapter 4 and Chapter 5, information acquisition from the environment at Node 2 is generally expensive in terms of system resources, e.g., time, bandwidth or energy. For instance, accessing a remote database as in Fig 6.1-(b) requires interfacing with a server by following the appropriate protocol, and activating sensors entails some energy expenditure. Therefore, data acquisition by Node 2 should be performed efficiently by adapting to the informational requirements of Node 1 and Node 2 , i.e., to the requirements on the calculation of $\hat{X}_1^n$ and $\hat{X}_2^n$, respectively.

[1]Integer n represents the dimension of the data.

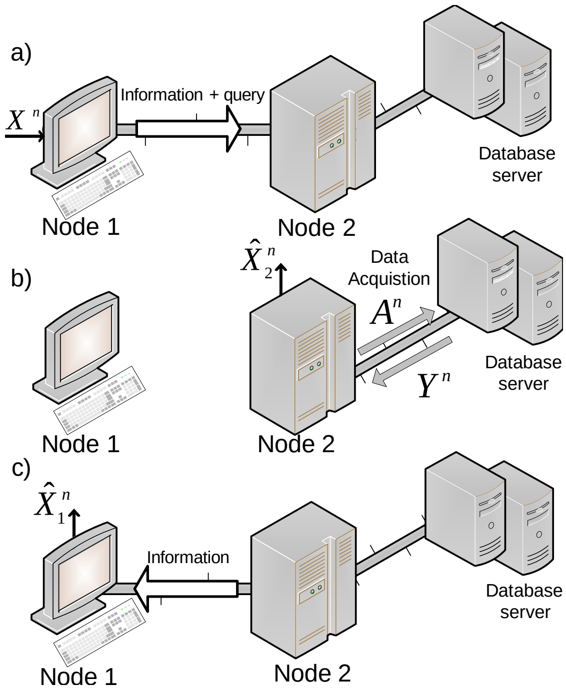

Figure 6.1 Two-way communication with adaptive data acquisition.

To summarize the discussion above, in the system of Figure 6.1 the forward communication from Node 1 to Node 2 serves three integrated purposes: $i)$ *Data exchange*: Node 1 provides Node 2 with the information necessary about data X^n for the latter to compute the desired quantity ; $ii)$ *Query*: Node 1 informs Node 2 about its own informational requirements with regard to $\hat{X}_1^n$, to be met via the backward link; $iii)$ *Control*: Node 1 instructs Node 2 on the most effective way to perform data acquisition (i.e., acquisition of Y^n) from the environment in order to satisfy Node 1's query and to allow Node 2 to perform the desired computation (of $\hat{X}_2^n$).

This chapter sets out to analyze the setting in Figure 6.1 from a fundamental theoretical standpoint via information theory. Specifically, the problem is formulated within the context of network rate-distortion theory, and the optimal communication strategy, involving the elements of data exchange, query and control, is identified. Examples are worked out to illustrate the relevance of the developed theory. Finally, the issue of robustness is tackled by assuming that, unbeknownst to Node 1, Node

2 may be unable to acquire information from the environment, due to, e.g., energy shortages or malfunctioning. The optimal robust strategy is derived and the examples are extended to account for this generalized model. It should be mentioned that the problem of characterizing the rate-distortion region for a two-way source coding models, with conventional action-independent side information sequences at Node 2 has been addressed in [40, 41, 42] and references therein.

6.1.1 Contributions and Organization of the Chapter

This chapter studies the model in Figure 6.1, which is detailed in terms of a block diagram in Figure 6.2. The system model is introduced in Section 6.2. The optimal trade-off between the rates of the bidirectional communication, the distortions of the reconstructions of the desired quantities at the two nodes, and the budget for information acquisition at Node 2 is derived in Section 6.3. An example that illustrates the application of the developed theory is discussed in Section 6.4. Finally, in Section 6.5, the results are extended to the scenario in Figure 6.5 in which, unbeknownst to Node 1, Node 2 may be unable to perform information acquisition.

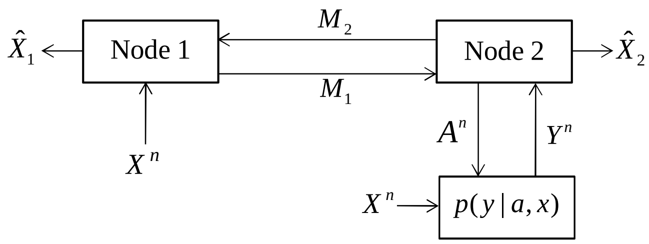

Figure 6.2 Two-way source coding with a side information vending machine at Node 2.

6.2 System Model

The two-way source coding problem of interest, sketched in Figure 3.1, is formally defined by the probability mass functions (pmfs) $p_X(x)$ and $p_{Y|A,X}(y|a,x)$, and by the discrete alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{A}, \hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2$, along with distortion and cost metrics to be discussed below. The source sequence $X^n = (X_1, ..., X_n) \in \mathcal{X}^n$ consists of n independent and identically distributed (i.i.d.) entries X_i for $i \in [1, n]$ with pmf $p_X(x)$. Node 1 measures sequence X^n and encodes it in a message M_1 of nR_1 bits, which is delivered to Node 2. Node 2 wishes to estimate a sequence $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ within given distortion requirements. To this end, Node 2 receives message M_1 and based on this, it selects an action sequence A^n, where $A^n \in \mathcal{A}^n$.

The action sequence affects the quality of the measurement Y^n of sequence X^n obtained at Node 2. Specifically, given A^n and X^n, the sequence Y^n is distributed as $p(y^n|a^n, x^n) = \prod_{i=1}^n p_{Y|A,X}(y_i|a_i, x_i)$. The cost of the action sequence is defined by a cost function Λ: $\mathcal{A} \rightarrow [0, \Lambda_{\max}]$ with $0 \leq \Lambda_{\max} < \infty$, as $\Lambda(a^n) = \sum_{i=1}^n \Lambda(a_i)$. The estimated sequence $\hat{X}_2^n$ with $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ is then obtained as a function of M_1 and Y^n. Upon reception on the forward link, Node 2 maps the message M_1 received from Node 1 and the locally available sequence Y^n in a message M_2 of nR_2 bits, which is delivered back to Node 1. Node 1 estimates a sequence $\hat{X}_1^n \in \hat{\mathcal{X}}_1^n$ as a function of M_2 and X^n within given distortion requirements.

The quality of the estimated sequence $\hat{X}_j^n$ is assessed in terms of the distortion metrics $d_j(x, y, \hat{x}_j)$: $\mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{X}}_j \rightarrow \mathbb{R}_+ \cup \{\infty\}$ for $j = 1, 2$, respectively. Note that this implies that $\hat{X}_j^n$ is allowed to be a lossy version of any function of the source and side information sequences. A more general model is studied in Section 6.3.1. It is assumed that the distortion accrued in the absence of measurements and communication is finite, i.e.,

$$D_{j,max} = \min_{\hat{x}_j \in \hat{\mathcal{X}}_j} \mathrm{E}[d(X, Y, \hat{X}_j)] < \infty \quad \text{for} \quad j = 1, 2. \tag{6.1}$$

A formal description of the operations at encoder and decoder follows.

Definition 6.1. An $(n, R_1, R_2, D_1, D_2, \Gamma, \epsilon)$ code for the set-up of Figure 3.1 consists of a source encoder for Node 1

$$\mathrm{g}_1\colon\ \mathcal{X}^n \rightarrow [1, 2^{nR_1}], \tag{6.2}$$

which maps the sequence X^n into a message M_1; an "action" function

$$\ell\colon\ [1, 2^{nR_1}] \times \mathcal{Y}^{i-1} \rightarrow \mathcal{A}, \tag{6.3}$$

which maps the message M_1 and the previously observed measurements into an action sequence A^n; a source encoder for Node 2

$$\mathrm{g}_2\colon\ \mathcal{Y}^n \times [1, 2^{nR_1}] \rightarrow [1, 2^{nR_2}], \tag{6.4}$$

which maps the sequence Y^n and message M_1 into a message M_2; two decoders, namely

$$\mathrm{h}_1\colon\ [1, 2^{nR_2}] \times \mathcal{X}^n \rightarrow \hat{\mathcal{X}}_1^n, \tag{6.5}$$

which maps the message M_2 and the sequence X^n into the estimated sequence $\hat{X}_1^n$;

$$\mathrm{h}_2\colon\ [1, 2^{nR_1}] \times \mathcal{Y}^n \rightarrow \hat{\mathcal{X}}_2^n, \tag{6.6}$$

which maps the message M_1 and the sequence Y^n into the estimated sequence $\hat{X}_2^n$; such that the action cost constraint Γ and distortion constraints D_j for $j = 1, 2$ are satisfied, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\Lambda(A_i)\right] \leq \Gamma \tag{6.7}$$

$$\text{and } \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d_j(X_i, Y_i, \hat{X}_{ji})\right] \leq D_j \text{ for } j = 1, 2. \tag{6.8}$$

Definition 6.2. Given a distortion-cost tuple (D_1, D_2, Γ), a rate tuple (R_1, R_2) is said to be achievable if, for any $\epsilon > 0$, and sufficiently large n, there exists a $(n, R_1, R_2, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)$ code.

Definition 6.3. The *rate-distortion-cost region* $\mathcal{R}(D_1, D_2, \Gamma)$ is defined as the closure of all rate tuples (R_1, R_2) that are achievable given the distortion-cost tuple (D_1, D_2, Γ).

Remark 6.1. For the special case in which the side information Y is independent of the action A given X, i.e., for $p(y|a,x) = p(y|x)$, the rate-distortion region $\mathcal{R}(D_1, D_2, \Gamma)$ has been derived in [40]. Instead, if $D_1 = D_{1,max}$, the set of all achievable rates R_1 was characterized in [4].

Remark 6.2. The definition (8.2) of an action encoder allows for adaptation of the actions to the previously observed values of the side information Y. This possibility was studied in [26] for the point-to-point one-way model, which is obtained by setting $R_2 = 0$ in the setting of Figure 3.1.

6.3 Rate-Distortion-Cost Region

In this section, a single-letter characterization of the rate-distortion-cost region is derived.

Proposition 6.1. *The rate-distortion-cost region $\mathcal{R}(D_1, D_2, \Gamma)$ for the two-way source coding problem illustrated in Figure 3.1 is given by the union of all rate pairs (R_1, R_2) that satisfy the conditions*

$$\begin{aligned} R_1 &\geq I(X;A) + I(X;U|A,Y) \qquad (6.9a) \\ \text{and } \mathrm{R}_2 &\geq I(Y;V|A,X,U), \qquad (6.9b) \end{aligned}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x,y,a,u,v) = p(x)p(a,u|x)p(y|a,x)p(v|a,u,y), \qquad (6.10)$$

for some pmfs $p(a,u|x)$ and $p(v|a,u,y)$ such that the inequalities

$$\begin{aligned} \mathrm{E}[\mathrm{d}_1(\mathrm{X},\mathrm{Y},\mathrm{f}_1(\mathrm{V},\mathrm{X}))] &\leq D_1, \qquad (6.11a) \\ \mathrm{E}[\mathrm{d}_2(\mathrm{X},\mathrm{Y},\mathrm{f}_2(\mathrm{U},\mathrm{Y}))] &\leq D_2, \qquad (6.11b) \\ \text{and } \mathrm{E}[\Lambda(\mathrm{A})] &\leq \Gamma, \qquad (6.11c) \end{aligned}$$

are satisfied for some function f_1*:* $\mathcal{V} \times \mathcal{X} \rightarrow \hat{\mathcal{X}}_1$ *and* f_2*:* $\mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}_2$*. Finally,* U *and* V *are auxiliary random variables whose alphabet cardinality can be constrained as* $|\mathcal{U}| \leq |\mathcal{X}||\mathcal{A}| + 4$ *and* $|\mathcal{V}| \leq |\mathcal{U}||\mathcal{Y}||\mathcal{A}| + 1$ *without loss of optimality.*

Remark 6.3. For the special case in which the side information Y is independent of the action A given X, i.e., for $p(y|a, x) = p(y|x)$, the rate-distortion region $\mathcal{R}(D_1, D_2, \Gamma)$ in Proposition 6.1 reduces to that derived in [40, 41]. Instead, if $D_1 = D_{1,max}$, the result reduces to that in [4].

The proof of the converse is provided in Appendix O. The achievability follows as a combination of the techniques proposed in [4] and [40], and requires the forward link to be used, in an integrated manner, for data exchange, query and control. Specifically, for the forward link, similar to [4] (see Chapter 2 and Figure 2.5), Node 1 uses a successive refinement codebook. Accordingly, the base layer is used by Node 1 to instruct Node 2 on which actions are best tailored to fulfill the informational requirements of both Node 1 and Node 2. This base layer thus represents control information that also serves the purpose of querying Node 2 in view of the backward communication. Node 1 selects this base layer as a function of the source X^n, thus allowing Node 2 to adapt its actions for information acquisition to the current realization of the source X^n. The refinement layer of the code used by Node 1 is leveraged, instead, to provide additional information to Node 2 in order to meet Node 2's distortion requirement. Node 2 then employs standard Wyner-Ziv coding (i.e., binning) [1] for the backward link to satisfy Node 1's distortion requirement.

The main technical aspects of the achievability proof is briefly outlined here, since the details follow from standard arguments and do not require further elaboration here. To be more precise and with reference to Figure 2.5, Node 1 first maps sequence X^n into the action sequence A^n using the standard joint typicality criterion. This mapping requires a codebook of rate $I(X; A)$ (see, e.g., [1, pp. 62-63]). Given the sequence A^n, the description of sequence X^n is further refined through mapping to a sequence U^n. This requires a codebook of size $I(X;U|A,Y)$ for each action sequence A^n using Wyner-Ziv binning with respect to side information Y^n [1, pp. 62-63]. In the reverse link, Node 2 employs Wyner-Ziv coding for the sequence Y^n by leveraging the side information X^n available at Node 1 and conditioned on the sequences U^n and A^n, which are known to both Node 1 and Node 2 as a result of the communication on the forward link. This requires a rate equal to the right-hand side of (6.9b). Finally, Node 1 and Node 2 produce the estimates $\hat{X}_1^n$ and $\hat{X}_2^n$ as the symbol-by-symbol functions $\hat{X}_{1i} = \mathrm{f}_1(V_i, X_i)$ and $\hat{X}_{2i} = \mathrm{f}_2(U_i, Y_i)$ for $i \in [1, n]$, respectively.

Remark 6.4. The achievability scheme discussed above uses actions that do not adapt to the previous values of the side information Y. The fact that this scheme attains the optimal performance characterized in Proposition 6.1 shows that, as demonstrated in [26] for the one-way model with $R_2 = 0$, adaptive actions do not improve the rate-distortion performance.

6.3.1 Indirect Rate-Distortion-Cost Region

In this section, a more general model in which Node 1 observes only a noisy version of the source X^n, as depicted in Figure 6.3 is considered. Following [43], this setting is referred to as posing an indirect source coding problem. The example studied in Section 6.4 illustrates the relevance of this generalization. The system model is as defined in Section 6.2 with the following differences. The source encoder for Node 1

$$\mathrm{g}_1\colon\ \mathcal{Z}^n \to [1, 2^{nR_1}], \tag{6.12}$$

maps the sequence Z^n into a message M_1; the decoder for Node 1

$$\mathrm{h}_1\colon\ [1, 2^{nR_2}] \times \mathcal{Z}^n \to \hat{\mathcal{X}}_1^n, \tag{6.13}$$

maps the message M_2 and the sequence Z^n into the estimated sequence $\hat{X}_1^n$; given (X^n, A^n, Z^n), the side information Y^n is distributed as $p(y^n|a^n, x^n, z^n) = \prod_{i=1}^n p_{Y|A,X,Z}(y_i|a_i, x_i, z_i)$ and the distortion constraints are given as

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d_j(X_i, Y_i, Z_i, \hat{X}_{ji})\right] \leq D_j \text{ for } j = 1, 2, \tag{6.14}$$

for some distortion metrics $d_j(x, y, z, \hat{x}_j) : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \hat{\mathcal{X}}_j \to \mathbb{R}_+ \cup \{\infty\}$, for $j = 1, 2$. The next proposition derives a single-letter characterization of the rate-distortion-cost region.

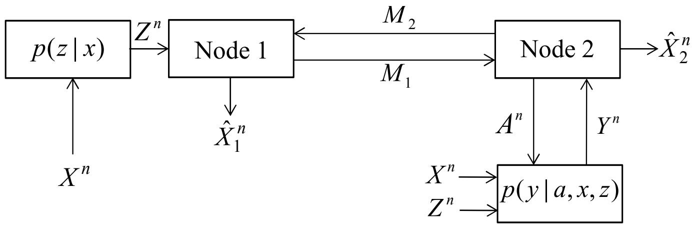

Figure 6.3 Indirect two-way source coding with a side information vending machine at Node 2.

Proposition 6.2. *The rate-distortion-cost region $\mathcal{R}(D_1, D_2, \Gamma)$ for the indirect two-way source coding problem illustrated in Figure 6.3 is given by the union of all rate pairs (R_1, R_2) that satisfy the conditions*

$$R_1 \geq I(Z; A) + I(Z; U|A, Y) \tag{6.15a}$$

$$\text{and } \mathrm{R}_2 \geq I(Y; V|A, Z, U), \tag{6.15b}$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x, y, z, a, u, v) = p(x, z)p(a,u|z)p(y|a,x,z)p(v|a,u,y), \tag{6.16}$$

for some pmfs $p(a,u|x)$ and $p(v|a,u,y)$ such that the inequalities

$$\begin{aligned} \mathrm{E}[\mathrm{d}_1(\mathrm{X},\mathrm{Y},\mathrm{Z},\mathrm{f}_1(\mathrm{V},\mathrm{Z}))] &\leq D_1, && (6.17a)\\ \mathrm{E}[\mathrm{d}_2(\mathrm{X},\mathrm{Y},\mathrm{Z},\mathrm{f}_2(\mathrm{U},\mathrm{Y}))] &\leq D_2, && (6.17b)\\ \text{and } \mathrm{E}[\Lambda(\mathrm{A})] &\leq \Gamma, && (6.17c) \end{aligned}$$

are satisfied for some function f_1: $\mathcal{V} \times \mathcal{Z} \to \hat{\mathcal{X}}_1$ and f_2: $\mathcal{U} \times \mathcal{Y} \to \hat{\mathcal{X}}_2$. Finally, U and V are auxiliary random variables whose alphabet cardinality can be constrained as $|\mathcal{U}| \leq |\mathcal{Z}||\mathcal{A}| + 3$ and $|\mathcal{V}| \leq |\mathcal{U}||\mathcal{Y}||\mathcal{A}| + 1$ without loss of optimality.

The proof of the achievability and converse follows with slight modifications from that of Proposition 6.1. Specifically, in the achievability the sequence X^n is replaced by its noisy version, i.e., the sequence Z^n, and the rest of the proof remains essentially unchanged. The proof of the converse is provided in Appendix O.

6.4 Case Study and Numerical Results

In this section, an example is considered for the set-up in Figure 6.3 in order to illustrate the main aspects of the problem and the relevance of the theoretical results derived above. Consider a sensor network consisting of two sensors deployed to monitor a given phenomenon of interest (i.e., the concentration of a given chemical). Assume that the state of the observed phenomenon is described by a random source $X \sim \mathrm{Bern}(0.5)$ (e.g., $X = 0$ indicates a low concentration of the chemical and $X = 1$ a high concentration). Due to malfunctioning or environmental causes, at each time i, Node 1 measures X_i with probability $1-\epsilon$, and reports instead an unusual event e (referred to as "erasure") with probability ϵ. This implies that $Z_i = \mathrm{e}$ with probability ϵ, and $Z_i = X_i$ with probability $1-\epsilon$, for $i \in [1,n]$.

Node 2 has the double purpose of monitoring the operation of Node 1 and of assisting Node 1 in case of measurement failures (erasures). To this end, if necessary, Node 2 can measure the phenomenon X_i by investing a unit of energy. This is modelled by assuming that the vending machine at Node 2 operates as follows:

$$Y = \begin{cases} X & \text{for } A = 1 \\ \phi & \text{for } A = 0 \end{cases}, \tag{6.18}$$

with cost constraint $\Lambda(a) = a$, for $a \in \{0, 1\}$, where ϕ is a dummy symbol representing the case in which no useful information is acquired by Node 2. This model implies that a cost budget of Γ limits the average number of samples of the sequence Y that can be measured by Node 2 to around $n\Gamma$ given the constraint (8.5).

Node 1 wishes to reconstruct the source X^n, while Node 2 is interested in recovering Z^n in order to monitor the operation of Node 1. The distortion functions are the Hamming metrics $d_1(x, \hat{x}_1) = 1_{\{x \neq \hat{x}_1\}}$ and $d_2(z, \hat{x}_2) = 1_{\{z \neq \hat{x}_2\}}$. Therefore, the maximum distortions (6.1) are easily seen to be given by $D_{1,max} = 0.5$ and $D_{2,max} = 1 - \max\{\epsilon, (1-\epsilon)/2\}$.

To obtain analytical insight into the rate-distortion-cost region, in the following on a number of special cases are considered.

6.4.1 $D_1 = D_{1,max}$ and $D_2 = 0$

Consider the distortion requirements $D_1 = D_{1,max}$ and $D_2 = 0$. As a result, Node 1 requires no backward communication from Node 2, while Node 2 wishes to recover Z^n losslessly. In the context of the example, here the only functionality of the network is the monitoring of the operation of Node 1 by Node 2. For the given distortions, the rate-cost region in Proposition 6.2 can be evaluated as

$$R_1 \geq H_2(\epsilon) + (1 - \epsilon - \Gamma)^+ \tag{6.19a}$$

$$\text{and } R_2 \geq 0, \tag{6.19b}$$

for any cost budget $\Gamma \geq 0$, where $H_2(\alpha) = -\alpha\log_2\alpha - (1-\alpha)\log_2(1-\alpha)$ is the binary entropy function.

A formal proof of this result can be found in Appendix P. The rate region (6.19) shows that, as the cost budget Γ for information acquisition increases, the required rate R_1 decreases down to the rate $H_2(\epsilon)$ that is required to describe only the erasures process E^n with $E_i = 1_{\{Z_i=\mathrm{e}\}}$, $i = 1, ..., n$, losslessly to Node 2. This can be explained by noting that the time-sharing strategy discussed next achieves region (6.19) and is thus optimal.

In the time-sharing strategy, Node 1 describes the process E^n losslessly to Node 2 with $H_2(\epsilon)$ bits per symbol. In addition to E^n, in order to obtain a lossless reconstruction of Z^n, Node 2 needs to be informed about $Z_i = X_i$ for all i in which $E_i = 0$. Note that, there are around $n(1-\epsilon)$ such samples of Z_i. Node 1 describes $Z_i = X_i$ for $n(1-\epsilon-\Gamma)^+$ of these samples, while the remaining $n\min(\Gamma, 1-\epsilon)$ are measured by Node 2 through the vending machine. Note that, in the strategy just described, sequence E^n can be interpreted as control data that is used by Node 2 to adapt its information acquisition process. An alternative strategy based directly on Proposition 6.2 can be found in Appendix P.

Figure 6.4 illustrates the rate R_1 in (6.19a) versus the cost budget Γ for $\epsilon = 0.2$ (line with circles). The first observation is that for $\Gamma = 0$, since there is no side information available at Node 2, $R_1 = H_2(\epsilon) + 1 - \epsilon = 1.52$, which is the rate required to transmit the sequence Z^n losslessly to Node 2. Moreover, as mentioned, as the cost budget Γ increases, the rate R decreases down to the value $H_2(\epsilon) = 0.72$ needed to describe only the sequence E^n. Finally, it is observed that if $\Gamma \geq 1-\epsilon = 0.8$ no further improvement of the rate is possible since Node 2 only needs to measure a fraction $(1-\epsilon)$ of values of X^n in order to recover Z^n losslessly.

6.4.2 $D_1 = 0$ and $D_2 = D_{2,max}$

Here, the dual case is considered in which Node 1 wishes to reconstruct sequence X^n losslessly ($D_1 = 0$), while Node 2 does not have any distortion requirements

($D_2 = D_{2,max}$). In the context of the example, the network thus operates with the aim of allowing Node 1 to measure the phenomenon of interest X^n reliably. As shown in Appendix P, if $\Gamma \geq \epsilon$, the rate-cost region is given by the union of all rate pairs (R_1, R_2) such that

$$R_1 \geq H_2(\epsilon) - \Gamma H\left(\frac{\epsilon}{\Gamma}\right) \tag{6.20a}$$

$$\text{and } R_2 \geq \epsilon. \tag{6.20b}$$

Moreover, for $\Gamma < \epsilon$, the region is empty as the lossless reconstruction of X at Node 1 is not feasible.

A proof of this result based on Proposition 6.2 can be found in Appendix P. In the following, it is argued that a natural time-sharing strategy, akin to that used for the case $D_1 = D_{1,max}, D_2 = 0$ above, would be suboptimal, implying that the optimal strategy requires a more sophisticated approach based on the successive refinement code presented in Section 6.3.

A natural time-sharing strategy would be the following. Node 1 describes $n\eta$ samples of the erasure process E^n, for some $0 \leq \eta \leq 1$, losslessly to Node 2, using rate $R_1 = \eta H_2(\epsilon)$. This information is used by Node 1 to query Node 2 about the desired information. Specifically, Node 2 sets $A_i = 1$ if $E_i = 1$, thus observing around $n\eta\epsilon$ samples $Y_i = X_i$ from the vending machine. These samples are needed to fulfill the distortion requirements of Node 1. For all the remaining $n(1-\eta)$ samples, for which Node 2 does not have control information from Node 1, Node 2 sets $A_i = 1$, thus acquiring all the side information samples. Again, this is necessary given Node 1's requirements. Node 2 conveys losslessly the $n\eta\epsilon$ samples $Y_i = X_i$ obtained when $E_i = 1$, which requires $\eta\epsilon$ bits per sample, along with the $n(1-\eta)$ samples Y_i in the second set, which amount instead to $(1-\eta)H(X|Z)$ bits per sample. Note that the rate is $H(X|Z)$ by the Slepian-Wolf theorem [1, Chapter 10], since Node 1 has side information Z_i for the second set of samples. Overall, $R_2 = \eta\epsilon + (1-\eta)\epsilon = \epsilon$ bits/source symbol. This entails a cost budget of $\Gamma = \eta\epsilon + 1 - \eta$, and thus $\eta = \frac{(1-\Gamma)}{(1-\epsilon)}$.

Figure 6.4 compares the rate R_1 as in (6.20a) (line with squared markers) with the corresponding rate obtained via time-sharing (dashed line), for $\epsilon = 0.2$. As seen, in this second case the time-sharing strategy is strictly suboptimal (except for the two extreme case of $\Gamma = 0$ and $\Gamma = 1$). Moreover, as discussed above, achieving $D_1 = 0$ is impossible for $\Gamma \leq \epsilon$, since Node 2 must obtain the fraction ϵ of samples of X^n that Node 1 fails to measure in order to guarantee lossless reconstruction at Node 1. Finally, for $\Gamma = 1$, there is no need for Node 1 to send any information to Node 2, as Node 2 is able to acquire the sequence X^n and send back to Node 1 on the backward link (using rate $R_2 = \epsilon$ by the Slepian-Wolf theorem).

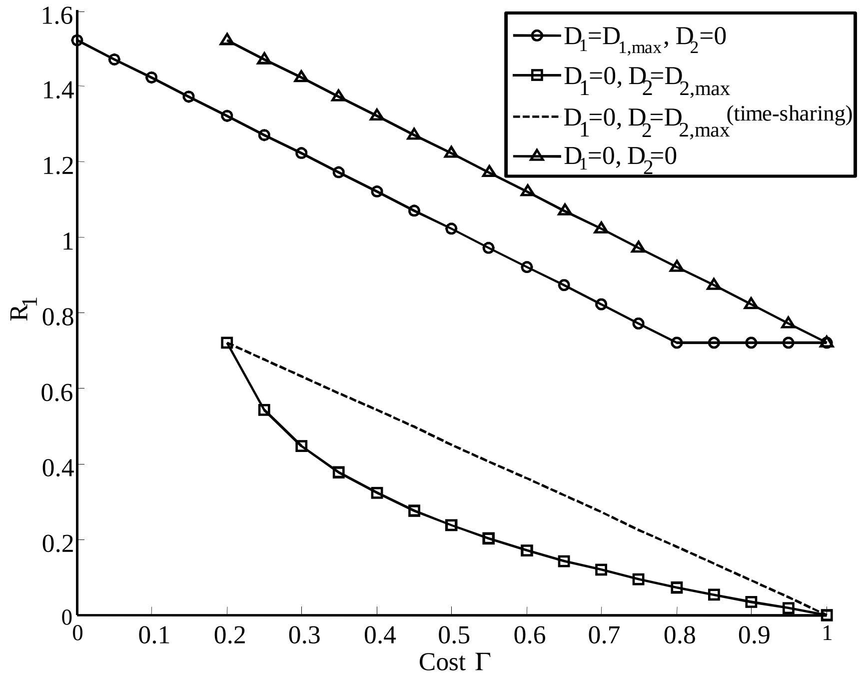

Figure 6.4 Rate R_1 versus cost Γ for the examples in Section 6.4 with $\epsilon = 0.2$.

6.4.3 $D_1 = D_2 = 0$

The case in which both nodes wish to achieve lossless reconstruction, i.e., $D_1 = D_2 = 0$ is considered. In this case, in the context of the example, both measurement of Node 1 and monitoring at Node 2 are required to operate correctly. As seen in the previous case, achieving $D_1 = 0$ is not possible if $\Gamma < \epsilon$ and thus this is a fortiori true for $D_1 = D_2 = 0$. For $\Gamma \geq \epsilon$, the rate-cost region is given by

$$R_1 \geq H_2(\epsilon) + (1 - \Gamma) \tag{6.21a}$$

$$\text{and } R_2 \geq \epsilon, \tag{6.21b}$$

as shown in Appendix P.

A time-sharing strategy that achieves (6.21) is as follows. Node 1 describes the process E^n losslessly to Node 2 with $H_2(\epsilon)$ bits per symbol. This information serves the functions of query and control for Node 2. In order to satisfy its distortion requirement, Node 2 now needs to be informed about $Z_i = X_i$ for all i in which $E_i = 0$. Note that there are $n(1 - \epsilon)$ such samples of Z_i. Node 1 describes $Z_i = X_i$ for $n(1 - \Gamma) \leq n(1 - \epsilon)$ of these samples, while the remaining $n(\Gamma - \epsilon)$ are measured by Node 2 through the vending machine. Node 2 compresses losslessly the sequence of around $n\epsilon$ samples of X_i with i such that $E_i = 1$ which requires $R_2 = \epsilon$ bits per sample.

Figure 6.4 illustrates the rate R_1 in (6.21a) versus the cost budget Γ for $\epsilon = 0.2$ (line with triangular markers). As discussed above, achieving $D_1 = 0$ is impossible for $\Gamma \leq \epsilon$. Moreover, for $\Gamma = 1$ the performance of system is identical to that with $D_1 = D_{1,max}$ and $D_2 = 0$, since in this case the informational requirements of both Node 1 and Node 2 are satisfied if Node 1 conveys the locations of the erasures to Node 2 (which requires rate $R_1 = H_2(\epsilon) = 0.72$).

6.5 When the Side Information May Be Absent

In this section, the results of the previous section are generalized to the scenario in Figure 6.5 in which, unbeknownst to Node 1, Node 2 may be unable to perform information acquisition due, e.g., to energy shortage or malfunctioning. The setup of the general model is illustrated in Figure 6.5.

6.5.1 System Model

The formal description of an $(n, R_1, R_2, D_1, D_2, D_3, \Gamma, \epsilon)$ code for the set-up of Figure 6.5 is given as in Section 6.3.1 (which generalizes the model in Section 6.2) with the addition of Node 3. This added node, which has no access to side information, models the case in which the receiver is not able to acquire the side information due to, e.g., malfunctioning. Note that the same message M_1 from Node 1 is received by both Node 2 and Node 3. This captures the fact that the information about whether or not the recipient is able to access the side information is not available to Node 1. The model in Figure 6.5 is a generalization of the so called Heegard-Berger problem [3, 9].

Formally, Node 3 is defined by the decoding function

$$\mathrm{h}_3\colon\ [1, 2^{nR_1}] \rightarrow \hat{\mathcal{X}}_3^n, \tag{6.22}$$

which maps the message M_1 into the estimated sequence $\hat{X}_3^n$; and the additional distortion constraint

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[d_3(X_i, Y_i, Z_i, \hat{X}_{3i})\right] \leq D_3. \tag{6.23}$$

It is remarked that adding a link between Node 3 and Node 1 cannot improve the system performance because Node 3 does not have access to any additional information. Therefore, there is no advantage in having a backward link from Node 3 to Node 1 because the information available at Node 3 is a subset of the information available at Node 1. Therefore, this link is not included in the model.

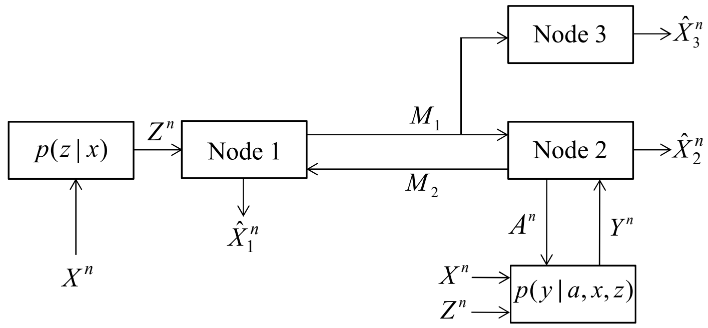

Figure 6.5 Indirect two-way source coding when the side information vending machine may be absent at the recipient of the message from Node 1.

6.5.2 Rate-Distortion-Cost Region

In this section, a single-letter characterization of the rate-distortion-cost region is derived for the set-up in Figure 6.5.

Proposition 6.3. *The rate-distortion-cost region* $\mathcal{R}(D_1, D_2, D_3, \Gamma)$ *for the two-way source coding problem illustrated in Figure 6.5 is given by the union of all rate pairs* (R_1, R_2) *that satisfy the conditions*

$$R_1 \geq I(Z;A) + I(Z;\hat{X}_3|A) + I(Z;U|A,Y,\hat{X}_3) \quad (6.24a)$$

$$\text{and } R_2 \geq I(Y;V|A,Z,U,\hat{X}_3), \quad (6.24b)$$

where the mutual information terms are evaluated with respect to the joint pmf

$$p(x,y,z,a,u,v) = p(x,z)p(a,u,\hat{x}_3|z)p(y|a,x,z)p(v|a,u,y,\hat{x}_3), \quad (6.25)$$

for some pmfs $p(a,u,\hat{x}_3|z)$ *and* $p(v|a,u,y)$ *such that the inequalities*

$$\mathrm{E}[d_1(X,Y,Z,\mathrm{f}_1(V,Z))] \leq D_1, \quad (6.26a)$$

$$\mathrm{E}[d_2(X,Y,Z,\mathrm{f}_2(U,Y))] \leq D_2, \quad (6.26b)$$

$$\mathrm{E}[d_3(X,Y,Z,\hat{X}_3)] \leq D_3, \quad (6.26c)$$

$$\text{and } \mathrm{E}[\Lambda(A)] \leq \Gamma, \quad (6.26d)$$

are satisfied for some function f_1*:* $\mathcal{V} \times \mathcal{Z} \rightarrow \hat{\mathcal{X}}_1$ *and* f_2*:* $\mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}_2$*. Finally,* U *and* V *are auxiliary random variables whose alphabet cardinality can be constrained as* $|\mathcal{U}| \leq |\mathcal{Z}||\mathcal{A}| + 3$ *and* $|\mathcal{V}| \leq |\mathcal{U}||\mathcal{Y}||\mathcal{A}| + 1$ *without loss of optimality.*

The proof of the converse is provided in Appendix Q. The achievable rate (6.24a) can be interpreted as follows. Node 1 uses a successive refinement code with three layers. The first layer is defined as for Section 6.3 and carries query and control information. The second and third layers are designed as in the optimal Heegard-Berger scheme [3]. Specifically, the second layer is destined to both Node 2 and Node 3, while the third layer targets only Node 2, which has enhanced decoding capabilities due to the availability of side information.

To provide further details, as for Proposition 6.1, the encoder first maps the input sequence Z^n into an action sequence A^n so that the two sequences are jointly typical, which requires $I(Z;A)$ bits/source sample. Then, it maps Z^n into the estimate $\hat{X}_3^n$ for Node 3 using a conditional codebook with rate $I(Z;\hat{X}_3|A)$. Finally, it maps Z^n into another sequence U^n using the fact that Node 2 has the action sequence A^n, the estimate $\hat{X}_3^n$ and the measurement Y^n. Using conditional codebooks (with respect to $\hat{X}_3^n$ and A^n) and from the Wyner-Ziv theorem, this requires $I(Z;U|A,Y,\hat{X}_3)$ bit/source sample (see Chapter 2 and Figure 2.5). As for the rate (6.24b), Node 2 employs Wyner-Ziv coding for the sequence Y^n by leveraging the side information Z^n available at Node 1 and conditioned on the sequences U^n, A^n and $\hat{X}_3^n$, which are known to both Node 1 and Node 2 as a result of the forward communication. This requires a rate equal to the right-hand side of (6.24b) (see Chapter 2 and Figure 2.5). Finally, Node 1 and Node 2 produce the estimates $\hat{X}_1^n$ and $\hat{X}_2^n$ as the symbol-by-symbol functions $\hat{X}_{1i} = \mathrm{f}_1(V_i, Z_i)$ and $\hat{X}_{2i} = \mathrm{f}_2(U_i, Y_i)$ for $i \in [1,n]$, respectively.

E \ $\hat{X}_3$	0	1	*
0	0	∞	1
1	∞	0	1

Table 6.1 Erasure Distortion for Reconstruction at Node 3.

6.5.3 Case Study and Numerical Results

In this section, the binary example of Section 6.4 is extended to the set-up in Figure 6.5. Specifically, the same setting as in Section 6.4 is considered, with the addition of Node 3. For the latter, a ternary reconstruction alphabet $\hat{\mathcal{X}}_3 = \{0, 1, *\}$ and the distortion metric $d_3(x, z, \hat{x}_3) = d_3(\mathbf{1}_{\{Z=\mathrm{e}\}}, \hat{x}_3)$ in Table 6.1 are assumed, where $E_i = \mathbf{1}_{\{Z_i=\mathrm{e}\}}$ is the erasure process[2]. Accordingly, Node 3 is interested in recovering the erasure process E^n under an erasure distortion metric (see, e.g., [20]) , where "*" represents the "don't care" or erasure reproduction symbol

It is first observed that for cases 1) and 3) in Section 6.4 the distortion requirements of Node 3 do not change the rate-distortion function. This is because, as discussed in Section 6.4, the requirement that D_2 be equal to zero entails that the erasure process E^n be communicated losslessly to Node 2 without leveraging the side information from the vending machine (which cannot provide information about the erasure process). It follows that one can achieve $D_3 = 0$ at no additional rate cost. Therefore, let's focus on the case 2) in Section 6.4, namely $D_1 = 0$ and $D_2 = D_{2,max} = 1 - \max\{\epsilon, (1-\epsilon)/2\}$.

In the case at hand, Node 1 wishes to recover X^n losslessly, Node 2 has no distortion requirements and Node 3 wants to recover E^n with distortion D_3. As explained in Section 6.4.2, in order to reconstruct X^n losslessly at Node 1 $\Gamma \geq \epsilon$ and $\Pr(A = 1|Z = \mathrm{e}) = 1$ have to be satisfied. Moreover, due to symmetry of the problem

[2]Infinity in Table 1 means that the corresponding reconstruction is unacceptable at the decoder and is thus measured to be infinity. This is a standard approach to account for errors that are not allowed by design.

with respect to $Z = 0$ and $Z = 1$, one can set $\Pr(A = 1|Z = 0) = \Pr(A = 1|Z = 1) = \frac{\gamma-\epsilon}{1-\epsilon}$, for some $0 \leq \gamma \leq \Gamma$. To evaluate the rate-distortion-cost region (6.24), let's define $\Pr(\hat{X}_3 = *|A = 1, Z = \mathrm{e}) \triangleq p_1$, $\Pr(\hat{X}_3 = *|A = 0, Z = 0) \triangleq p_2$ and $\Pr(\hat{X}_3 = *|A = 1, Z = 0) \triangleq p_3$. Thus, the rate-distortion-cost region is given by

$$
\begin{aligned}
R_1 \geq\ & H_2(\epsilon) + 1 - \epsilon - (1-\Gamma)(1-p_2) - (\Gamma-\epsilon) \\
& (1-p_3) - (1-\Gamma)p_2 - \Big(\epsilon p_1 + (\Gamma-\epsilon)p_3\Big) \\
& \Big(H_2\big(\frac{\epsilon p_1}{\epsilon p_1 + (\Gamma-\epsilon)p_3}\big) + \frac{(\Gamma-\epsilon)p_3}{\epsilon p_1 + (\Gamma-\epsilon)p_3}\Big)
\end{aligned} \tag{6.27a}
$$

$$
\text{and } R_2 \geq \epsilon, \tag{6.27b}
$$

where parameters $p_1, p_2, p_3 \in [0, 1]$ must be selected so as to satisfy the distortion constraint of Node 3, namely $D_3 \geq \epsilon p_1 + (1-\Gamma)p_2 + (\Gamma-\epsilon)p_3$.

Figure 6.6 illustrates the rate R_1 in (6.27a), minimized over p_1, p_2 and p_3 under the constraints mentioned above versus the cost budget Γ for $\epsilon = 0.2$ and different values of D_3, namely $D_3 = 0.4$, 0.6, 0.8 and $D_3 = D_{3,max} = 1$. Note that for $D_3 = D_{3,max} = 1$ the rate in (6.20a) is obtained. As it can be seen, for $\Gamma \leq D_3$, the rate decreases with increasing cost Γ, but for $\Gamma \geq D_3$ the rate remains constant while increasing Γ. The reason is that for the latter region, i.e., $\Gamma \geq D_3$, the performance of the system is dominated by the distortion requirement of Node 3 and thus increasing the cost budget Γ does not improve the rate. Instead, for $\Gamma \leq D_3$, it is sufficient to cater only to Node 2, and Node 3 is able to recover E with distortion $D_3 = \Gamma$ at no additional rate cost.

6.6 Concluding Remarks

For applications such as complex communication networks for cloud computing or machine-to-machine communication, the bits exchanged by two parties serve a number of integrated functions, including data transmission, control and query.

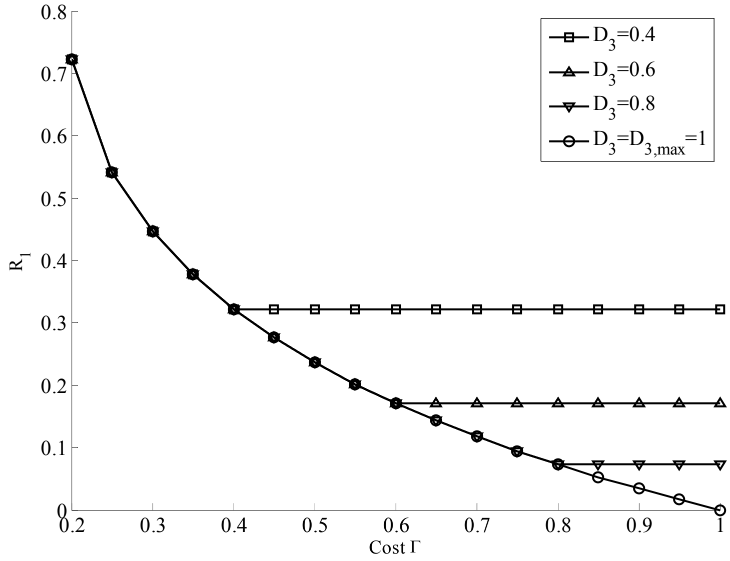

Figure 6.6 Rate R_1 versus cost Γ for the examples in Section 6.5.3 with $\epsilon = 0.2$, $D_1 = 0$ and $D_2 = D_{2,max}$.

In this chapter, a baseline two-way communication scenario that captures some of these aspects have been considered. The problem is addressed from a fundamental theoretical standpoint using an information theoretic formulation. The analysis reveals the structure of optimal communication strategies and can be applied to elaborate on specific examples, as illustrated in this chapter. This work opens a number of possible avenues for future research, including the analysis of scenarios in which more than one round of interactive communication is possible, set-ups in which there are multiple sources communicating to a single receiver in an interactive fashion, or, dually, multiple receivers, each connected to its own vending machine.

CHAPTER 7

ON CHANNELS WITH ACTION-DEPENDENT STATES

7.1 Introduction

In [44], the framework of action-dependent channels was introduced as a means to model scenarios in which transmission takes place in two successive phases. In the first phase, the encoder selects an "action" sequence, with the twofold aim of conveying information to the receiver and of affecting in a desired way the state of the channel to be used in the second phase. In the second phase, communication takes place in the presence the mentioned action-dependent state. With a cost constraint on the actions in the first phase and on the channel input in the second phase, reference [44] derived the capacity-cost-trade-off under the assumption that the channel state is available either causally or non-causally at the encoder in the second phase.

A number of applications and extensions of the results in [44] have been reported since then. In [45], the result in [44] is leveraged to study a model in which encoder and decoder can "probe" the channel state to obtain partial state information during the first communication phase. In [46], unlike [44] the decoder is required to decode both the transmitted message and channel input reliably. Finally, in [47], the decoder is interested in estimating not only the message but also the state sequence, and the latter is available strictly causally at the encoder in the second transmission phase.

In this chapter, two further extensions of the original action-dependent channel are studied. In the first, similar to [47], the decoder is interested in estimating not only the message but also the state sequence within given average per-letter distortion constraints (see Figure 7.1). Unlike [47], it is assumed *non-causal* state knowledge in the second phase, and, under the constraint of common reconstruction (CR) (i.e., the decoder's estimate of the state must be recoverable also at the encoder with high probability [8]), a single-letter characterization of the achievable rate-distortion-cost trade-off is obtained. It is remarked that, for conventional state-dependent states without actions, the problem of joint estimation of message and state with non-causal state information at the encoder without the CR constraint is open (see, e.g., [48]), while with the CR constraint the problem has been solved in [8]. In the second extension, illustrated in Figure 7.2, an action-dependent degraded broadcast channel is studied. Under the assumption that the encoder knows the state sequence causally in the second phase, the capacity-cost region is identified.[1] The corresponding result for action-independent states was derived in [51] (see also [52]), while it is recalled that with non-causal state information the problem is open (see [53]). Various examples, including Gaussian channels and a model with a "probing" encoder, are also provided throughout to show the advantage of a proper joint design of the two communication phases.

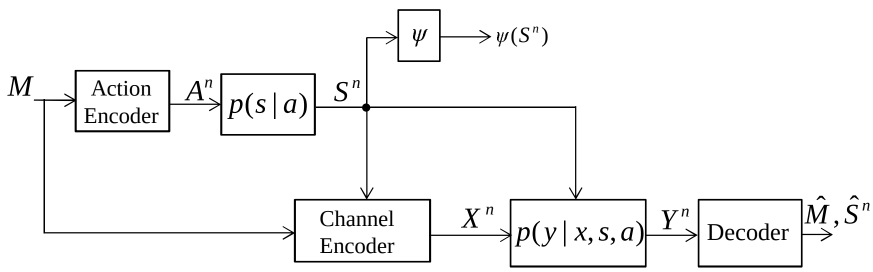

Figure 7.1 Channel with action-dependent state in which the decoder estimates both message and state, and there is a a common reconstruction (CR) constraint on the state reconstruction. The state is known non-causally at the channel encoder.

7.2 Transmission of Data and Action-Dependent State with Common Reconstruction Constraint

In this section, the setting illustrated in Figure 7.1 of a channel with action-dependent state is studied in which the decoder estimates both message and state. First, the

[1] After submitting [49], the authors are informed of the reference [50], where the problem illustrated in Figure 7.2 has also been solved.

system model is detailed in Section 8.4.1. Next, the characterization of the trade-off between the achievable data rate and state reconstruction distortion is derived in Section 8

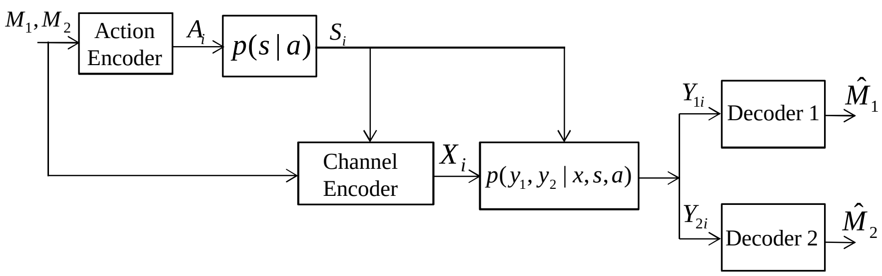

Figure 7.2 Broadcast channel with action-dependent states known causally to the encoder (i.e., the ith transmitted symbol X_i is a function of messages M_1, M_2 and the state symbols up to time i, S^i).

7.2.1 System Model

In this section, the system model is detailed. The system is defined by the probability mass functions (pmfs) $p(x)$, $p(y|x,s,a)$, $p(s|a)$ and discrete alphabets $\mathcal{X}, \mathcal{A}$, $\mathcal{S}$, $\hat{\mathcal{S}}$, and $\mathcal{Y}$ as follows. Given the message M, selected randomly from the set $\mathcal{M} = [1, 2^{nR}]$, an action sequence $A^n \in \mathcal{A}^n$ is selected. As a result of this selection, the state sequence $S^n \in \mathcal{S}^n$ is generated as the output of a memoryless channel $p(s|a)$ so that $p(s^n|a^n) = \prod_{i=1}^n p(s_i|a_i)$ holds for an action sequence $A^n = a^n$. The input sequence $X^n \in \mathcal{X}^n$ is selected based on both message M and state sequence S^n. The action sequence A^n and the input X^n have to satisfy an average cost constraint defined by a function $\gamma : \mathcal{A} \times \mathcal{X} \to [0, \infty)$, so that the cost for the input sequences a^n and x^n is given by $\gamma(a^n, x^n) = \frac{1}{n}\sum_{i=1}^n \gamma(a_i, x_i)$. Given $X^n = x^n$, $S^n = s^n$ and $A^n = a^n$, the received signal is distributed as $p(y^n|x^n, s^n, a^n) = \prod_{i=1}^n p(y_i|x_i, s_i, a_i)$. The decoder, based on the received signal Y^n, estimates the message M and the sequences $S^n \in \mathcal{S}^n$. The estimate $\hat{S}^n \in \hat{\mathcal{S}}^n$ is constrained to satisfy a distortion criterion defined by a per-symbol distortion metric $d(s, \hat{s}) : \mathcal{S} \times \hat{\mathcal{S}} \to [0, D_{max}]$ with $0 < D_{max} < \infty$. Based on the given distortion metric, the overall distortion for the estimated state sequences $\hat{s}^n$ is defined as $d^n(s^n, \hat{s}^n) = \frac{1}{n}\sum_{i=1}^n d(s_i, \hat{s}_i)$. The reconstructions $\hat{S}^n$ is also required to satisfy the CR constraint, which imposes that the state estimate be also reproducible at the encoder with high probability, as formalized below.

Definition 7.1. An $(n, R, D, \Gamma, \epsilon)$ code for the model in Figure 7.1 consists of an action encoder

$$g_1\colon \mathcal{M} \to \mathcal{A}^n, \tag{7.1}$$

which maps message M into an action sequence A^n; a channel encoder

$$g_2\colon \mathcal{M} \times \mathcal{S}^n \to \mathcal{X}^n, \tag{7.2}$$

which maps message M and the state sequence S^n into the sequence X^n; two decoding functions,

$$h_1\colon \mathcal{Y}^n \to \mathcal{M}, \tag{7.3}$$

$$\text{and } h_2\colon \mathcal{Y}^n \to \hat{\mathcal{S}}^n, \tag{7.4}$$

which map the sequence Y_1^n into the estimated message $\hat{M}$ and into the estimated sequence $\hat{S}^n$, respectively; and a reconstruction function

$$\psi\colon \mathcal{S}^n \to \hat{\mathcal{S}}^n, \tag{7.5}$$

which maps the state sequence into the estimated state sequence at the encoder; such that the probability of error in decoding the message M is small

$$\Pr[\hat{M} \neq M] \quad \leq \quad \epsilon, \tag{7.6}$$

the distortion and cost constraints are satisfied, i.e.,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d(S_i, h_{2i}(Y^n))\right] &\leq D+\epsilon && (7.7)\\ \text{and } \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\gamma(A_i, X_i)\right] &\leq \Gamma+\epsilon, && (7.8) \end{aligned}$$

where $h_{2i}(Y^n) \in \hat{\mathcal{S}}$ is the ith symbol of the sequence $h_2(Y^n)$, and the CR requirement is verified, namely,

$$\Pr\left[\psi(S^n) \neq h_2(Y^n)\right] \leq \epsilon. \tag{7.9}$$

It is noted that, given the definition above, the pmf of the random variables (M, A^n, S^n, X^n, Y^n) factorizes as

$$\begin{aligned} p(m, a^n, s^n, x^n, y^n) &= \frac{1}{2^{nR}}\delta[a^n - g_1(m)]\left\{\prod_{i=1}^{n} p(s_i|a_i)\right\}\delta[x^n - g_2(m, s^n)] \\ &\cdot\left\{\prod_{i=1}^{n} p(y_i|x_i, s_i, a_i)\right\}, \end{aligned} \tag{7.10}$$

where $\delta[\cdot]$ is the Kronecker delta function (i.e., $\delta[x] = 1$ if $x = 0$ and $\delta[x] = 0$ otherwise) and the arguments of the pmf range in the alphabets of the corresponding random variables.

Given a cost-distortion pair (D, Γ), a rate R is said to be achievable if, for any $\epsilon > 0$ and sufficiently large n, there a exists a $(n, R, D, \Gamma, \epsilon)$ code. The goal is to characterize the capacity-distortion-cost trade-off function $C(D, \Gamma) = \inf\{R :$ the triple (R, D, Γ) is achievable$\}$.

7.2.2 Capacity-Distortion-Cost Function

In this section, a single-letter characterization of the capacity-distortion-cost function is derived.

Proposition 7.1. *The capacity-distortion-cost function for the system in Figure 7.1 is given by*

$$C(D,\Gamma) = \max I(U;Y) - I(U;S|A) \tag{7.11}$$

where the mutual informations are evaluated with respect to the joint pmf

$$p(a,u,s,x,y) = p(a)p(s|a)p(u|s,a)p(x|u,s)p(y|x,s,a), \tag{7.12}$$

and minimization is done with respect to the pmfs $p(a)$, $p(u|s,a)$ and $p(x|u,s)$ under the constraint that there exists a deterministic function $\phi : \mathcal{U} \to \hat{\mathcal{S}}$ such that the inequalities

$$\mathrm{E}[d(S,\phi(U))] \leq D \tag{7.13a}$$

$$\text{and } \mathrm{E}[\gamma(A,X)] \leq \Gamma \tag{7.13b}$$

are satisfied. Finally, U is an auxiliary random variable whose alphabet cardinality can be bounded as $|\mathcal{U}| \leq |\mathcal{A}||\mathcal{S}||\mathcal{X}| + 2$.

Remark 7.1. If $D \geq D_{max}$, the result above recovers Theorem 1 of [44]. If instead, $p(s|a) = p(s)$, so that the channel is not action-dependent, Theorem 1 in [8] is recovered.

The proof of achievability follows using the same arguments as in [44] with the difference that here U is also used to estimate the state S via a function $\phi(U)$. The proof of the converse can be found in Appendix R.

7.2.3 A Gaussian Example

In this section, a continous-alphabet version of the model of Figure 7.1 is considered in which the actions and the channel input are subject to the cost constraints $1/n \sum_{i=1}^{n} \mathrm{E}[A^2] \leq P_A$ and $\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[X_i^2\right] \leq P_X$, respectively; the action channel is given by

$$S = A + W, \tag{7.14}$$

where $W \sim \mathcal{N}(0, \sigma_W^2)$ and the transmission channel is given by

$$Y = X + S + Z, \tag{7.15}$$

where $Z \sim \mathcal{N}(0, \sigma_Z^2)$ is independent of W. The rate $I(U;Y) - I(U;S|A)$ in (7.11) is evaluated by assuming the variables (A, S, U, X, Y) to be jointly Gaussian without claiming the optimality of this choice. Specifically, similar to [44, Section VI], let's choose $A \sim \mathcal{N}(0, P_A)$,

$$X = \alpha A + \gamma W + G \tag{7.16a}$$

$$\text{and } U = \delta X + A + \beta W, \tag{7.16b}$$

with $G \sim \mathcal{N}(0, P_X - (\alpha^2 P_A + \gamma^2 \sigma_W^2))$, where the constraint $P_X \geq (\alpha^2 P_A + \gamma^2 \sigma_W^2)$ is enforced, and the variables (A, W, G, Z) are all independent of each other. Then the rate $I(U;Y) - I(U;S|A)$ as in [44] is evaluated, with the additional constraint (7.7) on the state estimate $\hat{S}$. Assuming the quadratic distortion metric $d(s, \hat{s}) = (s - \hat{s})^2$, $\hat{S}$ is chosen to be the MMSE estimate of S given U and A.[2] This leads to the constraint

$$D \geq E[(S - \hat{S})^2] = var(S|U, A) = \sigma_W^2 - \frac{(E[W(U-A)])^2}{E[(U-A)^2]}, \tag{7.17}$$

where $E[W(U - A)] = (\delta\gamma + \beta)\sigma_W^2$ and $E[(U - A)^2] = \delta^2 P_G + (\delta\gamma + \beta)^2 \sigma_W^2$. The rate $I(U;Y) - I(U;S|A)$ optimized over parameters $(\alpha, \beta, \delta, \gamma)$ under the constraint (7.17) for different values of the distortion D for $P_A = P_X = \sigma_W^2 = \sigma_Z^2 = 1$ in Figure (7.3). Moreover for reference, Figure 7.3 shows also the rate achievable if distribution (8.8) is designed to be optimal for message transmission only as in [44, eq. (95)],

[2]Note that U in the characterization of Proposition 2 can be always redefined to include also A without loss of performance, and hence $\hat{S}$ can be made to be a function of U and A.

and the rate achievable, if A is selected to be independent of the message, namely, $\max I(U;Y|A) - I(U;S|A)$, where the mutual information terms are evaluated with respect to the joint Gaussian distribution given above in (7.16) under the constraint (7.17). The performance gains attainable by designing the transmission strategy jointly in the two phases and by accounting for the constraint (7.17) are apparent.

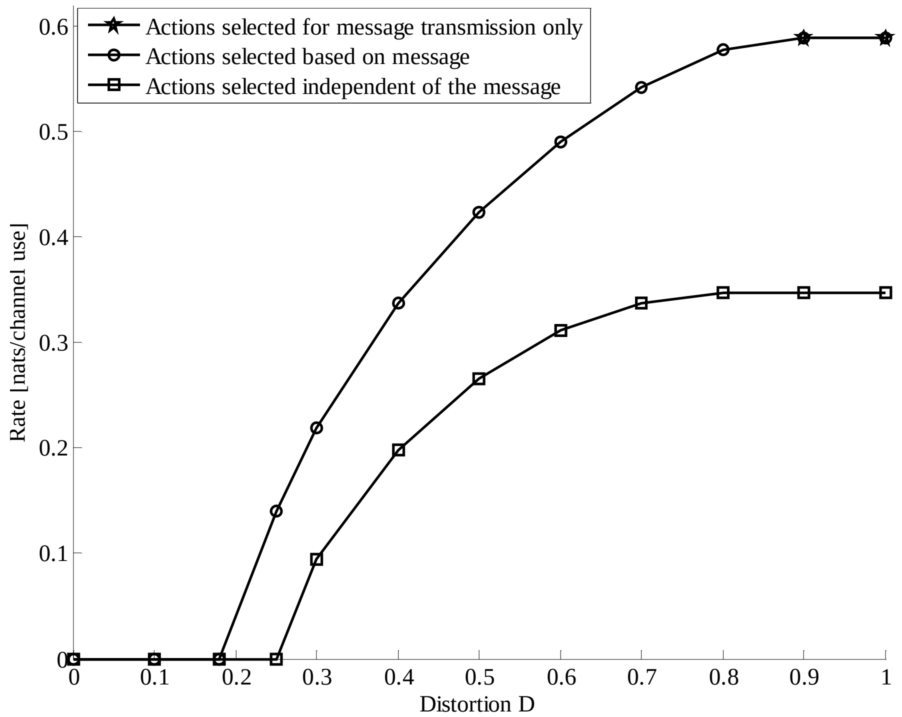

Figure 7.3 Achievable rates (constrained to a Gaussian joint distribution, see (7.16)) for the Gaussian model (7.14)-(7.15) versus distortion D for $P_A = P_X = \sigma_W^2 = \sigma_Z^2 = 1$.

7.3 Degraded Broadcast Channels with Action-Dependent States

In this section, the problem illustrated in Figure 7.2 of a broadcast channel with action-dependent states known causally to the encoder is studied. First, the system model is detailed in Section 7.3.1. Next, the characterization of the capacity region for physically degraded broadcast channels is given in Section 7.3.2. In Section 7.3.4, the special case of a broadcast channel with a probing encoder in the sense of [45] is studied.

7.3.1 System Model

In this section, the system model is detailed. The system is defined by the pmfs $p(x)$, $p(y_1, y_2|x,\ s, a)$, $p(s|a)$ and discrete alphabets $\mathcal{X}$, $\mathcal{A}$, $\mathcal{S}$, and $\mathcal{Y}$ as follows. Given the messages M_1 and M_2, selected randomly from the sets $\mathcal{M}_1 = [1, 2^{nR_1}]$ and $\mathcal{M}_2 = [1, 2^{nR_2}]$, respectively, an action sequence $A^n \in \mathcal{A}^n$ is selected. As a result of this selection, the state sequence $S^n \in \mathcal{S}^n$ is generated as in the previous section. The action sequence A^n and the input X^n have to satisfy the average cost constraint (8.38). Given the transmitted signal $X^n = x^n$, the state sequence $S^n = s^n$, and the action sequence $A^n = a^n$, the received signals are distributed as $p(y_1^n, y_2^n|x^n, s^n, a^n) = \prod_{i=1}^n p(y_{1i}, y_{2i}|x_i, s_i, a_i)$. The decoders, based on the received signals Y_1^n and Y_2^n , estimate the messages M_1 and M_2, respectively.

Definition 7.2. An $(n, R_1, R_2, \Gamma, \epsilon)$ code for the model in Figure 7.2 consists of an action encoder

$$g_1\colon\ \mathcal{M}_1 \times \mathcal{M}_2 \to \mathcal{A}^n, \tag{7.18}$$

which maps messages M_1 and M_2 into an action sequence A^n; a sequence of channel encoders

$$g_{2i}\colon\ \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{S}^i \to \mathcal{X}, \tag{7.19}$$

for $i \in [1, n]$ which map messages M_1 and M_2 and the first i samples of the state sequence S^i into the ith symbol X_i; two decoding functions,

$$h_1\colon\ \mathcal{Y}_1^n \to \mathcal{M}_1, \tag{7.20}$$

$$\text{and } h_2\colon\ \mathcal{Y}_2^n \to \mathcal{M}_2, \tag{7.21}$$

which map the received sequences Y_1^n and Y_2^n into the estimated messages $\hat{M}_1$ and $\hat{M}_2$, respectively; such that the probability of error in decoding the messages M_1 and M_2 is small

$$\Pr[\hat{M}_j \neq M_j] \quad \leq \quad \epsilon \text{ for } j = 1, 2, \tag{7.22}$$

and the cost constraint (8.38) is satisfied.

It is noted that, given the definitions above, the distribution of the random variables $(M_1, M_2, A^n,\ S^n, X^n, Y_1^n, Y_2^n)$ factorizes as

$$p(m_1, m_2, a^n, s^n, x^n, y_1^n, y_2^n) = \\ \frac{1}{2^{n(R_1+R_2)}}\delta[a^n - g_1(m_1, m_2)] \cdot \left\{\prod_{i=1}^{n} p(s_i|a_i)\delta[x_i - g_{2i}(m_1, m_2, s^i)]p(y_{1i}, y_{2i}|x_i, s_i, a_i)\right\}, \tag{7.23}$$

where the arguments of the pmf range in the alphabets of the corresponding random variables.

Given a cost Γ, a rate pair (R_1, R_2) is said to be achievable if, for any $\epsilon > 0$ and sufficiently large n, there a exists a $(n, R_1, R_2, \Gamma, \epsilon)$ code. The capacity region $\mathcal{C}(\Gamma)$ is defined as the closure of all rate pairs (R_1, R_2) that are achievable given the cost Γ.

7.3.2 Capacity-Cost Region

In this section, a single-letter characterization of the capacity region is derived for the special case in which the channel is physically degraded in the sense that the condition below is satisfied

$$p(y_1, y_2|x, s, a) = p(y_1|x, s, a)p(y_2|y_1), \tag{7.24}$$

or equivalently, the Markov chain $(X_i, S_i, A_i) - Y_{1i} - Y_{2i}$ holds for all $i \in [1, n]$.

Proposition 7.2. *The capacity region of the system in Figure 7.2 under the degradedness condition (7.24) is given by the union of the rate pairs* (R_1, R_2) *satisfying*

$$R_1 \leq I(U_1; Y_1|U_2) \tag{7.25a}$$

$$\text{and } R_2 \leq I(U_2; Y_2), \tag{7.25b}$$

where the mutual informations are evaluated with respect to the joint pmf

$$p(a, u_1, u_2, s, x, y_1, y_2) = p(u_1, u_2)$$
$$\cdot\, \delta[a - f_a(u_1, u_2)]p(s|a)\delta[x - f_x(u_1, u_2, s)]p(y_1|x, s, a)p(y_2|y_1), \tag{7.26}$$

for some pmfs $p(u_1, u_2)$ *and deterministic functions* f_a*:* $\mathcal{U}_1 \times \mathcal{U}_2 \to \mathcal{A}$ *and* f_x*:* $\mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{S} \to \mathcal{X}$ *such that the inequality* $\mathrm{E}[\gamma(A, X)] \leq \Gamma$ *is satisfied. Auxiliary random variables* U_1 *and* U_2 *have finite alphabets.*

The proof of achievability can be sketched as follows. The codewords $u_2^n(m_2)$, encoding message $m_2 \in [1, 2^{nR_2}]$, are generated independently and i.i.d. according to the pmf $p(u_2)$. Then, superimposed on each codeword $u_2^n(m_2)$, 2^{nR_1} codewords $u_1^n(m_1, m_2)$ are generated independently according to the distribution $\prod_{i=1}^n p(u_{1i}|u_{2i}(m_2))$. To encode messages (M_1, M_2), the action sequence A^n is obtained as a deterministic function of $u_{1i}(M_1, M_2)$ and $u_{2i}(M_2)$ such that $A_i = f_a(u_{1i}(M_1, M_2), u_{2i}(M_2))$ for all $i \in [1, n]$. The transmitted symbol X_i is obtained instead as a function of $u_{1i}(M_1, M_2)$, $u_{2i}(M_2)$, and of the ith state symbol S_i as $X_i = f_x(u_{1i}(M_1, M_2), u_{2i}(M_2), S_i)$. Decoder 2 decodes the codeword $u_2^n(m_2)$, while decoder 1 decodes both codewords $u_2^n(m_2)$ and $u_1^n(m_1, m_2)$. Using standard arguments, the rates (7.25) are easily shown to be achievable. The proof of the converse can be found in Appendix S.

Remark 7.2. If $p(s|a) = p(s)$ so that the channel is not action-dependent, Proposition 6 recovers Proposition 4 of [51] (see also [52]).

7.3.3 A Binary Example

In this section, let's consider a special case of the model in Figure 7.2 in which the action channel $p(s|a)$ is binary and given by

$$S = A \oplus B, \tag{7.27}$$

where the action A is binary, $B \sim \text{Ber}(b)$ and the transmission channels are given by

$$\begin{aligned} Y_1 &= X \oplus S \oplus Z_1, && (7.28\text{a}) \\ \text{and } Y_2 &= Y_1 \oplus \widetilde{Z}_2, && (7.28\text{b}) \end{aligned}$$

where $Z_1 \sim \text{Ber}(N_1)$ and $\widetilde{Z}_2 \sim \text{Ber}(\widetilde{N}_2)$ are independent of each other and of B. The cost metric is selected as $\gamma(a, x) = x$. Moreover, $N_2 = N_1 * \widetilde{N}_2 = N_1(1 - \widetilde{N}_2) + \widetilde{N}_2(1 - N_1)$.

As a first remark, consider the ideal system with $b = 0$ (i.e., no interference) and no cost constraint (i.e., $\Gamma = 1/2$). The system reduces to a standard physical degraded binary symmetric broadcast channel, and thus the capacity region is given by the union over $\alpha \in [0, 0.5]$ of the rate pairs satisfying the inequalities [1, p. 115]

$$\begin{aligned} R_1 &\leq H(\alpha * N_1) - H(N_1) && (7.29\text{a}) \\ \text{and } R_2 &\leq 1 - H(\alpha * N_2). && (7.29\text{b}) \end{aligned}$$

It is observed that, by construction, this rate region sets an outer bound on the rate achievable in the system at hand. The outer bound above is in fact achievable by setting $X = B$, $U_2 \sim \text{Ber}(1/2)$, $U_1 = U_2 \oplus \widetilde{U}_1$ with $\widetilde{U}_1 \sim \text{Ber}(\alpha)$, and $A = U_1$ in (7.25), where U_2 and $\widetilde{U}_1$ are independent. This entails that, by leveraging the actions, the interference-free capacity region (7.29) is obtained for all cost constraints $\Gamma \geq b$. It can be instead seen that, if one is forced to set A to be constant, achieving the rate region (7.29) requires a cost $\Gamma = 1/2$, since X needs to be distributed $\text{Ber}(1/2)$. This example illustrates the advantage of being able to affect the state via actions selected as a function of the messages.

7.3.4 Probing Capacity of Degraded Broadcast Channels

In this section, the setting of probing capacity introduced in [45] is applied to the degraded broadcast channel. Following [45], the state sequence S^n is thus assumed to be generated i.i.d. according to a pmf $p(s)$. Moreover, based on the messages (M_1, M_2), the encoder selects an action sequence as in (7.18). However, here, through the choice of actions, the encoder affects the state *information* available at the encoder and the decoders, and not the state sequence S^n. Specifically, the encoder obtains partial state information $S_{e,i} = b_e(S_i, A_i)$, and the decoders obtain partial state informations $S_{d_{1,i}} = b_{d_1}(S_i, A_i)$ and $S_{d_{2,i}} = b_{d_2}(S_{d_{1,i}})$, respectively, where $i \in [1, n]$, and $b_e : \mathcal{S} \times \mathcal{A} \to \mathcal{S}_e$, $b_{d_1} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}_{d_1}$ and $b_{d_2} : \mathcal{S}_{d_1} \to \mathcal{S}_{d_2}$ are deterministic functions for given alphabets $\mathcal{S}_e$, $\mathcal{S}_{d_1}$ and $\mathcal{S}_{d_2}$. Note that the state information available at decoder 2 is degraded with respect to that of decoder 1 (i.e., it is a function of the latter). As in [45], it is assumed that the state information at the encoder is characterized as

$$S_{e,i} = b_e(S_i, A_i) \quad = \quad \begin{cases} S_i & \text{if} \quad A_i = 1 \\ * & \text{if} \quad A_i = 0 \end{cases}, \tag{7.30}$$

where $*$ represents the absence of channel state information at the encoder. Moreover, the state information $S_{e,i}$ is assumed to be available causally at the encoder so that the encoding function is g_{2i}: $\mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{S}_e^i \to \mathcal{X}$ (cf. (7.19)). The rest of the code definition is similar to Definition 7.2 with the caveat that the decoder 1 and 2 have available also the information sequences $S_{d_1}^n$ and $S_{d_2}^n$, respectively.

It is noted that, given the definitions above, the distribution of the random variables $(M_1, M_2, A^n,\ S^n, X^n, Y_1^n, Y_2^n)$ factorizes as

$$p(m_1, m_2, a^n, s^n, s_e^n, s_{d_1}^n, s_{d_2}^n, x^n, y_1^n, y_2^n) = \frac{1}{2^{n(R_1+R_2)}}\delta[a^n - g_1(m_1, m_2)]$$
$$\cdot \left\{ \prod_{i=1}^{n} p(s_i)\delta[s_{e,i} - b_e(s_i, a_i)]\delta[s_{d_{1,i}} - b_{d_1}(s_i, a_i)]\delta[x_i - g_2(m_1, m_2, s_e^i) \right\}$$
$$\cdot \left\{ \prod_{i=1}^{n} \delta[s_{d_{2,i}} - b_{d_2}(s_{d_{1,i}})]p(y_{1i}, y_{2i}|x_i, s_i, a_i) \right\} \tag{7.31}$$

where the arguments of the pmf range in the alphabets of the corresponding random variables.

As discussed below, the setting at hand, which is referred to as having a probing encoder, is a special case of the one studied in Section 8.4.1. Therefore, one can leverage Proposition 7.2 to obtain the following result.

Proposition 7.3. *The capacity region of the system in Figure 7.2 under the degradedness condition (7.24) and with a probing encoder is given by the union of the rate pairs (R_1, R_2) satisfying*

$$R_1 \leq I(U_1; Y_1, S_{d_1}|U_2) \tag{7.32a}$$
$$\textit{and } R_2 \leq I(U_2; Y_2, S_{d_2}) \tag{7.32b}$$

where the mutual informations are evaluated with respect to the joint pmf

$$p(a, u_1, u_2, s, s_e, s_{d_1}, s_{d_2}, x, y_1, y_2) = p(s)p(u_1, u_2)$$
$$\delta[a - f_a(u_1, u_2)]\delta[s_e - b_e(s, a)] \cdot \delta[s_{d_1} - b_{d_1}(s, a)]$$
$$\delta[s_{d_2} - b_{d_2}(s_{d_1})] \cdot \delta[x - f_x(u_1, u_2, s_e)]p(y_1|x, s, a)p(y_2|y_1), \tag{7.33}$$

for some pmf $p(u_1, u_2)$ and deterministic functions f_a: $\mathcal{U}_1 \times \mathcal{U}_2 \to \mathcal{A}$ and f_x: $\mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{S}_e \to \mathcal{X}$ such that the inequality $\mathrm{E}\left[\gamma(A, X)\right] \leq \Gamma$ is satisfied.

Proof. The result is obtained by noticing that the setting described above is a special case of the one described in Section 7.3.1 by making the following substitutions

$$
\begin{aligned}
S &\rightarrow S_e && \text{(7.34a)}\\
\text{and } Y_j &\rightarrow (Y_j, S_{d_j}) \text{ for } j=1,2. && \text{(7.34b)}
\end{aligned}
$$

To see this, it will be shown that the pmf (7.31) reduces to (7.23) under the given substitutions. Specifically, by marginalizing (7.31) over S^n it is true that

$$
\frac{1}{2^{n(R_1+R_2)}}\delta[a^n - g_1(m_1,m_2)]\left\{\prod_{i=1}^{n}\delta[x_i - g_2(m_1,m_2,s_e^i)]\delta[s_{d_{2,i}} - b_{d_2}(s_{d_{1,i}})]p(y_{2i}|y_{1i})\right\}
\prod_{i=1}^{n}\sum_{s_i\in\mathcal{S}}\{p(s_i)\delta[s_{e,i}-b_e(s_i,a_i)]\delta[s_{d_{1,i}}-b_{d_1}(s_i,a_i)]p(y_{1i}|x_i,s_i,a_i)\}\quad(7.35)
$$

The terms outside the summation in (7.35) are equal to the corresponding terms in (7.23) under the substitutions (7.34). For the remaining terms, it is observed that, for $a_i = 0$, $S_{e,i} = *$, and thus $p(s_{e,i}|a_i) = \delta[s_{e,i} - *]$ and

$$
p(y_{1i}, s_{d_{1i}}|x_i, s_{e,i}, a_i) = \sum_{s_i\in\mathcal{S}} p(s_i)\delta[s_{d_{1,i}} - b_{d_1}(s_i,a_i)]p(y_{1i}|x_i,s_i,a_i);
$$

instead, for $a_i = 1$, $S_{e,i} = S_i$, and thus $p(s_{e,i}|a_i) = \Pr[S_i = s_{e,i}]$ and

$$
p(y_{1i}, s_{d_{1i}}|x_i, s_{e,i}, a_i) = \delta[s_{d_{1,i}} - b_{d_1}(s_{e,i},a_i)]p(y_{1i}|x_i,s_{e,i},a_i),
$$

which completes the proof. □

7.4 Concluding Remarks

Action-dependent channels are useful abstractions of two-phase communication scenarios. This chapter has reported on two variations on this theme, namely the problem of message and state transmission in an action-dependent channel and the degraded action-dependent broadcast channel. Under given assumptions, the information-theoretic performance of these systems is characterized. The analytical results, and specific examples, emphasize the importance of jointly designing the transmission strategy across the two communication phases.

CHAPTER 8

INFORMATION EMBEDDING ON ACTIONS

8.1 Introduction

As reviewed in Chapter 4, the recent works [4, 44] study the problem of *optimal actuation for source and channel coding* for resource-constrained systems. Specifically, in [4], an extension of the Wyner-Ziv source coding problem is considered in which the decoder or the encoder can take actions that affect the quality of the side information available at the decoder's side. When the actions are taken by the decoder, the latter operates in two stages. In the first stage, based on the message received from the encoder, the decoder selects cost-constrained actions A that affect the measurement of the side information Y. This effect is modeled by a channel $p_{Y|X,A}(y|x,a)$, where X represents the source available at the encoder. In the second stage, the decoder produces an estimate of source X based on the side information Y as in the standard Wyner-Ziv problem (see, e.g., [1]). A similar formulation also applies when the actions are taken at the encoder's side. This model can account, as an example, for computer networks in which the acquisition of side information from remote data bases is costly in terms of system resources and thus should be done efficiently. This class of problems are referred to as having *actions for side information acquisition.*

As discussed in Chapter 7, in [44], a related channel coding problem is studied in which the encoder in a point-to-point channel can take actions to affect the state of a channel. The encoder operates in two stages. In the first stage, based on the message to be conveyed to the decoder, cost-constrained actions A are selected by the encoder that affect the channel state S of the channel $p_{Y|X,S}(y|x,s)$ used for communication to the decoder in the second stage. In the second stage, the channel $p_{Y|X,S}(y|x,s)$ is used in a standard way based on the available information about the state S (which can be non-causal or causal, see, e.g., [1]). This problem is referred to as having *actions for channel state control*. As shown in [45], this model can be used to account for an encoder that in the first stage *probes* the channel to acquire state information.

8.1.1 Information Embedding on Actions

As discussed above, *optimal actuation for channel and source coding*, as proposed in [4, 44], prescribes the selection of the actions A towards the goal of improving the performance of the resource-constrained communication link between encoder and decoder. This can be done by acquiring side information in an efficient way for source coding problems, and by controlling or probing effectively the channel state for channel coding problems.

This work starts from the observations that the actions A often entail the use of physical resources for communication within the system encompassing the link under study. For instance, acquiring information from a data base requires the receiver to exchange control signals with a server, and probing the congestion state of a network (modelled as a channel) requires transmission of training packets to the closest router. In all these cases, the "recipient" of the actions, e.g., the server or a router in the examples above, may request to obtain partial information about the source or message being communicated on the link. To illustrate this point, the server in the data base application might need to acquire some explicit information about the file being transmitted in the link before granting access to the server. Similarly, the router might need to obtain the header of the packet (message) that the transmitter intends to deliver to the end receiver.

In the scenarios discussed above, the action A thus serves a double purpose: on the one hand, it should be designed to improve the performance of the communication link at hand as in [4, 44, 45], and, on the other, it should provide explicit information about source or message for a separate decoder (the server or router in the examples above). A relevant question thus is: *How much information can be embedded in the actions A without affecting the performance of the link*? Or, to turn the question around, *what is the performance loss for the link as a function of the amount of information that is encoded in the actions A*? This work aims at answering these questions for both the source and channel coding scenarios discussed above (see Figure 8.1, Figure 8.2, Figure 8.3 and Section 8.1.3).

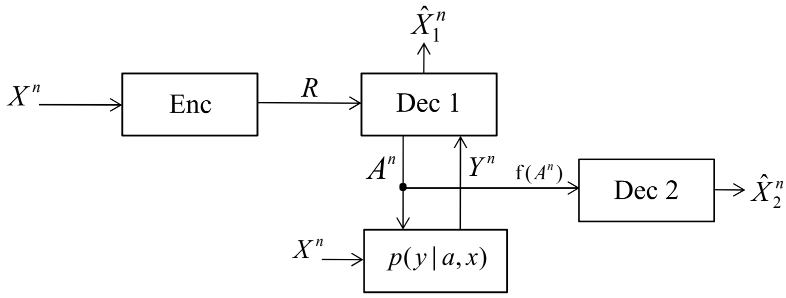

Figure 8.1 Source coding with decoder-side actions for information acquisition and with information embedding on actions. A function of the actions $\mathrm{f}(A^n) = (\mathrm{f}(A_1), ..., \mathrm{f}(A_n))$ is observed in full ("non-causally") by Decoder 2 before decoding. See Figure 8.4 and Figure 8.5 for the corresponding models with strictly causal and causal observation of the actions at Decoder 2, respectively.

8.1.2 Related Work

The interplay between communication and actuation, or control, is recognized to arise at different levels. As mentioned, the main theme in the papers [4, 44, 45] is "*control for communication*": As reviewed in Chapter 4, Chapter 5 and Chapter 6,in [4, 44, 45], actuation is instrumental in improving the performance of a resource-constrained communication system. Extensions of this research direction include models with additional design constraints [24, 28], with adaptive actions [26], with multiple terminals [23], [54] (see Chapter 4) and [55] (see Chapter 5) and with memory [56, 28]. Somewhat related, but distinct, is the line of work including [57]-[58], in which control-theoretic tools are leveraged to design effective communication schemes. An altogether different theme is instead central in work such as [59, 60] that can be referred to as "*communication for control*". In fact, in a reversed way, in [59, 60] (and references therein), communication is instrumental in carrying out control tasks such as stabilization of a plant. For instance, [60] shows that an implicit message communicated between two controllers can greatly improve the performance of the control task.

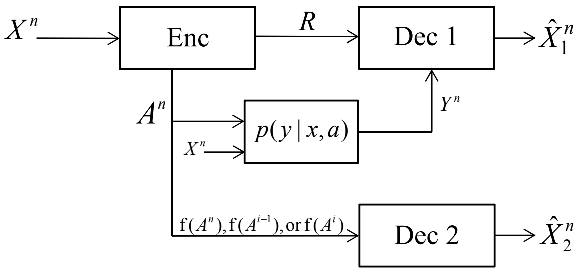

Figure 8.2 Source coding with encoder-side actions for information acquisition and with information embedding on actions.

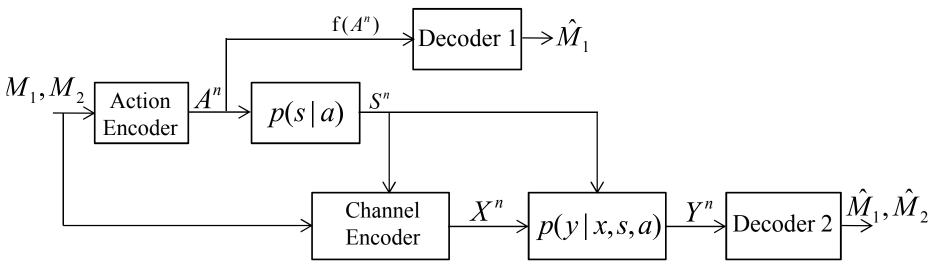

Figure 8.3 Channel coding with actions for channel state control and with information embedding on actions.

The idea of embedding information in the actions is related to the classical problem of information hiding (see, e.g., [61] and references therein). In information hiding, a message is embedded in a host data under distortion constraints. The message is then retrieved by a decoder that observes the host signal through a noisy channel. Note that the (host) signal onto which the message is embedded is a given process. Instead, in the set-up of information embedding on actions considered here, the (action) signal on which information is embedded is designed to optimize the given communication task.

The set-up at hand is also related to the source coding model of [62], in which an encoder communicates to two decoders and one of the decoders is able to observe the source estimate produced by the other. For its duality with the classical channel coding model studied in [63], the operation of the first decoder was referred to in [62] as *cribbing*. Although the problem of interest here (in the source coding part) is significantly different, in that the recipient of the embedded information is a decoder "cribbing" the actions and not the estimates of another decoder, the solutions of the two problems turn out to be related, as it will be discussed.

8.1.3 Contributions and Chapter Organization

The main contributions of this chapter are as follows.

- **Decoder-side actions for side information acquisition:** First, the model in Figure 8.1 is considered, in which the problem of source coding with actions taken at the decoder (Decoder 1) [4] is generalized by including an additional decoder (Decoder 2). Decoder 2 is the recipient of a function of the action sequence and is interested in reconstructing a lossy version of the source measured at the encoder. A single-letter characterization of the set of all achievable tuples of rate, distortions at the two decoders and action cost is derived in Section 8.2 under the assumption that Decoder 2 observes a function of the actions non-causally (Section 8.2.2), strictly causally (Section 8.2.3) or causally (Section 8.2.4). An example is provided to shed light into the effect of information embedding on actions in Section 8.2.5;

- **Encoder-side actions for side information acquisition:** Next, the set-up in Figure 8.2 is considered, in which an additional decoder observing the actions is added to the problem of source coding with actions taken at the encoder [4].

Section 8.3 derives the achievable rate-distortion-cost region in the special case in which the channel $p_{Y|X,A}(y|x,a)$ with source and action (X, A) as inputs and side information Y as output is such that Y is a deterministic function of A;

- **Actions for channel control and probing:** Finally, the impact of information embedding on actions for channel control is considered by studying the set-up in Figure 8.3, which generalizes [44]. Specifically, a decoder (Decoder 1) is added to the model in [44], that observes a function of the actions taken by the encoder and wishes to decode part of the message that is intended for the channel decoder (Decoder 2). A single-letter characterization of the achievable capacity-cost region is obtained in Section 8.4. Finally, the special case of actions for channel probing [45] is elaborated on with an example in Section 8.4.3.

8.2 Decoder-Side Actions for Side Information Acquisition

In this section, first, the system model is described for the set-up illustrated in Figure 8.1, Figure 8.4 and Figure 8.5 of source coding with decoder-side actions. Then, a single letter characterization of the set of all achievable tuples of rate, distortions at the two decoders and action cost is derived under the assumption that Decoder 2 observes the actions fully (non-causally) in Section 8.2.2, strictly causally in Section 8.2.3 and causally in Section 8.2.4. An example is provided in Section 8.2.5.

8.2.1 System Model

Here the problem corresponding to full observation of a function of the actions is presented as per Figure 8.1. This model is referred to as having non-causal action observation. The changes necessary to account for causal or strictly causal as illustrated in Figure 8.4 and Figure 8.5 will be discussed in the appropriate sections later. It is remarked that this definition does not entail any non-causal operation, but only a larger estimation delay for Decoder 2 as compared to the causal cases in Figure 8.4 and Figure 8.5. The model is defined by the probability mass functions (pmfs) $p_X(x)$ and $p_{Y|AX}(y|a,x)$, by the function f: $\mathcal{A} \rightarrow \mathcal{B}$, and by discrete alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}, \hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2$, as follows. The source sequence X^n is such that $X_i \in \mathcal{X}$ for $i \in [1,n]$ is independent and identically distributed (i.i.d.) with pmf $p_X(x)$. The Encoder measures sequence X^n and encodes it in a message M of nR bits, which is delivered to Decoder 1. Decoder 1 receives message M and selects an action sequence A^n, where $A^n \in \mathcal{A}^n$. The action sequence affects the quality of the measurement Y^n of sequence X^n obtained at the Decoder 1. Specifically, given $A^n = a^n$ and $X^n = x^n$, the sequence Y^n is distributed as $p(y^n|a^n,x^n) = \prod_{i=1}^n p_{Y|A,X}(y_i|a_i,x_i)$. The cost of the action sequence is defined by a cost function Λ: $\mathcal{A} \rightarrow [0,\Lambda_{\max}]$ with $0 \leq \Lambda_{\max} < \infty$, as $\Lambda(a^n) = \sum_{i=1}^n \Lambda(a_i)$. The estimated sequence $\hat{X}_1^n \in \hat{\mathcal{X}}_1^n$ is then obtained as a function of M and Y^n. Decoder 2 observes a function of the action sequence A^n, thus obtaining $\mathrm{f}(A^n) = (\mathrm{f}(A_1), ..., \mathrm{f}(A_n)) \in \mathcal{B}^n$. Based on $\mathrm{f}(A^n)$, Decoder 2 obtains an estimate $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ within given distortion requirements. The estimated sequences $\hat{X}_j^n$ for $j = 1,2$ must satisfy distortion constraints defined by functions $d_j(x,\hat{x}_j)$: $\mathcal{X} \times \hat{\mathcal{X}}_j \rightarrow [0, D_{j,\max}]$ with $0 \leq D_{j,\max} < \infty$ for $j = 1,2$, respectively. A formal description of the operations at encoder and decoder follows.

Definition 8.1. An (n, R, D_1, D_2, Γ) code for the set-up of Figure 8.1 consists of a source encoder

$$\mathrm{h}^{(e)}\colon \mathcal{X}^n \rightarrow [1, 2^{nR}], \tag{8.1}$$

which maps the sequence X^n into a message M; an "action" function

$$\mathrm{h}^{(a)}\colon [1, 2^{nR}] \rightarrow \mathcal{A}^n, \tag{8.2}$$

which maps the message M into an action sequence A^n; two decoders, namely

$$\mathrm{h}_1^{(d)}\colon [1, 2^{nR}] \times \mathcal{Y}^n \rightarrow \hat{\mathcal{X}}_1^n, \tag{8.3}$$

which maps the message M and the measured sequence Y^n into the estimated sequence $\hat{X}_1^n$;

$$\mathrm{h}_2^{(d)}\colon\ \mathcal{B}^n \to \hat{\mathcal{X}}_2^n, \tag{8.4}$$

which maps the observed sequence $\mathrm{f}(A^n)$ into the the estimated sequence $\hat{X}_2^n$; such that the action cost constraint Γ and distortion constraints D_j for $j = 1, 2$ are satisfied, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\Lambda(A_i)\right] \leq \Gamma \tag{8.5}$$

$$\text{and } \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d_j(X_i, \hat{X}_{ji})\right] \leq D_j \text{ for } j = 1, 2. \tag{8.6}$$

Definition 8.2. Given a distortion-cost tuple (D_1, D_2, Γ), a rate R is said to be achievable if, for any $\epsilon > 0$, and sufficiently large n, there exists a $(n, R, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)$ code.

Definition 8.3. The *rate-distortion-cost function* $R(D_1, D_2, \Gamma)$ is defined as $R(D_1, D_2, \Gamma) = \inf\{R :$ the tuple $(R, D_1, D_2, \Gamma)$is achievable$\}$.

8.2.2 Non-Causal Action Observation

In this section, a single-letter characterization of the rate-distortion region is derived for the set-up in Figure 8.1 in which Decoder 1 observes the entire sequence $\mathrm{f}^n(A^n)$ prior to decoding.

Proposition 8.1. *The rate-distortion-cost function $R(D_1, D_2, \Gamma)$ for the source coding problem with decoder-side actions and non-causal observation of the actions at Decoder 2 illustrated in Figure 8.1 is given by*

$$R(D_1, D_2, \Gamma) = \min_{p(\hat{x}_2, a, u|x),\ \mathrm{g}(U,Y)} I(X; \hat{X}_2, A) + I(X; U|\hat{X}_2, A, Y), \tag{8.7}$$

where the mutual information is evaluated with respect to the joint pmf

$$p(x, y, a, \hat{x}_2, u) = p(x)p(\hat{x}_2, a, u|x)p(y|x, a), \tag{8.8}$$

for some pmf $p(\hat{x}_2, a, u|x)$ *such that the inequalities*

$$\begin{aligned} \mathrm{E}[d_j(X, \hat{X}_j)] &\leq D_j, \text{ for } j = 1, 2, & (8.9a) \\ \mathrm{E}[\Lambda(A)] &\leq \Gamma, & (8.9b) \\ \text{and } I(X; \hat{X}_2, \mathrm{f}(A)) &\leq H(\mathrm{f}(A)) & (8.9c) \end{aligned}$$

are satisfied for $\hat{X}_1 = \mathrm{g}(U, Y)$ *for some function* $\mathrm{g}\colon \mathcal{U} \times \mathcal{Y} \to \hat{\mathcal{X}}_1$. *Finally,* U *is an auxiliary random variable whose alphabet cardinality can be constrained as* $|\mathcal{U}| \leq |\mathcal{X}||\hat{\mathcal{X}}_2||\mathcal{A}| + 1$ *without loss of optimality.*

At an intuitive level, in (8.7), the term $I(X; \hat{X}_2, A)$ accounts for the rate needed to instruct Decoder 1 about the actions A to be taken for the acquisition of the side information Y, which are selected on the basis of the source X, and, at the same time, to communicate the reconstruction $\hat{X}_2$ to Decoder 2. The additional rate $I(X; U|\hat{X}_2, A, Y)$ is instead required to refine the description of the source X provided via $(\hat{X}_2, A)$ using an auxiliary codebook U for Decoder 1. Note that this rate is conditioned on the side information Y, thanks to the rate saving obtained through Wyner-Ziv binning. The condition (8.9c) ensures that, based on the observation of $\mathrm{f}(A)$, Decoder 2 is able to reconstruct $\hat{X}_2$. The details of achievability follow as a combination of the techniques proposed in [4] and [64, 62]. Below, the main ideas are briefly outlined, since the technical details follow from standard arguments. The proof of the converse is provided in Appendix T.

Sketch of the achievability proof: Fix a pmf (8.8) and define a random variable $B = \mathrm{f}(A)$. The joint pmf $p(x, y, a, \hat{x}_2, u, b)$ of variables $(X, Y, A, \hat{X}_2, U, B)$ is obtained by multiplying the right-hand side of (8.8) by the term[1] $\mathbf{1}_{\{b=\mathrm{f}(a)\}}$. In the scheme at

[1]The notation $\mathbf{1}_{\{S\}}$ is used for the indicator function of the event S.

hand, the Encoder first maps sequence X^n into a sequence $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ using the joint typicality criterion with respect to the joint pmf $p(x, \hat{x}_2)$. This mapping requires a codebook of rate $I(X; \hat{X}_2)$ (see, e.g., [1, pp. 62-63]). Given the sequence $\hat{X}_2^n$, the sequence X^n is further mapped into a sequence $B^n \in \mathcal{B}^n$ using the joint typicality criterion with respect to the joint pmf $p(x, b|\hat{x}_2)$ where $B = \mathrm{f}(A)$, which requires a codebook of rate $I(X; \mathrm{f}(A)|\hat{X}_2)$ for each sequence $\hat{X}_2^n$. For later reference, every such codebook is referred to as a bin in the following. Note that there exists one bin for every sequence $\hat{X}_2^n$. For each pair $(\hat{X}_2^n, B^n)$, the sequence X^n is mapped into an action sequence A^n using joint typicality with respect to the joint pmf $p(x, a|\hat{x}_2, b)$, which requires a codebook of rate $I(X; A|\hat{X}_2, \mathrm{f}(A))$. Note that, by construction, $B^n = \mathrm{f}(A^n)$ for each generated A^n. Finally, the source sequence X^n is mapped into a sequence U^n using the joint typicality criterion with respect to the joint pmf $p(x, u|\hat{x}_2, a)$, which requires a codebook of rate $I(X; U|\hat{X}_2, A)$ for each pair $(\hat{X}_2^n, A^n)$.

The indices of codewords $\hat{X}_2^n$, B^n and A^n are sent to Decoder 1, along with the index for the codeword U^n. For the latter, by leveraging the side information Y^n available at Decoder 1, the rate can be reduced to $I(X; U|\hat{X}_2, A, Y)$ by the Wyner-Ziv theorem [1, p. 280]. Decoder 2 estimates the sequence $\hat{X}_2^n$ from the observed sequence $\mathrm{f}(A^n)$ as follows: if there is only one bin containing the observed sequence $\mathrm{f}(A^n)$, then $\hat{X}_2^n$ equals the sequence corresponding to such bin (recall that each bin corresponds to one sequence $\hat{X}_2^n$). Otherwise, an error is decoded. To obtain a vanishing probability of error, the sequence $\mathrm{f}^n(A^n)$ should thus not lie within more than one bin with high probability. The probability of the latter event can be upper bounded by $2^{n(I(X;\hat{X}_2,\mathrm{f}(A))-H(\mathrm{f}(A))}$ since each sequence B^n is generated with probability approximately $2^{-nH(\mathrm{f}(A))}$ and there are $2^{nI(X;\hat{X}_2,\mathrm{f}(A))}$ sequences B^n [62]. Therefore, as long as $I(X; \hat{X}_2, \mathrm{f}(A)) \leq H(\mathrm{f}(A))$, Decoder 1 is able to infer the conveyed bin index with high probability. Finally, Decoder 1 produces the estimate $\hat{X}_1^n$ through a symbol-by-symbol function as $\hat{X}_{1i} = \mathrm{g}(U_i, Y_i)$ for $i \in [1, n]$. $\square$

Remark 8.1. Assume that the action A_i is allowed to be a function, not only of the message M as per (8.2), but also of the previous values of the side information Y^{i-1}, which is referred to as adaptive actions. Then, the rate-distortion-cost function derived in Proposition 8.2 can generally be improved. This can be seen by considering the case in which $R = 0$. In this case, if the actions were selected as per (8.2), then the distortion at Decoder 2 would be forced to be maximal, i.e., $D_2 = D_{2,\max}$, since the actions A cannot depend in any way on the source X. Instead, by selecting A as a function of the previously observed values of Y, Decoder 1 can provide Decoder 2 with information about X, thus decreasing the distortion D_2. It is noted that the usefulness of adaptive actions in this setting contrasts with the known fact that, in the absence of Decoder 2, adaptive actions do not decrease the rate-distortion function [26].

8.2.3 Strictly Causal Action Observation

The system model for the set-up in Figure 8.4, is similar to the one described in Section 8.2.1 with the only difference the decoding function for Decoder 2 a time i is given as

$$\mathrm{h}_{2i}^{(d)}\colon \mathcal{B}^{i-1} \to \hat{\mathcal{X}}_2, \tag{8.10}$$

which maps the strictly causally observed sequence $\mathrm{f}(A^{i-1}) = (\mathrm{f}(A_1), ..., \mathrm{f}(A_{i-1}))$ into the ith estimated symbol $\hat{X}_{2i}$.

Proposition 8.2. *The rate-distortion-cost function $R(D_1, D_2, \Gamma)$ for the source coding problem with decoder-side actions and strictly causal observation of the actions at Decoder 1 as illustrated in Figure 8.4 is given by*

$$R(D_1, D_2, \Gamma) = \min_{p(\hat{x}_2,a,u|x),\ \mathrm{g}(U,Y)} I(X;\hat{X}_2, A) + I(X;U|\hat{X}_2, A, Y), \tag{8.11}$$

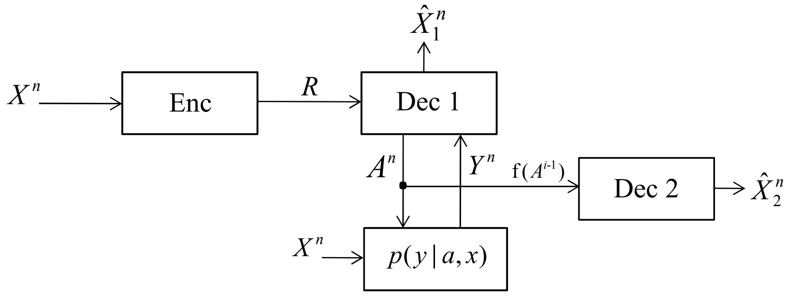

Figure 8.4 Source coding with decoder-side actions for information acquisition and with information embedding on actions. At time i, Decoder 2 has available the samples $\mathrm{f}(A^{i-1}) = (\mathrm{f}(A_1), ..., \mathrm{f}(A_{i-1}))$ in a strictly causal fashion.

where the mutual information is evaluated with respect to the joint pmf

$$p(x, y, a, \hat{x}_2, u) = p(x)p(\hat{x}_2, a, u|x)p(y|x, a), \tag{8.12}$$

for some pmf $p(\hat{x}_2, a, u|x)$ *such that the inequalities*

$$\mathrm{E}[d_j(X, \hat{X}_j)] \leq D_j, \text{ for } j = 1, 2, \tag{8.13a}$$

$$\mathrm{E}[\Lambda(A)] \leq \Gamma, \tag{8.13b}$$

$$\text{and } I(X; \hat{X}_2, \mathrm{f}(A)) \leq H(\mathrm{f}(A)|\hat{X}_2) \tag{8.13c}$$

are satisfied for $\hat{X}_1 = \mathrm{g}(U, Y)$ *for some function* $\mathrm{g}\colon \mathcal{U} \times \mathcal{Y} \to \hat{\mathcal{X}}_1$. *Finally,* U *is an auxiliary random variable whose alphabet cardinality can be constrained as* $|\mathcal{U}| \leq |\mathcal{X}||\hat{\mathcal{X}}_2||\mathcal{A}| + 1$ *without loss of optimality.*

The only difference between the rate-distortion-cost function of Proposition 8.1 with non-causal action observation with respect to the case with strictly causal action observation of Proposition 8.2 is the constraint (8.13c). Recall that the latter is needed to ensure that Decoder 2 is able to recover the reconstruction $\hat{X}_2$. As detailed below, the strict causality of the observation of the action at Decoder 2 calls for a block-based encoding in which the actions carries information about the source sequence as observed in two different blocks, namely the current block for Decoder 1 and the future block for Decoder 2. This additional requirement causes the conditioning on $\hat{X}_2$ in (8.13c), which generally increases the rate (8.11) with respect to the counterpart (8.20) achievable with non-causal action observation. A sketch of the achievability proof is provided below and is based on the techniques proposed in [64, 62] (see also [65]). The proof of the converse is provided in Appendix U.

Sketch of the achievability proof: Fix a pmf (8.12) and define a random variable $B = \mathrm{f}(A)$. The joint pmf $p(x, y, a, \hat{x}_2, u, b)$ of variables $(X, Y, A, \hat{X}_2, U, B)$ is obtained by multiplying the right-hand side of (8.12) by the term $1_{\{b=\mathrm{f}(a)\}}$. The "Forward Encoding" and "Block Markov Decoding" strategy of [64, 62] (see also [65]) along with the coding scheme of [4] is used. The scheme operates over multiple blocks and $X^n(l)$ is the portion of the source sequence encoded in block l. The sequence $\mathrm{f}(A^n(l))$ observed during block l is used in block $l+1$ by Decoder 2 due to the strict causality constraint. To this end, the action sequence $\mathrm{f}(A^n(l))$ produced in block l must carry information about the source sequence $X^n(l+1)$ corresponding to the next block $l+1$. Note that this is possible since encoder knows the entire sequence X^n. At the same time, sequence $A^n(l)$ should also perform well as an action sequence to be used by Decoder 1 to estimate sequence $X^n(l)$ for the current block. This is accomplished as follows; In each block l, $2^{nI(X;\hat{X}_2)}$ codewords $\hat{X}_2^n \in \hat{\mathcal{X}}_2^n$ are generated according to the pmf $p(\hat{x}_2)$. Next, $2^{nI(X;\hat{X}_2)}$ bins are assigned to each codeword $\hat{X}_2^n$, where each bin contains $2^{nI(X;\mathrm{f}(A)|\hat{X}_2)}$ codewords $B^n \in \mathcal{B}^n$, generated according to pmf $p(b|\hat{x}_2)$. For each pair $(\hat{X}_2^n, B^n)$, a codebook of $2^{nI(X;A|\hat{X}_2,\mathrm{f}(A))}$ codewords $A^n \in \mathcal{A}^n$ is generated according to the joint pmf $p(x, a|\hat{x}_2, b)$. Finally, a codebook of $2^{nI(X;U|\hat{X}_2,A)}$ codewords $U^n \in \mathcal{U}^n$ is generated according to the joint pmf $p(x, u|\hat{x}_2, a)$. The latter codebook is further binned into a codebook of rate $I(X; U|\hat{X}_2, A, Y)$ to leverage the side information Y^n available at Decoder 1 via the Wyner-Ziv theorem [1, p. 280].

For encoding, in each block l, a sequence $\hat{X}_2^n$ is selected from the $\hat{X}_2-$codebook of block l to be jointly typical with the source sequence $X^n(l)$ in the current block.

Instead, the bin index describes a $\hat{X}_2^n$ sequence in the $\hat{X}_2-$codebook of block $(l+1)$th that is jointly typical with the source sequence $X^n(l+1)$ of the $(l+1)$th block. Moreover, given $\hat{X}_2^n$ and the bin index, a sequence A^n is chosen such that $(A^n, X^n(l))$ are jointly typical. Similarly, a sequence U^n is selected for block l to be jointly typical with the sequence of $X^n(l)$ of block l.

Thanks to the observation of the actions, at block $l+1$ Decoder 2 knows the functions $\mathrm{f}(A^n(l))$, and aims to find the bin index in which the corresponding codeword B^n lies. As shown in [62], this is possible with vanishing probability of error, if $I(X;\hat{X}_2,\mathrm{f}(A)) \leq H(\mathrm{f}(A)|\hat{X}_2)$. Note that the conditioning in the right-hand side is due to the fact that the sequences B^n are generated conditioned on the sequence $\hat{X}_2^n$ representing a compressed version of the source for the current block l. The latter does not bring any information regarding the desired sequence $X^n(l+1)$. □

Remark 8.2. From the proof of the converse in Appendix U, it follows, similarly to [26], that, adaptive actions (see Remark 8.1) do not increase the rate-distortion-cost function derived in Proposition 8.2.

8.2.4 Causal Action Observation

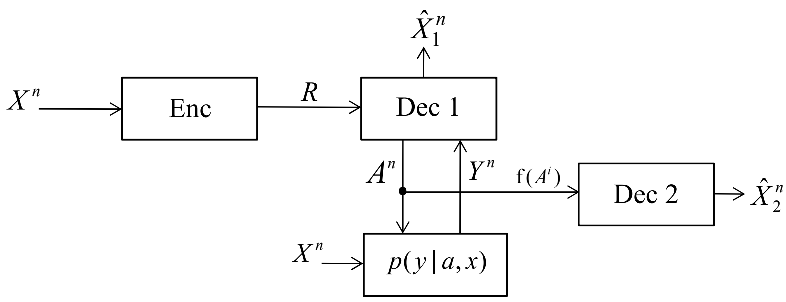

Figure 8.5 Source coding with decoder-side actions for information acquisition and with information embedding on actions. At time i, Decoder 2 has available the samples $\mathrm{f}(A^i) = (\mathrm{f}(A_1), ..., \mathrm{f}(A_i))$ in a causal fashion.

The system model for the set-up in Figure 8.5, is similar to the one described in Section 8.2.1 with the only difference the decoding function for Decoder 2 is

$$\mathrm{h}_{2i}^{(d)}\colon \mathcal{B}^i \to \hat{\mathcal{X}}_2, \tag{8.14}$$

which maps the causally observed sequence $\mathrm{f}(A^i) = (\mathrm{f}(A_1), ..., \mathrm{f}(A_i))$ into the ith estimated symbol $\hat{X}_{2i}$.

Proposition 8.3. *The rate-distortion-cost function $R(D_1, D_2, \Gamma)$ for the source coding problem with decoder-side actions and causal observation of the actions illustrated in Figure 8.5 is given by*

$$R(D_1, D_2, \Gamma) = \min_{p(v,a,u|x),\ \mathrm{g}_1(U,Y),\ \mathrm{g}_2(V,\mathrm{f}(A))} I(X; V, A) + I(X; U|V, A, Y), \tag{8.15}$$

where the mutual information is evaluated with respect to the joint pmf

$$p(x, y, a, u, v) = p(x)p(v, a, u|x)p(y|x, a), \tag{8.16}$$

for some pmf $p(v, a, u|x)$ such that the inequalities

$$\begin{aligned} \mathrm{E}[d_j(X, \hat{X}_j)] &\leq D_j, \ \textit{for } j = 1, 2, && (8.17a)\\ \mathrm{E}[\Lambda(A)] &\leq \Gamma, && (8.17b)\\ \text{and } I(X; V, \mathrm{f}(A)) &\leq H(\mathrm{f}(A)|V) && (8.17c) \end{aligned}$$

are satisfied for $\hat{X}_1 = \mathrm{g}_1(U, Y)$ and $\hat{X}_2 = \mathrm{g}_2(V, \mathrm{f}(A))$ with some functions $\mathrm{g}_1\colon \mathcal{U} \times \mathcal{Y} \to \hat{\mathcal{X}}_1$ and $\mathrm{g}_2\colon \mathcal{V} \times \mathcal{B} \to \hat{\mathcal{X}}_2$, respectively. Finally, U and V are auxiliary random variables whose alphabet cardinalities can be constrained as $|\mathcal{U}| \leq |\mathcal{X}||\mathcal{V}||\mathcal{A}| + 1$ and $|\mathcal{V}| \leq |\mathcal{X}| + 3$, respectively, without loss of optimality.

The difference between the rate-distortion-cost function above with causal and strictly causal action observation is given by the fact that, with causal action observation, Decoder 2 can use the current value of the function $\mathrm{f}(A)$ for the estimate of $\hat{X}_2$. This is captured by the fact that the Encoder provides Decoder 2 with an auxiliary source description V, which is then combined with $\mathrm{f}(A)$ via a function $\hat{X}_2 = \mathrm{g}_2(V, \mathrm{f}(A))$ to obtain $\hat{X}_2$. The rate (8.15) and the constraint (8.17c) are changed accordingly. The proof of the converse is given in Appendix U.

Remark 8.3. As seen in Appendix U, with adaptive actions, the rate-distortion-cost function derived in Proposition 8.3 remains unchanged.

8.2.5 Binary Example

In this section, an example is provided to illustrate the effect of the communication requirements of the additional decoder (Decoder 2) that observes a function of the actions on the system performance. Binary alphabets as $\mathcal{X} = \mathcal{A} = \mathcal{Y} = \{0, 1\}$ and a source distribution $X \sim \mathrm{Bern}(\frac{1}{2})$ are assumed. The distortion metrics are assumed to be Hamming, i.e., $d_j(x, \hat{x}_j) = 0$ if $x = \hat{x}_j$ and $d_j(x, \hat{x}_j) = 1$ otherwise for $j = 1, 2$. Moreover, as shown in Figure 8.6, the side information Y at Decoder 1 is observed through a Z-channel for $A = 0$ or an S-channel for $A = 1$. No cost constraint on the actions is assumed taken by Decoder 1 (which can be enforced by choosing $\Lambda(A) = A$ and $\Gamma = 1$), and $\mathrm{f}(A) = A$. The example extends that of [4, Section II-D] to a set-up with the additional Decoder 2. Under the requirement of lossless reconstruction at Decoder 1, i.e., $D_1 = 0$, the rate-distortion-cost function $R(0, D_2, \Gamma = 1)$ with non-causal action observation is obtained from Proposition 8.1 by setting $U = \hat{X}_1 = X$, obtaining

$$R(0, D_2, 1) = \min_{p(\hat{x}_2, a|x)} I(X; \hat{X}_2, A) + H(X|\hat{X}_2, A, Y), \tag{8.18}$$

where the minimization is done under the constraints $\mathrm{E}[d_2(X, \hat{X}_2)] \leq D_2$ and $I(X; \hat{X}_2, A) \leq H(A)$.

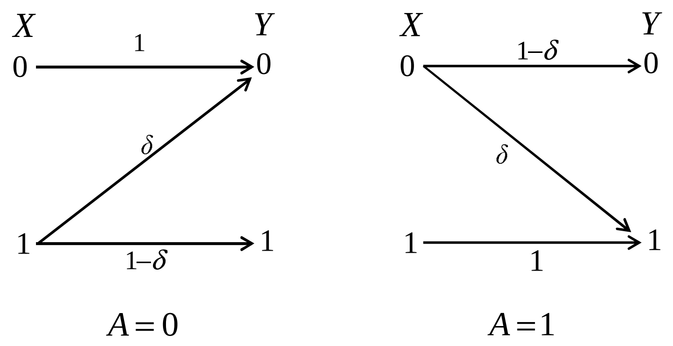

Figure 8.6 The side information channel $p(y|x,a)$ used in the example of Section 8.2.5.

The minimization in (8.18) can be done over the parameters $p(a=1,\hat{x}_2=0|x=0) \triangleq \alpha_1$, $p(a=1,\hat{x}_2=1|x=0) \triangleq \alpha_2$ and $p(a=0,\hat{x}_2=1|x=0) \triangleq \alpha_3$ with $\sum_{i=1}^{3}\alpha_i \leq 1$ and $\alpha_i \geq 0$ for $i=1,2,3$, since by symmetry, one can set $p(a=0,\hat{x}_2=1|x=1)=\alpha_1$, $p(a=0,\hat{x}_2=0|x=1)=\alpha_2$ and $p(a=1,\hat{x}_2=0|x=1)=\alpha_3$ without loss of optimality. Explicit expressions can be easily found and have been optimized numerically.

Figure 8.7 depicts the rate-distortion function versus the distortion D_2 of Decoder 2 for values of $\delta=0.2$, $\delta=0.5$ and $\delta=0.8$. It can be seen that if the distortion D_2 tolerated by Decoder 2 is sufficiently large (e.g., $D_2 \geq 0.4$ for $\delta=0.5$), then the communication requirements of Decoder 2 do not increase the required rate. This can be observed by comparing the rate $R(0,D_2,\Gamma)$ with rate $R(0,0.5,\Gamma)$ corresponding to a distortion level $D_2=0.5$, which requires no communication to Decoder 2. The smallest distortion D_2 that does not affect the rate can be found as $D_2=\alpha_{2,opt}+\alpha_{3,opt}$, where $\alpha_{2,opt}$ and $\alpha_{3,opt}$ are the optimal values for problem (8.18) with $D_2=0.5$ that minimizes $\alpha_2+\alpha_3$.

Now, the performance between non-causal action observation, as considered above, and strictly causal action observation are compared. The performance in the latter case can be obtained from Proposition 8.2 and leads to (8.18) with the more restrictive constraint (8.13c). Figure 8.8 plots the difference between rate-distortion

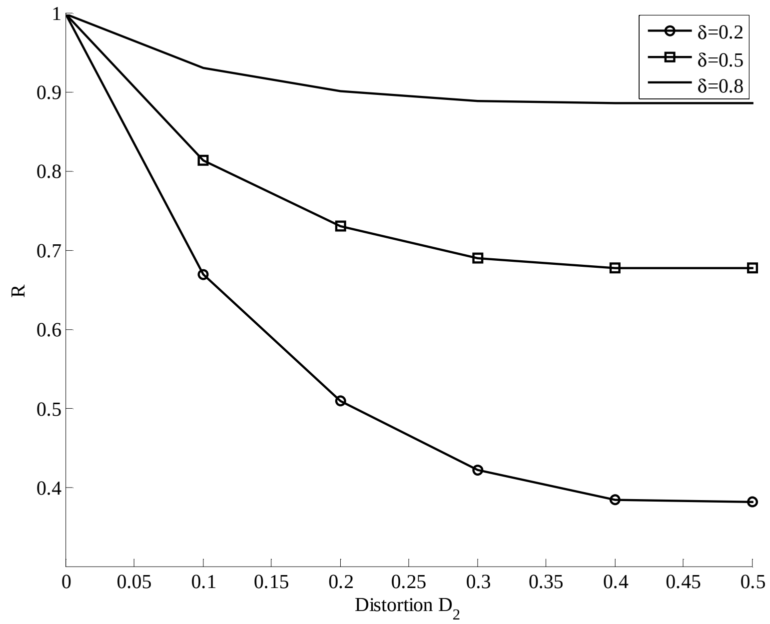

Figure 8.7 Rate-distortion function $R(0, D_2, 1)$ in (8.18) versus distortion D_2 with the side information channel in Figure 8.6 (non-causal side information).

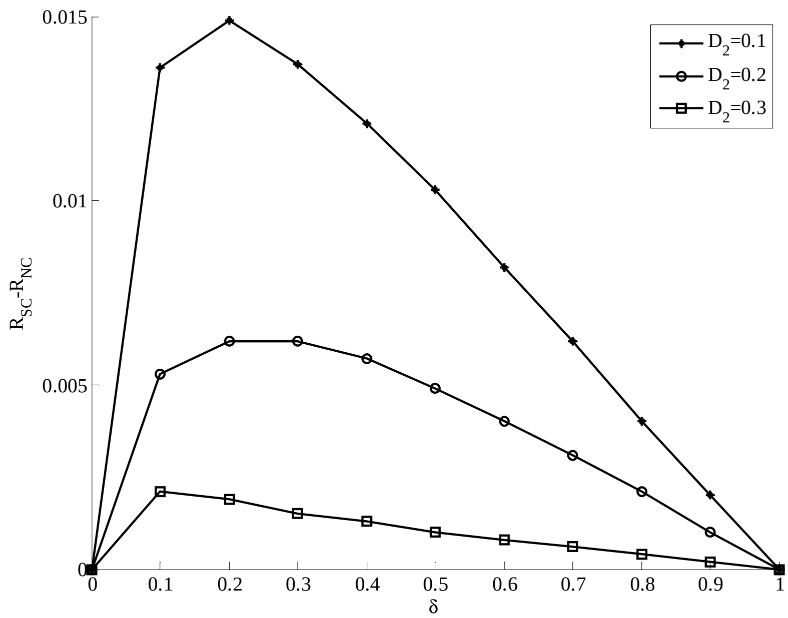

Figure 8.8 Difference between the rate-distortion function (8.18) with non-causal (NC) and strictly causal (SC) action observation versus δ for values of distortion $D_2 = 0.1$, 0.2 and 0.3.

function (8.18) for the case of non-causal and strictly causal action observation versus δ for three values of distortion, namely $D_2 = 0.1, 0.2, 0.3$. As shown, irrespective of the value of distortion D_2, for values of $\delta = 0$ and $\delta = 1$, the performance with non-causal action observation is equal to that with strictly causal observation. This is due to the facts that: *i*) for $\delta = 0$, the side information Y is a noiseless measure of the source sequence X for both $A = 0$ and $A = 1$ and thus there is no gain in making the actions at Decoder 1 to be dependent of X, and thus $\hat{X}_2$; *ii*) for $\delta = 1$, the side information Y is independent of the source sequence X given both $A = 0$ and $A = 1$, and thus it is without loss of optimality to choose actions at Decoder 1 to be independent of X and $\hat{X}_2$. It is can be concluded that for both $\delta = 0$ and $\delta = 1$, causal action observation, and in fact even selecting A to be independent of X, does not entail any performance loss. Instead, for values $0 < \delta < 1$, it is generally advantageous for Decoder 1 to select actions correlated with the source X, and hence some performance loss is observed with strictly causal action observation owing to the more restrictive constraint (8.13c). This reflects the need to cater to both Decoder 1 and Decoder 2 when selecting actions A, which requires description of two different source blocks. Following similar arguments, it is also noted that, as the communication requirements for Decoder 2 become more pronounced, i.e., as D_2 decreases, the difference between the rate-distortion function with non-causal and strictly-causal action observation increases. The performance with causal action observation is intermediate between full and strictly causal observation, and it is not shown here.

8.3 Encoder-Side Actions for Side Information Acquisition

In the previous section, the actions controlling the quality and availability of the side information were taken by the decoder. In this section, following [4, Section III], instead scenarios are considered in which the encoder takes the actions affecting the side information of Decoder 1, as shown in Figure 8.2. Specifically, the encoder takes actions $A^n \in \mathcal{A}^n$, thus influencing the side information available to the Decoder 1 through a discrete memoryless channel $p(y|x, a)$. Decoder 2 observes the action sequence to obtain the deterministic function $\mathrm{f}(A^n) = (\mathrm{f}(A_1), ..., \mathrm{f}(A_n))$, or the corresponding causal and strictly causal function, which is used to estimate the source sequence subject to a distortion constraint.

An (n, R, D_1, D_2, Γ) code is defined similar to the previous sections with the difference that the action encoder (8.2) maps directly the source sequence X^n into the action sequence A^n, i.e.,

$$\mathrm{h}^{(a)}\colon \mathcal{X}^n \to \mathcal{A}^n. \tag{8.19}$$

As discussed in [4], even in the absence of Decoder 2, the problem at hand is challenging. Therefore, focus is on certain special cases, first the special case in which the side information channel $p(y|x, a)$ is such that Y is a deterministic function of A, i.e., $Y = \mathrm{f}_Y(A)$, and $\mathrm{f}(A) = A$. This is solved in Proposition 8.4, and generalized by the following remark to the case of all deterministic function f for which $H(\mathrm{f}_Y(A)|\mathrm{f}(A)) = 0$. Following Proposition 8.4 is Proposition 8.5, where the case $H(\mathrm{f}(Y)|\mathrm{f}_Y(A)) = 0$ is solved.

Proposition 8.4. *The rate-distortion-cost function $R(D_1, D_2, \Gamma)$ for the source coding problem with encoder-side actions, non-causal, causal or strictly causal observation of the actions illustrated in Figure 8.2 with $\mathrm{f}(A) = A$ and $Y = \mathrm{f}_Y(A)$ is given by*

$$R(D_1, D_2, \Gamma) = \min_{p(u|x),\ p(\hat{x}_1|u,x),\ p(\hat{x}_2|u,x),\ p(a)} \{I(X; \hat{X}_1, U) - H(\mathrm{f}_Y(A))\}^+, \tag{8.20}$$

where the information measures are evaluated with respect to the joint pmf

$$p(x, u, \hat{x}_1, \hat{x}_2, a) = p(x)p(u|x)p(\hat{x}_1|u, x)p(\hat{x}_2|u, x)p(a), \tag{8.21}$$

for some pmfs $p(u|x)$, $p(\hat{x}_1|u,x)$, $p(\hat{x}_2|u,x)$ and $p(a)$ such that the inequalities

$$\begin{aligned}
\mathrm{E}[d_j(X,\hat{X}_j)] &\leq D_j, \ \textit{for } j=1,2, && \text{(8.22a)}\\
\mathrm{E}[\Lambda(A)] &\leq \Gamma, && \text{(8.22b)}\\
I(X;U) &\leq H(\mathrm{f}_Y(A)) && \text{(8.22c)}\\
\text{and } I(X;\hat{X}_2|U) &\leq H(A|\mathrm{f}_Y(A)) && \text{(8.22d)}
\end{aligned}$$

are satisfied. Finally, U is an auxiliary random variable whose alphabet cardinality can be constrained as $|\mathcal{U}| \leq |\mathcal{X}||\hat{\mathcal{X}}_1||\hat{\mathcal{X}}_2|+3$ without loss of optimality.

Remark 8.4. The results above generalizes a number of known single-letter characterizations. Notably, if $D_2 = D_{2,\max}$, so that the distortion requirements of Decoder 2 are immaterial to the system performance, the result reduces to [4, Theorem 7]. Moreover, in the special case in which $A=(A_0,A_2)$, $Y=A_0$, $R=R_1$, $|\mathcal{A}_0|=2^{R_0}$, $|\mathcal{A}_2|=2^{R_2}$, the model coincides with the lossy Gray-Wyner problem [66][2].

As detailed in the proof below, Proposition 8.4 establishes the optimality of separate source-channel coding for the set-up in Figure 8.2 under the stated conditions. In particular, the encoder compresses using a standard successive refinement source code in which U represents the coarse description and $\hat{X}_1, \hat{X}_2$ two independent refinements. The indices of the coarse description U and of the refined description $\hat{X}_2$ are sent on the degraded (deterministic) broadcast channel with input A and outputs $(A, \mathrm{t}(A))$ using superposition coding. Reliable compression and communication is guaranteed by the two bounds (8.22c)-(8.22d). A further refined description $\hat{X}_1$ is produced for Decoder 1, and the corresponding index is sent partly over the mentioned broadcast channel and partly over the link of rate R, leading to the rate (8.20). Details of the achievability proof can be found below, while the proof of the converse is given in Appendix V.

[2]Note that here 2^{R_0} and 2^{R_2} are constrained to be integers.

Remark 8.5. Following the discussion above, specializing Proposition 8.4 to the case $R = 0$ shows the optimality of source-channel coding separation for the lossy transmission of a source over a deterministic degraded broadcast channel (see [1, Chapter 14] for a review of scenarios in which the optimality of separation holds for lossless transmission over a broadcast channel).

Sketch of the achievability proof: As anticipated above, achievability uses the ideas of a source-channel coding separation, successive refinement and superposition coding. Only the outline is described, as the rigorous details can be derived based on standard techniques [1]. Starting with the case of non-causal action observation at Decoder 2, note that the deterministic channel with input A and outputs A (to Decoder 2) and $\mathrm{f}_Y(A)$ (to Decoder 1) is not only deterministic but also degraded [1, Chapter 5]. This channel is used to send a common source description of rate $\tilde{R}_1$ to both the decoders and a refined description of rate $\tilde{R}_2$ to Decoder 2 only. To elaborate, fix the pmfs $p(u|x)$, $p(\hat{x}_1|u,x)$, $p(\hat{x}_2|u,x)$ and $p(a)$. Generate a codebook of $2^{nI(X;U)}$ sequences U^n i.i.d. with the pmf $p(u)$ and, for each U^n sequence, generate a codebook of $2^{nI(X;\hat{X}_1|U)}$ $\hat{X}_1^n$ sequences i.i.d. with pmf $p(\hat{x}_1|u)$ and a codebook of $2^{nI(X;\hat{X}_2|U)}$ sequences $\hat{X}_2^n$ i.i.d. with pmf $p(\hat{x}_2|u)$. Given a source sequence X^n, the encoder finds a jointly typical U^n codeword, and then a codeword $\hat{X}_1^n$ jointly typical with (X^n, U^n) and similarly for $\hat{X}_2^n$. Using source-channel separation on the broadcast "action" channel described above, the index from the U-codebook and a part of the index from the $\hat{X}_1$-codebook, of rate r, is described to both decoders, and the index from $\hat{X}_2$-codebook is described to Decoder 2 as its private information. Thus, the inequalities below hold

$$\tilde{R}_1 \geq I(X;U) + r \tag{8.23a}$$

$$\text{and } \tilde{R}_2 \geq I(X;\hat{X}_2|U). \tag{8.23b}$$

The capacity region of the broadcast channel is given by the conditions [1, Chapter 9], $\tilde{R}_1 \leq H(\mathrm{f}_Y(A))$ and $\tilde{R}_1 + \tilde{R}_2 \leq H(A)$, and thus the following rates are achievable

$$\tilde{R}_1 \leq H(\mathrm{f}_Y(A)) \tag{8.24a}$$

$$\text{and } \tilde{R}_2 \leq H(A|\mathrm{f}_Y(A)). \tag{8.24b}$$

Finally the remaining part of the index of codeword $\hat{X}_1^n$ is sent through the direct rate R, leading to the condition

$$R \geq I(X; \hat{X}_1|U) - r. \tag{8.25}$$

Combining (8.23), (8.24) and (8.25), and using Fourier-Motzkin elimination, one can obtain

$$R \geq I(X; \hat{X}_1, U) - H(\mathrm{f}_Y(A)) \tag{8.26}$$

and (8.29c)-(8.22d). The distortion and cost constraints are handled in a standard manner and hence the details are omitted.

It remains to discuss how to handle the case of causal or strictly causal action observation. Given the converse result in Appendix V, it is enough to show that (8.20)-(8.22) is achievable also with strictly causal and causal action observation. This can be simply accomplished by encoding in blocks as per achievability of Proposition 8.2 and Proposition 8.3. Specifically, in each block the encoder compresses the source sequence corresponding to the next block. Decoder 2 then operates as above, while Decoder 1 can recover all source blocks at the end of all blocks. □

Remark 8.6. The scenario solved above is when the action observation is perfect, i.e., $\mathrm{f}(A) = A$. The result also carries verbatim for the more general case where $\mathrm{f}(A)$ is a generic function as long as $H(\mathrm{f}_Y(A)|\mathrm{f}(A)) = 0$. The expressions of the rate region remain the same as in the proposition above except that A is replaced by $\mathrm{f}(A)$.

Proposition 8.4 characterizes the optimal performance for the case when Decoder 2 has a better information about the actions taken by the encoder than Decoder 1 in the sense that $H(\mathrm{f}_Y(A)—\mathrm{f}(A)) = 0$. It is noted here that a similar characterization can be given also for the dual setting in which $H(\mathrm{f}(A)—\mathrm{f}_Y(A)) = 0$ so that Decoder 1 has the better observation about the actions.

Proposition 8.5. *The rate-distortion-cost function $R(D_1, D_2, \Gamma)$ for the source coding problem with encoder-side actions, non-causal, causal or strictly causal observation of the actions illustrated in Figure 8.2 with $H(\mathrm{f}(A)|\mathrm{f}_Y(A)) = 0$, is given by*

$$R(D_1, D_2, \Gamma) = \min_{p(a),\ p(\hat{x}_1, \hat{x}_2|x)} \{I(X; \hat{X}_1, \hat{X}_2) - H(\mathrm{f}_Y(A))\}^+, \tag{8.27}$$

where the information measures are evaluated with respect to the joint pmf

$$p(x, \hat{x}_1, \hat{x}_2, a) = p(x)p(\hat{x}_1, \hat{x}_2|x)p(a), \tag{8.28}$$

such that the following inequalities are satisfied,

$$\begin{aligned} \mathrm{E}[d_j(X, \hat{X}_j)] &\leq D_j, \text{ for } j = 1, 2, && (8.29\text{a}) \\ \mathrm{E}[\Lambda(A)] &\leq \Gamma, && (8.29\text{b}) \\ I(X; \hat{X}_2) &\leq H(\mathrm{f}(A)). && (8.29\text{c}) \end{aligned}$$

The converse follows similarly as that for Proposition 8.4 where instead of U in the converse, $\hat{X}_2$ is used, as knowing $Y^n = \mathrm{f}_Y(A^n)$ implies knowing $\mathrm{f}(A^n)$, due to the assumption $H(\mathrm{f}(A)|\mathrm{f}_Y(A)) = 0$. The achievability is outlined for only for the non-causal case (the achievability for strictly causal and causal case uses block coding ideas as in Proposition 8.4). A successive refinement codebook is generated by drawing $2^{nI(X;\hat{X}_2)}$ codewords $\hat{X}_2^n$, and, for each codeword $\hat{X}_2^n$, a number $2^{nI(X;\hat{X}_1|\hat{X}_2)}$ of codewords $\hat{X}_1^n$. As for Proposition 8.4, the indices of these two codebooks obtained via standard joint typicality encoding are sent through the degraded broadcast channel $p(y,b|a) = \mathbf{1}_{\{y=\mathrm{f}_Y(A), b=\mathrm{f}(A)\}}$. Splitting the rate for the index of codeword $\hat{X}_1^n$ so that a rate R is sent over the direct link to Decoder 1, reliability of compression and communication over the "action" broadcast channel is guaranteed if

$$I(X;\hat{X}_2) \leq H(\mathrm{f}(A)) \tag{8.30}$$

$$I(X;\hat{X}_2) + I(X;\hat{X}_1|\hat{X}_2) - R \leq H(\mathrm{f}(A), \mathrm{f}_Y(A)) = H(\mathrm{f}_Y(A)), \tag{8.31}$$

where the latter inequality implies $R \geq I(X;\hat{X}_1,\hat{X}_2) - H(\mathrm{f}_Y(A))$. The proof is concluded using the usual steps. □

8.4 Actions for Channel State Control and Probing

In this section, the impact of information embedding on actions for the set-up of channel coding with actions of [44] is considered. To this end, consider the model in Figure 8.3, in which Decoder 1, based on the observation of a deterministic function of the actions, wishes to retrieve part of the information destined to Decoder 2. Note that for simplicity of notation here the additional encoder that observes the actions is denoted as Decoder 1, rather than Decoder 2 as done above. Also, it is emphasized that in the original set-up of [44], Decoder 1 was not present.

8.4.1 System Model

The system is defined by the pmfs $p(x)$, $p(y|x,s,a)$, $p(s|a)$, function f: $\mathcal{A} \to \mathcal{B}$ and by discrete alphabets $\mathcal{X}, \mathcal{A}$, $\mathcal{B}$, $\mathcal{S}$, and $\mathcal{Y}$. Given the messages (M_1, M_2), selected randomly from the set $\mathcal{M}_1 \times \mathcal{M}_2 = [1, 2^{nR_1}] \times [1, 2^{nR_2}]$, an action sequence $A^n \in \mathcal{A}^n$ is selected by the Encoder. Decoder 1 observes the signal $B^n = \mathrm{f}(A^n)$ as a deterministic function of the actions, and estimates message M_1. Note that the notation here implies a "non-causal" observation of the actions, but it is easy to see that the results below hold also with causal and strictly causal observation of the actions. Moreover, the state sequence $S^n \in \mathcal{S}^n$ is generated as the output of a memoryless channel $p(s|a)$ and $p(b^n, s^n|a^n) = \prod_{i=1}^n p(s_i|a_i)\mathbf{1}_{\{b_i=\mathrm{f}(a_i)\}}$ holds for an action sequence $A^n = a^n$. The input sequence $X^n \in \mathcal{X}^n$ is selected on the basis of both messages (M_1, M_2) and of the state sequence S^n by the Encoder. The action sequence A^n and the input X^n have to satisfy an average cost constraint defined by a function $\gamma : \mathcal{A} \times \mathcal{X} \to [0, \infty)$, so that the cost for the input sequences a^n and x^n is given by $\gamma(a^n, x^n) = \frac{1}{n}\sum_{i=1}^n \gamma(a_i, x_i)$. Given $X^n = x^n$, $S^n = s^n$ and $A^n = a^n$, the received signal is distributed as $p(y^n|x^n, s^n, a^n) = \prod_{i=1}^n p(y_i|x_i, s_i, a_i)$. Decoder 2, having received the signal Y^n, estimates both messages (M_1, M_2).

The setting includes the semi-deterministic broadcast channel with degraded message sets [67] (see also [1, Ch. 8]) as a special case by setting X to be constant and $Y = S$, and the channel with action-dependent states studied in [44] for $R_1 = 0$.

Definition 8.4. An $(n, R_0, R_1, \Gamma, \epsilon)$ code for the model in Figure 8.3 consists of an action encoder

$$\mathrm{h}^{(a)}\colon \mathcal{M}_1 \times \mathcal{M}_2 \to \mathcal{A}^n, \tag{8.32}$$

which maps message (M_1, M_2) into an action sequence A^n; a channel encoder

$$\mathrm{h}^{(e)}\colon \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{S}^n \to \mathcal{X}^n, \tag{8.33}$$

which maps message (M_1, M_2) and the state sequence S^n into the sequence X^n; two decoding functions

$$\mathrm{h}_1^{(d)}\colon \mathcal{B}^n \to \mathcal{M}_1, \tag{8.34}$$

$$\text{and } \mathrm{h}_2^{(d)}\colon \mathcal{Y}^n \to \mathcal{M}_1 \times \mathcal{M}_2, \tag{8.35}$$

which map the sequences B^n and Y^n into the estimated messages $\hat{M}_1$ and $(\hat{M}_1, \hat{M}_2)$, respectively; such that the probability of error in decoding the messages (M_1, M_2) is small,

$$\Pr[\mathrm{h}_1^{(d)}(B^n) \neq M_1] \leq \epsilon, \tag{8.36}$$

$$\text{and } \Pr[\mathrm{h}_2^{(d)}(Y^n) \neq (M_1, M_2)] \leq \epsilon, \tag{8.37}$$

and the cost constraint is satisfied, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\gamma(A_i, X_i)\right] \;\; \leq \;\; \Gamma + \epsilon. \tag{8.38}$$

Given a cost Γ, a rate pair (R_1, R_2) is said to be achievable for a cost-constraint Γ if, for any $\epsilon > 0$ and sufficiently large n, there a exists a $(n, R_1, R_2, \Gamma, \epsilon)$ code. The goal is to characterize the capacity-cost region $\mathcal{C}(\Gamma)$, which is the closure of all achievable rate pairs (R_1, R_2) for the given cost Γ.

8.4.2 Capacity-Cost Region

In this section, a single-letter characterization of the capacity-cost region is derived.

Proposition 8.6. *The capacity-cost region $\mathcal{C}(\Gamma)$ for the system in Figure 8.3 is given by the union of all rate pairs (R_1, R_2) such that the inequalities*

$$R_1 \;\; \leq \;\; H(\mathrm{f}(A)) \tag{8.39a}$$

$$\text{and } R_1 + R_2 \;\; \leq \;\; I(A, U; Y) - I(U; S|A), \tag{8.39b}$$

are satisfied, where the mutual informations are evaluated with respect to the joint pmf

$$p(a, s, u, x, y) = p(a)p(s|a)p(u|s, a)\mathbf{1}_{\{x=\mathrm{g}(u,s)\}}p(y|x, s, a), \tag{8.40}$$

for some pmfs $p(a)$, $p(u|s, a)$ and function g*: $\mathcal{U} \times \mathcal{S} \rightarrow \mathcal{X}$ such that*

$$\mathrm{E}[\gamma(A, X)] \;\; \leq \;\; \Gamma. \tag{8.41}$$

Finally, one can set $|\mathcal{U}| \leq |\mathcal{X}||\mathcal{S}||\mathcal{A}| + 1$ *without loss of optimality.*

The proof of converse is an immediate consequence of cut-set arguments and of the proof of the upper bound obtained in [44, Theorem 1]. Specifically, inequality (8.39a) follows by considering the cut around Decoder 1, while the inequality (8.39b) coincides with the bound derived in [44, Theorem 1] on the rate that can be communicated between the Encoder and Decoder 2 with no regards for Decoder 1[3]. The achievability requires rate splitting, superposition coding and the coding strategy proposed in [44, Theorem 1]. A sketch of proof of the achievability is relegated to Appendix W.

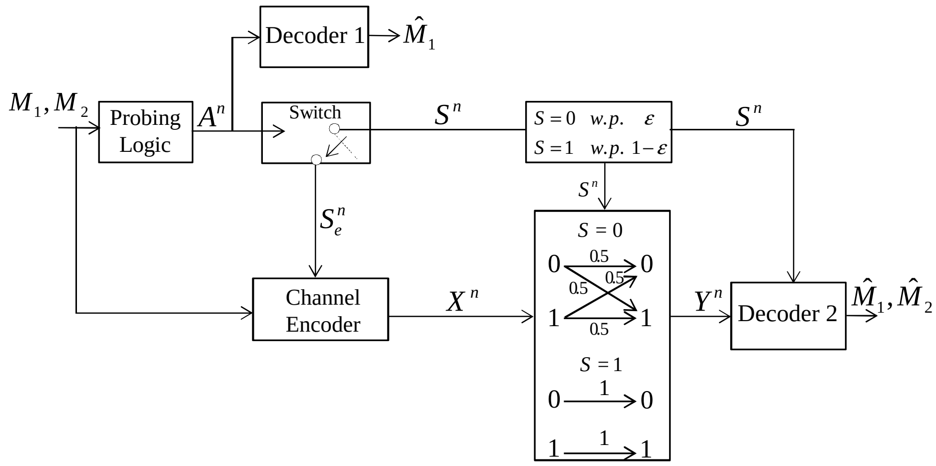

Figure 8.9 Channel coding with actions for channel state probing and with information embedding on actions.

8.4.3 Probing Capacity

Here, an example is provided to illustrate the effect of the communication requirements of the action-cribbing decoder on the system performance. Consider the communication system shown in Figure 8.9, where the states is known to Decoder 2. It is further assumed that actions are binary, such that, if $A = 1$, the channel encoder observes the state S, and if $A = 0$, it does not obtain any information about S.

[3]The cardinality constraints follow from [44, Theorem 1]

This problem is modeled by defining the state information available at the encoder as $S_e = \mathrm{u}(S, A)$, where $\mathrm{u}(S, 1) = S$ and $\mathrm{u}(S, 0) = \mathrm{e}$, where represents as "erasure" symbol. Following [45], this problem is referred to as having a "probing" encoder.

The channel encoder maps the state information S_e^n and messages M_1, M_2 into a codeword X^n (see Figure 8.9). Moreover, two cost constraints, namely $\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\gamma_a(A_i)\right] \leq \Gamma_A$ and $\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\gamma_x(X_i)\right] \leq \Gamma_X$ are imposed for given action input cost functions $\gamma_a : \mathcal{A} \rightarrow [0, \Lambda_{a,max}]$ and $\gamma_x : \mathcal{X} \rightarrow [0, \Lambda_{x,max}]$ with $0 \leq \Lambda_{a,max} < \infty$ and $0 \leq \Lambda_{x,max} < \infty$, respectively. In [45, Theorem 1], a correspondence was proved between the set-up of a probing encoder and that of action dependent states. Using [45, Theorem 1] and Proposition 8.6, one can easily obtain that the capacity-cost region $\mathcal{C}(\Gamma_A, \Gamma_X)$ for the system in Figure 8.9 is given by the union of all rate pairs (R_1, R_2) such that the inequalities

$$R_1 \leq H(A|Q) \tag{8.42a}$$

$$\text{and } R_1 + R_2 \leq I(X; Y|S, Q), \tag{8.42b}$$

are satisfied, where the mutual informations are evaluated with respect to the joint pmf

$$p(q, a, s, s_e, x, y) = p(q)p(a|q)p(s)\mathbf{1}_{\{s_e=\mathrm{u}(s,a)\}}p(x|s_e, a, q)p(y|x, s), \tag{8.43}$$

for some pmfs $p(q)$, $p(a|q)$, $p(x|s_e, a, q)$ such that $\mathrm{E}[\gamma_a(A)] \leq \Gamma_A$ and $\mathrm{E}[\gamma_x(X)] \leq \Gamma_X$.

Now, (8.42a)-(8.42b) are applied to the channel shown in Figure 8.9 in which alphabets are binary $\mathcal{X} = \mathcal{Y} = \mathcal{S} = \{0, 1\}$, S is a Bern$(1-\epsilon)$ variable for $0 \leq \epsilon \leq 1$ and the channel is a binary symmetric with flipping probability 0.5 if $S = 0$ ("bad" channel state) and 0 if $S = 1$ ("good" channel state).

To evaluate the maximum achievable sum-rate $R_1 + R_2$ for a given rate R_1, let's define $\Pr[A = 1] = \gamma$, $\Pr[X = 1|S_e = 1, A = 1] = p_1$ and $\Pr[X = 1|S_e = \mathrm{e}, A = 0] = p_2$, and set $\Pr[X = 1|S_e = 0, A = 1] = 0$ without loss of optimality. The maximum sum-rate $R_1 + R_2$ for a given rate R_1 is then obtained from (8.42b) by solving the problem

$$R_1 + R_2 = \max_{0 \leq p_1, p_2, \gamma \leq 1} \gamma(1-\epsilon)H(p_1) + (1-\gamma)(1-\epsilon)H(p_2), \tag{8.44}$$

under the constraint $\mathrm{E}[X] = p_1\gamma(1-\epsilon) + p_2(1-\gamma) \leq \Gamma_X$, $\mathrm{E}[A] = \gamma \leq \Gamma_A$ and $H(A) = H(\gamma) \geq R_1$. Note that the last constraint imposes that the rate achievable by the Decoder 1 is larger than R_1 as per (8.42a).

The sum-rate in (8.44) is shown in Figure 8.10 for $\epsilon = 0.5$, $\Gamma_A = 1$ and different values of R_1. It can be seen that, for sufficiently small values of the cost constraint Γ_X, increasing the communication requirements, i.e., R_1, of the Decoder 1, reduces the achievable sum-rate $R_1 + R_2$. This is due to the fact that increasing R_1 requires to encode more information in the action sequence, which in turn reduces the portion of the actions that can be set to $A = 1$, i.e., $\Pr[A = 1]$. As a result, the encoder is less informed about the state sequence and thus bound to waste some power on bad channel states.

Remark 8.7. The communication requirements of Decoder 1 need not necessarily affect the system performance. For instance, consider the example 1 in [45, Section V.A], which includes a probing encoder as in Figure 8.9 but transmitting over a different channel. There, it turns out that it is sufficient to have $\Pr[A = 1] \gtrsim 0.2$ in order to achieve the same performance that can be achieved with full encoder channel state information. Therefore, the additional constraint on the rate of Decoder 1 (8.42a), namely $R_1 \geq H(A)$, does not affect the sum-rate achievable in this example for any rate $R_1 \in [0, 1]$.

8.5 Concluding Remarks

There is a profound interplay between actuation and communication in that both actuation can be instrumental to improve the efficiency of communication, and, vice

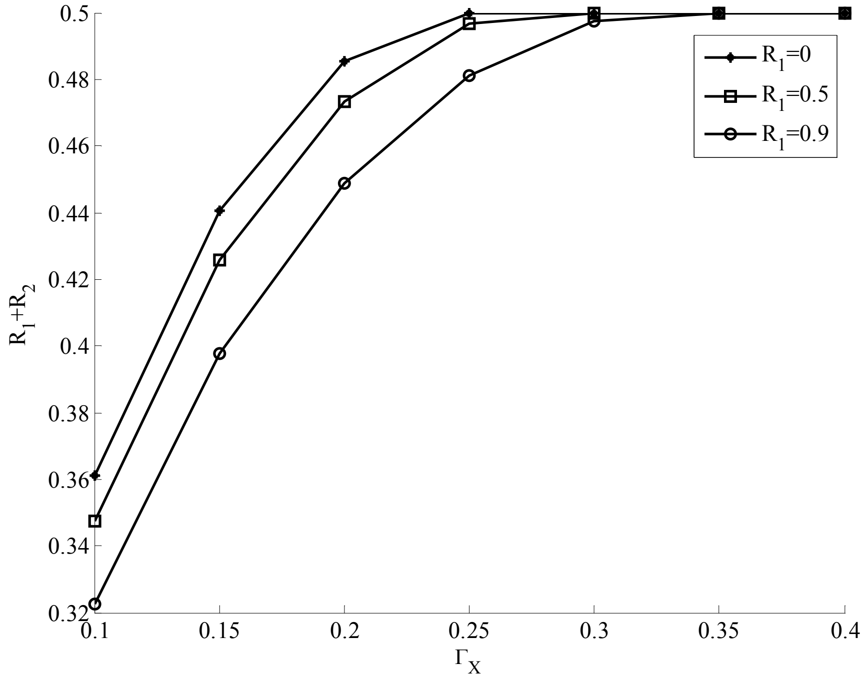

Figure 8.10 Sum-rate $R_1 + R_2$ versus the input cost constraint Γ_X for values of $R_1 = 0$, $R_1 = 0.5$ and $R_1 = 0.9$.

versa, communication, implicit or explicit, can provide an essential tool to improve control tasks. This work has focused on the first type of interplay, and has investigated the implications of embedding information directly in the actions for the aim of communicating with a separate decoder. The communication requirements of this decoder are generally in conflict with the goal of improving the efficiency of the given communication link. This performance trade-off has been studied here for both source and channel coding. The results provided in this chapter allow to give a quantitative answer to the questions posed in Section 8.1.1 regarding the impact of the requirements of action information embedding on the system performance. They also shed light into the structure of optimal embedding strategies, which turns out to be related, for the source coding model, with the strategies studied in [62, 64].

The investigation on the theme of information embedding on actions can be further developed in a number of directions, including models with memory [56, 28] and with multiple terminals [23, 55, 27]. It is also noted that results akin to the ones reported here can be developed assuming causal state information at the decoder for source coding problems or causal state information at the transmitter.

APPENDIX A

PROOF OF PROPOSITION 3.1

First, it is observed that from Definition 3.1, since distortion and CR constraints (3.8) and (3.9) depend only on the marginal pmfs $p(x, y_1)$ and $p(x, y_2)$, so does the rate-distortion function. Therefore, in the proof, one can assume, without loss of generality, that the joint pmf $p(x, y_1, y_2)$ satisfies the Markov chain condition $X - Y_2 - Y_1$ so that it factorizes as (cf. (3.10))

$$p(x, y_1, y_2) = p(x, y_2)\tilde{p}(y_1|y_2). \tag{A.1}$$

Consider an $(n, R, D_1 + \epsilon, D_2 + \epsilon, \epsilon)$ code, whose existence is required for achievability by Definition 3.1. By the CR requirements (3.9), first it is observed that the following Fano inequalities hold

$$H(\psi_j(X^n)|h_j(g(X^n), Y_j^n)) \leq n\delta(\epsilon), \text{ for } j = 1, 2, \tag{A.2}$$

for n sufficiently large, where $\delta(\epsilon) = n\epsilon\log|\mathcal{X}| + H_b(\epsilon)$. Moreover, one can write

$$\begin{aligned} nR = H(J) &\geq H(J|Y_1^n) && \text{(A.3a)} \\ &\overset{(a)}{=} H(J|Y_1^n Y_2^n) + I(J; Y_2^n|Y_1^n), && \text{(A.3b)} \end{aligned}$$

where (a) follows by the definition of mutual information. From now on, to simplify notation, explicit the dependence of ψ_j, g_j and h_j on X^n and (J, Y_j^n), respectively, is not made. Also, ψ_{ji} is defined as the ith symbol of the sequence ψ_j so that $\psi_j = (\psi_{j1}, ..., \psi_{jn})$.

The first term in (A.3b), $H(J|Y_1^n Y_2^n)$, can be treated as in [8, Section V.A.], or, more simply, one can proceed as follows:

$$\begin{aligned} H(J|Y_1^n Y_2^n) &\overset{(a)}{=} I(J; X^n|Y_1^n Y_2^n) && \text{(A.4a)} \\ &\overset{(b)}{\geq} I(h_1 h_2; X^n|Y_1^n Y_2^n) && \text{(A.4b)} \\ &= I(h_1 h_2 \psi_1 \psi_2; X^n|Y_1^n Y_2^n) - I(\psi_1 \psi_2; X^n|Y_1^n Y_2^n h_1 h_2) && \text{(A.4c)} \\ &\overset{(c)}{\geq} I(\psi_1 \psi_2; X^n|Y_1^n Y_2^n) - I(\psi_1 \psi_2; X^n|Y_1^n Y_2^n h_1 h_2) && \text{(A.4d)} \\ &\overset{(d)}{=} I(\psi_1 \psi_2; X^n|Y_2^n) - H(\psi_1 \psi_2|Y_1^n Y_2^n h_1 h_2) \\ &\quad + H(\psi_1 \psi_2|Y_1^n Y_2^n h_1 h_2 X^n) && \text{(A.4e)} \\ &\overset{(e)}{\geq} I(\psi_1 \psi_2; X^n|Y_2^n) - n\delta(\epsilon) && \text{(A.4f)} \\ &\overset{(f)}{\geq} \sum_{i=1}^{n} I(\psi_{1i} \psi_{2i}; X_i|Y_{2i}) - n\delta(\epsilon), && \text{(A.4g)} \end{aligned}$$

where (a) follows because J is a function of X^n; (b) follows since h_1 and h_2 are functions of (J, Y_1^n) and (J, Y_2^n), respectively ; (c) follows by using the Markov chain (ψ_1, ψ_1, X^n)—Y_2^n—Y_1^n; (d) follows by the chain rule of mutual information and since mutual information is non-negative; (e) follows by (R.2a) and since entropy is non-negative; and (f) follows by the chain rule for entropy, since X^n and Y_2^n are i.i.d., and due to the fact conditioning decreases entropy.

Similarly, the second term in (A.3b), namely, $I(J; Y_2^n|Y_1^n)$, leads to

$$\begin{aligned}
I(J; Y_2^n|Y_1^n) &\overset{(a)}{\geq} I(\mathrm{h}_1; Y_2^n|Y_1^n) && \text{(A.5a)}\\
&= I(\mathrm{h}_1\psi_1; Y_2^n|Y_1^n) - I(\psi_1; Y_2^n|Y_1^n h_1) && \text{(A.5b)}\\
&\overset{(b)}{\geq} I(\psi_1; Y_2^n|Y_1^n) - H(\psi_1|Y_1^n h_1) + H(\psi_1|Y_1^n Y_2^n h_1) && \text{(A.5c)}\\
&\overset{(c)}{\geq} I(\psi_1; Y_2^n|Y_1^n) - n\delta(\epsilon) && \text{(A.5d)}\\
&\overset{(d)}{\geq} \sum_{i=1}^{n} I(\psi_{1i}; Y_{2i}|Y_{1i}) - n\delta(\epsilon), && \text{(A.5e)}
\end{aligned}$$

where (a) follows because h_1 is a function of J and Y_1^n; (b) follows by the chain rule of mutual information and since mutual information is non-negative; (c) follows by (R.2a) and since entropy is non-negative; and (d) follows by the chain rule for entropy, since Y_2^n and Y_1^n are i.i.d., and due to the fact conditioning decreases entropy. From (A.3b), (A.4g), and (A.5e), it is also true that

$$\begin{aligned}
nR &\geq \sum_{i=1}^{n} I(\psi_{1i}\psi_{2i}; X_i|Y_{2i}) + I(\psi_{1i}; Y_{2i}|Y_{1i}) - n\delta(\epsilon) && \text{(A.6a)}\\
&\overset{(a)}{=} \sum_{i=1}^{n} I(X_i; \psi_{1i}|Y_{1i}) + I(X_i; \psi_{2i}|Y_{2i}\psi_{1i}) - n\delta(\epsilon), && \text{(A.6b)}
\end{aligned}$$

where (a) follows because of the Markov chain relationship $(\psi_{1i}, \psi_{2i}) - X_i - Y_{2i} - Y_{1i}$, for $i = 1, ..., n$. By defining $\hat{X}_{ji} = \psi_{ji}$ with $j = 1, 2$ and $i = 1, ..., n$, the proof is concluded as in [8].

APPENDIX B

PROOF OF PROPOSITION 3.2

As explained in the text, it is only required to focus on the case where $D_2 \leq D_1 \leq \sigma_x^2$. As per the discussion in Appendix A, one can assume, without loss of generality, that the Markov chain relationship $X - Y_2 - Y_1$ holds, so that

$$\begin{aligned} Y_2 &= X + Z_2 & \text{(B.1a)} \\ \text{and } Y_1 &= Y_2 + \tilde{Z}_1, & \text{(B.1b)} \end{aligned}$$

where $\tilde{Z}_1 \sim \mathcal{N}(0, N_1)$ is independent of (X, Z_2).

First a converse is proved. Calculating the rate-distortion function in (3.14) requires minimization over the pmf $p(\hat{x}_1, \hat{x}_2|x)$ under the constraint (3.13). A minimizing $p(\hat{x}_1, \hat{x}_2|x)$ exists by the Weierstrass theorem due to the continuity of the mutual information and the compactness of the set of pmfs defined by the constraint (3.13)[68]. Fixing one such optimizing $p(\hat{x}_1, \hat{x}_2|x)$, the rate-distortion function (3.14) can be written as

$$R_{HB}^{CR}(D_1, D_2) = I(X; \hat{X}_2|Y_2) + I(\hat{X}_1; Y_2|Y_1). \tag{B.2}$$

The first term in (B.2), i.e., $I(X; \hat{X}_2|Y_2)$, can be easily bounded using the approach in [8, p. 5007]. Specifically, the following holds

$$\begin{aligned} I(X; \hat{X}_2|Y_2) &= h(X|Y_2) - h(X|\hat{X}_2 Y_2) \\ &= h(X|X+Z_2) - h(X-\hat{X}_2|\hat{X}_2, \hat{X}_2 + (X-\hat{X}_2) + Z_2) \\ &= h(X|X+Z_2) - h(X-\hat{X}_2|\hat{X}_2, (X-\hat{X}_2) + Z_2) \\ &\overset{(a)}{\geq} h(X|X+Z_2) - h(X-\hat{X}_2|(X-\hat{X}_2) + Z_2) \\ &\overset{(b)}{\geq} \frac{1}{2}\log_2\left(2\pi e \frac{\sigma_x^2}{1+\frac{\sigma_x^2}{N_2}}\right) - \frac{1}{2}\log_2\left(2\pi e \frac{D_2}{1+\frac{D_2}{N_2}}\right) \\ &= \frac{1}{2}\log_2\left(\frac{\sigma_x^2}{\sigma_x^2+N_2} \cdot \frac{D_2+N_2}{D_2}\right), \end{aligned} \tag{B.3}$$

where (a) follows because conditioning decreases entropy; and (b) follows from the maximum conditional entropy lemma [1, p. 21], which implies that $h(E|E+Z_2) \leq \frac{1}{2}\log_2(2\pi e \sigma^2_{E|E+Z_2})$ with $E = X - \hat{X}_2$. In fact, it is true that $\sigma^2_{E|E+Z_2} \leq \frac{D_2}{1+\frac{D_2}{N_2}}$, since the conditional variance $\sigma^2_{E|E+Z_2}$ is upper bounded by the linear minimum mean square error of the estimate of E given $E + Z_2$. This mean square error is given by $\frac{D_2}{1+\frac{D_2}{N_2}}$, since $\mathrm{E}[E^2] \leq D_2$ and since Z_2 is independent of E due to the factorization (3.12) and to the independence of X and Z_2. For the second term in (B.2), the following holds:

$$\begin{aligned} I(\hat{X}_1;Y_2|Y_1) &= h(Y_2|Y_1) - h(Y_2|Y_1,\hat{X}_1) \\ &= \frac{1}{2}\log_2\left(2\pi e\frac{N_1(N_2+\sigma_x^2)}{N_1+N_2+\sigma_x^2}\right) - h(Y_2|Y_1,\hat{X}_1). \end{aligned} \tag{B.4}$$

Moreover, it can be evaluated as

$$\begin{aligned} h(Y_2|Y_1,\hat{X}_1) &= h(Y_2,Y_1|\hat{X}_1) - h(Y_1|\hat{X}_1) \\ &= h(Y_2|\hat{X}_1) + h(Y_1|Y_2,\hat{X}_1) - h(Y_1|\hat{X}_1) \\ &= h(Y_2|\hat{X}_1) - h(Y_2+\tilde{Z}_1|\hat{X}_1) + h(Y_2+\tilde{Z}_1|Y_2,\hat{X}_1) \\ &\overset{(a)}{=} h(Y_2|\hat{X}_1) - h(Y_2+\tilde{Z}_1|\hat{X}_1) + \frac{1}{2}\log_2(2\pi e N_1), \end{aligned} \tag{B.5}$$

where (a) follows because $\tilde{Z}_1$ is independent of Y_2 and of $\hat{X}_1$, due to the factorization (3.12) and due to the independence of $\tilde{Z}_1$ and X. Next, a lower bound is obtained on the term $h(Y_2+\tilde{Z}_1|\hat{X}_1)$ in (B.5) as a function of $h(Y_2|\hat{X}_1)$ by using the entropy power inequality (EPI) [1, p. 22]. Specifically, by using the conditional version of EPI [1, p. 22], it holds that

$$\begin{aligned} 2^{2h(Y_2+\tilde{Z}_1|\hat{X}_1)} &\geq 2^{2h(Y_2|\hat{X}_1)} + 2^{2h(\tilde{Z}_1|\hat{X}_1)} \\ &\overset{(a)}{=} 2^{2h(Y_2|\hat{X}_1)} + 2^{2h(\tilde{Z}_1)} \\ &= 2^{2h(Y_2|\hat{X}_1)} + 2\pi e N_1, \end{aligned} \tag{B.6}$$

where (a) follows because $\tilde{Z}_1$ is independent of $\hat{X}_1$ as explained above. The first two terms in (B.5) can thus be bounded as

$$\begin{aligned} h(Y_2|\hat{X}_1) - h(Y_2+\tilde{Z}_1|\hat{X}_1) &\leq h(Y_2|\hat{X}_1) - \frac{1}{2}\log(2^{2h(Y_2|\hat{X}_1)} + 2\pi e N_1) \\ &= \frac{1}{2}\log_2\left(\frac{2^{2h(Y_2|\hat{X}_1)}}{2^{2h(Y_2|\hat{X}_1)} + 2\pi e N_1}\right) \\ &\overset{(a)}{\leq} \log_2\left(\frac{2\pi e(D_1+N_2)}{2\pi e(D_1+N_2) + 2\pi e N_1}\right), \end{aligned} \tag{B.7}$$

where (a) follows because $\log_2\left(\frac{2^{2h(Y_2|\hat{X}_1)}}{2^{2h(Y_2|\hat{X}_1)}+2\pi e N_1}\right)$ is an increasing function of $h(Y_2|\hat{X}_1)$ and $h(Y_2|\hat{X}_1) \leq \frac{1}{2}\log_2(2\pi e(D_1+N_2))$, as can be proved by using the same approach used for the bounds (a) and (b) in (B.3). By substituting (B.7) into (B.5), and using the result in (B.4), one can obtain

$$\begin{aligned} I(\hat{X}_1;Y_2|Y_1) &\geq \frac{1}{2}\log_2\left(2\pi e\frac{N_1(N_2+\sigma_x^2)}{N_1+N_2+\sigma_x^2}\right) - \frac{1}{2}\log_2\left(2\pi e\frac{N_1(D_1+N_2)}{D_1+N_2+N_1}\right) \\ &= \frac{1}{2}\log_2\left(\frac{(N_2+\sigma_x^2)(D_1+N_2+N_1)}{(N_1+N_2+\sigma_x^2)(D_1+N_2)}\right). \end{aligned} \tag{B.8}$$

Finally, by substituting (B.3) and (B.8) into (B.2), the lower bound is obtained as

$$\begin{aligned}
R_{HB}^{CR}(D_1, D_2) &\geq \frac{1}{2}\log_2\left(\frac{\sigma_x^2}{\sigma_x^2+N_2}\cdot\frac{D_2+N_2}{D_2}\right)+\frac{1}{2}\log_2\left(\frac{(N_2+\sigma_x^2)(D_1+N_2+N_1)}{(N_1+N_2+\sigma_x^2)(D_1+N_2)}\right) \\
&= \frac{1}{2}\log_2\left(\frac{\sigma_x^2}{(\sigma_x^2+N_1+N_2)}\cdot\frac{(D_1+N_1+N_2)(D_2+N_2)}{(D_1+N_2)D_2}\right). \qquad \text{(B.9)}
\end{aligned}$$

For achievability, (3.14) is calculated with $X = \hat{X}_2 + Q_2$ and $\hat{X}_2 = \hat{X}_1 + Q_1$, where $Q_1 \sim \mathcal{N}(0, D_1 - D_2)$ and $Q_2 \sim \mathcal{N}(0, D_2)$ are independent of each other and of $(\hat{X}_1,\tilde{Z}_1, Z_2)$. This leads to the upper bound

$$\begin{aligned}
R_{HB}^{CR}(D_1, D_2) &\leq I(X;\hat{X}_1\hat{X}_2|Y_2) + I(\hat{X}_1;Y_2|Y_1) \\
&= I(X;\hat{X}_2|Y_2) + I(\hat{X}_1;Y_2|Y_1) \\
&= h(X|Y_2) - h(X|Y_2,\hat{X}_2) + h(Y_2|Y_1) - h(Y_2|Y_1,\hat{X}_1) \\
&= h(X|X+Z_2) - h(\hat{X}_2+Q_2|\hat{X}_2+Q_2+Z_2,\hat{X}_2) \\
&\quad +h(X+Z_2|X+Z_2+\tilde{Z}_1) \\
&\quad -h(\hat{X}_1+Q_1+Q_2+Z_2|\hat{X}_1+Q_1+Q_2+Z_2+\tilde{Z}_1,\hat{X}_1) \\
&= h(X|X+Z_2) - h(Q_2|Q_2+Z_2) + h(X+Z_2|X+Z_2+\tilde{Z}_1) \\
&\quad -h(Q_1+Q_2+Z_2|Q_1+Q_2+Z_2+\tilde{Z}_1) \\
&\overset{(a)}{=} \frac{1}{2}\log_2\left(2\pi e\frac{\sigma_x^2}{1+\frac{\sigma_x^2}{N_2}}\right) - \frac{1}{2}\log_2\left(2\pi e\frac{D_2}{1+\frac{D_2}{N_2}}\right) \\
&\quad +\frac{1}{2}\log_2\left(2\pi e\frac{\sigma_x^2+N_2}{1+\frac{\sigma_x^2+N_2}{N_1}}\right) - \frac{1}{2}\log_2\left(2\pi e\frac{D_1+N_2}{1+\frac{D_1+N_2}{N_1}}\right) \\
&= \frac{1}{2}\log_2\left(\frac{\sigma_x^2}{(\sigma_x^2+N_1+N_2)}\cdot\frac{(D_1+N_1+N_2)(D_2+N_2)}{(D_1+N_2)D_2}\right), \qquad \text{(B.10)}
\end{aligned}$$

where (a) follows using $h(A|A+B) = \frac{1}{2}\log_2\left(2\pi e\frac{S_A}{1+\frac{S_A}{S_B}}\right)$, for A and B being independent Gaussian sources with $A \sim \mathcal{N}(0, S_A)$ and $B \sim \mathcal{N}(0, S_B)$. By comparing (B.9) with (B.10), the proof is completed.

APPENDIX C

PROOF OF (3.23)

Here, it is proved that (3.2) equals (3.23) for the given sources. For the converse, the following holds

$$\begin{aligned}
I(X;\hat{X}|Y) &= H(X|Y) - H(X|\hat{X},Y) \\
&= p - pH(X|\hat{X}, Y \neq X) - (1-p_2)H(X|\hat{X}, Y = X) \\
&= p - pH(X|\hat{X}, Y \neq X) \\
&= p - pH(X|\hat{X}) \\
&= p - pH(X \oplus \hat{X}|\hat{X}) \\
&\overset{(a)}{\geq} p - pH(X \oplus \hat{X}) \\
&\geq p - pH(D) \\
&= p(1 - H(D)),
\end{aligned} \tag{C.1}$$

where (a) follows because conditioning decreases entropy. Achievability follows by calculating (3.2) with $X = \hat{X} \oplus Q$ where $Q \sim \text{Ber}(D)$.

APPENDIX D

PROOF OF PROPOSITION 3.3

As explained in the text, it is only required to focus on the case where $D_2 \leq D_1 \leq 1/2$. As for Appendix A and Appendix B, one can assume, without loss of generality, that the joint pmf of (x, y_1, y_2) factorizes as (A.1) as shown Figure 3.7. First, a converse is proved. Similar to (B.2), the rate-distortion function (3.14) can be written as

$$R_{HB}^{CR}(D_1, D_2) = I(X; \hat{X}_2|Y_2) + I(\hat{X}_1; Y_2|Y_1), \tag{D.1}$$

where the mutual information terms are calculated with a distribution $p(\hat{x}_1, \hat{x}_2|x)$ minimizing (3.14) under the constraint (3.13). The first term in (D.1), i.e., $I(X; \hat{X}_2|Y_2)$, can be easily bounded by following the same steps used in the derivation of (C.1), leading to

$$I(X; \hat{X}_2|Y_2) \geq p_2(1 - H(D_2)). \tag{D.2}$$

For the second term in (D.1), instead the following holds

$$\begin{aligned}
I(\hat{X}_1; Y_2|Y_1) &= H(Y_2|Y_1) - H(Y_2|Y_1, \hat{X}_1) \\
&= H(Y_2|Y_1) - H(Y_2, Y_1|\hat{X}_1) + H(Y_1|\hat{X}_1) && \text{(D.3)} \\
&= H(Y_2|Y_1) - H(Y_2|\hat{X}_1) - H(Y_1|\hat{X}_1, Y_2) + H(Y_1|\hat{X}_1) && \text{(D.4)} \\
&\overset{(a)}{=} H(Y_2|Y_1) - H(Y_2|\hat{X}_1) - H(Y_1|Y_2) + H(Y_1|\hat{X}_1), && \text{(D.5)}
\end{aligned}$$

where (a) follows because of the Markov chain condition $Y_1 - Y_2 - \hat{X}_1$. The second term in the right-hand side of (D.5) can be evaluated as

$$\begin{aligned}
H(Y_2|\hat{X}_1) &= H(Y_2, X|\hat{X}_1) - H(X|Y_2, \hat{X}_1) \\
&= H(X|\hat{X}_1) + H(Y_2|X, \hat{X}_1) - H(X|Y_2, \hat{X}_1) \\
&= H(X|\hat{X}_1) + H(Y_2|X) - p_2 H(X|Y_2 \neq X, \hat{X}_1) \\
&\quad -(1 - p_2) H(X|Y_2 = X, \hat{X}_1) \\
&\overset{(a)}{=} H(X|\hat{X}_1) + H(p_2) - p_2 H(X|\hat{X}_1) \\
&= H(p_2) + (1 - p_2) H(X|\hat{X}_1) && \text{(D.6)}
\end{aligned}$$

where (a) follows because $H(Y_2|X) = H(p_2)$. The fourth term in the right-hand side of (D.5) can similarly be evaluated as

$$H(Y_1|\hat{X}_1) = H(p_1) + (1 - p_1) H(X|\hat{X}_1). \tag{D.7}$$

Substituting (D.6) and (D.7) in (D.5), it holds that

$$\begin{aligned}
I(\hat{X}_1; Y_2|Y_1) &= H(p_1) + (1-p_1)H(X|\hat{X}_1) - (H(p_2) + (1-p_2)H(X|\hat{X}_1)) \\
&\quad + H(Y_2|Y_1) - H(Y_1|Y_2) \\
&= H(p_1) - H(p_2) - (p_1 - p_2)H(X|\hat{X}_1) + H(Y_2) - H(Y_1) \\
&\overset{(a)}{\geq} (p_1 - p_2) - (p_1 - p_2)H(D_1) \qquad \text{(D.8)}
\end{aligned}$$

where (a) follows since $H(Y_2) = H(p_2) + (1-p_2)$ and $H(Y_1) = H(p_1) + (1-p_1)$ and due to the inequality $H(X|\hat{X}_1) \leq H(D_1)$. Substituting (D.8) and (D.2) into (D.1), it is true that

$$\begin{aligned}
R_{HB}^{CR}(D_1, D_2) &\geq p_2(1 - H(D_2)) + (p_1 - p_2)(1 - H(D_1)) \\
&= p_1(1 - H(D_1)) + p_2(H(D_1) - H(D_2)). \qquad \text{(D.9)}
\end{aligned}$$

For achievability, (3.14) is calculated with $X = \hat{X}_2 \oplus Q_2$ and $\hat{X}_2 = \hat{X}_1 \oplus Q_1$, where $Q_1 \sim \text{Ber}(D_1 * D_2)$ and $Q_2 \sim \text{Ber}(D_2)$ are independent of each other and of $(\hat{X}_1, E_1, E_2)$ where $E_j = \mathbf{1}\{Y_j = \mathrm{e}\}$ for $j = 1, 2$. This leads to the upper bound

$$\begin{aligned}
R_{HB}^{CR}(D_1, D_2) &\leq I(X; \hat{X}_1\hat{X}_2|Y_2) + I(\hat{X}_1; Y_2|Y_1) \\
&= I(X; \hat{X}_2|Y_2) + I(\hat{X}_1; Y_2|Y_1) \\
&= H(X|Y_2) - H(X|Y_2, \hat{X}_2) + H(Y_2|Y_1) - H(Y_2|Y_1, \hat{X}_1) \\
&\overset{(a)}{=} p_2 - p_2 H(X|\hat{X}_2, Y_2 \neq X) - (1-p_2)H(X|\hat{X}_2, Y_2 = X) + p_1 H\left(\frac{p_2}{p_1}\right) \\
&\quad + \tilde{p}_1(1-p_2) - p_1 H(Y_2|\hat{X}_1, Y_1 = e) - (1-p_1)H(Y_2|\hat{X}_1, Y_1 = X) \\
&\overset{(b)}{=} p_2 - p_2 H(X|\hat{X}_2) + p_1 H\left(\frac{p_2}{p_1}\right) + \tilde{p}_1(1-p_2) \\
&\quad - p_1\left(H\left(\frac{p_2}{p_1}\right) + \frac{\tilde{p}_1(1-p_2)}{p_1} H(X|\hat{X}_1)\right) \\
&\overset{(c)}{=} p_2 - p_2 H(\hat{X}_2 \oplus Q_2|\hat{X}_2) + \tilde{p}_1(1-p_2) \\
&\quad - \tilde{p}_1(1-p_2)H(\hat{X}_1 \oplus Q_1 \oplus Q_2|\hat{X}_1) \\
&\overset{(d)}{=} p_2 - p_2 H(D_2 + \tilde{p}_1(1-p_2) - \tilde{p}_1(1-p_2)H(D_1) \\
&= p_1(1 - H(D_1)) + p_2(H(D_1) - H(D_2)), \qquad \text{(D.10)}
\end{aligned}$$

where (a) follows because $H(Y_2|Y_1) = p_1 H\left(\frac{p_2}{p_1}\right) + \tilde{p}_1(1-p_2)$; (b) follows because $H(Y_2|\hat{X}_1, Y_1 = X) = H(X|\hat{X}_1, Y_1 = X) = 0$ and $H(Y_2|\hat{X}_1, Y_1 = e) = H\left(\frac{p_2}{p_1}\right) + \frac{\tilde{p}_1(1-p_2)}{p_1} H(X|\hat{X}_1)$; (c) follows by using the inverse test channels $X = \hat{X}_2 \oplus Q_2$ and $\hat{X}_2 = \hat{X}_1 \oplus Q_1$; and (d) follows because $Q_2 \sim \text{Ber}(D_2)$ and $Q_1 \oplus Q_2 \sim \text{Ber}(D_1)$. By comparing (D.9) with (D.10), the proof is completed.

APPENDIX E

PROOF OF PROPOSITION 3.9

Here the proof of Proposition 3.9 is provided. To this end, it is proved that for any pair (D_1, D_2) there exists a joint distribution $p(\hat{x}_1, \hat{x}_2|x)$ such that (3.13) is satisfied and the conditions (3.43a) and (3.43b) coincide with (3.42a) and (3.42b), respectively. This entails that the inner and outer bounds of Proposition 3.7 and Proposition 3.8 coincide.

The four region in the (D_1, D_2) plane depicted in Figure 3.5 are distinguished. If $D_1 \geq \sigma_x^2$ and $D_2 \geq \sigma_x^2$, it is enough to set $\hat{X}_1 = \hat{X}_2 = 0$ in (3.43) to prove. For $D_1 \leq \sigma_x^2$ and $D_2 \geq \min(D_1, \sigma_x^2)$, instead one can set $\hat{X}_1 = \hat{X}_2$ and $X = \hat{X}_1 + Q_1$ in (3.43), where $Q_1 \sim \mathcal{N}(0, D_1)$ is independent of $\hat{X}_1$. Following the discussion in Section 3.2.3, it is easy to see that this choice is such that (3.43) coincides with (3.42). Next, in the sub-region where $D_1 \geq \sigma_x^2$ and $D_2 \leq \sigma_x^2$, one can select $\hat{X}_1 = 0$ and $X = \hat{X}_2 + Q_2$ in (3.43), where $Q_2 \sim \mathcal{N}(0, D_2)$ is independent of $\hat{X}_2$. Finally, for the region in Figure 3.5, for which $D_2 \leq D_1 \leq \sigma_x^2$, One can choose $X = \hat{X}_2 + Q_2$ and $\hat{X}_2 = \hat{X}_1 + Q_1$, where $Q_1 \sim \mathcal{N}(0, D_1 - D_2)$ and $Q_2 \sim \mathcal{N}(0, D_2)$ are independent of each other and of $(\hat{X}_1, E_1, E_2)$. With this choice, following the derivations in Appendix B, it is concluded that condition (3.43a) coincides with (3.42a). As for (3.43b), the following holds

$$\begin{aligned}
I(X; \hat{X}_1|Y_2) + I(X; \hat{X}_2|\hat{X}_1 Y_2) &= I(X; \hat{X}_1 \hat{X}_2|Y_2) \\
&= h(X|Y_2) - h(X|\hat{X}_1, \hat{X}_2, Y_2) \\
&= h(X|X + Z_2) \\
&\quad - h(\hat{X}_1 + Q_1 + Q_2|\hat{X}_1, \hat{X}_1 + Q_1, \hat{X}_1 + Q_1 + Q_2 + Z_2) \\
&= h(X|X + Z_2) - h(Q_1 + Q_2|Q_1, Q_1 + Q_2 + Z_2) \\
&= h(X|X + Z_2) - h(Q_2|Q_2 + Z_2) \\
&= \frac{1}{2}\log_2\left(\frac{\sigma_x^2}{1 + \frac{\sigma_x^2}{N_2}}\right) - \frac{1}{2}\log_2\left(\frac{D_2}{1 + \frac{D_2}{N_2}}\right) \\
&= \frac{1}{2}\log_2\left(\frac{\sigma_x^2}{\sigma_x^2 + N_2}\frac{D_2 + N_2}{D_2}\right) \\
&= R_{\mathcal{G}}^{CR}(D_2, N_2), \qquad \text{(E.1)}
\end{aligned}$$

which concludes the proof.

APPENDIX F

PROOF OF PROPOSITION 3.10

Here the proof of Proposition 3.10 is provided. Following similar steps as in Appendix E, it is proved that for any pair (D_1, D_2) there exists a joint distribution $p(\hat{x}_1, \hat{x}_2|x)$ such that (3.13) is satisfied and the conditions (3.43a) and (3.43b) coincide with (3.42a) and (3.42b), respectively. This entails that the inner and outer bounds of Proposition 3.7 and Proposition 3.8 coincide.

The four region in the (D_1, D_2) plane depicted in Figure 3.8 are distinguished. If $D_1 \geq 1/2$ and $D_2 \geq 1/2$, it is enough to set $\hat{X}_1 = \hat{X}_2 = 0$ in (3.43) to prove the desired result. For $D_1 \leq 1/2$ and $D_2 \geq \min(D_1, 1/2)$, one can instead set $\hat{X}_1 = \hat{X}_2$ and $X = \hat{X}_1 \oplus Q_1$ in (3.43), where $Q_1 \sim \mathrm{Ber}(D_1)$ is independent of $\hat{X}_1$. Following the discussion in Section 3.2.4, it is easy to see that this choice is such that (3.43) coincides with (3.42). Next, in the sub-region where $D_1 \geq 1/2$ and $D_2 \leq 1/2$, one can select $\hat{X}_1 = 0$ and $X = \hat{X}_2 \oplus Q_2$ in (3.43), where $Q_2 \sim \mathrm{Ber}(D_2)$ is independent of $\hat{X}_2$. Finally, for the region in Figure 3.8, for which $D_2 \leq D_1 \leq 1/2$, one can choose $X = \hat{X}_2 \oplus Q_2$ and $\hat{X}_2 = \hat{X}_1 \oplus Q_1$, where $Q_1 \sim \mathrm{Ber}(D_1 * D_2)$ and $Q_2 \sim \mathrm{Ber}(D_2)$ are independent of each other and of $(\hat{X}_1, E_1, E_2)$. With this choice, following the derivations in Appendix D, it is concluded that condition (3.43a) coincides with (3.42a). As for (3.43b), the following hold

$$
\begin{aligned}
I(X; \hat{X}_1 \hat{X}_2 | Y_2) &= H(X|Y_2) - H(X|\hat{X}_1, \hat{X}_2, Y_2) \\
&= p_2 - p_2 H(X|\hat{X}_1, \hat{X}_2, Y_2 \neq X) - (1 - p_2) H(X|\hat{X}_1, \hat{X}_2, Y_2 = X) \\
&= p_2 - p_2 H(X|\hat{X}_1, \hat{X}_2) \\
&\overset{(a)}{=} p_2 - p_2 H(X|\hat{X}_2) \\
&= p_2 - p_2 H(\hat{X}_2 \oplus Q_2 | \hat{X}_2) \\
&= p_2 - p_2 H(D_2) \\
&= R_B^{CR}(D_2, p_2), \qquad \text{(F.1)}
\end{aligned}
$$

where (a) follows by the Markov chain relationship $X - \hat{X}_2 - \hat{X}_1$. This completes the proof.

APPENDIX G

PROOF OF PROPOSITION 3.11

The proof of the achievability follows from standard arguments, similar to [3]. For the converse, following the proof of [3, Theorem 3] It is true that for any $(R, D_{e,1}+\epsilon, D_{e,2}+\epsilon, D_1+\epsilon, D_2+\epsilon)$ code, the following inequality holds:

$$nR \geq \sum_{i=1}^{n} I(X_i;U_{1i}|Y_{1i}) + I(X_i;U_{2i}|Y_{2i}), \tag{G.1}$$

with the definitions $U_{ji} \triangleq (J, Y_j^{n\backslash i})$, for $j=1,2$, with $Y_j^{n\backslash i} = [Y_{j1}^{i-1}, Y_{j(i+1)}^{n}]$. Note that with the given definition of U_{ji}, the ith element of the decoding functions (3.5)-(3.6) can be written as $h_{ji}(J, Y_j^n) = \hat{\mathrm{x}}_{ji}(U_{ji}, Y_{ji})$ for all $i=1,...,n$ and $j=1,2$. Now, defining $D_{e,ji} \triangleq \mathrm{E}[d_{e,j}(h_{ji}^n(M, Y_j^n), \psi_{ji}(X^n)]$, The following chain of inequalities for the code at hand and $j=1,2$ hold:

$$\begin{aligned}
D_{e,ji} &= \mathrm{E}_{X^nY_j^n}[d_{e,j}(h_{ji}^n(J,Y_j^n),\psi_{ji}(X^n))] && \text{(G.2a)}\\
&\overset{(a)}{=} \mathrm{E}_{X^nU_{ji}Y_{ji}}[d_{e,j}(\hat{\mathrm{x}}_{ji}(U_{ji},Y_{ji}),\psi_{ji}(X^n))] && \text{(G.2b)}\\
&= \mathrm{E}_{X^nU_{ji}}\mathrm{E}_{Y_{ji}}[d_{e,j}(\hat{\mathrm{x}}_{ji}(U_{ji},Y_{ji}),\psi_{ji}(X_i,X^{n\backslash i}))|X^nU_{ji}] && \text{(G.2c)}\\
&= \sum_{x^n\in\mathcal{X}^n,u_{ji}\in\mathcal{U}}^{n} p(x^n,u_{ji}) && \text{(G.2d)}\\
&\quad \mathrm{E}_{Y_{ji}}[d_{e,j}(\hat{\mathrm{x}}_{ji}(U_{ji},Y_{ji}),\psi_{ji}(X_i,X^{n\backslash i}))|X_i=x_i,X^{n\backslash i}=x^{n\backslash i},U_{ji}=u_{ji}]\\
&\overset{(b)}{\geq} \sum_{x^n\in\mathcal{X}^n,u_{ji}\in\mathcal{U}}^{n} p(x^n,u_{ji}) && \text{(G.2e)}\\
&\quad \mathrm{E}_{Y_{ji}}[d_{e,j}(\hat{\mathrm{x}}_{ji}(U_{ji},Y_{ji}),\psi_{ji}(X_i,X^{n\backslash i}))|X_i=x_i,X^{n\backslash i}=\mathrm{x}^{*n\backslash i}(x_i,u_{ji}),U_{ji}=u_{ji}]\\
&\overset{(c)}{=} \sum_{x^n\in\mathcal{X}^n,u_{ji}\in\mathcal{U}}^{n} p(x^n,u_{ji}) && \text{(G.2f)}\\
&\quad \mathrm{E}_{Y_{ji}}[d_{e,j}(\hat{\mathrm{x}}_{ji}(U_{ji},Y_{ji}),\hat{\mathrm{x}}_{e,ji}(U_{ji},X_i))|X_i=x_i,U_{ji}=u_{ji}]\\
&= \mathrm{E}_{X_iU_{ji}Y_{ji}}[d_{e,j}(\hat{\mathrm{x}}_{ji}(U_{ji},Y_{ji}),\hat{\mathrm{x}}_{e,ji}(U_{ji},X_i))], && \text{(G.2g)}
\end{aligned}$$

where (a) follows by using the definition of random variables $U_j = (J, Y_j^{n\backslash i})$; (b) follows by selecting $\mathrm{x}^{*n\backslash i}(x_i,u_{ji})$ as

$$\mathrm{x}^{*n\backslash i}(x_i,u_{ji}) \in \mathrm{argmin}_{x^{n\backslash i}\in\mathcal{X}^{n\backslash i}} \\ \mathrm{E}_{Y_{ji}}[d_{e,j}(\hat{\mathrm{x}}_{ji}(U_{ji},Y_{ji}),\psi_{ji}(X_i,X_i^{n\backslash i}))|X_i=x_i,X^{n\backslash i}=x^{n\backslash i},U_{ji}=u_{ji}];$$

and (c) follows from the Markov chain relationship $Y_{ji} - (X_i, U_{ji}) - X^{n\backslash i}$ and from the definition $\hat{\mathrm{x}}_{e,ji}(U_{ji}, X_i) = \psi_{ji}(X_i, \mathrm{x}^{*n\backslash i}(X_i, U_{ji}))$. Let Q be a uniform random variable over the interval $[1, n]$ and independent of the variables $(X^n, Y_1^n, Y_2^n, U_1^n, U_2^n, \hat{X}_1^n, \hat{X}_2^n$ $, \hat{X}_{e,1}^n, \hat{X}_{e,2}^n)$ and define the random variables $U_j \triangleq (Q, U_{jQ})$, $X \triangleq X_Q$, $Y_j \triangleq Y_{jQ}$, $\hat{X}_j \triangleq \hat{X}_{jQ}$, and $\hat{X}_{e,j} \triangleq \hat{X}_{e,jQ}$ for $j = 1, 2$. Moreover, note that $\hat{X}_j$ is a deterministic function of U_{ji} and Y_{ji}, and $\hat{X}_{e,j}$ is a deterministic function of U_{ji} and X_i for $j = 1, 2$. The proof is completed by using (3.45) and the fact that the term $I(X_i; U_{1i}|Y_{1i}) + I(X_i; U_{2i}|Y_{2i})$ in (G.1) is convex with respect to the pmf $p(u_{1i}, u_{2i}|x_i)$, using standard steps (see, e.g., [13]).

APPENDIX H

CARDINALITY BOUNDS

Using standard inequalities, it can be seen that the rate region (4.12) evaluated with a constant Q is a contra-polymatroid, as the Berger-Tung region (4.17) (see e.g., [69]). Moreover, the role of the variable Q is that of performing the convexification of the union of all regions of tuples $(R_1, R_2, D_1, D_2, \Gamma)$ that satisfy (4.12) and (4.14) for some fixed Q. It follows from [69] that every extreme point of region of achievable tuples $(R_1, R_2, D_1, D_2, \Gamma)$ satisfies the equations

$$\begin{aligned} R_1 &= I(X_1; V_1|V_2) + I(X_1; U_1|U_2, V_1, V_2, Y) && \text{(H.1a)} \\ R_2 &= I(X_2; V_2) + I(X_2; U_2|V_1, V_2, Y) && \text{(H.1b)} \end{aligned}$$

along with (4.14), where both relationships are satisfied with equality, or

$$\begin{aligned} R_1 &= I(X_1; V_1) + I(X_1; U_1|V_1, V_2, Y) && \text{(H.2a)} \\ R_2 &= I(X_2; V_2|V_1) + I(X_2; U_2|U_1, V_1, V_2, Y) && \text{(H.2b)} \end{aligned}$$

along with (4.14) satisfied with equality. Applying the Fenchel–Eggleston–Caratheodory theorem to the right-hand side of the equations above and to (4.14) concludes the proof (See [1, Appendix C] and [29]).

APPENDIX I

PROOF OF THE CONVERSE FOR PROPOSITION 4.3

In this section, the proof of converse for Proposition 4.3 is given. For any $(n, R_1, R_2, D_1+\epsilon, D_2+\epsilon, \Gamma+\epsilon)$ code, the following inequalities hold

$$\begin{aligned}
nR_1 &\geq H(M_1) \geq H(M_1|M_2) \\
&\stackrel{(a)}{=} I(M_1; X_1^n, X_2^n|M_2) \\
&= \sum_{i=1}^{n} H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_2) - H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_1, M_2) \\
&\stackrel{(b)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_2) - H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_1, M_2, Y^{i-1}) \\
&\stackrel{(c)}{\geq} \sum_{i=1}^{n} H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_2, Y^{i-1}) - H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_1, M_2, Y^{i-1}) \\
&\stackrel{(d)}{=} \sum_{i=1}^{n} I(X_{1i}, X_{2i}; U_{1i}|U_{2i}),
\end{aligned}$$

where (a) follows because M_1 is a function of (X_1^n, X_2^n) given that X_2^n is a function of X_1^n by assumption; (b) follows since (X_{1i}, X_{2i})—$(X_1^{i-1}, X_2^{i-1}, M_1, M_2)$—$Y^{i-1}$forms a Markov chain; (c) follows by the fact that conditioning decreases entropy; and (d) follows by defining $U_{ji} = (X_1^{i-1}, X_2^{i-1}, Y^{i-1}, M_j)$ for $j = 1, 2$. A similar chain of inequalities hold for R_2. As for the sum-rate $R_1 + R_2$, it is true that

$$\begin{aligned}
n(R_1 + R_2) &\geq H(M_1, M_2) \\
&\stackrel{(a)}{=} I(M_1, M_2; X_1^n, X_2^n) \\
&= \sum_{i=1}^{n} H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}) - H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_1, M_2) \\
&\stackrel{(b)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}) - H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_1, M_2, Y^{i-1}) \\
&\stackrel{(c)}{\geq} \sum_{i=1}^{n} I(X_{1i}, X_{2i}; U_{1i}, U_{2i}),
\end{aligned}$$

where (a) follows because (M_1, M_2) are functions of (X_1^n, X_2^n); (b) follows since (X_{1i}, X_{2i})— $(X_1^{i-1}, X_2^{i-1}, M_1, M_2)$—$Y^{i-1}$ forms a Markov chain; and (c) follows using the definition of U_{ji} for $j = 1, 2$. Next, let Q be a uniform random variable over the interval $[1, n]$ and independent of $(X_1^n, X_2^n, U_1^n, U_2^n, Y^n)$ and define $U_j \triangleq (Q, U_{jQ})$, for $j = 1, 2$, $X_1 \triangleq X_{1Q}$, $X_2 \triangleq X_{2Q}$, $Y \triangleq Y_Q$. Note that $\hat{X}_j$ is a function of U_1, U_2 and Y for $j = 1, 2$. Moreover, from (4.8) and (7.7), it holds that

$$\Gamma + \epsilon \geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\Lambda(A_i)\right] = \mathrm{E}[\Lambda(A)] \tag{I.1}$$

$$\text{and } D_j + \epsilon \geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[d_j(X_{1i}, X_{2i}, Y_i.\hat{X}_{ji},)\right] = \mathrm{E}[d_1(X_1, X_2, Y, \hat{X}_j)], \text{ for } j = 1, 2. \tag{I.2}$$

APPENDIX J

PROOF OF THE CONVERSE FOR PROPOSITION 4.4

In this section, the proof of converse for Proposition 4.4 is given. Fix a code $(n, R_1, R_2, D_1 + \epsilon, \epsilon, \Gamma)$ for an $\epsilon > 0$, whose existence for all sufficiently large n is required by the definition of achievability.

From the distortion constraint for $\hat{X}_2$, the following inequality holds

$$\epsilon \geq \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[d_H(X_{2i}, \hat{X}_{2i})] \overset{(a)}{=} \frac{1}{n}\sum_{i=1}^{n}p_{e,2i}, \tag{J.1}$$

where the definition $p_{e,2i} = \Pr[X_{2i} \neq \hat{X}_{2i}]$ is used, and (a) follows from the definition of the metric $d_H(x, \hat{x})$ as the Hamming distortion. Moreover, the following chain of inequalities hold

$$\begin{aligned}
H(X_2^n|\hat{X}_2^n) &\overset{(a)}{\leq} \sum_{i=1}^{n}H(X_{2i}|\hat{X}_{2i}) \overset{(b)}{\leq} \sum_{i=1}^{n}H(p_{e,i}) + p_{e,i}\log\left|\hat{\mathcal{X}}_{2i}\right| \\
&\overset{(c)}{\leq} nH\left(\frac{1}{n}\sum_{i=1}^{n}p_{e,i}\right) + n\left(\frac{1}{n}\sum_{i=1}^{n}p_{e,i}\right)\log\left|\hat{\mathcal{X}}_{2i}\right| \\
&\overset{(d)}{\leq} nH(\epsilon) + n\epsilon\log\left|\hat{\mathcal{X}}_{ji}\right| \\
&\triangleq n\delta(\epsilon),
\end{aligned} \tag{J.2}$$

where (a) follows by conditioning reduces entropy; (b) follows by Fano's inequality; (c) follows by Jensen's inequality; and (d) follows by (J.1), where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. Note that, in the following, the convention in [1, Chapter 3] of defining as $\delta(\epsilon)$ any function such that $\delta(\epsilon) \to 0$ as $\epsilon \to 0$ is used.

For rate R_1, then the following series of inequalities hold

$$\begin{aligned}
nR_1 &\geq H(M_1) \overset{(a)}{=} H(M_1, A^n) \\
&= H(A^n) + H(M_1|A^n) \\
&\overset{(b)}{\geq} H(A^n) - H(A^n|X_1^n, X_2^n) + H(M_1|A^n, Y^n, X_2^n) - H(M_1|A^n, Y^n, X_1^n, X_2^n) \\
&= I(A^n; X_1^n, X_2^n) + I(M_1; X_1^n|A^n, Y^n, X_2^n) \\
&= I(A^n; X_1^n, X_2^n) + H(X_1^n|A^n, Y^n, X_2^n) - H(X_1^n|A^n, Y^n, X_2^n, M_1) \\
&= H(X_1^n, X_2^n) - H(X_1^n, X_2^n|A^n) + H(X_1^n, X_2^n, Y^n|A^n) - H(Y^n, X_2^n|A^n) \\
&\quad - H(X_1^n|A^n, Y^n, X_2^n, M_1) \\
&= H(X_1^n, X_2^n) + H(Y^n|A^n, X_1^n, X_2^n) - H(Y^n, X_2^n|A^n) \\
&\quad - H(X_1^n|A^n, Y^n, X_2^n, M_1),
\end{aligned} \tag{J.3}$$

where (a) follows because A^n is a function of M_1 and (b) follows because entropy is non-negative and conditioning decreases entropy. For the first three terms in (J.3) it is also true that

$$
\begin{aligned}
&H(X_1^n, X_2^n) + H(Y^n|A^n, X_1^n, X_2^n) - H(Y^n, X_2^n|A^n) \\
&= H(X_1^n, X_2^n) + H(Y^n|A^n, X_1^n, X_2^n) - H(Y^n|A^n) - H(X_2^n|A^n, Y^n) \\
&\overset{(a)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) + H(Y_i|Y^{i-1}, A^n, X_1^n, X_2^n) - H(Y_i|Y^{i-1}, A^n) - H(X_{2i}|X_2^{i-1}, A^n, Y^n) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) + H(Y_i|A_i, X_{1i}, X_{2i}) - H(Y_i|A_i) - H(X_{2i}|A_i, Y_i) \\
&= \sum_{i=1}^{n} H(X_{1i}, X_{2i}) - I(Y_i; X_{1i}, X_{2i}|A_i) - H(X_{2i}|A_i, Y_i) \\
&= \sum_{i=1}^{n} H(X_{1i}, X_{2i}) - H(X_{1i}, X_{2i}|A_i) + H(X_{1i}, X_{2i}|A_i, Y_i) - H(X_{2i}|A_i, Y_i) \\
&= \sum_{i=1}^{n} I(X_{1i}, X_{2i}; A_i) + H(X_{1i}|A_i, Y_i, X_{2i}),
\end{aligned} \tag{J.4}
$$

where (a) follows by the chain rule for entropy and the fact that X_1^n, X_2^n are i.i.d. and (b) follows since Y_i—(A_i, X_{1i}, X_{2i})—$(Y^{i-1}, A^{n\backslash i}, X_1^{n\backslash i}, X_2^{n\backslash i})$ forms a Markov chain, by the definition of problem, and since conditioning reduces entropy.

Combining (J.3) and (J.4), and defining $U_{1i} = (A^{n\backslash i}, Y^{n\backslash i}, X_2^{n\backslash i}, M_1)$, it holds that

$$
\begin{aligned}
nR_1 &\overset{(a)}{\geq} \sum_{i=1}^{n} I(X_{1i}, X_{2i}; A_i) + H(X_{1i}|A_i, Y_i, X_{2i}) \\
&- H(X_{1i}|X_1^{i-1}, A^n, Y^n, X_2^n, M_1) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} I(X_{1i}; A_i) + H(X_{1i}|A_i, Y_i, X_{2i}) - H(X_{1i}|A^n, Y^n, X_2^n, M_1) \\
&\overset{(c)}{=} \sum_{i=1}^{n} (X_{1i}; A_i) + I(X_{1i}; U_{1i}|A_i, Y_i, X_{2i}),
\end{aligned} \tag{J.5}
$$

where (a) follows by the chain rule for entropy; (b) follows because mutual information is non-negative and due to the fact that conditioning decreases entropy; and (c) follows by the definition of mutual information and definition of U_{1i}.

Next, the rate R_2 is considered.

$$\begin{aligned}
nR_2 &\geq H(M_2) \geq H(M_2|A^n, Y^n, M_1) - H(M_2|A^n, Y^n, M_1, X_2^n) \\
&= I(M_2; X_2^n|A^n, Y^n, M_1) \\
&= H(X_2^n|A^n, Y^n, M_1) - H(X_2^n|A^n, Y^n, M_1, M_2) \\
&\overset{(a)}{\geq} H(X_2^n|A^n, Y^n, M_1) - n\delta(\epsilon) \\
&= \sum_{i=1}^{n} H(X_{2i}|X_2^{i-1}, A^n, Y^n, M_1) - n\delta(\epsilon) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} H(X_{2i}|A_i, Y_i, U_{1i}) - n\delta(\epsilon),
\end{aligned} \tag{J.6}$$

where (a) follows because from (J.2), $H(X_2^n|A^n, Y^n, M_1, M_2) \leq H(X_2^n|\hat{X}_2^n) \leq n\delta(\epsilon)$, given that $\hat{X}_2^n$ is a function of M_1, M_2 and $Y^n$and (b) follows using the definition of U_{1i} and due to the fact that conditioning decreases entropy. For the sum-rate R_1+R_2, also the following series of inequalities hold

$$\begin{aligned}
n(R_1 + R_2) &\geq H(M_1, M_2) \overset{(a)}{=} H(M_1, M_2, A^n) \\
&= H(A^n) + H(M_1, M_2|A^n) \\
&\geq H(A^n) - H(A^n|X_1^n, X_2^n) + H(M_1, M_2|A^n, Y^n) \\
&- H(M_1, M_2|A^n, Y^n, X_1^n, X_2^n) \\
&= I(A^n; X_1^n, X_2^n) + I(M_1, M_2; X_1^n, X_2^n|A^n, Y^n) \\
&= I(A^n; X_1^n, X_2^n) + H(X_1^n, X_2^n|A^n, Y^n) - H(X_1^n, X_2^n|A^n, Y^n, M_1, M_2) \\
&= H(X_1^n, X_2^n) - H(X_1^n, X_2^n|A^n) + H(X_1^n, X_2^n, Y^n|A^n) - H(Y^n|A^n) \\
&- H(X_2^n|A^n, Y^n, M_1, M_2) - H(X_1^n|A^n, Y^n, X_2^n, M_1, M_2) \\
&\overset{(b)}{\geq} H(X_1^n, X_2^n) + H(Y^n|A^n, X_1^n, X_2^n) - H(Y^n|A^n) \\
&- H(X_1^n|A^n, Y^n, X_2^n, M_1, M_2) - n\delta(\epsilon),
\end{aligned} \tag{J.7}$$

where (a) follows because A^n is a function of M_1; and (b) follows as in (a) of (J.6). For the first three terms in (J.7), it is true that

$$
\begin{aligned}
& H(X_1^n, X_2^n) + H(Y^n|A^n, X_1^n, X_2^n) - H(Y^n|A^n) \\
& \stackrel{(a)}{=} \sum_{i=1}^n H(X_{1i}, X_{2i}) + H(Y_i|Y^{i-1}, A^n, X_1^n, X_2^n) - H(Y_i|Y^{i-1}, A^n) \\
& \stackrel{(b)}{\geq} \sum_{i=1}^n H(X_{1i}, X_{2i}) + H(Y_i|A_i, X_{1i}, X_{2i}) - H(Y_i|A_i) \\
& = \sum_{i=1}^n H(X_{1i}, X_{2i}) - I(Y_i; X_{1i}, X_{2i}|A_i) \\
& = \sum_{i=1}^n H(X_{1i}, X_{2i}) - H(X_{1i}, X_{2i}|A_i) + H(X_{1i}, X_{2i}|A_i, Y_i) \\
& = \sum_{i=1}^n I(X_{1i}, X_{2i}; A_i) + H(X_{2i}|A_i, Y_i) + H(X_{1i}|A_i, Y_i, X_{2i}), \qquad \text{(J.8)}
\end{aligned}
$$

where (a) follows from the chain rule for entropy and by the chain rule for entropy and the fact that (X_1^n, X_2^n) are i.i.d.; and (b) follows since Y_i—$(A_i, X_{\{1,2\}i})$—$(Y^{i-1}, A^{n\backslash i}, X_1^{n\backslash i}, X_2^{n\backslash i})$ forms a Markov chain, by the definition of problem, and since conditioning reduces entropy. Combining (J.7) and (J.8), and using the definition of U_{1i}, it holds that

$$
\begin{aligned}
n(R_1 + R_2) & \stackrel{(a)}{\geq} \sum_{i=1}^n I(X_{1i}, X_{2i}; A_i) + H(X_{2i}|A_i, Y_i) + H(X_{1i}|A_i, Y_i, X_{2i}) \\
& - H(X_{1i}|X_1^{i-1}, A^n, Y^n, X_2^n, M_1, M_2) - n\delta(\epsilon) \\
& \stackrel{(b)}{\geq} \sum_{i=1}^n I(X_{1i}; A_i) + H(X_{2i}|A_i, Y_i) + H(X_{1i}|A_i, Y_i, X_{2i}) \\
& - H(X_{1i}|A^n, Y^n, X_2^n, M_1) - n\delta(\epsilon) \\
& \stackrel{(c)}{\geq} \sum_{i=1}^n (X_{1i}; A_i) + H(X_{2i}|A_i, Y_i) + I(X_{1i}; U_{1i}|A_i, Y_i, X_{2i}) - n\delta(\epsilon), \qquad \text{(J.9)}
\end{aligned}
$$

where (a) follows by the chain rule for entropy; (b) follows because mutual information is non-negative and due to the fact that conditioning decreases entropy; and (c) follows by the definition of mutual information and definition of U_{1i} and the fact that conditioning decreases entropy.

Moreover, $(X_{2i}, Y_i) - (X_{1i}, A_i) - U_{1i}$ forms a Markov chain. This can be seen by using the principle of d-separation [70, Section A.9] from Figure J.1, which represents the joint distribution of all the variables at hand.

Let Q be a uniform random variable over the interval $[1, n]$ and independent of $(X_1^n, X_2^n, A^n, U_1^n, Y^n, \hat{X}_1^n)$ and define $U_1 \triangleq (Q, U_{1Q})$, $X_1 \triangleq X_{1Q}$, $X_2 \triangleq X_{2Q}$, $Y \triangleq Y_Q$, $A \triangleq A_Q$, and $\hat{X}_1 \triangleq \hat{X}_{1Q}$. Note that $\hat{X}_1$ is a function of U_1, X_2 and Y. Moreover, from

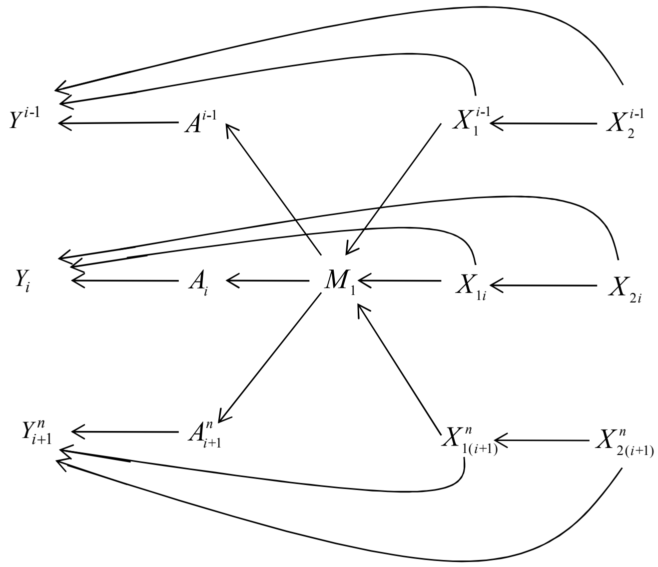

Figure J.1 Bayesian network representing the joint pmf of variables $(M_1, X_1^n, X_2^n, A^n, Y^n)$ for the model in Figure 4.2.

(4.8) and (7.7), the following holds

$$\Gamma + \epsilon \geq \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\Lambda(A_i)\right] = \mathrm{E}[\Lambda(A)]$$

$$\text{and } D_1 + \epsilon \geq \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d_1(X_{1i}, X_{2i}, Y_i.\hat{X}_{1i})\right] = \mathrm{E}[d_1(X_1, X_2, Y, \hat{X}_1)]. \tag{J.10}$$

Finally, since (J.5), (J.6) and (J.9) are convex with respect to $p(a, u_1|x_1, q)$ for fixed $p(q)$, $p(x_1, x_2)$, and $p(y|a, x_1, x_2)$, inequalities (4.20) hold, which completes the proof of (4.20a)-(4.22b). The cardinality bounds are proved by using the Fenchel–Eggleston–Caratheodory theorem in the standard way.

APPENDIX K

GREEDY ACTIONS ARE OPTIMAL WITH SUM SIDE INFORMATION

Here, the equality below is proved.

$$R^{\oplus}_{sum,\ greedy}(\Gamma) = R^{\oplus}_{sum}(\Gamma). \tag{K.1}$$

which shows that no gain is accrued by choosing the actions based only on message M_1 with the sum side information. Fix the pmf $p(a|x_1)$ that achieves the minimum in the sum-rate obtained from (4.20c), namely

$$R^{\oplus}_{sum}(\Gamma) \quad = \quad \min I(X_1; A) + H(X_1, X_2|A, Y),$$

where the mutual information is calculated with respect to the distribution

$$p(x_1, x_2, y, a) = p(x_1, x_2)p(a|x_1)p(y|a, x_1, x_2), \tag{K.2}$$

and the minimum is taken over all distributions $p(a|x_1)$ such that $\mathrm{E}[\Lambda(A)] = \mathrm{E}[A] \leq \Gamma$. Note that for such a pmf $p(a|x_1)$, it is true that $\mathrm{E}[A] = p(a) = \Gamma$, as it can be easily seen. Next, following series of equalities hold

$$\begin{aligned}
&R^{\oplus}_{sum,\ greedy}(\Gamma) - R^{\oplus}_{sum}(\Gamma) \\
&\overset{(a)}{=} \Gamma H(X_1, X_2|X_1 \oplus X_2) + (1-\Gamma)H(X_1, X_2) \\
&- H(X_1, X_2|A, X_1 \oplus X_2) - I(X_1; A) \\
&\overset{(b)}{=} \Gamma H(X_1|X_1 \oplus X_2) + (1-\Gamma)(1 + H(p)) - \Gamma H(X_1, X_2|A = 1, X_1 \oplus X_2) \\
&- (1-\Gamma)H(X_1, X_2|A = 0) - I(X_1; A) \\
&\overset{(c)}{=} \Gamma H(X_1) + (1-\Gamma)(1 + H(p)) - \Gamma H(X_1|A = 1) - (1-\Gamma)H(X_1|A = 0) \\
&- (1-\Gamma)H(X_2|X_1, A = 0) - I(X_1; A) \\
&\overset{(d)}{=} \Gamma + (1-\Gamma)(1 + H(p)) - H(X_1|A) - (1-\Gamma)H(X_2|X_1) - I(X_1; A) \\
&= \Gamma + (1-\Gamma)(1 + H(p)) - H(X_1|A) - (1-\Gamma)H(p) - 1 + H(X_1|A) = 0,
\end{aligned}$$

where (a) follows by the definition (4.27); (b) follows using the chain rule for entropy and from the definition of conditional entropy; (c) follows by the crypto-lemma [37, Lemma 2]; (d) follows from the fact that $X_2 - X_1 - A$ forms a Markov chain.

APPENDIX L

PROOF OF THE CONVERSE FOR PROPOSITION 4.5

In this section, the proof of converse for Proposition 4.5 is provided. For any $(n, R_{12}, R_{23}, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)$ code, the following inequalities hold:

$$
\begin{aligned}
nR_{12} &\geq H(M_{12}) \geq H(M_{12}|X_2^n) \stackrel{(a)}{=} H(M_{12}, M_{23}|X_2^n) \\
&\stackrel{(b)}{=} I(X_1^n; M_{12}, M_{23}|X_2^n) \\
&= \sum_{i=1}^{n} H(X_{1i}|X_1^{i-1}, X_2^n) - H(X_{1i}|X_1^{i-1}, X_2^n, M_{12}, M_{23}) \\
&\stackrel{(c)}{=} \sum_{i=1}^{n} H(X_{1i}|X_{2i}) - H(X_{1i}|X_1^{i-1}, X_2^n, A^n, M_{12}, M_{23}) \\
&\stackrel{(d)}{=} \sum_{i=1}^{n} H(X_{1i}|X_{2i}) - H(X_{1i}|X_1^{i-1}, X_2^n, Y^{i-1}, M_{12}, M_{23}, A^n, \hat{X}_1^n) \\
&\stackrel{(e)}{\geq} \sum_{i=1}^{n} H(X_{1i}|X_{2i}) - H(X_{1i}|X_{2i}, A_i, U_i, \hat{X}_{1i}) \\
&= \sum_{i=1}^{n} I(X_{1i}; A_i, U_i, \hat{X}_{1i}|X_{2i}),
\end{aligned} \tag{L.1}
$$

where (a) follows because M_{23} is a function of (M_{12}, X_2^n); (b) follows by definition of mutual information and since M_{12} and M_{23} are functions of X_1^n and X_2^n; (c) follows because X_1^n and X_2^n are i.i.d and since A^n is a function of M_{23}; (d) follows because $Y^{i-1} - (X_1^{i-1}, X_2^n, A^n, M_{12}, M_{23}) - X_{1i}$ forms a Markov chain and since $\hat{X}_1^n$ is a function of M_{12} and X_2^n; and (e) follows by defining $U_i = (X_1^{i-1}, X_2^{i-1}, Y^{i-1}, A^{n\backslash i}, M_{23})$ and since conditioning decreases entropy.

It is also true that

$$\begin{aligned}
nR_{23} \geq H(M_{23}) &\overset{(a)}{=} I(X_1^n, X_2^n; M_{23}) \\
&\overset{(b)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) - H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, M_{23}) \\
&\overset{(c)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) - H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, A^n, M_{23}) \\
&\overset{(d)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) - H(X_{1i}, X_{2i}|X_1^{i-1}, X_2^{i-1}, Y^{i-1}, A^n, M_{23}) \\
&\overset{(e)}{=} \sum_{i=1}^{n} H(X_{1i}, X_{2i}) - H(X_{1i}, X_{2i}|A_i, U_i) \\
&= \sum_{i=1}^{n} I(X_{1i}, X_{2i}; A_i, U_i),
\end{aligned} \tag{L.2}$$

where (a) follows because M_{23} is a function of X_1^n and X_2^n; (b) follows by the definition of mutual information and the chain rule for entropy and since X_1^n and X_2^n are i.i.d; (c) follows because A^n is a function of M_{23}; (d) follows because $Y^{i-1} - (X_1^{i-1}, X_2^{i-1}, A^n, M_{23}) - (X_{1i}, X_{2i})$ forms a Markov chain; and (e) follows by the definition of U_i.

Let Q be a uniform random variable over $[1, n]$ and independent of $(X_1^n, X_2^n, Y^n, A^n, U^n, \hat{X}_1^n)$ and define $U \triangleq (Q, U_Q)$, $X_1 \triangleq X_{1Q}$, $X_2 \triangleq X_{2Q}$, $Y \triangleq Y_Q$, $A \triangleq A_Q$, $\hat{X}_1 \triangleq \hat{X}_{1Q}$, and $\hat{X}_2 \triangleq \hat{X}_{2Q}$. Note that $\hat{X}_2$ is a function of U and Y. Moreover, from (8.5) and (8.6), the following holds

$$\Gamma + \epsilon \geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\Lambda(A_i)\right] = \mathrm{E}[\Lambda(A)] \tag{L.3}$$

$$\text{and } D_j + \epsilon \geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[d_j(X_{1i}, X_{2i}, Y_i, \hat{X}_{ji})\right] = \mathrm{E}[d_j(X_1, X_2, Y, \hat{X}_j)] \text{ for } j = 1, 2. \tag{L.4}$$

Finally, since (L.1) and (L.2) are convex with respect to $p(a, u, \hat{x}_1|x_1, x_2)$ for fixed $p(x_1, x_2)$ and $p(y|a, x_1, x_2)$, from (L.1) and (L.2) it can be concluded that inequalities (4.35) hold. The cardinality bounds are proved by using the Fenchel-Eggleston-Caratheodory theorem in the standard way.

APPENDIX M

CONVERSE PROOF FOR PROPOSITION 5.1 AND 5.4

Here, the converse part of Proposition 5.4 is proved. Since the setting of Proposition 5.1 is more restrictive, as it does not allow for adaptive actions, the converse proof for Proposition 5.1 follows immediately. For any $(n, R_1, R_2, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)$ code, it is true that

$$\begin{aligned}
nR_1 &\geq H(M_1) \\
&\geq H(M_1|Y^n) \\
&\overset{(a)}{=} I(M_1; X^n, Z^n|Y^n) \\
&= H(X^n, Z^n|Y^n) - H(X^n, Z^n|M_1, Y^n) \\
&= H(X^n|Y^n) + H(Z^n|X^n, Y^n) - H(Z^n|Y^n, M_1) - H(X^n|Z^n, Y^n, M_1) \\
&\overset{(a,b)}{=} H(X^n|Y^n) + H(Z^n|X^n, Y^n, M_1, M_2) - H(Z^n|Y^n, M_1, M_2) \\
&- H(X^n|Z^n, Y^n, M_1, M_2) \\
&\overset{(c)}{=} H(X^n|Y^n) - H(X^n|Z^n, Y^n, M_1, M_2, A^n, \hat{X}_1^n) \\
&+ \sum_{i=1}^{n} H(Z_i|Z^{i-1}, X^n, Y^n, M_1, M_2) - H(Z_i|Z^{i-1}, Y^n, M_1, M_2) \\
&\overset{(c,d)}{\geq} \sum_{i=1}^{n} (H(X_i|Y_i) - H(X_i|X^{i-1}, Y^i, M_2, A^i, Z^n, \hat{X}_{1i})) \\
&+ \sum_{i=1}^{n} H(Z_i|Z^{i-1}, X^n, Y^n, M_1, M_2, A_i) - H(Z_i|Z^{i-1}, Y^n, M_1, M_2, A_i) \\
&\overset{(e)}{=} \sum_{i=1}^{n} I(X_i; \hat{X}_{1i}, A_i, U_i|Y_i) + H(Z_i|Y_i, A_i) - H(Z_i|Y_i, A_i) \\
&= \sum_{i=1}^{n} I(X_i; \hat{X}_{1i}, A_i, U_i|Y_i), \qquad \text{(M.1)}
\end{aligned}$$

where (a) follows since M_1 is a function of (X^n, Y^n); (b) follows since M_2 is a function of (M_1, Y^n); (c) follows since A_i is a function of (M_2, Z^{i-1}) and since $\hat{X}_1^n$ is a function of (M_1, Y^n); (d) follows since conditioning decreases entropy and since X^n and Y^n are i.i.d.; and (e) follows by defining $U_i = (M_2, X^{i-1}, Y^{i-1}, A^{i-1}, Z^{n\backslash i})$ and since $(Z^{i-1}, X^n, Y^{n\backslash i}, M_1, M_2)$— (A_i, Y_i)—Z_i form a Markov chain by construction. It is also true that

$$\begin{aligned}
nR_2 &\geq H(M_2) \\
&= I(M_2; X^n, Y^n, Z^n)
\end{aligned}$$

$$
\begin{aligned}
&= H(X^n, Y^n, Z^n) - H(X^n, Y^n, Z^n|M_2) \\
&= H(X^n, Y^n) + H(Z^n|X^n, Y^n) - H(Z^n|M_2) - H(X^n, Y^n|M_2, Z^n) \\
&= \sum_{i=1}^{n} H(X_i, Y_i) + H(Z_i|Z^{i-1}, X^n, Y^n) - H(Z_i|Z^{i-1}, M_2) \\
&\quad -H(X_i, Y_i|X^{i-1}, Y^{i-1}, M_2, Z^n) \\
&\overset{(a)}{=} \sum_{i=1}^{n} H(X_i, Y_i) + H(Z_i|Z^{i-1}, X^n, Y^n, M_2, A_i) - H(Z_i|Z^{i-1}, M_2, A_i) \\
&\quad -H(X_i, Y_i|X^{i-1}, Y^{i-1}, M_2, Z^n, A^i) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} H(X_i, Y_i) + H(Z_i|X_i, Y_i, A_i) - H(Z_i|A_i) - H(X_i, Y_i|U_i, A_i, Z_i), \quad \text{(M.2)}
\end{aligned}
$$

where (a) follows because M_2 is a function of (M_1, Y^n) and thus of (X^n, Y^n) and because A^i is a function of (M_2, Z^{i-1}) and (b) follows since conditioning decreases entropy, since the Markov chain relationship Z_i—(X_i, Y_i, A_i)— $(X^{n\backslash i}, Y^{n\backslash i}, M_2)$ holds and by using the definition of U_i.

Defining Q to be a random variable uniformly distributed over $[1, n]$ and independent of all the other random variables and with $X \triangleq X_Q$, $Y \triangleq Y_Q$, $Z \triangleq Z_Q$, $A \triangleq A_Q$, $\hat{X}_1 \triangleq \hat{X}_{1Q}$, $\hat{X}_2 \triangleq \hat{X}_{2Q}$ and $U \triangleq (U_Q, Q)$, from (M.1), the following holds

$$
nR_1 \geq I(X; \hat{X}_1, A, U|Y, Q) \overset{(a)}{\geq} H(X|Y) - H(X|\hat{X}_1, A, U, Y) = I(X; \hat{X}_1, A, U|Y),
$$

where in (a) follows due to the fact that (X^n, Y^n) are i.i.d and conditioning reduces entropy. Moreover, from (M.2) it follows that

$$
\begin{aligned}
nR_2 &\geq H(X, Y|Q) + H(Z|X, Y, A, Q) - H(Z|A, Q) - H(X, Y|U, A, Z, Q) \\
&\overset{(a)}{\geq} H(XY) + H(Z|X, Y, A) - H(Z|A) - H(X, Y|U, A, Z) \\
&= I(XY; U, A, Z) - I(Z; X, Y|A) \\
&= I(XY; A) + I(X, Y; U|A, Z),
\end{aligned}
$$

where (a) follows since (X^n, Y^n) are i.i.d, since conditioning decreases entropy, by the definition of U and by the problem definition. It is noted that the defined random variables factorize as (8.8) since the Markov chain relationship X—(A, Y)—Z holds by the problem definition and that $\hat{X}_2$ is a function f(U, Z) of U and Z by the definition of U. Moreover, from the cost and distortion constraints (8.5)-(8.6), it is true that

$$
D_j + \epsilon \geq \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[d_j(X_i, \hat{X}_{ji})] = \mathrm{E}[d_j(X, \hat{X}_j)], \text{ for } j = 1, 2, \quad \text{(M.3a)}
$$

$$
\text{and } \Gamma + \epsilon \geq \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left[\Lambda(A_i)\right] = \mathrm{E}\left[\Lambda(A)\right]. \quad \text{(M.3b)}
$$

To bound the cardinality of auxiliary random variable U, $p(z|y,a)$ is fixed and the joint pmf $p(x,y,z,a,u,\hat{x}_1)$ is factorized as

$$p(x,y,z,a,u,\hat{x}_1) \quad = \quad p(u)p(\hat{x}_1,a,x,y|u)p(z|y,a).$$

Therefore, for fixed $p(z|y,a)$, the quantities (8.39a)-(8.9b) can be expressed in terms of integrals given by $\int g_j(p(\hat{x}_1,a,x,y|u))dF(u)$, for $j=1,...,|\mathcal{X}||\mathcal{Y}||\mathcal{A}|+3$, of functions $g_j(\cdot)$ that are continuous on the space of probabilities over alphabet $\mathcal{X}\times\mathcal{Y}\times\mathcal{A}\times\hat{\mathcal{X}}_1$. Specifically, g_j for $j=1,...,|\mathcal{X}||\mathcal{Y}||\mathcal{A}|-1$, are given by the pmf $p(a,x,y)$ for all values of $x\in\mathcal{X}$, $y\in\mathcal{Y}$ and $a\in\mathcal{A}$, (except one), $g_{|\mathcal{X}||\mathcal{Y}||\mathcal{A}|}=H(X|A,Y,\hat{X}_1,U=u)$, $g_{|\mathcal{X}||\mathcal{Y}||\mathcal{A}|+1}=H(X,Y|A,Z,U=u)$, and $g_{|\mathcal{X}||\mathcal{Y}||\mathcal{A}|+1+j}=\mathrm{E}[d_j(X,\hat{X}_j)|U=u]$, for $j=1,2$. The proof in concluded by invoking the Fenchel–Eggleston–Caratheodory theorem [1, Appendix C].

APPENDIX N

PROOF OF PROPOSITION 5.3

Here, the converse parts of Proposition 5.3 and Proposition 5.5 are proved. Let's start by proving Proposition 5.3. The proof of Proposition 5.5 will follow by setting $Z = \emptyset$, and noting that in the proof below the action A_i can be made to be a function of Y^{i-1}, in addition to being a function of M_b, without modifying any steps of the proof. By the CR requirements (5.26), first it is observed that for any $(n, R_1, R_2, R_b, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)$ code, the Fano inequalities below hold

$$\begin{aligned} H(\psi_1(X^n)|\mathrm{h}_1(M_1, M_b, Y^n)) &\leq n\delta(\epsilon), && \text{(N.1a)} \\ \text{and } H(\psi_2(X^n)|\mathrm{h}_2(M_2, M_b, Z^n)) &\leq n\delta(\epsilon), && \text{(N.1b)} \end{aligned}$$

where $\delta(\epsilon)$ denotes any function such that $\delta(\epsilon) \to 0$ if $\epsilon \to 0$. Next, it is also true that

$$\begin{aligned} nR_b &\geq H(M_b) \\ &\overset{(a)}{=} I(M_b; X^n, Y^n) \\ &= H(X^n, Y^n) - H(X^n, Y^n|M_b) \\ &\overset{(a)}{=} H(X^n) + H(Y^n|X^n, M_b) - H(X^n, Y^n|M_b) \\ &\overset{(b)}{=} \sum_{i=1}^{n} H(X_i) + H(Y_i|Y^{i-1}, X^n, M_b, A_i) - H(X_i, Y_i|X^{i-1}, Y^{i-1}, M_b, A_i) \\ &= \sum_{i=1}^{n} H(X_i) + H(Y_i|Y^{i-1}, X^n, M_b, A_i) - H(X_i|X^{i-1}, Y^{i-1}, M_b, A_i) \\ &- H(Y_i|X^i, Y^{i-1}, M_b, A_i) \\ &\overset{(c)}{=} \sum_{i=1}^{n} H(X_i) + H(Y_i|X_i, A_i) - H(X_i|X^{i-1}, Y^{i-1}, M_b, A_i) - H(Y_i|X_i, A_i) \\ &\overset{(d)}{\geq} \sum_{i=1}^{n} H(X_i) - H(X_i|A_i), \end{aligned} \tag{N.2}$$

where (a) follows since M_b is a function of X^n; (b) follows since A_i is a function of M_b and since X^n is i.i.d.; (c) follows since $(Y^{i-1}, X^{n\backslash i}, M_b)$—$(A_i, X_i)$—$Y_i$ forms a Markov chain by problem definition; and (d) follows conditioning reduces entropy. In the following, for simplicity of notation, $\mathrm{h}_1, \mathrm{h}_2, \psi_1, \psi_2$ is used for the values of corresponding functions in Section 5.3.4. Next, the following holds

$$
\begin{aligned}
&n(R_1 + R_b)\\
&\geq H(M_1, M_b)\\
&\overset{(a)}{=} I(M_1, M_b; X^n, Y^n, Z^n)\\
&= H(X^n, Y^n, Z^n) - H(X^n, Y^n, Z^n|M_1, M_b)\\
&= H(X^n) + H(Y^n, Z^n|X^n) - H(Y^n, Z^n|M_1, M_b) - H(X^n|Y^n, Z^n, M_1, M_b)\\
&\overset{(b)}{=} H(X^n) + H(Y^n, Z^n|X^n, M_b) - H(Y^n|M_1, M_b)\\
&- H(Z^n|M_1, M_b, Y^n, A^n) - H(X^n|Y^n, Z^n, M_1, M_b, M_2, A^n)\\
&\overset{(b,c)}{=} \sum_{i=1}^{n} H(X_i) + H(Y_i, Z_i|X_i, A_i) - H(Y_i|Y^{i-1}, M_1, M_b, A_i)\\
&- H(Z_i|Z^{i-1}, M_1, M_b, Y^n, A^n) - H(X_i|X^{i-1}, Y^n, Z^n, M_1, M_b, A^n, M_2, \mathrm{h}_1, \mathrm{h}_2)\\
&\overset{(d)}{\geq} \sum_{i=1}^{n} H(X_i) + H(Y_i|X_i, A_i) + H(Z_i|Y_i, A_i) - H(Y_i|A_i) - H(Z_i|Y_i, A_i)\\
&- H(X_i|Y_i, A_i, \mathrm{h}_1, \mathrm{h}_2)\\
&\sum_{i=1}^{n} I(X_i; Y_i, A_i, \mathrm{h}_1, \mathrm{h}_2) - I(Y_i; X_i|A_i)\\
&\sum_{i=1}^{n} I(X_i; Y_i, A_i, \mathrm{h}_1, \mathrm{h}_2, \psi_1, \psi_2) - I(X_i; \psi_1, \psi_2|Y_i, A_i, \mathrm{h}_1, \mathrm{h}_2) - I(Y_i; X_i|A_i)\\
&\overset{(e)}{\geq} \sum_{i=1}^{n} I(X_i; Y_i, A_i, \psi_1, \psi_2) - H(\psi_1, \psi_2|Y_i, A_i, \mathrm{h}_1, \mathrm{h}_2)\\
&+ H(\psi_1, \psi_2|Y_i, A_i, \mathrm{h}_1, \mathrm{h}_2, X_i) - I(Y_i; X_i|A_i)\\
&\overset{(f)}{\geq} \sum_{i=1}^{n} I(X_i; Y_i, A_i, \psi_1, \psi_2) - I(Y_i; X_i|A_i) + n\delta(\epsilon)\\
&= \sum_{i=1}^{n} I(X_i; A_i) + I(X_i; \psi_1, \psi_2|Y_i, A_i) + n\delta(\epsilon),
\end{aligned} \tag{N.3}
$$

where (a) follows because (M_1, M_b) is a function of X^n; (b) follows because M_b is a function of X^n, A^n is a function of M_b and M_2 is a function of (M_1, M_b, Y^n); (c) follows since $H(Y^n, Z^n|X^n, M_b) = \sum_{i=1}^{n} H(Y_i, Z_i| \ Y^{i-1}, Z^{i-1}, X^n, M_b, A_i) = \sum_{i=1}^{n} H(Y_i, Z_i|X_i, A_i)$ and since h_1 and h_2 are functions of (M_1, M_b, Y^n) and (M_2, M_b, Z^n), respectively and because (Y_i, Z_i)—(X_i, A_i)— $(X^{n\backslash i}, Y^{i-1}, Z^{i-1}, M_b)$ forms a Markov chain; (d) follows since conditioning reduces entropy, since side information VM follows $p(y^n, z^n|a^n, x^n) = \prod_{i=1}^{n} p_{Y|A,X}(y_i|a_i, x_i)\, p_{Z|A,Y}(z_i|a_i, y_i)$ from (7.24) and because Z_i—(Y_i, A_i)— $(Y^{n\backslash i}, Z^{i-1}, M_1, M_b)$ forms a Markov chain; (e) follows by the chain rule for mutual information and the fact that mutual information is non-negative; and (f) follows by the Fano inequality (R.2) and because entropy is non-negative. It holds that

$$
\begin{aligned}
n(R_2+R_b) &\geq H(M_2,M_b)\\
&\overset{(a)}{=} I(M_2,M_b;X^n,Y^n,Z^n)\\
&= H(X^n,Y^n,Z^n) - H(X^n,Y^n,Z^n|M_2,M_b)\\
&\overset{(a)}{=} H(X^n) + H(Y^n,Z^n|X^n,M_b) - H(Z^n|M_2,M_b)\\
&- H(X^n,Y^n|Z^n,M_2,M_b)\\
&\overset{(b)}{=} \sum_{i=1}^{n} H(X_i) + H(Y_i,Z_i|Y^{i-1},Z^{i-1},X^n,M_b,A_i)\\
&- H(Z_i|Z^{i-1},M_2,M_b,A_i) - H(X_i,Y_i|X^{i-1},Y^{i-1},M_2,M_b,Z^n,A_i)\\
&= \sum_{i=1}^{n} H(X_i,Y_i) - H(Y_i|X_i) + H(Y_i,Z_i|Y^{i-1},Z^{i-1},X^n,M_b,A_i)\\
&- H(Z_i|Z^{i-1},M_2,M_b,A_i) - H(X_i,Y_i|X^{i-1},Y^{i-1},M_2,M_b,Z^n,A_i)\\
&\overset{(c)}{=} \sum_{i=1}^{n} H(X_i,Y_i) - H(Y_i|X_i) + H(Y_i|X_i,A_i) + H(Z_i|A_i,Y_i,X_i)\\
&- H(Z_i|Z^{i-1},M_2,M_b,A_i) - H(X_i,Y_i|X^{i-1},Y^{i-1},M_2,M_b,Z^n,A_i)\\
&\overset{(d)}{=} \sum_{i=1}^{n} H(X_i,Y_i) - I(Y_i;A_i|X_i) + H(Z_i|A_i,Y_i,X_i)\\
&- H(Z_i|Z^{i-1},M_2,M_b,A_i) - H(X_i,Y_i|X^{i-1},Y^{i-1},M_2,M_b,\mathrm{h}_2,Z^n,A_i)\\
&\overset{(e)}{\geq} \sum_{i=1}^{n} H(X_i,Y_i) + I(X_i;A_i) - I(Y_i,X_i;A_i) + H(Z_i|A_i,Y_i,X_i)\\
&- H(Z_i|A_i) - H(X_i,Y_i|\mathrm{h}_2,A_i,Z_i)\\
&= \sum_{i=1}^{n} I(X_i,Y_i;\mathrm{h}_2,A_i,Z_i,\psi_{2i}) - I(X_i,Y_i;\psi_{2i}|\mathrm{h}_2,A_i,Z_i) + I(X_i;A_i)\\
&- I(Y_i,X_i;A_i) - I(X_i,Y_i;Z_i|A_i)\\
&\geq \sum_{i=1}^{n} I(X_i,Y_i;A_i,Z_i,\psi_{2i}) - H(\psi_{2i}|\mathrm{h}_2,A_i,Z_i) + H(\psi_{2i}|\mathrm{h}_2,A_i,X_i,Y_i,Z_i)\\
&+ I(X_i;A_i) - I(X_i,Y_i;Z_i,A_i)\\
&\overset{(f)}{\geq} \sum_{i=1}^{n} I(X_i;A_i) + I(X_i,Y_i;\psi_{2i}|A_i,Z_i) + n\delta(\epsilon),
\end{aligned}
\tag{N.4}
$$

where (a) follows since M_b is a function of X^n and because M_2 is a function of (M_1,M_b,Y^n) and thus of (X^n,Y^n); (b) follows since A_i is a function of M_b and since X^n is i.i.d.; (c) follows since (Y_i,Z_i)—(X_i,A_i)— $(X^{n\backslash i},Y^{i-1},Z^{i-1},M_b)$ forms a Markov chain and since $p(y^n,z^n|a^n,x^n)=\prod_{i=1}^{n}p_{Y|A,X}(y_i|a_i,x_i)p_{Z|A,Y}(z_i|a_i,y_i)$; (d) follows since h_2 is a function of (M_2,M_b,Z^n); (e) follows since conditioning reduces entropy; and (f) follows since entropy is non-negative and using the Fanos inequality. Moreover, with the definition $M=(M_1,M_2,M_b)$, the following chain of inequalities hold

$$
\begin{aligned}
&n(R_1+R_2+R_b) \geq H(M)\\
&\overset{(a)}{=} I(M;X^n,Y^n,Z^n)\\
&= H(X^n,Y^n,Z^n) - H(X^n,Y^n,Z^n|M)\\
&\overset{(a)}{=} H(X^n) + H(Y^n,Z^n|X^n,M_b) - H(X^n,Y^n,Z^n|M)\\
&= I(X^n;A^n) + H(Y^n,Z^n|X^n,M_b) - H(Y^n,Z^n|M)\\
&- H(X^n|Y^n,Z^n,M) + H(X^n|A^n)\\
&= I(X^n;A^n) + H(Y^n,Z^n|X^n,M_b) - H(Y^n,Z^n|M) + I(X^n;Y^n,Z^n,M|A^n)\\
&= I(X^n;A^n) + I(M;X^n|Y^n,A^n,Z^n) + H(Y^n,Z^n|X^n,M_b)\\
&- H(Y^n,Z^n|M) + I(X^n;Y^n,Z^n|A^n)\\
&\overset{(b)}{=} H(X^n) - H(X^n|A^n) + H(X^n|Y^n,A^n,Z^n) - H(X^n|Y^n,A^n,Z^n,M)\\
&- H(Y^n,Z^n|M) + H(Y^n,Z^n|A^n)\\
&= H(X^n) - H(X^n|A^n) + H(X^n,Y^n,Z^n|A^n) - H(X^n|Y^n,A^n,Z^n,M)\\
&- H(Y^n,Z^n|M)\\
&= H(X^n) + H(Y^n,Z^n|A^n,X^n) - H(X^n|Y^n,A^n,Z^n,M) - H(Y^n,Z^n|M)\\
&\overset{(c)}{=} \sum_{i=1}^{n} H(X_i) + H(Y_i|A_i,X_i) + H(Z_i|A_i,Y_i) - H(X_i|X^{i-1},Y^n,A^n,Z^n,M)\\
&- H(Z_i|Z^{i-1},M,A_i) - H(Y_i|Y^{i-1},Z^n,M,A_i)\\
&\overset{(d)}{=} \sum_{i=1}^{n} H(X_i)+H(Y_i|A_i,X_i)+H(Z_i|A_i,Y_i)-H(X_i|X^{i-1},Y^n,A^n,Z^n,M,\mathrm{h}_1,\mathrm{h}_2)\\
&- H(Z_i|Z^{i-1},M,A_i) - H(Y_i|Y^{i-1},Z^n,M,A_i,\mathrm{h}_2)\\
&\geq \sum_{i=1}^{n} H(X_i) + H(Y_i|A_i,X_i) + H(Z_i|A_i,Y_i) - H(X_i|Y_i,A_i,\mathrm{h}_1,\mathrm{h}_2)\\
&- H(Z_i|A_i) - H(Y_i|Z_i,A_i,\mathrm{h}_2)\\
&\overset{(e)}{\geq} I(X_i;A_i,Y_i,\psi_1,\psi_2) + H(Y_i|A_i,X_i) + H(Z_i|A_i,Y_i)\\
&- H(Z_i|A_i) - H(Y_i|Z_i,A_i,\psi_2) - n\delta(\epsilon),
\end{aligned}
\tag{N.5}
$$

where (a) follows since (M_1,M_b) is a function of X^n and M_2 is a function of (M_1,M_b,Y^n); (b) follows since $H(Y^n,Z^n|X^n,M_b) = \sum_{i=1}^{n} H(Y_i,Z_i|\ Y^{i-1},Z^{i-1},X^n, M_b,A_i) = \sum_{i=1}^{n} H(Y_i,Z_i|\ X_i,A_i) = H(Y^n,Z^n|X^n,A^n)$; (c) follows since A_i is a function of M_b; (d) follows since $\mathrm{h}_1,\mathrm{h}_2$ are functions of (M,Y^n) and (M,Z^n), respectively; and (e) follows since entropy is non-negative and by Fano's inequality. Next, from (N.5) it is true that

$$
\begin{aligned}
&n(R_1+R_2+R_b) \geq I(X_i;A_i,Y_i,\psi_1,\psi_2)+H(Y_i|A_i,X_i)+H(Z_i|A_i,Y_i)-H(Z_i|A_i)\\
&-H(Y_i,Z_i|A_i,\psi_2)+H(Z_i|A_i,\psi_2)-n\delta(\epsilon)\\
&=I(X_i;A_i,Y_i,\psi_1,\psi_2)+H(Y_i|A_i,X_i)-H(Z_i|A_i)-H(Y_i|A_i,\psi_2)\\
&+H(Z_i|A_i,\psi_2)-n\delta(\epsilon)\\
&=I(X_i;A_i,Y_i,\psi_1,\psi_2)-I(X_i;Y_i|A_i,\psi_2)-I(Z_i;\psi_2|A_i)-n\delta(\epsilon)\\
&\overset{(a)}{=} I(X_i;A_i,Y_i,\psi_1,\psi_2)-I(X_i;Y_i|A_i,\psi_2)-I(Y_i;A_i|X_i)-I(Z_i;Y_i|A_i)\\
&+I(Y_i;A_i,\psi_2|X_i)+I(Z_i;Y_i|\psi_2,A_i)-n\delta(\epsilon)\\
&\overset{(b)}{=} I(X_i;A_i,Y_i,\psi_1,\psi_2)-I(X_i;Y_i|A_i,\psi_2)+I(X_i;A_i)-I(Y_i,X_i;A_i)\\
&-I(Z_i;X_i,Y_i|A_i)+I(X_i,Y_i;A_i,\psi_2)+I(Z_i;X_i,Y_i|\psi_2,A_i)-I(X_i;A_i,\psi_2)-n\delta(\epsilon)\\
&=I(X_i;A_i)+I(X_i;A_i,Y_i,\psi_1,\psi_2)+I(X_i,Y_i;A_i,\psi_2,Z_i)-I(A_i,Z_i;X_i,Y_i)\\
&-I(X_i;Y_i,A_i,\psi_2)-n\delta(\epsilon)\\
&=I(X_i;A_i)+I(X_i;A_i,Y_i,\psi_1,\psi_2)+I(X_i,Y_i;\psi_2|A_i,Z_i)-I(X_i;Y_i,A_i,\psi_2)-n\delta(\epsilon)\\
&=I(X_i;A_i)+I(X_i,Y_i;\psi_2|A_i,Z_i)+I(X_i;\psi_1|A_i,Y_i,\psi_2)-n\delta(\epsilon),
\end{aligned}
\tag{N.6}
$$

where (a) is true since

$$
\begin{aligned}
&I(Y_i;A_i|X_i)+I(Z_i;Y_i|A_i)-I(Y_i;A_i,\psi_2|X_i)-I(Z_i;Y_i|\psi_2,A_i)\\
&=H(Y_i|X_i)-H(Y_i|X_i,A_i)+H(Z_i|A_i)-H(Z_i|A_i,Y_i)-H(Y_i|X_i)+H(Y_i|X_i,A_i)\\
&-H(Z_i|\psi_2,A_i)+H(Z_i|A_i,Y_i)\\
&=H(Z_i|A_i)-H(Z_i|\psi_2,A_i);
\end{aligned}
$$

(b) follows because $I(Z_i;X_i,Y_i|A_i)=I(Z_i;Y_i|A_i)$ and $I(Z_i;X_i,Y_i|A_i,\psi_2)=I(Z_i;Y_i|A_i,\psi_2)$.

Next, define $\hat{X}_{ji}=\psi_{ji}(X^n)$ for $j=1,2$ and $i=1,2,...,n$ and let Q be a random variable uniformly distributed over $[1,n]$ and independent of all the other random variables and with $X\triangleq X_Q$, $Y\triangleq Y_Q$, $A\triangleq A_Q$, from (N.2), it holds that

$$
nR_b \geq H(X|Q)-H(X|A,Q)\overset{(a)}{\geq} H(X)-H(X|A)=I(X;A),
$$

where (a) follows since X^n is i.i.d. and since conditioning decreases entropy. Next, from (N.3), the following holds

$$
\begin{aligned}
n(R_1+R_b) &\geq I(X;A|Q)+I(X;\hat{X}_1,\hat{X}_2|Y,A,Q)\\
&\overset{(a)}{\geq} I(X;A)+I(X;\hat{X}_1,\hat{X}_2|Y,A),
\end{aligned}
$$

where (a) follows since X^n is i.i.d., since conditioning decreases entropy and by the problem definition. From (N.4), it is also true that

$$\begin{aligned}
n(R_2+R_b) &\geq I(X;A|Q)+I(X,Y;\hat{X}_2|A,Z,Q)\\
&\overset{(a)}{\geq} I(X;A)+H(X,Y|A,Z,Q)-H(X,Y|A,Z,\hat{X}_2)\\
&\overset{(b)}{=} I(X;A)+H(Y|A,Z)+H(X|A,Y,Z)-H(X,Y|A,Z,\hat{X}_2)\\
&= I(X;A)+I(X,Y;\hat{X}_2|A,Z)\\
&\geq I(X;A)+I(X;\hat{X}_2|A,Z)
\end{aligned}$$

where (a) follows since X^n is i.i.d. and by conditioning reduces entropy; and (b) follows by the problem definition. Finally, from (N.6), the following holds

$$\begin{aligned}
&n(R_1+R_2+R_b) \geq I(X,A|Q)+I(X,Y;\hat{X}_2|A,Z,Q)+I(X;\hat{X}_1|A,Y,\hat{X}_2,Q)\\
&\overset{(a)}{\geq} I(X,A)+H(X,Y|A,Z,Q)-H(X,Y|A,Z,\hat{X}_2)+I(X;\hat{X}_1|A,Y,\hat{X}_2)\\
&\overset{(b)}{=} I(X;A)+H(Y|A,Z)+H(X|A,Y,Z)-H(X,Y|A,Z,\hat{X}_2)+I(X;\hat{X}_1|A,Y,\hat{X}_2)\\
&= I(X;A)+I(X,Y;\hat{X}_2|A,Z)+I(X;\hat{X}_1|A,Y,\hat{X}_2)\\
&\geq I(X;A)+I(X;\hat{X}_2|A,Z)+I(X;\hat{X}_1|A,Y,\hat{X}_2) \qquad \text{(N.7)}
\end{aligned}$$

where (a) follows since X^n is i.i.d, since conditioning decreases entropy, and by the problem definition; and (b) follows by the problem definition. From cost constraint (8.5), it holds

$$\Gamma+\epsilon \geq \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[\Lambda(A_i)\right]=\mathrm{E}\left[\Lambda(A)\right]. \qquad \text{(N.8)}$$

Moreover, let $\mathcal{B}$ be the event $\mathcal{B}=\{(\psi_1(X^n)\neq h_1(M_1,M_b,Y^n))\wedge(\psi_2(X^n)\neq h_2(M_2,M_b))\}$. Using the CR requirement (5.26), $\Pr(\mathcal{B})\leq\epsilon$. For $j=1,2$, it is true that

$$\begin{aligned}
\mathrm{E}\left[d(X_j,\hat{X}_j)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d(X_{ji},\hat{X}_{ji})\right]\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Big[d(X_{ji},\hat{X}_{ji})\Big|\mathcal{B}\Big]\Pr(\mathcal{B})+\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Big[d(X_{ji},\hat{X}_{ji})\Big|\mathcal{B}^c\Big]\Pr(\mathcal{B}^c)\\
&\overset{(a)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d(X_{ji},\hat{X}_{ji})\Big|\mathcal{B}^c\right]\Pr(\mathcal{B}^c)+\epsilon D_{max}\\
&\overset{(b)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d(X_{ji},h_{ji})\right]+\epsilon D_{max}\\
&\overset{(c)}{\leq} D_j+\epsilon D_{max}, \qquad \text{(N.9)}
\end{aligned}$$

where (a) follows using the fact that $\Pr(\mathcal{B}) \leq \epsilon$ and that the distortion is upper bounded by D_{max}; (b) follows by the definition of $\hat{X}_{ji}$ and $\mathcal{B}$; and (c) follows by (8.6).

APPENDIX O

CONVERSE PROOF FOR PROPOSITION 6.1 AND 6.2

Here, the converse part of Proposition 6.1 is proved. For any $(n, R_1, R_2, D_1+\epsilon, D_2+\epsilon, \Gamma+\epsilon)$ code, the series of inequalities below hold

$$\begin{aligned}
nR_1 &\geq H(M_1) \\
&\overset{(a)}{=} I(M_1; X^n, Y^n) \\
&= H(X^n) + H(Y^n|X^n) \\
&- H(Y^n|M_1) - H(X^n|Y^n, M_1) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} H(X_i) - H(X_i|X_{i+1}^n, Y^n, M_1, A^i) \\
&+ H(Y_i|Y^{i-1}, X^n, M_1, A_i) - H(Y_i|Y^{i-1}, M_1, A_i) \\
&\overset{(c)}{\geq} \sum_{i=1}^{n} H(X_i) - H(X_i|A_i, Y_i, U_i) \\
&+ H(Y_i|X_i, A_i) - H(Y_i|A_i),
\end{aligned} \tag{O.1}$$

where (a) follows since M_1 is a function of X^n and since conditioning reduces entropy; (b) follows since A_i is a function of (M_1, Y^{i-1}) and M_1 is a function of X^n and (c) follows since conditioning decreases entropy, by defining $U_i = (M_1, X_{i+1}^n, A^{i-1}, Y^{i-1})$ and using the fact that the vending machine is memoryless. It is also true that

$$\begin{aligned}
nR_2 &\geq H(M_2) \\
&\geq H(M_2|X^n, M_1) \\
&\overset{(a)}{=} I(M_2; Y^n|X^n, M_1) \\
&\overset{(b)}{=} \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X^n, M_1, A^i) \\
&- H(Y_i|Y^{i-1}, X^n, M_1, M_2, A^i) \\
&\overset{(c)}{\geq} \sum_{i=1}^{n} H(Y_i|X_i, A_i, U_i) - H(Y_i|X_i, A_i, U_i, V_i),
\end{aligned} \tag{O.2}$$

where (a) follows since M_2 is a function of (M_1, Y^n), (b) follows since A_i is a function of (M_1, Y^{i-1}) and (c) follows since the Markov chain Y_i—(X_i, U_i, A_i)—X^{i-1} holds by the problem definition (the validity of the Markov chain can be verified using d-separation on the Bayesian network representation of the joint distribution of the variables at hand as induced by the system model in Figure O.1, see, e.g., [70, Section

A.9]), since conditioning reduces entropy and by defining $V_i = M_2$. It is recalled that d-separation [70, Section A.9] is a standard procedure that allows to test whether a set of variables X is independent of another set Y, when conditioning on a third set Z. The procedure operates on the Bayesian network that describes the joint distribution of all the variables.

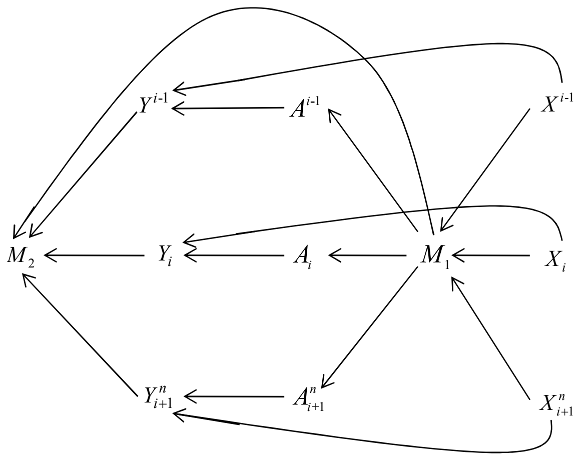

Figure O.1 Bayesian network representing the joint pmf of variables $(M_1, M_2, X^n, Y^n, A^n)$ for the two-way source coding problem with a vending machine in Figure 6.2.

Defining Q to be a random variable uniformly distributed over $[1, n]$ and independent of all the other random variables and with $X \triangleq X_Q$, $Y \triangleq Y_Q$, $A \triangleq A_Q$, $\hat{X}_1 \triangleq \hat{X}_{1Q}$, $\hat{X}_2 \triangleq \hat{X}_{2Q}$, $V \triangleq (V_Q, Q)$ and $U \triangleq (U_Q, Q)$, from (O.1) it holds that

$$\begin{aligned} nR_1 &\geq H(X|Q) - H(X|A, Y, U, Q) \\ &\quad + H(Y|X, A, Q) - H(Y|A, Q) \\ &\overset{(a)}{\geq} H(X) - H(X|A, Y, U) \\ &\quad + H(Y|X, A) - H(Y|A) \\ &= I(X; A) + I(X; U|A, Y), \end{aligned} \tag{O.3}$$

where (a) follows by the fact that source X^n and side information vending machine are memoryless and since conditioning decreases entropy. Next, from (O.2), the following holds

$$\begin{aligned} nR_2 &\geq H(Y|X, A, U) - H(Y|X, A, U, V) \\ &= I(Y; V|X, A, U). \end{aligned} \tag{O.4}$$

Moreover, from Figure O.1 and using d-separation, it can be seen that Markov chains U_i—(X_i, A_i)—Y_i and V_i—(A_i, U_i, Y_i)—X_i hold. This implies that the random variables (X, Y, A, U, V) factorize as in (8.8).

The next step is to show that the estimates $\hat{X}_1$ and $\hat{X}_2$ can be taken to be functions of (V, X) and (U, Y), respectively. To this end, recall that, by the problem definition, the reconstruction $\hat{X}_{1i}$ is a function of (M_2, X^n) and thus of (X_i, U_i, V_i, X^{i-1}). Moreover, $\hat{X}_{1i}$ can be taken to be a function of (X_i, U_i, V_i) only without loss of optimality, due to the Markov chain relationship Y_i—(X_i, U_i, V_i)—X^{i-1}, which can be again proved by d-separation using Figure O.1. This implies that the distortion $d_1(X_i, Y_i, X_1^i)$ cannot be reduced by including also X^{i-1} in the functional dependence of X^i. Similarly, the reconstruction $\hat{X}_{2i}$ is a function of (M_1, Y^n) by the problem definition, and can be taken to be a function of (U_i, Y_i) only without loss of optimality, since the Markov chain relationship X_i—(Y_i, A_i, U_i)—Y_{i+1}^n holds. These arguments and the fact that the definition of V and U includes the time-sharing variable Q allow us to conclude that one can take $\hat{X}_1$ to be a function of (U, V, X) and $\hat{X}_2$ of (U, Y). It is finally observed that V is arbitrarily correlated with U as per (8.8) and thus it is possible without loss of generality to set $\hat{X}_1$ to be a function of (V, X) only. The bounds (6.11) follow immediately from the discussion above and the constraints (8.5)-(8.6).

To bound the cardinality of auxiliary random variable U, it is observed that (8.8) factorizes as

$$p(x, y, a, u, v) = p(u)p(a, x|u)p(y|a, x)p(v|a, u, y). \tag{O.5}$$

Therefore, for fixed $p(y|a, x)$, $p(a, u|x)$ and $p(v|a, u, y)$ the characterization in Proposition 6.1 can be expressed in terms of integrals $\int g_j(\cdot)dF(u)$, for $j = 1, ..., |\mathcal{X}|\,|\mathcal{A}| + 3$, of functions $g_j(\cdot)$ of the given fixed pmfs. Specifically, g_j for $j = 1, ..., |\mathcal{X}_1|\,|\mathcal{X}_2| - 1$, are given by $p(a, x|u)$ for all values of $x \in \mathcal{X}$ and $a \in \mathcal{A}$ (except one); $g_{|\mathcal{X}_1||\mathcal{X}_2|} = H(X|A, Y, U = u)$; $g_{|\mathcal{X}_1||\mathcal{X}_2|+1} = I(Y; V|A, X, U = u)$; $g_{|\mathcal{X}_1||\mathcal{X}_2|+2} = \mathrm{E}[d_1(X, Y, \mathrm{f}_1(V, X))|U = u]$ and $g_{|\mathcal{X}_1||\mathcal{X}_2|+3} = \mathrm{E}[d_2(X, Y, \mathrm{f}_2(U, Y))|U = u]$. The proof is concluded by invoking Caratheodory Theorem.

To bound the cardinality of auxiliary random variable V, it is noted that (8.8) can be factorized as

$$p(x, y, a, u, v) = p(v)p(a, y, u|v)p(x|a, u, y), \tag{O.6}$$

so that, for fixed $p(x|a, u, y)$, the characterization in Proposition 6.1 can be expressed in terms of integrals $\int g_j(p(a, u, y|v))dF(v)$, for $j = 1, ..., |\mathcal{A}||\mathcal{U}||\mathcal{Y}| + 1$, of functions $g_j(\cdot)$ that are continuous on the space of probabilities over alphabet $|\mathcal{A}| \times |\mathcal{U}| \times |\mathcal{Y}|$. Specifically, g_j for $j = 1, ..., |\mathcal{A}||\mathcal{U}||\mathcal{Y}| - 1$, are given by $p(a, u, y)$ for all values of $a \in \mathcal{A}$, $u \in \mathcal{U}$ and $y \in \mathcal{Y}$ (except one); $g_{|\mathcal{A}||\mathcal{U}||\mathcal{Y}|} = H(Y|A, X, U, V = v)$; and $g_{|\mathcal{A}||\mathcal{U}||\mathcal{Y}|+1} = \mathrm{E}[d_1(X, Y, \mathrm{f}_1(V, X))|V = v]$. The proof is concluded by invoking Fenchel–Eggleston–Caratheodory Theorem [1, Appendix C].

The converse for Proposition 6.2 follows similar steps as above with the only difference that here the following inequalities hold

$$\begin{aligned} nR_1 &\overset{(a)}{\geq} \sum_{i=1}^{n} H(Z_i) - H(Z_i|Z_{i+1}^n, Y^n, M_1, A^i) \\ &+ H(Y_i|Y^{i-1}, Z^n, M_1, A_i) - H(Y_i|Y^{i-1}, M_1, A_i) \\ &\overset{(b)}{\geq} \sum_{i=1}^{n} H(Z_i) - H(Z_i|A_i, Y_i, U_i) + H(Y_i|Z_i, A_i) - H(Y_i|A_i), \end{aligned} \tag{O.7}$$

where (a) follows as in (a)-(b) of (O.1); and (b) follows since Markov chain relationship Y_i—(Z_i, A_i)—$(Y^{i-1}, Z^{n\backslash i}, M_1)$ holds. The rest of the proof is as above.

APPENDIX P

PROOFS FOR THE EXAMPLE IN SECTION 6.4

1) $D_1 = D_{1,max}$ and $D_2 = 0$:

Here, it is proved the rate-cost region in Proposition 6.2 is given by (6.19) for $D_1 = D_{1,max}$ and $D_2 = 0$. Let's begin with the converse part. Starting from (6.15a), it is true that

$$\begin{aligned} R_1 &\overset{(a)}{\geq} I(A;Z) + H(Z|A,Y) \\ &= H(Z) - I(Z;Y|A) \\ &\overset{(b)}{\geq} H(Z) - \Gamma I(Z;X|A=1) && \text{(P.1)} \\ &\overset{(c)}{\geq} H(Z) - \Gamma H(X|A=1) \\ &\overset{(d)}{\geq} H(Z) - \Gamma \\ &\overset{(e)}{=} H_2(\epsilon) + 1 - \epsilon - \Gamma, && \text{(P.2)} \end{aligned}$$

where (a) follows from (6.15a) and since Z has to be recovered losslessly at Node 2; (b) follows since $\Pr[A=1] = \mathrm{E}[\Lambda(A)] \leq \Gamma$; (c) follows because entropy is non-negative; (d) follows since $H(X|A=1) \leq 1$; and (e) follows because $H(Z) = H_2(\epsilon) + 1 - \epsilon$. Achievability follows by setting $U = Z$, $V = \emptyset$, $\Pr(A=1|Z=0) = \Pr(A=1|Z=1) = \Gamma/(1-\epsilon)$ and $\Pr(A=0|Z=\mathrm{e}) = 1$ in (6.15).

2) $D_1 = 0$ and $D_2 = D_{2,max}$:

Here, the case $D_1 = 0$ and $D_2 = D_{2,max}$ is evaluated. Let's start with the converse. Since X is to be reconstructed losslessly at Node 1, the requirement $H(X|V,Z) = 0$ is resulted from (6.17a). It is easy to see that this requires that the equalities $A = 1$ and $V = Y = X$ are met if $Z = \mathrm{e}$. In fact, otherwise, X could not be a function of (V,Z) as required by the equality $H(X|V,Z) = 0$. The condition that $A = 1$ if $Z = \mathrm{e}$ requires that the pmf $p(a|z)$ be such that $\Pr(A=1|Z=\mathrm{e}) = 1$, which entails $\Gamma = \Pr[A=1] \geq \Pr[Z=\mathrm{e}] = \epsilon$. Moreover, one can set $\Pr(A=1|Z=0) = \Pr(A=1|Z=1) = (\gamma-\epsilon)/(1-\epsilon)$, for some $0 \leq \gamma \leq \Gamma$, by leveraging the symmetry of the problem on the selection of the actions given $Z = 0$

and $Z = 1$[1]. Starting from (6.15a), it holds that

$$\begin{aligned}
R_1 &\overset{(a)}{\geq} I(Z;A) \\
&= H(Z) - H(Z|A) \\
&= H_2(\epsilon) + 1 - \epsilon - \gamma H(Z|A=1) - (1-\gamma)H(Z|A=0) \\
&\overset{(b)}{=} H_2(\epsilon) + 1 - \epsilon - \gamma H\Big(\frac{\epsilon}{\gamma}, \frac{\gamma-\epsilon}{2\gamma}, \frac{\gamma-\epsilon}{2\gamma}\Big) - (1-\gamma) \\
&= H_2(\epsilon) - \gamma H_2\big(\frac{\epsilon}{\gamma}\big) \\
&\overset{(a)}{\geq} H_2(\epsilon) - \Gamma H_2\big(\frac{\epsilon}{\Gamma}\big), \qquad \text{(P.3)}
\end{aligned}$$

where (a) follows from (6.15a) and since there is no distortion requirement at Node 2; (b) follows by direct calculation; and (c) follows since $H_2(\epsilon) - \gamma H_2(\frac{\epsilon}{\gamma})$ is minimized at $\gamma = \Gamma$ over all $0 \leq \gamma \leq \Gamma$.

The bound (6.20b) follows immediately by providing Node 2 with the sequence X^n and then using the bound $R_2 \geq H(X|Z) = \epsilon$.

Achievability follows by setting $U = \emptyset$ and the pmf $p(a|z)$ be such that $\Pr(A = 1|Z = \mathrm{e}) = 1$ and $\Pr(A = 1|Z = 0) = \Pr(A = 1|Z = 1) = \frac{\Gamma-\epsilon}{1-\epsilon}$. Moreover, let $V = Y = X$ if $Z = \mathrm{e}$ and $V = Y = \phi$ otherwise. Evaluating (6.15) with these choices leads to (6.20).

3) $D_1 = D_2 = 0$:

Here, the rate-cost region (6.21) is proved for the case $D_1 = D_2 = 0$. Starting from (6.15a), it is true that

$$\begin{aligned}
R_1 &\overset{(a)}{\geq} H(Z) - \Gamma I(Z;X|A=1) \\
&\overset{(b)}{=} H(Z) - \Gamma H(X|A=1) + \Gamma H(X|A=1, Z=\mathrm{e})\Pr(Z=\mathrm{e}|A=1) \\
&\overset{(c)}{\geq} H(Z) - \Gamma + \Gamma . \frac{\epsilon}{\Gamma} \\
&= H_2(\epsilon) + 1 - \Gamma, \qquad \text{(P.4)}
\end{aligned}$$

where (a) follows as in (P.1); (b) follows because $H(X|A = 1, Z = 0) = H(X|A = 1, Z = 1) = 0$; (c) follows since $H(X|A = 1) \leq 1$, $H(X|A = 1, Z = \mathrm{e}) = 1$ and because $p(Z = \mathrm{e}|A = 1) = \frac{\epsilon}{\Gamma}$, where latter follows from the requirement $H(X|V, Z) = 0$ as per discussion provided in the previous section.

For the achievability, let $U = Z$, $\Pr(A = 1|Z = \mathrm{e}) = 1$ and $\Pr(A = 1|Z = 0) = \Pr(A = 1|Z = 1) = \frac{\Gamma-\epsilon}{1-\epsilon}$. Moreover, let $V = Y = X$ if $Z = \mathrm{e}$ and $V = Y = \emptyset$ otherwise. Evaluating (6.15) with these choices leads to (6.21).

---

[1]This is due to the fact that, by the problem definition, the events $Z = 0$ and $Z = 1$ are statistically equivalent, and hence, there is no advantage in mapping $Z = 0$ to $A = 1$ with higher probability than mapping $Z = 1$ to $A = 1$ and vice versa.

APPENDIX Q

CONVERSE PROOF FOR PROPOSITION 6.3

Here, the converse part of Proposition 6.3 is proved. For any $(n, R_1, R_2, D_1+\epsilon, D_2+\epsilon, D_3+\epsilon, \Gamma+\epsilon)$ code, the series of inequalities below hold

$$\begin{aligned}
nR_1 &\geq H(M_1) \\
&\overset{(a)}{\geq} \sum_{i=1}^{n} H(Z_i) - H(Z_i|Z_{i+1}^n, Y^n, M_1, A^i, \hat{X}_{3i}) \\
&+H(Y_i|Y^{i-1}, X^n, M_1, A_i, \hat{X}_{3i}) - H(Y_i|Y^{i-1}, M_1, A_i, \hat{X}_{3i}) \\
&\overset{(b)}{\geq} \sum_{i=1}^{n} H(Z_i) - H(Z_i|A_i, Y_i, U_i, \hat{X}_{3i}) + H(Y_i|Z_i, A_i, \hat{X}_{3i}) - H(Y_i|A_i, \hat{X}_{3i}),
\end{aligned} \tag{Q.1}$$

where (a) follows from (a) in (O.7) by noting that $\hat{X}_{3i}$ is a function of M and (b) follows since conditioning decreases entropy, by defining $U_i = (M_1, X_{i+1}^n, A^{i-1}, Y^{i-1})$ and using the Markov chain relationship Y_i—$(Z_i, A_i, \hat{X}_{3i})$—$(Y^{i-1}, X^{n\backslash i}, M_1)$. The following series of inequalities hold

$$\begin{aligned}
nR_2 &\geq H(M_2) \\
&\overset{(a)}{\geq} \sum_{i=1}^{n} H(Y_i|Z_i, A_i, U_i, \hat{X}_{3i}) - H(Y_i|Z_i, A_i, U_i, V_i, \hat{X}_{3i}),
\end{aligned} \tag{Q.2}$$

where (a) follows from (O.2), by replacing sequence X^n with the sequence Z^n and by observing that $\hat{X}_{3i}$ is a function of M_1. Defining Q as in Appendix A, along with $\hat{X}_3 \triangleq \hat{X}_{3Q}$, from (Q.1) it is true that

$$\begin{aligned}
nR_1 &\geq H(Z|Q) - H(Z|A, Y, U, \hat{X}_3, Q) + H(Y|Z, A, \hat{X}_3, Q) - H(Y|A, \hat{X}_3, Q) \\
&\overset{(a)}{\geq} H(Z) - H(Z|A, Y, U, \hat{X}_3) + H(Y|Z, A, \hat{X}_3) - H(Y|A, \hat{X}_3) \\
&= I(Z; A) + I(Z; \hat{X}_3|A) + I(Z; U|A, Y, \hat{X}_3),
\end{aligned}$$

where (a) follows by the fact that source Z^n and side information vending machine are memoryless and since conditioning decreases entropy. Next, from (Q.2), it holds that

$$\begin{aligned}
nR_2 &\geq H(Y|Z, A, U, \hat{X}_3) - H(Y|Z, A, U, V, \hat{X}_3) \\
&= I(Y; V|Z, A, U, \hat{X}_3).
\end{aligned} \tag{Q.3}$$

Moreover, by just adding $\hat{X}_3^n$ to the Bayesian graph in Figure O.1 and using d-separation, it can be seen that Markov chains U_i—(Z_i, A_i)—Y_i and V_i—$(A_i, U_i, Y_i, \hat{X}_3)$—$Z_i$ hold, which implies that the random variables $(X, Y, Z, A, U, V, \hat{X}_3)$ factorize as in (6.25). Based on the discussion in the converse proof in Appendix A, it is easy to see that the estimates $\hat{X}_1$ and $\hat{X}_2$ are functions of (V, X) and (U, Y), respectively. The bounds (6.24) follow immediately from the discussion above and the constraints (8.5)-(8.6) and (6.23).

APPENDIX R

PROOF OF PROPOSITION 7.1

First, it is observed that given the probability of error constraint (7.6) the Fano inequality holds as follows

$$H(M|Y^n) \;\; \leq \;\; n\delta(\epsilon), \tag{R.1}$$

where the notation $\delta(\epsilon)$ represents any function such that $\delta(\epsilon) \to 0$ as $\epsilon \to 0$, and that given the CR constraint (7.9), the Fano inequality holds as follows

$$H(\psi|Y^n) \leq n\delta(\epsilon). \tag{R.2a}$$

It also holds

$$\begin{aligned}
nR &= H(M) \overset{(a)}{\leq} I(M;Y^n) + n\delta(\epsilon) \\
&\overset{(b)}{=} I(M;Y^n) - I(M;S^n|A^n) + n\delta(\epsilon) \\
&= I(\psi, M;Y^n) - I(\psi;Y^n|M) - I(\psi M;S^n|A^n) + I(\psi;S^n|A^n, M) + n\delta(\epsilon) \\
&= I(\psi, M;Y^n) - H(\psi|M) + H(\psi|M, Y^n) - I(\psi, M;S^n|A^n) + H(\psi|A^n, M) \\
&\quad -H(\psi|A^n, M, S^n) + n\delta(\epsilon) \\
&= I(\psi, M;Y^n) - I(\psi;A^n|M) + H(\psi|M, Y^n) - I(\psi, M;S^n|A^n) \\
&\quad -H(\psi|A^n, M, S^n) + n\delta(\epsilon) \\
&\overset{(c)}{\leq} I(\psi, M;Y^n) - I(\psi, M;S^n|A^n) + n\delta(\epsilon) \\
&= \sum_{i=1}^{n} I(\psi, M;Y_i|Y^{i-1}) - I(\psi, M;S_i|S_{i+1}^n, A^n) + n\delta(\epsilon) \\
&\overset{(d)}{\leq} \sum_{i=1}^{n} H(Y_i) - H(Y_i|Y^{i-1}, \psi, M, S_{i+1}^n, A^n) - H(S_i|S_{i+1}^n, A^n) \\
&\quad +H(S_i|Y^{i-1}, \psi, M, S_{i+1}^n, A^n) + n\delta(\epsilon) \\
&\overset{(e)}{=} \sum_{i=1}^{n} H(Y_i) - H(Y_i|U_i) - H(S_i|A_i) + H(S_i|U_i, A_i) + n\delta(\epsilon) \\
&= \sum_{i=1}^{n} I(U_i;Y_i) - I(U_i;S_i|A_i) + n\delta(\epsilon)
\end{aligned} \tag{R.3}$$

where (a) follows due to Fano's inequality as in (R.1); (b) follows using the Markov chain $M - A^n - S^n$; (c) follows by (R.2a) and since mutual information is non-negative (recall that by definition $2n\delta(\epsilon) = n\delta(\epsilon)$); (d) follows using the same steps provided in the proof of Theorem 1 in [44, eq. (9)-(12)] by substituting M with (M,ψ) ; and (e) follows by defining $U_i \triangleq (Y^{i-1},\psi,M,S_{i+1}^n,A^{n\backslash i})$ and because the Markov relation $S_i - A_i - (S_{i+1}^n, A^{n\backslash i})$ holds.

Defining Q to be a random variable uniformly distributed over $[1,n]$ and independent of (A^n,S^n,U^n,X^n,Y^n), and with $A \triangleq A_Q$, $S \triangleq S_Q$, $X \triangleq X_Q$, $Y \triangleq Y_Q$ and $U \triangleq (U_Q,Q)$, from (R.3) the following holds

$$\begin{aligned} R &\leq I(U;Y|Q) - I(U;S|A,Q) + \delta(\epsilon) \\ &\overset{(a)}{=} H(Y|Q) - H(Y|U) - H(S|A,Q) + H(S|A,U) + \delta(\epsilon) \\ &\overset{(b)}{\leq} H(Y) - H(Y|U) - H(S|A) + H(S|A,U) + \delta(\epsilon) \\ &= I(U;Y) - I(U;S|A) + \delta(\epsilon) \end{aligned} \tag{R.4}$$

where (a) follows using the definition of U and (b) follows because conditioning reduces entropy. Moreover, from (8.38), it is true that

$$\Gamma + \epsilon \geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\gamma(A_i,X_i)\right] = \mathrm{E}\left[\gamma(A,X)\right]. \tag{R.5}$$

Next, define $\hat{S}_i = \psi_i(S^n)$ and $\hat{S} = \hat{S}_Q$, where $\psi_i(S^n)$ represents the ith symbol of $\psi(S^n)$. Moreover, let $\mathcal{B}$ be the event $\mathcal{B} = \{\psi(S^n) \neq h_2(Y^n)\}$. Using the CR requirement (7.9), $\Pr(\mathcal{B}) \leq \epsilon$. Then the distortion can be calculated as (the dependence of h_{2i} on Y^n is dropped for simplicity of notation)

$$\begin{aligned} \mathrm{E}\left[d(S,\hat{S})\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d(S_i,\hat{S}_i)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d(S_i,\hat{S}_i)\Big|\mathcal{B}\right]\Pr(\mathcal{B}) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d(S_i,\hat{S}_i)\Big|\mathcal{B}^c\right]\Pr(\mathcal{B}^c) \\ &\overset{(a)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d(S_i,\hat{S}_i)\Big|\mathcal{B}^c\right]\Pr(\mathcal{B}^c) + \epsilon D_{max} \\ &\overset{(b)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[d(S_i,h_{2i})\right] + \epsilon D_{max} \\ &\overset{(c)}{\leq} D + \epsilon D_{max}, \end{aligned} \tag{R.6}$$

where (a) follows using the fact that $\Pr(\mathcal{B}) \leq \epsilon$ and that the distortion is upper bounded by D_{max}; (b) follows by the definition of $\hat{S}_i$ and $\mathcal{B}$; and (c) follows by (7.7).

To bound the cardinality of auxiliary random variable U, first, it is observed that the distribution of the variables $(A,U,S,X,Y,\hat{S})$ identified above factorizes as

$$p(a,u,s,x,y) = p(u)p(a,s,x|u)p(y|x,s,a), \tag{R.7}$$

and $\hat{S}$ is a deterministic function $\phi(U)$. Therefore, for fixed $p(y|x,s,a)$, the characterization in Proposition 8.6 can be expressed in terms of integrals $\int g_j(p(a,s,x|u))dF(u)$ for $j = 1,...,|\mathcal{A}| \times |\mathcal{S}| \times |\mathcal{X}| + 2$, of functions $g_j(.)$ that are continuous over pmf on the alphabet $|\mathcal{A}| \times |\mathcal{S}| \times |\mathcal{X}|$. Specifically, g_j for $j = 1,...,|\mathcal{A}| \times |\mathcal{S}| \times |\mathcal{X}| - 1$ are given by $p(a,s,x)$ for all values of $a \in \mathcal{A}$, $s \in \mathcal{S}$, and $x \in \mathcal{X}$ (except one); $g_{|\mathcal{A}|\times|\mathcal{S}|\times|\mathcal{X}|} = H(Y|U=u)$; $g_{|\mathcal{A}|\times|\mathcal{S}|\times|\mathcal{X}|+2} = H(S|A,U=u)$; and $g_{|\mathcal{A}|\times|\mathcal{S}|\times|\mathcal{X}|+1} = \mathrm{E}\left[d(S,\hat{S})|U=u\right]$. The cardinality bound follows by invoking Fenchel-Eggleston-Caratheodory Theorem [1, Appendix C]. Finally, it is observed that the joint distribution (R.7) can be written as (8.8) without loss of generality, since U can always contain A without reducing rate (7.11).

APPENDIX S

PROOF OF PROPOSITION 7.2

The proof is similar to that given in [51] (see also [52]), although care must be taken to properly account for the presence of the actions. First, it is observed that given the probability of error constraint (7.22), the Fano inequality below holds $H(M_j|Y_j^n) \leq n\delta(\epsilon)$ for $j = 1, 2$. The following holds

$$\begin{aligned}
nR_2 &= H(M_2) \overset{(a)}{\leq} I(M_2; Y_2^n) + n\delta(\epsilon) && \text{(S.1)}\\
&\overset{(b)}{=} \sum_{i=1}^{n} I(M_2; Y_{2i}|Y_2^{i-1}) + n\delta(\epsilon) && \text{(S.2)}\\
&\overset{(c)}{\leq} \sum_{i=1}^{n} I(M_2, Y_2^{i-1}; Y_{2i}) + n\delta(\epsilon) && \text{(S.3)}\\
&\overset{(d)}{\leq} \sum_{i=1}^{n} I(M_2, Y_2^{i-1}, Y_1^{i-1}; Y_{2i}) + n\delta(\epsilon) && \text{(S.4)}\\
&\overset{(e)}{=} \sum_{i=1}^{n} I(U_{2i}; Y_{2i}) + n\delta(\epsilon), && \text{(S.5)}
\end{aligned}$$

where (a) follows due to Fano's inequality as in (R.1); (b) follows by using the chain rule for mutual information; (c) and (d) follow because conditioning increases entropy; and (e) follows by defining $U_{2i} \triangleq (M_2, Y_1^{i-1})$ and noting the Markov relation $Y_2^{i-1} - (Y_1^{i-1}, M_2) - Y_{2i}$ due to the degradedness property (7.24). It is also true that

$$\begin{aligned}
nR_1 &= H(M_1) \overset{(a)}{\leq} I(M_1; Y_1^n) + n\delta(\epsilon) && \text{(S.6)}\\
&\overset{(b)}{\leq} I(M_1; Y_1^n|M_2) + n\delta(\epsilon) && \text{(S.7)}\\
&\overset{(c)}{=} \sum_{i=1}^{n} I(M_1; Y_{1i}|Y_1^{i-1}, M_2) + n\delta(\epsilon) && \text{(S.8)}\\
&\overset{(d)}{\leq} \sum_{i=1}^{n} I(M_1, Y_1^{i-1}, S^{i-1}; Y_{1i}|Y_1^{i-1}, M_2) + n\delta(\epsilon) && \text{(S.9)}\\
&\overset{(d)}{\leq} \sum_{i=1}^{n} I(U_{1i}; Y_{1i}|U_{2i}) + n\delta(\epsilon), && \text{(S.10)}
\end{aligned}$$

where (a) follows due to Fano's inequality as in (R.1); (b) follows because M_1 and M_2 are independent and since conditioning reduces entropy; (c) follows using the chain rule for mutual information; (d) follows since conditioning decreases entropy; and (e) follows by defining $U_{1i} \triangleq (M_1, Y_1^{i-1}, S^{i-1})$. Let Q be a random variable uniformly distributed over $[1, n]$ and independent of $(A^n, S^n, U_1^n, U_2^n, X^n, Y_1^n, Y_2^n)$ and define $A \triangleq A_Q$, $S \triangleq S_Q$, $X \triangleq X_Q$, $Y_1 \triangleq Y_{1Q}$, $Y_2 \triangleq Y_{2Q}$, $U_1 \triangleq (U_{1Q}, Q)$, and $U_2 \triangleq (U_{2Q}, Q)$. It can be easily seen that, with these definitions, the sum (S.5) is upper bounded by $I(U_2; Y_2)$, and (S.10) equals $I(U_1; Y_1|U_2)$. Moreover, note that, from the definitions above, X is a function of U_1, U_2 and S, given the encoding function (7.19). Similarly, A is a function of (U_1, U_2) given (7.18). Also, the Markov relationship $(U_1, U_2) - A - S$ holds as it can be easily checked by using the d-separation principle [70]. Finally, from (8.38), it holds that

$$\Gamma + \epsilon \geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[\gamma(A_i, X_i)\right] = \mathrm{E}\left[\gamma(A, X)\right],$$

which completes the proof.

APPENDIX T

PROOF OF PROPOSITION 8.1

Here, the converse part of Proposition 8.1 is proved. For any $(n, R, D_1+\epsilon, D_2+\epsilon, \Gamma+\epsilon)$ code, It holds that

$$
\begin{aligned}
&nR \geq H(M)\\
&= I(M; X^n, Y^n)\\
&= H(X^n, Y^n) - H(X^n, Y^n|M)\\
&= H(X^n) + H(Y^n|X^n) - H(Y^n|M) - H(X^n|M, Y^n)\\
&= \sum_{i=1}^{n} H(X_i) + H(Y_i|Y^{i-1}, X^n) - H(Y_i|Y^{i-1}, M) - H(X_i|X^{i-1}, M, Y^n)\\
&\overset{(a)}{\geq} \sum_{i=1}^{n} H(X_i) + H(Y_i|Y^{i-1}, X^n, A^n) - H(Y_i|Y^{i-1}, M, A^n) - H(X_i|X^{i-1}, M, Y^n, A^n) \qquad \text{(T.1)}\\
&\overset{(b)}{=} \sum_{i=1}^{n} H(X_i) + H(Y_i|Y^{i-1}, X^n, A^n, \hat{X}_{2i}) - H(Y_i|Y^{i-1}, M, A^n, \hat{X}_{2i})\\
&- H(X_i|X^{i-1}, M, Y^n, A^n, \hat{X}_{2i}) \qquad \text{(T.2)}\\
&\overset{(c)}{\geq} \sum_{i=1}^{n} H(X_i) + H(Y_i|X_i, A_i, \hat{X}_{2i}) - H(Y_i|A_i, \hat{X}_{2i}) - H(X_i|U_i, Y_i, A_i, \hat{X}_{2i}), \qquad \text{(T.3)}
\end{aligned}
$$

where (a) because A^n is a function of M and since conditioning reduces entropy; (b) follows since $\hat{X}_{2i}$ is a function of A^n; and (c) follows because the Markov relation Y_i—$(X_i, A_i, \hat{X}_{2i})$—$(X^{n\backslash i}, A^{n\backslash i})$ holds, by defining $U_i = (M, X^{i-1}, Y^{n\backslash i}, A^{i-1})$ and since conditioning decreases entropy.

Defining Q to be a random variable uniformly distributed over $[1, n]$ and independent of all the other random variables and with $X \triangleq X_Q$, $Y \triangleq Y_Q$, $A \triangleq A_Q$, $\hat{X}_1 \triangleq \hat{X}_{1Q}$, $\hat{X}_2 \triangleq \hat{X}_{2Q}$ and $U \triangleq (U_Q, Q)$, from (T.3) The following holds

$$
\begin{aligned}
nR &\geq H(X|Q) + H(Y|X, A, \hat{X}_2, Q) - H(Y|A, \hat{X}_2, Q) - H(X|U, Y, A, \hat{X}_2, Q)\\
&\overset{(a)}{\geq} H(X) + H(Y|X, A, \hat{X}_2) - H(Y|A, \hat{X}_2) - H(X|U, Y, A, \hat{X}_2)\\
&= I(X; U, Y, A, \hat{X}_2) - I(Y; X|A, \hat{X}_2)\\
&= I(X; A, \hat{X}_2) + I(X; U|Y, A, \hat{X}_2),
\end{aligned}
$$

where in (a) holds due to the fact that X^n is i.i.d., conditioning reduces entropy and by the problem definition. Moreover, the following chain of inequalities hold

$$H(\mathrm{f}(A^n)) \leq \sum_{i=1}^{n} H(\mathrm{f}(A_i)) = nH(\mathrm{f}(A)|Q) \leq nH(\mathrm{f}(A)), \tag{T.4}$$

where the last inequality follows since conditioning reduces entropy, and

$$\begin{aligned}
H(\mathrm{f}(A^n)) &\geq I(\mathrm{f}(A^n); X^n) \\
&= \sum_{i=1}^{n} I(\mathrm{f}(A^n); X_i|X^{i-1}) \\
&= \sum_{i=1}^{n} I(\mathrm{f}(A^n), \hat{X}_{2i}; X_i|X^{i-1}) \\
&= \sum_{i=1}^{n} I(\mathrm{f}(A^n), \hat{X}_{2i}, X^{i-1}; X_i) \\
&\overset{(a)}{\geq} \sum_{i=1}^{n} I(\mathrm{f}(A_i), \hat{X}_{2i}; X_i) \\
&= n(H(X|Q) - H(X|\mathrm{f}(A), \hat{X}_2, Q)) \\
&\overset{(b)}{\geq} n(H(X) - H(X|\mathrm{f}(A), \hat{X}_2)) \\
&= n(I(X; \mathrm{f}(A), \hat{X}_2)),
\end{aligned} \tag{T.5}$$

where (a) follows by the chain rule for mutual information and since mutual information is non-negative; and (b) follows since X^n is i.i.d. and due to the fact that conditioning decreases entropy. Combining (T.5) and (T.4), the inequality below obtained

$$I(X; \mathrm{f}(A), \hat{X}_2) \leq H(\mathrm{f}(A)). \tag{T.6}$$

It is noted that the defined random variables factorizes as (8.8) since the Markov chain relationship $(\hat{X}_1, \hat{X}_2, U)$—$(A, X)$—$Y$ holds by the problem definition and that $\hat{X}_2$ is a function $\mathrm{g}(U, Y)$ of U and Y by the definition of U. Moreover, from cost and distortion constraints (8.5)-(8.6), it is true that

$$D_j + \epsilon \geq \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}[d_j(X_i, \hat{X}_{ji})] = \mathrm{E}[d_j(X, \hat{X}_j)], \text{ for } j = 1, 2, \tag{T.7a}$$

$$\text{and } \Gamma + \epsilon \geq \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\left[\Lambda(A_i)\right] = \mathrm{E}\left[\Lambda(A)\right]. \tag{T.7b}$$

The cardinality constraint on the auxiliary random variable U is obtained as follows using Caratheodory's theorem as in [1, Appendix C]. Note that one can write $I(X; \hat{X}_2, A) + I(X; U|\hat{X}_2, A, Y) = H(X) - H(X|\hat{X}_2, A) + H(X|\hat{X}_2, A, Y) -$

$H(X|\hat{X}_2, A, Y, U)$. Now, to preserve the joint distribution of variables (X,$\hat{X}_2$,A), and thus the distribution of all variables (X,$\hat{X}_2$,A,Y) and the terms $H(X)$, $H(X|\hat{X}_2, A)$ and $H(X|\hat{X}_2, A, Y)$ (since $p(y|x, a)$ is fixed), the set $\mathcal{U}$ should have $|\mathcal{X}||\hat{\mathcal{X}}_2||\mathcal{A}| - 1$ elements; moreover, one further element is required to preserve the conditional entropy $H(X|\hat{X}_2, A, Y, U)$ and one for the distortion $E[d_1(X, \hat{X}_1)]$.

APPENDIX U

PROOF OF PROPOSITION 8.2 AND PROPOSITION 8.3

First, the converse part of Proposition 8.2 is proved and then describe the different steps needed to prove Proposition 8.3. The first part of the converse follows the same steps as in Appendix A. However, it is noted that in (T.1) and (T.2), A^i can be written instead of A^n, without changing the following steps. This is due to the strictly causal dependence of $\hat{X}_{2i}$ on the action sequence which is used in (T.2). This allows to validate the claim in Remark 8.2. To prove the constraint in (8.13c), the following chain of inequalities hold

$$H(\mathrm{f}(A^n)) \;=\; \sum_{i=1}^{n} H(\mathrm{f}(A_i)|\mathrm{f}(A^{i-1})) = \sum_{i=1}^{n} H(\mathrm{f}(A_i)|\hat{X}_{2i}) \stackrel{(a)}{\leq} nH(\mathrm{f}(A)|\hat{X}_2). \quad \text{(U.1)}$$

Moreover, it is also true that

$$\begin{aligned}
H(\mathrm{f}(A^n)) &\geq I(\mathrm{f}(A^n); X^n) \\
&= \sum_{i=1}^{n} I(\mathrm{f}(A^n); X_i|X^{i-1}) \\
&= \sum_{i=1}^{n} I(\mathrm{f}(A^n), \hat{X}_{2i}; X_i|X^{i-1}) \\
&= \sum_{i=1}^{n} I(\mathrm{f}(A^n), \hat{X}_{2i}, X^{i-1}; X_i) \\
&\stackrel{(a)}{\geq} \sum_{i=1}^{n} I(\mathrm{f}(A_i), \hat{X}_{2i}; X_i) \\
&= n(H(X|Q) - H(X|\mathrm{f}(A), \hat{X}_2, Q)) \\
&\stackrel{(b)}{\geq} n(H(X) - H(X|\mathrm{f}(A), \hat{X}_2)) \\
&= n(I(X; \mathrm{f}(A), \hat{X}_2)), \qquad \text{(U.2)}
\end{aligned}$$

where (a) follows by the chain for mutual information and since mutual information is non-negative; and (b) follows since X^n is i.i.d. and due to the fact that conditioning decreases entropy. Combining (U.2) and (U.1), the inequality (8.13c) is obtained. It is noted that the joint pmf of the defined random variables factorizes as (8.12) since the Markov chain relationship $(\hat{X}_1, \hat{X}_2, U)$—$(A, X)$—$Y$ holds by the problem definition and that $\hat{X}_1$ is a function $\mathrm{g}(U, Y)$ of U and Y by the definition of U as in Appendix A. The distortion, cost and cardinality constraint are obtained as in Appendix A.

The converse for Proposition 8.3 follows from similar steps by defining $V_i = \mathrm{f}(A^{i-1})$ and noting that $\hat{X}_{2i}$ is a function of V_i and $\mathrm{f}(A_i)$.

The cardinality of the auxiliary random variables U and V for Proposition 8.3 is bounded using [1, Appendix C]. The bounds for U in Proposition 8.2 follow in the same way. Note that one can write $I(X;V,A) + I(X;U|V,A,Y) = H(X) - H(X|V,A) + H(X|V,A,Y) - H(X|V,A,Y,U)$. Starting with V, the alphabet $\mathcal{V}$ should have $|\mathcal{X}| - 1$ elements to preserve the distribution $p(x)$ and hence $H(X)$, one element to preserve $-H(X|V,A) + H(X|V,A,Y)$, two elements to preserve the distortion constraints and and one more to preserve the condition $I(X;V,\mathrm{f}(A)) \leq H(\mathrm{f}(A)|V)$. As for U, just as in Appendix A, U should have $|\mathcal{X}||\mathcal{V}||\mathcal{A}| - 1$ elements to preserve the joint distribution $p(x,v,a)$ (which preserves the joint distribution $p(x,a,v,y)$ and hence $H(X)$, $H(X|V,A)$, $H(X|V,A,Y)$), one element to preserve $H(X|V,A,Y,U)$ and one more to preserve the distortion constraint of Decoder 1.

APPENDIX V

PROOF OF PROPOSITION 8.4

Here, the converse part of Proposition 8.4 is proved. To establish the converse, it is sufficient to consider the case of non-causal action observation, as done in the following. For any $(n, R, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)$ code, define the auxiliary variable $U_i = (Y^n, X^{i-1})$, and Q as a random time sharing variable uniformly distributed in the interval $[1, n]$ independent of $(X, U, \hat{X}_1, \hat{X}_2, A, Y)$. It is true that

$$\begin{aligned}
H(Y^n) &= \sum_{i=1}^{n} H(Y_i|Y^{i-1}) \\
&\leq \sum_{i=1}^{n} H(Y_i) \\
&= \sum_{i=1}^{n} H(\mathrm{f}_Y(A_i)) \\
&= nH(\mathrm{f}_Y(A_Q)|Q) \\
&\leq H(\mathrm{f}_Y(A_Q)). \qquad \text{(V.1)}
\end{aligned}$$

Also, the following holds

$$\begin{aligned}
H(Y^n) &\geq I(X^n; Y^n) \\
&= \sum_{i=1}^{n} I(X_i; Y^n|X^{i-1}) \\
&\overset{(a)}{=} \sum_{i=1}^{n} I(X_i; Y^n, X^{i-1}) \\
&= \sum_{i=1}^{n} I(X_i; U_i) \\
&= nI(X_Q; U_Q|Q) \\
&\overset{(b)}{=} I(X_Q; U_Q, Q), \qquad \text{(V.2)}
\end{aligned}$$

along with

$$\begin{aligned} H(A^n|Y^n) &= \sum_{i=1}^{n} H(A_i|Y^n, A^{i-1}) \\ &\leq \sum_{i=1}^{n} H(A_i|Y_i) \\ &= \sum_{i=1}^{n} H(A_i|\mathrm{f}_Y(A_i)) \\ &= n(H(A_Q)|\mathrm{f}_Y(A_Q), Q) \\ &\leq H(A_Q|\mathrm{f}_Y(A_Q)), \end{aligned} \tag{V.3}$$

and

$$\begin{aligned} H(A^n|Y^n) &\geq I(X^n; A^n|Y^n) \\ &\stackrel{(c)}{=} I(X^n; A^n, \hat{X}_2^n|Y^n) \\ &\geq \sum_{i=1}^{n} I(X_i; \hat{X}_{2,i}|Y^n, X^{i-1}) \\ &= \sum_{i=1}^{n} I(X_i; \hat{X}_{2,i}|U_i) \\ &= nI(X_Q; \hat{X}_{2,Q}|U_Q, Q). \end{aligned} \tag{V.4}$$

Furthermore, it also holds that

$$\begin{aligned} H(Y^n, M) &\leq \sum_{i=1}^{n} H(Y_i) + nR \\ &\leq nH(Y_Q) + nR, \end{aligned} \tag{V.5}$$

and

$$\begin{aligned} H(Y^n, M) &\geq I(X^n; Y^n, M) \\ &\stackrel{(d)}{=} I(X^n; \hat{X}_1^n, Y^n, M) \\ &\geq \sum_{i=1}^{n} I(X_i; \hat{X}_{1,i}, Y^n|X^{i-1}) \\ &= \sum_{i=1}^{n} I(X_i; \hat{X}_{1,i}, Y^n, X^{i-1}) \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^{n} I(X_i; \hat{X}_{1,i}, U_i) \\ &= n(X_Q; \hat{X}_{1,Q}, U_Q, Q), \end{aligned} \tag{V.6}$$

where (a) follows from the independence of X_i and X^{i-1}; (b) follows from the independence of Q from all other random variables; (c) follows from the fact that $\hat{X}_2^n$ is a function of A^n; and (d) follows from the fact that $\hat{X}_1^n$ is a function of (M, Y^n). Defining $U \triangleq (U_Q, Q)$ along with $X \triangleq X_Q$, $Y \triangleq Y_Q$, $A \triangleq A_Q$, $\hat{X}_1 \triangleq \hat{X}_{1Q}$, $\hat{X}_2 \triangleq \hat{X}_{2Q}$ and combining (V.1), (V.2), (V.3), (V.4), (V.5) and (V.6), the rate region inequalities are obtained as mentioned in the proposition. Note that the joint distribution of the random variables $(X, Y, A, \hat{X}_1, \hat{X}_2)$ established above factorizes as $p(x)p(u, \hat{x}_1, \hat{x}_2, a|x)$ but can be restricted only to pmfs factorizing as in (8.21). This is because the information measures in (8.20)-(8.22) only depends on the marginals $p(x, u, \hat{x}_1)$, $p(a)$ and $p(x, u, \hat{x}_2)$. Distortion and cost constraints are handled in the standard manner [1].

The cardinality of the auxiliary random variables U is bounded using [1, Appendix C]. The set $\mathcal{U}$ should have $|\mathcal{X}||\hat{\mathcal{X}}_1||\hat{\mathcal{X}}_2| - 1$ elements to preserve the joint distribution $p(x, \hat{x}_1, \hat{x}_2)$, one element to preserve the Markov chain $\hat{X}_1 - U - \hat{X}_2$, and three elements to preserve $H(X|\hat{X}_1, U)$, $H(X|\hat{X}_2, U)$ and $H(X|U)$.

APPENDIX W

SKETCH OF PROOF OF ACHIEVABILITY FOR PROPOSITION 8.6

Below, it is proved that the following rate region is achievable

$$\begin{aligned} R_1 &\leq H(\mathrm{f}(A)) && \text{(W.1a)} \\ R_1 + R_2 &\leq I(A, U; Y) - I(U; S|A), && \text{(W.1b)} \\ R_2 &\leq I(A; Y|\mathrm{f}(A)) + I(A, U; Y|A) - I(U; S|A), && \text{(W.1c)} \end{aligned}$$

for a given joint distribution as in (8.40). Assuming now that this rate region is achievable, it is shown that the rate region (8.39) is also achievable. Region (8.39) is larger than (W.1) owing to the absence of the inequality (W.1c). The two regions are illustrated in Figure W.1 for a given choice of the distribution (8.40), with region (W.1) in solid lines and (8.39) in dashed lines. Now, it is argued that the achievability of region (W.1) (solid lines) implies the achievability of region (8.39) (dashed lines) as well, by following the same arguments as in [67]. Specifically, it is observed that, if (R_1, R_2) is achievable with some scheme, then $(R_1 - t, R_2 + t)$ is also achievable for all $0 \leq t \leq R_1$. This is due to the fact that, if the rate pair (R_1, R_2) is achievable, then some of the rate of the common message M_1 can always be transferred to the private message M_2 for Decoder 2 to achieve $(R_1 - t, R_2 + t)$ if $0 \leq t \leq R_1$. It follows immediately that all the points on the dashed line in Figure W.1 are also achievable.

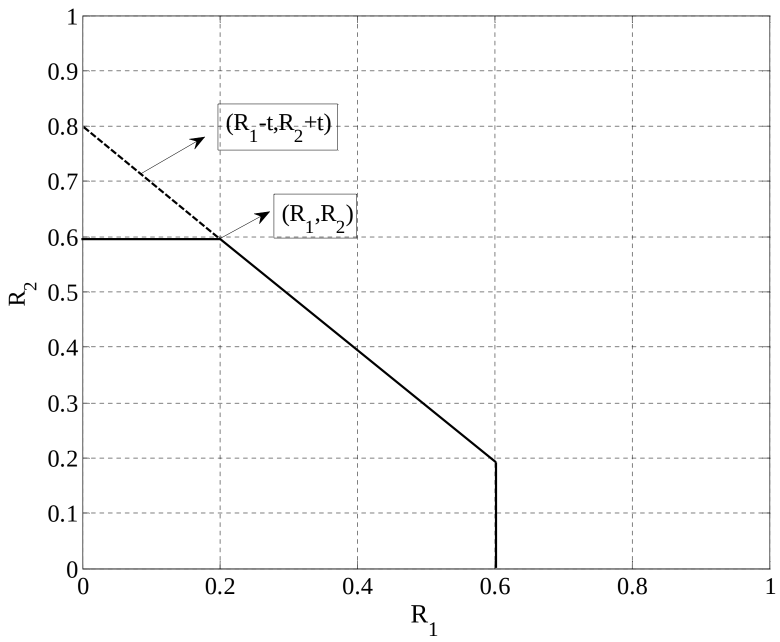

Figure W.1 Illustration of the rate regions (8.39) (dashed lines) and (W.1) (solid lines).

The discussion above allows us to conclude that concludes that, if region (W.1) is achievable, then the desired rate region (8.39) is also achievable. Let's now focus on proving the achievability of (W.1). To this end, superposition coding and the technique proposed in [44] are combined. Fix the joint distribution as in (8.40). First the codebook $b^n(m_1)$, $m_1 \in [1 : 2^{nR_1}]$ is generated i.i.d. with pmf $p(b)$. Next, a superimposed codebook is generated for each b^n of $a^n(m_1, m_2)$ codewords, $m_2 \in [1 : 2^{nR_2}]$, i.i.d. with pmf $p(a|b)$. For every a^n sequence, a codebook of $u^n(m_1, m_2, j)$ sequences is generated, $j \in [1 : 2^{n\tilde{R}}]$, i.i.d. with pmf $p(u|a)$.

To encode messages (m_1, m_2), Encoder selects the codeword $a^n(m_1, m_2)$, and chooses a u^n codeword jointly typical with action and state sequence, which requires $\tilde{R} \geq I(U; S|A)$. Then $x_i = g(u_i, s_i)$ is then sent through the channel. Decoder 1 decodes the message m_1 correctly if $R_1 \leq H(B)$. Decoder 2 looks for the unique pair of messages (m_1, m_2) such that the tuple $(y^n, b^n(m_1), a^n(m_1, m_2), u^n(m_1, m_2, j))$ is jointly typical for some $j \in [1 : 2^{n\tilde{R}}]$. This step is reliable if $R_1 + R_2 + \tilde{R} \leq I(A, U; Y)$ and $R_2 + \tilde{R} \leq I(U, A; Y|B) = I(A; Y|B) + I(U; Y|A)$. Using Fourier-Motzkin elimination to eliminate rate $\tilde{R}$ leads to the bounds (W.1).

BIBLIOGRAPHY

[1] A. El Gamal and Y. Kim, *Network Information Theory*, Cambridge University Press, New York, NY, Dec. 2011.

[2] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. 22, no. 1, pp. 1- 10, Jan. 1976.

[3] C. Heegard and T. Berger, "Rate distortion when side information may be absent," *IEEE Trans. Inform. Theory*, vol. 31, no. 6, pp. 727–734, Nov. 1985.

[4] H. Permuter and T. Weissman, "Source coding with a side information "vending machine"," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4530–4544, Jul. 2011.

[5] A. Aaron, R. Zhang and B. Girod, "Wyner-Ziv coding of motion video," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2002.

[6] Z. Xiong, A. D. Liveris and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 80- 94, Sept. 2004.

[7] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory,* vol. 19, no. 4, pp. 471- 480, Jul. 1973.

[8] Y. Steinberg, "Coding and common reconstruction," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, 2009.

[9] A. H. Kaspi, "Rate-distortion function when side-information may be present at the decoder," *IEEE Trans. Inform. Theory*, vol. 40, pp. 2031–2034, Nov. 1994.

[10] D. Vasudevan, "On the Heegard-Berger/Kaspi problem with decoder cooperation," in *Proc. IEEE Int'l Conf. on Telecommunications*, St. Petersburg, Russia, June 2008.

[11] H. Yamamoto, "Source coding theory for cascade and branching communication systems," *IEEE Trans. Inform. Theory*, vol. 27, no. 3, pp. 299–308, May 1981.

[12] D. Vasudevan, C. Tian, and S. N. Diggavi, "Lossy source coding for a cascade communication system with side-informations," In *Proc. 44th Annual Allerton Conference on Communications, Control and Computing*, Monticello, IL, September 2006.

[13] Y. K. Chia, H. Permuter and T. Weissman, "Cascade, triangular and two way source coding with degraded side information at the second user,"*IEEE Trans. Inform. Theory*, vol. 58, pp. 189–206, Jan. 2012.

[14] R. Tandon, S. Mohajer, and H. V. Poor, "Cascade source coding with erased side information," in *Proc. IEEE Symp. Inform. Theory*, St. Petersburg, Russia, Aug. 2011.

[15] A. Lapidoth, A. Malar, and M. Wigger, "Constrained Wyner-Ziv coding," in *Proc. IEEE Symp. Inform. Theory*, St. Petersburg, Russia, Aug. 2011.

[16] T. Weissman and A. El Gamal, "Source coding with limited-look-ahead side information at the decoder,"*IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5218–5239, Dec. 2006.

[17] R. Timo, A. Grant and G. Kramer, "Rate-distortion functions for source coding with complementary side information," in *Proc. IEEE Symp. Inform. Theory*, St. Petersburg, Russia, Aug. 2011.

[18] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder—II: General sources," *Inform. Contr.*, vol. 38, pp. 60–80, Jul. 1978.

[19] E. Perron, S. N. Diggavi, and E. Telatar, "Lossy source coding with Gaussian or erased side-information," in *Proc. IEEE Symp. Inform. Theory*, Seoul, South Korea, Jun. 28-July 3, 2009.

[20] T. Weissman and S. Verdu. "The information lost in erasures," *IEEE Trans. Inform. Theory*, vol. 54, no. 11, pp. 5030 – 5058, Nov. 2008.

[21] P. Cuff, H.-I. Su and A. El Gamal, "Cascade multiterminal source coding," *in Proc. IEEE Int. Symp. Inform. Theory*, Seoul, South Korea, Jun. 28-July 3, 2009.

[22] B. Ahmadi and O. Simeone, "Robust coding for lossy computing with receiver-side observation costs," in *Proc. IEEE International Symposium on Information Theory* (ISIT 2011), pp. 2939-2943, July 31-Aug. 5, Saint Petersburg, Russia, 2011 (see also arXiv:1108.1535).

[23] Y. Chia, H. Asnani, and T. Weissman, "Multi-terminal source coding with action dependent side information," in *Proc. IEEE International Symposium on Information Theory* (ISIT 2011), pp. 2035-2039, July 31-Aug. 5, Saint Petersburg, Russia, 2011.

[24] K. Kittichokechai, T. J. Oechtering and M. Skoglund, "Source coding with common reconstruction and action-dependent side information," in *Proc. IEEE Information Theory Workshop* (ITW 2010), pp. 1-5, Dublin, Ireland, Aug. 30-Sept. 3 2010.

[25] K. Kittichokechai, T. J. Oechtering and M. Skoglund, "Secure source coding with action-dependent side information," in *Proc. IEEE International Symposium on Information Theory* (ISIT 2011), pp. 1678-1682, July 31-Aug. 5, Saint Petersburg, Russia, 2011.

[26] C. Choudhuri and U. Mitra, "How useful is adaptive action?," in *Proc. IEEE Globecom,* Dec. 3-7, Anaheim, CA, 2012.

[27] L. Zhao, Y. K. Chia, and T. Weissman, "Compression with actions," in Allerton conference on communications, control and computing, Monticello, Illinois, September 2011.

[28] H. Asnani and T. Weissman, "On real time coding with limited lookahead," submitted, http://arxiv.org/abs/1105.5755, May. 2011.

[29] S.-Y. Tung, "Multiterminal source coding," Ph.D. Thesis, Cornell University, Ithaca, NY, 1978.

[30] M. Gastpar, "The Wyner–Ziv problem with multiple sources," *IEEE Trans. Inform. Theory*, vol. 50, no. 11, pp. 2762–2767, Nov. 2004.

[31] B. Ahmadi, C. Choudhuri, O. Simeone and U. Mitra, "Cascade source coding with a side information "Vending Machine"," http://arxiv.org/abs/1207.2793, Jul. 2012.

[32] S. Jana, "Alphabet sizes of auxiliary random variables in canonical inner bounds," in *Proc. Conf. Inf. Sci. Syst. (CISS)*, pp. 67-71, March. 18-20, Baltimore, MD, 2009.

[33] A. H. Kaspi and T. Berger, "Rate-distortion for correlated sources with partially separated encoders," *IEEE Trans. Inform. Theory*, vol. 28, no. 6, pp. 828–840, Nov. 1982.

[34] A. B. Wagner, B. G. Kelly, and Y. Altug, "The lossy one-helper conjecture is false," *in Proc. Allerton Conf. on Communications, Control, and Computing*, pp. 716–723, Sept. 30–Oct. 2, Monticello, IL, 2009.

[35] T. Berger and R. Yeung, "Multiterminal source encoding with one distortion criterion," *IEEE Trans. Inform. Theory*, vol. 35, no. 2, pp. 228–236, Mar 1989.

[36] B. Ahmadi and O. Simeone, "Distributed and cascade lossy source coding with a side information "Vending Machine"," to appear in *IEEE Trans. Inform. Theory*, http://arxiv.org/abs/1109.6665, Sept. 2011.

[37] G. D. Forney Jr., "On the role of MMSE estimation in approaching the information-theoretic limits of linear Gaussian channels: Shannon meets Wiener," in *Proc. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, pp. 430–439, Oct. 2003.

[38] Y.-K. Chia and T. Weissman, "Cascade and triangular source coding with causal side information," in *Proc. IEEE International Symposium on Information Theory* (ISIT 2011), pp. 1683–1687, July 31-Aug. 5, Saint Petersburg, Russia, 2011.

[39] B. Ahmadi, R. Tandon, O. Simeone, and H. V. Poor, "Heegard-Berger and cascade source coding problems with common reconstruction constraints," *IEEE Trans. Inform. Theory*, vol. 59, no. 3, pp. 1458,1474, Mar. 2013.

[40] A. H. Kaspi., "Two-way source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. 31, no. 6, pp. 735–740, Nov. 1985.

[41] H. Permuter, Y. Steinberg, and T. Weissman, "Two-way source coding with a helper," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2905–2919, Jun. 2010.

[42] N. Ma and P. Ishwar, "Some results on distributed source coding for interactive function computation," *IEEE Trans. Inform. Theory*, vol. 57, no. 9, pp. 6180–6195, Sep. 2011.

[43] H. S. Witsenhausen, "Indirect rate distortion problems," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 518–521, Sept. 1980.

[44] T. Weissman, "Capacity of channels with action-dependent states," *IEEE Trans. Inform. Theory*, vol. 56, no. 11, pp. 5396–5411, Nov. 2010.

[45] H. Asnani, H. Permuter, and T. Weissman, "Probing capacity," *IEEE Trans. Inform. Theory*, vol. 57, no. 11, pp. 7317–7332, Nov. 2011.

[46] K. Kittichokechai, T. Oechtering and M. Skoglund, "Coding with action-dependent side information and additional reconstruction requirements," http://arxiv.org/abs/1202.1484, Feb. 2012.

[47] C. Choudhuri and U. Mitra, "Action dependent strictly causal state communication," http://arxiv.org/abs/1202.0934, Feb. 2012.

[48] Y.-H. Kim, A. Sutivong, and T. M. Cover, "State amplification," *IEEE Trans. Inform. Theory,* vol. 54, no. 5, pp. 1850–1859, May 2008.

[49] B. Ahmadi and O. Simeone, "On channels with action-dependent states," http://arxiv.org/abs/1202.4438v1, Feb. 2012.

[50] Y. Steinberg and T. Weissman, "The degraded broadcast channel with action-dependent states," in *Proc. IEEE International Symposium on Information Theory* (ISIT 2012), Cambridge, MA, USA, July 1-6, 2012.

[51] Y. Steinberg, "Coding for the degraded broadcast channel with random parameters, with causal and noncausal side information," *IEEE Trans. Inform. Theory*, vol. 51, no. 8, pp. 2867-2877, Aug. 2005.

[52] S. Sigurjonsson and Y. H. Kim, "On multiple user channels with causal state information at the transmitters," in *Proc. IEEE Symp. Inform. Theory*, Adelaide, Australia, pp. 72-76, Sept. 2005.

[53] Y. Steinberg and S. Shamai, "Achievable rates for the broadcast channel with states known at the transmitter," in *Proc. IEEE Symp. Inform. Theory*, Adelaide, Australia, pp. 2184 - 2188 , Sept. 2005.

[54] B. Ahmadi and O. Simeone, "Distributed and Cascade Lossy Source Coding with a Side Information "vending machine"," in *Proc. IEEE International Symposium on Information Theory* (ISIT 2012), Cambridge, MA, USA, July 1-6, 2012.

[55] B. Ahmadi, O. Simeone, C. Choudhuri, and U. Mitra, "On cascade source coding with a side information"vending machine"," http://arxiv.org/abs/1204.1548, Apr. 2012.

[56] H. Asnani, H. Permuter, and T. Weissman, "To feed or not to feed back," submitted, http://arxiv.org/abs/1011.1607, Nov. 2010.

[57] N. Elia, "When Bode meets Shannon: Control-oriented feedback communication schemes," *IEEE Trans. Autom. Contr.*, vol. 49, no. 9, pp. 1477–1488, Sep. 2004.

[58] T. P. Coleman, "A stochastic control viewpoint on 'Posterior Matching'-style feedback communication schemes," in *Proc. IEEE International Symposium on Information Theory* (ISIT 2009), pp. 1520-1524, 2009.

[59] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Trans. Automatic Control*, vol. 49, no. 7, pp. 1056- 1068, July 2004.

[60] A. Sahai and P. Grover, "Demystifying the Witsenhausen counterexample," *IEEE Control System Magazine*, vol. 30, no. 6, pp. 20–24, Dec. 2010.

[61] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 563- 593, Mar. 2003.

[62] H. Asnani, H. Permuter and T. Weissman, "Successive refinement with decoder cooperation and its channel coding duals," submitted, http://arxiv.org/abs/1203.4865, Mar. 2012.

[63] F. Willems, "The discrete memoryless multiple access channel with partially cooperating encoders (Corresp.)," *IEEE Trans. Inform. Theory*, vol. 29, no. 3, pp. 441- 445, May 1983.

[64] P. Cuff and L. Zhao, "Coordination using implicit communication," in *Proc. IEEE Information Theory Workshop* (ITW 2011), pp. 467-471, Oct. 2011.

[65] S. S. Pradhan, "On the role of feedforward in Gaussian sources: point-to-point source coding and multiple description source coding," *IEEE Trans. Inform. Theory*, vol. 53, no. 1, pp. 331-349, Jan. 2007.

[66] R. M. Gray and A. D. Wyner, "Source coding for a simple network," *Bell System Tech. J.*, vol. 53, pp. 1681–1721, 1974.

[67] J. Korner and K. Marton, "General broadcast channels with degraded message sets," *IEEE Trans. on Information Theory*, vol.23, no.1, pp. 60- 64, Jan 1977.

[68] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.

[69] J. Chen, X. Zhang, T. Berger, and S. B. Wicker, "An upper bound on the sum rate distortion function and its corresponding rate allocation schemes for the CEO problem," *IEEE J. Select. Areas Commun.*, vol. 22, no. 6, pp. 977–987, Aug. 2004.

[70] G. Kramer, *Topics in Multi-User Information Theory*, Hanover, MA, Now Publishers, 2008.