Rate Distortion Theory for Causal Video Coding: Characterization, Computation Algorithm, Comparison, and Code Design

by

Lin Zheng

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2012

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Due to the sheer volume of data involved, video coding is an important application of lossy source coding, and has received wide industrial interest and support as evidenced by the development and success of a series of video coding standards. All MPEG-series and H-series video coding standards proposed so far are based upon a video coding paradigm called predictive video coding, where video source frames $X_1, X_2, \cdots, X_N$ are encoded in a frame by frame manner, the encoder and decoder for each frame X_k, $k = 1, 2, \cdots, N$, enlist help only from all previous encoded frames S_j, $j = 1, 2, \cdots, k-1$. In this thesis, we will look further beyond all existing and proposed video coding standards, and introduce a new coding paradigm called causal video coding, in which the encoder for each frame X_k can use all previous original frames X_j, $j = 1, 2, \cdots, k-1$, and all previous encoded frames S_j, while the corresponding decoder can use only all previous encoded frames. We consider all studies, comparisons, and designs on causal video coding from an information theoretic point of view by modeling each frame X_k itself as a source $X_k = \{X_k(i)\}_{i=1}^{\infty}$. Let $R_c^*(D_1, \cdots, D_N)$ ($R_p^*(D_1, \cdots, D_N)$, respectively) denote the minimum total rate required to achieve a given distortion level $D_1, \cdots, D_N > 0$ in causal video coding (predictive video coding, respectively).

A novel computation approach is proposed to analytically characterize, numerically compute, and compare the minimum total rate of causal video coding $R_c^*(D_1, \cdots, D_N)$ required to achieve a given distortion (quality) level $D_1, \cdots, D_N > 0$. Specifically, we first show that for jointly stationary and ergodic sources $X_1, X_2, \cdots, X_N$, $R_c^*(D_1, \cdots, D_N)$ is equal to the infimum of the nth order total rate distortion function $R_{c,n}(D_1, \cdots, D_N)$ over all n, where $R_{c,n}(D_1, \cdots, D_N)$ itself is given by the minimum of an information quantity over a set of auxiliary random variables. We then present an iterative algorithm for computing $R_{c,n}(D_1, \cdots, D_N)$ and demonstrate the convergence of the algorithm to the global

minimum. The global convergence of the algorithm further enables us to not only establish a single-letter characterization of $R_c^*(D_1, \cdots, D_N)$ in a novel way when the N sources are an independent and identically distributed (IID) vector source, but also demonstrate a somewhat surprising result (dubbed the more and less coding theorem)—under some conditions on source frames and distortion, the more frames need to be encoded and transmitted, the less amount of data after encoding has to be actually sent. With the help of the algorithm, it is also shown by example that $R_c^*(D_1, \cdots, D_N)$ is in general much smaller than the total rate offered by the traditional greedy coding method by which each frame is encoded in a local optimum manner based on all information available to the encoder of the frame. As a by-product, an extended Markov lemma is established for correlated ergodic sources.

From an information theoretic point of view, it is interesting to compare causal video coding and predictive video coding, which all existing video coding standards proposed so far are based upon. In this thesis, by fixing $N = 3$, we first derive a single-letter characterization of $R_p^*(D_1, D_2, D_3)$ for an IID vector source (X_1, X_2, X_3) where X_1 and X_2 are independent, and then demonstrate the existence of such X_1, X_2, X_3 for which $R_p^*(D_1, D_2, D_3) > R_c^*(D_1, D_2, D_3)$ under some conditions on source frames and distortion. This result makes causal video coding an attractive framework for future video coding systems and standards.

The design of causal video coding is also considered in the thesis from an information theoretic perspective by modeling each frame as a stationary information source. We first put forth a concept called causal scalar quantization, and then propose an algorithm for designing optimum fixed-rate causal scalar quantizers for causal video coding to minimize the total distortion among all sources. Simulation results show that in comparison with fixed-rate predictive scalar quantization, fixed-rate causal scalar quantization offers as large as 16% quality improvement (distortion reduction).

Acknowledgements

There are many people I wish to express my sincere gratitude and appreciation who have made this thesis successful and meaningful. First of all, I wish to express my sincere gratitude to my supervisor, Professor En-hui Yang, for his invaluable and constant guidance through out my graduate studies in University of Waterloo. During years of extensive training he provided me, he not only inspired me to do cutting-edge research, but also profoundly influenced my thinking to be precise, logic, and sharp.

I am grateful to have an examining committee with distinguished scholars. I would like to thank Professor Liangliang Xie, Professor Zhou Wang, and Professor Guangzhe Fan for their helpful comments and and insightful questions. I would like to thank Professor Zixiang Xiong from Texas A&M University for his commitment and serving as my external examining committee member.

I would like to thank all my friends and colleagues in the Multimedia Communications Laboratory, University of Waterloo. My sincere appreciation is extended to all friends and co-workers for their friendship and interaction. Special thanks go to Professor Zhen Zhang from University of Southern California, and Dr. Da-ke He from Research In Motion Limited for their great encouragement and constructive discussions along the way. I am also grateful to Dr. Haiquan Wang, Dr. Xiang Yu, Dr. Jing Wang, Dr. Shenghao Yang, Dr. Wei Sun, Dr. Guixing Wu, Dr. Xudong Ma, Jin Meng, Chang Sun, Yuhan Zhou, Fei Teng and Jie Zhang for their expertise and friendship.

Last, but not the least, I am much obliged to my parents, Chunjie Zheng and Jianhua Liao, for their unconditional love, understanding and support in my life.

To My Parents

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Data Compression

With the explosive growth of data volume, data compression has become one of the key driving forces to fill the data in any bandwidth/storage space available. The ultimate goal of data compression is to pursue the best possible compression efficiency.

Data compression, which is studied under the name source coding in information theory, is the science of transforming information into another "compressed" representation by using fewer bits than the original representation would use. The task for data compression consists of two components, an encoder that converts the original data stream into a compressed binary representation, and a decoder reconstructs a data stream from the binary representation. Any method that specifies how the encoder and decoder work is called a data compression algorithm. We distinguish between *lossless* data compression algorithm and *lossy* data compression algorithm in terms of whether the reconstructed data stream is identical to the original data stream. In lossless data compression algorithm, the original data stream is required to be reconstructed exactly; otherwise, the data compression algorithm is lossy.

Lossless data compression is typically used in cases where the original and the reconstructed data streams are required to be identical. For example, text files, executable programs and source code need to be compressed without deviations from the original data because no difference between original and reconstructed data is allowed. By contrast, lossy data compression allows an approximation of the original data to be reconstructed, in other words, it is a data encoding method that compresses data by discarding some of it, in exchange for better compression rates. Typically, data streams contain more information than is needed for a particular purpose, that is, a substantial amount of data can be discarded before the reconstruction is intolerable or sufficiently degraded to be noticed by users. In such practical applications, lossy data compression is usually favored.

Lossy data compression is commonly used to compress multimedia data (image, audio, and video, etc.), where in many applications, exact reconstruction of multimedia data is not necessary and information loss can be tolerated as long as a required perceptual quality can be achieved. The multimedia data after lossy compression can effectively reduce the bandwidth required for transmission via practical multimedia systems, and a widely used criterion that reflects rate saving extent at invisible quality loss is the compression ratio, i.e., the size of the compressed data compared to that of the uncompressed data, of multimedia data. For instance, the compression ratio for still image, audio and video can be achieved as high as $10:1, 10:1,$ and $100:1$ with imperceptible loss of quality, respectively. It is observed that the compression ratio of video is always far superior to that of the other two data types, and the emphasis of this thesis is on lossy video compression from an information theoretic point of view.

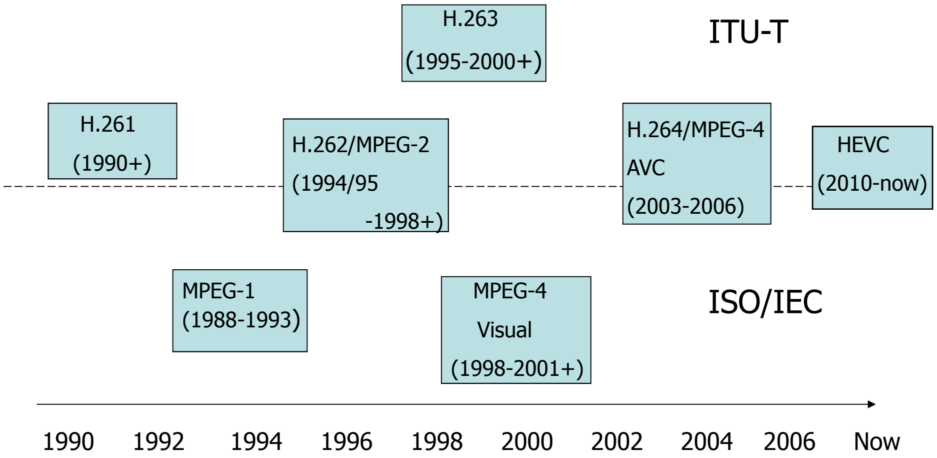

Figure 1.1: Evolution of video compression standards

1.2 Lossy Video Compression

1.2.1 Video Compression Standards Evolution

Due to the sheer volume of data involved, video coding is an important application of lossy source coding, and has received wide industrial interest and support as evidenced by the commercial success of MPEG-2 and H.264/AVC standards. MPEG-2 and H.264/AVC [32] are highlights of the MPEG-series and the H-series standards which have been evolving since MPEG-1/H.261 was ratified in 1988.

As shown in Fig 1.1, the development of MPEG-1 [15] standard began in May 1988 by ISO/IEC MPEG group, and the standard was published in August 1993. MPEG-1 was based on CD-ROM video applications, and also for making video CDs. It is a popular standard for video on the internet even nowadays. MPEG-2 [33] was developed soon after MPEG-1 to support digital television set-top boxes and DVD applications, and to efficiently process interlaced video to satisfy requirements of television applications. On the great success of MPEG-2, the MPEG working group started to work on a new standard MPEG-4 [32]. It was introduced in late 1998 and designated a standard for a group of audio and video coding formats and related technology. Uses of MPEG-4 include compression of data for web (streaming media) and CD distribution, voice (telephone, videophone) and broadcast television applications.

Besides ISO/IEC MPEG, ITU-T VCEG is another working group whose work is essentially focused on efficient video communications over telecommunication networks and computer networks, resulting in a series of standards from H.261 to H.264.

In 2001, a joint group was formed by ITU-T VCEG and ISO/IEC MPEG to work on MPEG-4 Advanced Video Coding (AVC) and MPEG-4 Part-10. This work finally led to the publication of the newest video coding standard H.264, which was intent to create a standard capable of providing good video quality at substantially lower bit rates than previous standards, without increasing the complexity of design so much that it would be impractical or excessively expensive to implement. An additional goal was to provide enough flexibility to allow the standard to be applied to a wide variety of applications on a wide variety of networks and systems.

As a successor to H.264/MPEG-4 AVC, since the formal joint call for proposals in January 2010, a draft video compression standard - High Efficiency Video Coding (HEVC) has been under development by a Joint Collaborative Team on Video Coding (JCT-VC) established based on ISO/IEC MPEG and ITU-T VCEG. HEVC aims to substantially improve coding efficiency compared to AVC High Profile, i.e. to reduce bitrate of compressed video by half at a comparable quality, probably at the expense of increased computational complexity.

1.2.2 Lossy Video Compression Paradigms

In view of all the existing video compression standards and proposals, they are all based on the principles of one specific coding model called *predictive video coding* (PVC) as depicted in Fig 1.2.

In Fig 1.2, source frames (pictures) are coded in a frame by frame manner. I-frame (or intra-frame) is encoded without reference to any other frames except itself, and is always utilized as reference for P-frames. P-frame (or inter-frame) is the predicted source frame, where the prediction is made from all previous reconstructed frames. Typically, I-frame requires more bits to encode than P-frame.

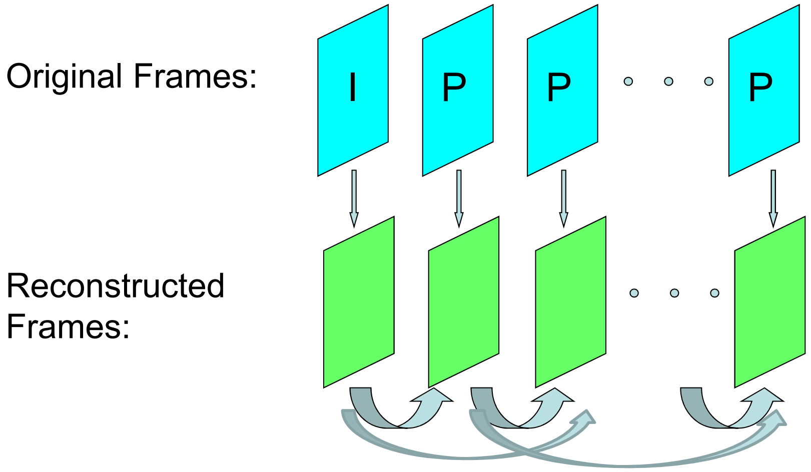

Figure 1.2: Video frames of predictive video coding

All MPEG-series and H-series video coding standards [32], [49] proposed so far fall into the above PVC model; the differences among these different video coding standards lie in how information available to the encoder of each source frame is used to generate its reconstructed frame.

Recently, a new video coding paradigm called *causal video coding* (CVC) was studied in [55], [53], [54]. As shown on the left of Fig.1.3, X_k, $k = 1, 2, \cdots, N$, represents a video frame, S_k and $\hat{X}_k$ represent respectively its encoded frame and reconstructed frame, all frames X_k, $k = 1, 2, \cdots, N$, are encoded in a frame by frame manner, and the encoder for X_k can use all previous frames X_j, $j = 1, 2, \cdots, k-1$, and all previous encoded frames S_j, $j = 1, 2, \cdots, k-1$, while the corresponding decoder can use only all previous encoded frames. The model is causal because the encoder for X_k is not allowed to access to future frames in the encoding order. In a special case where the encoder for each X_k is further restricted to enlist help only from all previous encoded frames S_j, $j = 1, 2, \cdots, k-1$, CVC reduces to PVC, as shown on the right of Fig.1.3.

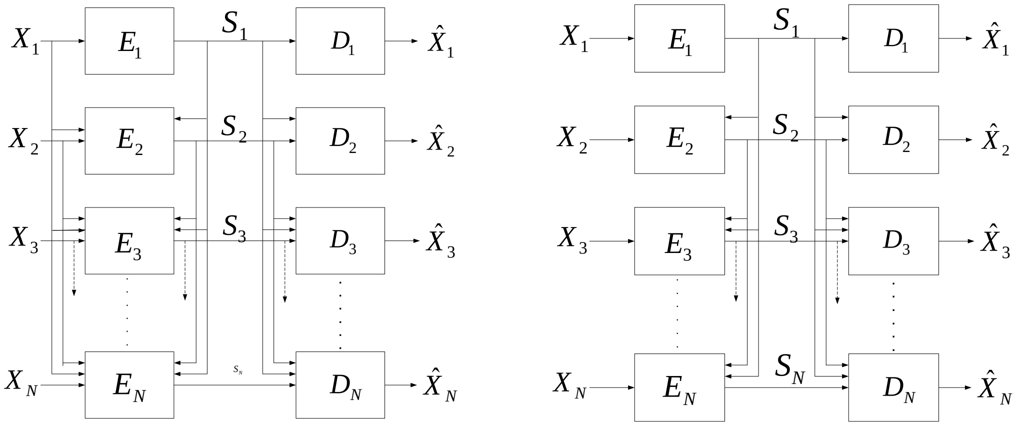

Figure 1.3: Causal video coding (left) versus Predictive video coding (right)

From their respective definitions, it follows that CVC includes PVC as a special case where the original video frames $X_1, \cdots, X_{k-1}$ are discarded at the encoder. It is expected that future video coding standards will continue to fall into the CVC model shown in Figure 1.3. This motivates us to investigate and gain deep insights into CVC to provide some design guidance for a future video coding standard.

1.3 Motivation

Causal video coding is investigated from an information theoretic point of view in this thesis. The first part, which is the main part of the thesis, studies rate-distortion theory for CVC, including characterization, computation algorithm, comparison and analysis; the code design of CVC is considered from an information theoretic perspective in the second part of this thesis. The motivation behind the research to be presented will be explained in this section.

Rate-distortion Theory for Causal Video Coding:

Ever since digital video was invented, video compression has become an essential part in all related applications, such as terrestrial broadcast, cable TV, video conferencing, and mobile communications, because of the enormous volume of video data. To effectively compress the video data with required video quality, a tradeoff between transmission bandwidth(rate), video quality(distortion), and computation cost need to be taken into consideration. In practice, video compression is usually categorized as lossy data compression. The theory that studies the theoretical limits for lossy data compression is called rate-distortion theory, and the fundamental tradeoff in video compression is its entire rate-distortion performance. In the case of CVC, its compression performance is analyzed by its minimum total rate $R_c^*(D_1, \cdots, D_N)$ required to achieve a given distortion (quality) level $D_1, \cdots, D_N > 0$.

It is expected that a future video coding standard will continue to fall into the CVC model shown in Figure 1.3. To provide some design guidance for a future video coding standard, in the first part of this thesis, we aim at investigating from an information theoretic point of view how each frame in the causal model should be encoded so that collectively the total rate is minimized subject to a given distortion (quality) level $D_1, \cdots, D_N \geq 0$.

To this end, the following questions naturally arise:

Q1 How shall we analytically characterize $R_c^*(D_1, \cdots, D_N)$?

After the existence of analytical characterization is proved, it normally requires future effort in finding an efficient way to compute the characterization, i.e., how each frame in the causal model should be encoded so that $R_c^*(D_1, \cdots, D_N)$ is achieved, such that it may have impact on practical video coding. Therefore, we are led to

Q2 Is there any algorithm to numerically compute $R_c^*(D_1, \cdots, D_N)$ such that this algorithm converges to an optimal solution that achieves $R_c^*(D_1, \cdots, D_N)$?

If we are lucky enough to find one, then with the help of this algorithm, the question is

Q3 What insights can we gain into CVC to guide practical video coding design?

In this thesis, we provide answers to all the above questions. A brief summary is as follows.

Characterization: Question Q1 is settled in three cases:

1. The vector source $(X_1, \cdots, X_N)$ is jointly stationary and totally ergodic[1] across samples (pixels);

2. The vector source $(X_1, \cdots, X_N)$ is general stationary ergodic[2] across samples (pixels);

[1] A vector source $(X_1, X_2, \cdots, X_N) = \{(X_1(i), X_2(i), \cdots, X_N(i))\}_{i=1}^{\infty}$ is said to be jointly stationary and totally ergodic if as a single process over the alphabet $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$, $\{(X_1(i), X_2(i), \cdots, X_N(i))\}_{i=1}^{\infty}$ is stationary and totally ergodic.

[2] A vector source $(X_1, X_2, \cdots, X_N) = \{(X_1(i), X_2(i), \cdots, X_N(i))\}_{i=1}^{\infty}$ is said to be general stationary ergodic if as a single process over the alphabet $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$, $\{(X_1(i), X_2(i), \cdots, X_N(i))\}_{i=1}^{\infty}$ is stationary and ergodic.

3. The vector source $(X_1, X_2, \cdots, X_N)$ is independent and identically distributed (IID)[3].

As we know, information-theoretic characterizations are usually critical to analyze the best coding efficiency that a compression method can achieve for a given information source or system. In the case of CVC, the best coding efficiency that an order-n causal video code (The formal definition of an order-n causal video code will be given in Chapter 3.) can achieve is characterized by the total rate-distortion function $R_{c,n}(D_1, \cdots, D_N)$, and the theoretic limit $R_c^*(D_1, \cdots, D_N)$ is equal to the infimum of $R_{c,n}(D_1, \cdots, D_N)$ over all n, where $R_{c,n}(D_1, \cdots, D_N)$ itself is given by the minimum of an information quantity over a set of auxiliary random variables. Having $R_c^*(D_1, \cdots, D_N)$, we are now clear about our objective function to analyze the rate-distortion performance.

The three cases, under which Q1 is settled, are different assumptions of sources in CVC. It is worthwhile to mention the motivation behind deriving the single-letter characterization of $R_c^*(D_1, \cdots, D_N)$ when $(X_1, X_2, \cdots, X_N)$ is IID (Case 3). Along the line of classic information theoretic research, one of the main motivations for deriving single-letter characterization is the hope that one day it can be computed by algorithms.

Computation Algorithm: Rate-distortion function characterization is what current classic information-theoretic research focus on, and the answer to Q1 is probably the best result one could hope for in terms of analytically characterizing $R_c^*(D_1, \cdots, D_N)$. However, its impact on practical video coding will be limited if the optimization problem involved can not be solved by an effective algorithm. To a large extent, this is also true even if $R_c^*(D_1, \cdots, D_N)$ admits a single-letter characterization, and true for many other multi-user information theoretic problems. Having single-letter characterization is only a tiny

[3] A vector source $(X_1, X_2, \cdots, X_N) = \{(X_1(i), X_2(i), \cdots, X_N(i))\}_{i=1}^{\infty}$ is said to be IID if as a single process over the alphabet $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$, $\{(X_1(i), X_2(i), \cdots, X_N(i))\}_{i=1}^{\infty}$ is IID. Note that the common joint distribution of each sample $(X_1(i), X_2(i), \cdots, X_N(i))$, $i \geq 1$, can be arbitrary even when the vector source $(X_1, X_2, \cdots, X_N)$ is IID.

step in the long path of pushing information theoretic results to have impact on practice. Even the single-letter characterization is successful, it normally does not offer any clue on how to compute the single-letter characterization. If the single-letter characterization can not be computed by algorithms with global convergence, its insights will hardly be gained, and its impact on practice will be further limited. In this thesis, we will develop an iterative algorithm to compute $R_{c,n}(D_1, \cdots, D_N)$, and establish its convergence to the global minimum.

The iterative algorithm is proposed for three purposes : first, it allows us to do numerical calculations; second, the global convergence of this algorithm provides a completely different approach to establish a single-letter characterization of $R_c^*(D_1, .., D_N)$ when the N sources are IID; and third, it allows us to do comparisons and gain deep insights into $R_c^*(D_1, .., D_N)$.

Comparisons and Analysis: To gain deep insights into CVC, equipped with our iterative algorithm, we first compare $R_c^*(D_1, .., D_N)$ among different values of N. Conventional wisdom extrapolates that the number of frames need to be coded and transmitted is always directly proportionate to the amount of data after encoding has to be sent. The iterative algorithm turns out that it is not always the case, and it indeed allows us to establish a somewhat surprising more and less coding theorem—under some conditions on source frames and distortion, the more frames need to be coded and transmitted, the less amount of data after encoding has to be sent!

All MPEG-series and H-series video coding standards [32], [49] proposed so far fall into PVC. From an information theoretic point of view, it is interesting to compare CVC and PVC which is widely used in practice. However, due to the limited information theoretic understanding of PVC (e.g., the information theoretic performance characterization of PVC is still unknown in general, let alone any algorithm to actually compute it.), the comparison between PVC and CVC is difficult and technically challenging, so it is better for us to address this comparison in a separate chapter (Chapter 5). Note that such comparison is of particular interest in the practice of video compression: if the rate-distortion performance of causal video coding strictly outperforms PVC, then it implies a possible paradigm shift from PVC to CVC in the design of future video coding systems and standards.

PVC is a coding methodology; there are many ways to do predictive coding, yielding different predictive codecs (or video coding standards in the context of video coding), some of which will be better than others. By using a technique called soft decision quantization [49], [47], [48], it has been demonstrated in a series of papers [49], [50], [46] that the greedy coding method (The formal definition of greedy coding method will be given in Section 4.4.) offers significant gains (ranging from 10% to 30% rate reduction at the same quality) over the respective reference codecs of these standards. The greedy coding method is a special form of PVC. It just happens that by applying the greedy coding method to different kinds of video coding standards, one always yields a codec better than the respective reference codec. All existing video coding standards and proposals still progress along the greedy coding method. Thus from a practical point of view, it is instructive to compare CVC and greedy coding.

The rate-distortion function characterization, computation algorithm, and all insights gained on CVC would provide a visionary guidance to future video coding standards.

Optimum Fixed-Rate Causal Scalar Quantization Design for Causal Video Coding:

To provide more specific guidelines and directions to further improve the rate-distortion performance of current video coding standard, we look at how to design specific codes for CVC following the line of the first part in this thesis by modeling each frame as a stationary source. As a starting point, we shall focus on the optimum fixed-rate scalar quantizer design for CVC (CSQ, which will be defined in Chapter 6.) by extending the classic Lloyd-Max algorithm for a single source to this multiple sources case. With the help of our proposed iterative algorithm for designing optimum fixed-rate CSQ, we demonstrate a significant quality improvement (distortion reduction) in comparison with the optimum fixed-rate predictive scalar quantizer for PVC. This result makes CVC more attractive, and suggests that one could explore to further improve the rate-distortion performance of current video coding standard on how quantization should be performed conditionally given previous frames and coded frames.

1.4 Organization and Main Contributions

The rest of the thesis is organized as follows. In Chapter 2, we introduce some basic concepts on information sources and rate-distortion theory. Some related works to this thesis are also talked there. Starting with a jointly stationary and totally ergodic vector source $(X_1, \cdots, X_N)$, in Chapter 3, we first analytically characterize the achievable rate region $\mathcal{R}_c^*$ and show that $R_c^*(D_1, \cdots, D_N)$ is equal to the infimum of the nth order total rate-distortion function $R_{c,n}(D_1, \cdots, D_N)$ over all n, where $R_{c,n}(D_1, \cdots, D_N)$ itself is given by the minimum of an information quantity over a set of auxiliary random variables. We further show that the analytical characterizations of $\mathcal{R}_c^*$ and $R_c^*(D_1, \cdots, D_N)$ remain valid for general stationary ergodic sources. As a by-product, an extended Markov lemma is established for correlated ergodic sources. Next, we develop an iterative algorithm in Chapter 4 to calculate $R_{c,n}(D_1, \cdots, D_N)$, and further show that this algorithm converges to an optimal solution that achieves $R_{c,n}(D_1, \cdots, D_N)$. The global convergence of the algorithm enables us to establish a single-letter characterization of $R_c^*(D_1, \cdots, D_N)$ in the case where the vector source $(X_1, X_2, \cdots, X_N)$ is IID, by comparing $R_{c,n}(D_1, \cdots, D_N)$

with $R_{c,1}(D_1, \cdots, D_N)$ through a novel application of the algorithm. With the help of the algorithm, we further demonstrate in Chapter 4 a somewhat surprising result dubbed the more and less coding theorem—under some conditions on source frames and distortion, the more frames need to be encoded and transmitted, the less amount of data after encoding has to be actually sent. The algorithm also gives an optimal solution for allocating bits to different frames. It is also shown that $R_c^*(D_1, \cdots, D_N)$ is in general much smaller than the total rate $R_g(D_1, \cdots, D_N)$ offered by the traditional greedy coding method by which each frame is encoded in a local optimum manner based on all information available to the encoder of the frame. We address the comparison between CVC and PVC in a separate chapter – Chapter 5, in which we start with the characterization of the minimum total rate $R_p^*(D_1, \cdots, D_N)$ of PVC required to achieve a given distortion level $D_1, \cdots, D_N$ for a general stationary ergodic vector source $(X_1, \cdots, X_N)$. It is then shown that if $X_1, \cdots, X_N$ are general stationary and ergodic, and also form a (first-order) Markov chain in the indicated order, then $R_c^*(D_1, \cdots, D_N) = R_p^*(D_1, \cdots, D_N)$. Moreover, we prove that in the case where $N = 3$, if X_1, X_2, X_3 do not form a (first-order) Markov chain, then $R_{c,n}(D_1, D_2, D_3) < R_{p,n}(D_1, D_2, D_3)$ for any finite $n \geq 1$ under mild conditions. A single-letter characterization of $R_p^*(D_1, D_2, D_3)$ is derived for an IID vector source (X_1, X_2, X_3) where X_1 and X_2 are independent, and we demonstrate the existence of such X_1, X_2, X_3 for which $R_c^*(D_1, D_2, D_3) < R_p^*(D_1, D_2, D_3)$. We consider the code design problem of CVC from an information theoretic perspective in Chapter 6 focusing on the optimum fixed-rate scalar quantization design, and finally in Chapter 7, we summarize the thesis and offer some future research directions.

To conclude this chapter, the main contributions of this thesis are listed as follows.

1. Propose an iterative algorithm with global convergence to calculate $R_{c,n}(D_1, \cdots, D_N)$.
   Given a distortion (quality) level $D_1, \cdots, D_N \geq 0$, we have proposed an algorithm to

compute $R_{c,n}(D_1, \cdots, D_N)$ by iteratively updating transition probability and probability functions until a stationary point is reached. We further show that from any initial point satisfying some mild condition, this algorithm converges to an optimal solution that achieves $R_{c,n}(D_1, \cdots, D_N)$. Finding optimal solutions to $R_{c,n}(D_1, \cdots, D_N)$ is actually a non-convex optimization problem. It is therefore kind of surprising to see the global convergence of our proposed iterative algorithm. Although there are many other ways to derive iterative procedures, however, it is not clear whether their global convergence can be guaranteed. Having algorithms with global convergence is important to not only numerical computation itself, but also single-letter characterization of performance.

2. Find a new approach (computational approach) to establish the single-letter characterization. One of the purposes in this thesis is to demonstrate for the first time that single-letter characterization of performance can also be established in a computational approach via algorithms with global convergence. In the computational approach, we first show that $R_{c,n}(D_1, \cdots, D_N) = R_{c,1}(D_1, \cdots, D_N)$ for any $n > 1$ by running our iterative algorithm, and then the single-letter characterization $R_c^*(D_1, \cdots, D_N) = R_{c,1}(D_1, \cdots, D_N)$ is implied following from the definition of $R_c^*(D_1, \cdots, D_N)$. On the other hand, in the classic approach, the converse proof is quite involved; coming up with auxiliary random variables with right Markov chain conditions is always challenging and sometimes seems impossible. Even the single-letter characterization is successful, the classic approach normally does not offer any clue on how to compute the single-letter characterization. In this sense, the computational approach goes beyond what the classic approach does—once it is successful, the computational approach not only establishes single-letter characterization, but also computes it numerically and offers additional insights into code design.

3. Demonstrate a somewhat surprising result dubbed the more and less coding theorem. The more and less coding theorem is really counter intuitive. It says that whenever some mild conditions on source frames and distortion are met, the more source frames need to be encoded and transmitted, the less amount of data after encoding has to be actually sent! If the cost of data transmission is proportional to the transmitted data volume, this translates literally into a scenario where the more frames you download, the less you would pay.

4. For the first time in the literature, obtain the single-letter characterization of $R_p^*(D_1, D_2, D_3)$ for a constructed IID vector source. Despite the fact that PVC is widely used in practice, the single-letter characterization of $R_p^*(D_1, \cdots, D_N)$ in the usual information theoretic sense, if any, is unknown in general, let alone any algorithm to actually compute it. In this thesis, a single-letter characterization of $R_p^*(D_1, D_2, D_3)$ is derived for an IID vector source (X_1, X_2, X_3) where X_1 and X_2 are independent, for the first time, which is critical to comparing the rate-distortion performance between CVC and PVC, and demonstrating the existence of such X_1, X_2, X_3 for which $R_c^*(D_1, D_2, D_3) < R_p^*(D_1, D_2, D_3)$.

5. Demonstrate the existence of such sources for which $R_c^*(D_1, D_2, D_3) < R_p^*(D_1, D_2, D_3)$. The main difficulty to compare $R_c^*(D_1, D_2, D_3)$ with $R_p^*(D_1, D_2, D_3)$ lies in the fact that when sources do not form a (first-order) Markov chain, there is no known algorithm to compute $R_p^*(D_1, D_2, D_3)$ even though it has a single-letter expression. To circumvent this problem, we instead look at the computation of $R_c^*(D_1, D_2, D_3)$. Our strategy is to show that any predictive code cannot be a stationary point in computing $R_c^*(D_1, D_2, D_3)$ by using the iterative algorithm. It is indeed an example to show the superiority of computational approach to classic approach, in the sense that

the strict inequality $R_c^*(D_1, D_2, D_3) < R_p^*(D_1, D_2, D_3)$ is shown by computational approach where the solution is still unknown by the classic approach.

6. For the first time in the literature, propose an algorithm for designing optimum fixed-rate causal scalar quantizers (CSQ). We first put forth a concept called causal scalar quantization, and then investigate how to design CSQ. By extending the classic Lloyd-Max algorithm for a single source to this multiple sources case, we propose an algorithm for designing optimum fixed-rate CSQ to minimize the total distortion among all sources. The proposed algorithm converges in the sense that the total distortion cost is monotonically decreasing until a stationary point is reached. Simulation results show that in comparison with fixed-rate predictive scalar quantization (PSQ), fixed-rate CSQ offers as large as 16% quality improvement (distortion reduction). Since PVC is what all previous and current video coding standards fall into, the rate-distortion performance gain of fixed-rate CSQ for CVC over fixed-rate PSQ for PVC would be instructive to practice.

7. We generated the Markov lemma to correlated ergodic sources. Since the vector source $(X_1, \cdots, X_N)$ now is not IID, but stationary and ergodic, the Markov lemma in its simple form as expressed in [11, Lemma 15.8.1, Chapter 15] is not valid any more. In this thesis, we extend the Markov lemma to be more general, from IID vector source to stationary and ergodic sources, which itself is useful for other multiterminal problems with stationary ergodic sources.

1.5 Notations and Acronyms

We model each frame X_k itself as a source $X_k = \{X_k(i)\}_{i=1}^{\infty}$ taking values in a finite alphabet $\mathcal{X}_k$. Together, the N frames then form a vector source $(X_1, X_2, \cdots, X_N) =$

$\{X_1(i), X_2(i), \cdots, X_N(i)\}_{i=1}^{\infty}$ taking values in the product alphabet $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$. The sources $X_1, X_2, \cdots, X_N$ are said to be (first-order) Markov if for any $1 < j \leq N$, X_j is the output of a memoryless channel in response to input X_{j-1}; in this case, we say $X_1 \to X_2 \to \cdots \to X_N$ forms a Markov chain. Let $\hat{X}_k = \{\hat{X}_k(i)\}_{i=1}^{\infty}$ denote the reconstruction of $X_k = \{X_k(i)\}_{i=1}^{\infty}$ drawn from a finite reproduction alphabet $\hat{\mathcal{X}}_k$. The distortion between $\hat{X}_k$ and X_k is measured by a single-letter distortion measure $d_k : \mathcal{X}_k \times \hat{\mathcal{X}}_k \to [0, \infty)$. Without loss of generality, we shall assume that

$$\max_{x \in \mathcal{X}_k} \min_{\hat{x} \in \hat{\mathcal{X}}_k} d_k(x, \hat{x}) = 0$$

for any $k = 1, 2, \cdots, N$. For convenience, we write $\{X_k(i)\}_{i=1}^{n}$ simply as $X_k(1;n)$ for any k and $n \geq 1$, and sometimes write a sequence $x_m x_{m+1} \cdots x_n$ as x_m^n unless specified otherwise, where $m \leq n$ are two integers, and x_1^n as x^n or simply x if n is clear from the context. For any N dimensional vector $V = (V_1, V_2, \cdots, V_N)$, denote $(V_1, \cdots, V_{t-1})$ by V_t^-, and $(V_{t+1}, \cdots, V_N)$ by V_t^+. As such, by $X_k^-(1;n)$ we shall mean that $X_k^-(1;n) = (X_1(1;n), \cdots, X_{k-1}(1;n))$. A similar convention will apply to reconstruction sequences and other vectors.

Chapter 2

Preliminaries and Related Works

2.1 Information Sources

In source coding theory, an information source is a sequence of discrete time random processes $X = \{X_i\}_{i=1}^{\infty}$ ranging over a finite alphabet $\mathcal{A}$ defined on a probability space $(\mathcal{A}^{\infty}, \mathcal{F}, P_X)$, where $\mathcal{F}$ is a σ-field generated by subsets of $\mathcal{A}^{\infty}$. All data sequences that have to be compressed are assumed to be emitted from some information sources. The i-th order joint distribution of the source X is the (probability) measure $P_X^{(i)}$ on $\mathcal{A}^i$ defined by

$$P_X^{(i)}(x_1^i) = Pr\{X_1^i = x_1^i\}, x_1^i \in \mathcal{A}^i.$$

Definition 1 *(**Stationary Source**) A source $X = \{X_i\}_{i=1}^{\infty}$ is* stationary *if*

$$Pr\{X_1^L = x_1^L\} = Pr\{X_{j+1}^{j+L} = x_1^L\}$$

for all lengths L, all integers j, and all sequences $x_1^L \in \mathcal{A}^L$.

Definition 2 *(**Ergodic Source**) A source $X = \{X_i\}_{i=1}^{\infty}$ is* ergodic *if every P-measurable, invariant set of sequences B has either probability one or probability zero, i.e., $T^{-1}B = B$ implies $P(B) = 0$ or $P(B) = 1$, where T denotes the right-shift transformation on $\mathcal{A}^{\infty}$.*

Definition 3 *(**Totally Ergodic Source**) A source $X = \{X_i\}_{i=1}^{\infty}$ is* totally ergodic *if every P-measurable set of sequences B such that $T^m B = B$ for any integer m has either probability one or probability zero, where T^m denotes the right-shift transformation on $\mathcal{A}^\infty$ by m positions.*

Definition 4 *(**Shannon Entropy**) The* Shannon entropy $H(X_1)$ *of a discrete random variable X_1 with alphabet $\mathcal{A}$ and probability mass function $p : \mathcal{A} \to [0, 1]$ is defined by*

$$H(X_1) = -\sum_{a \in \mathcal{A}} p(a) \log p(a).$$

The entropy of a random variable is a measure of the amount of information required on the average to describe the random variable. If the base of logarithm is 2, then the entropy is measured in bits. If the base of logarithm is e, then the entropy is measured in nats. We use the convention that $0 \log 0 = 0$, which is consistent with the limit $\lim_{p \to 0} p \log p = 0$.

Definition 5 *(**Shannon Entropy Rate**) The* Shannon entropy Rate $H(X)$ *of a source $X = \{X_i\}_{i=1}^{\infty}$ with alphabet $\mathcal{A}$ is defined by*

$$H(X) = \lim_{n \to \infty} \frac{1}{n} H(X_1 X_2 \cdots X_n) \tag{2.1}$$

when the limit exists.

The limit (2.1) exists if X is stationary and ergodic.

Definition 6 *(**Mutual Information**) The* mutual information $I(X; Y)$ *of two discrete random variables X and Y with alphabet $\mathcal{A}$ and $\mathcal{B}$ respectively, joint probability mass function: $p : \mathcal{A} \times \mathcal{B} \to [0, 1]$, and marginal probability mass functions: $p_1 : \mathcal{A} \to [0, 1]$, and $p_2 : \mathcal{B} \to [0, 1]$ can be defined by*

$$I(X; Y) = \textstyle\sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p(a, b) \log \frac{p(a,b)}{p_1(a) p_2(b)}$$

Mutual information is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other.

Definition 7 ***(Relative Entropy)*** *The* relative entropy *or* Kullback Leibler distance $D(p||q)$ *between two probability mass functions* $p : \mathcal{A} \rightarrow [0,1]$ *and* $q : \mathcal{A} \rightarrow [0,1]$ *is defined by*

$$D(p||q) \quad = \quad \sum_{a \in \mathcal{A}} p(a) \log \frac{p(a)}{q(a)}$$

The *relative entropy* is always non-negative and is zero if and only if $p = q$.

The most commonly used finite alphabet sources are independent, identically distributed (IID) sources.

Definition 8 ***(I.I.D Source)*** *An alphabet* $\mathcal{A}$ *source* $X = \{X_i\}_{i=1}^{\infty}$ *is an* IID source *if there exits a probability function* $p : \mathcal{A} \rightarrow [0,1]$, *i.e.,* p *satisfies*

$$\sum_{a \in \mathcal{A}} p(a) = 1,$$

such that

$$Pr\{X_n = x_n | X_1^{n-1} = x_1^{n-1}\} = p(x_n)$$

for all $n \geq 1$ *and* $x^n \in \mathcal{A}^n$.

Typicality is an important tool to prove coding theorems. Next we review the definition of typicality and some basic properties [11], [17].

Definition 9 *A sequence* $x^n \in \mathcal{A}^n$ *is said to be* ϵ*-strongly* typical *with respect to a distribution* $p : \mathcal{A} \rightarrow [0,1]$ *if*

1) for all $a \in \mathcal{A}$ with $p(a) > 0$, we have

$$|\frac{1}{n}N(a|x^n) - p(a)| < \frac{\epsilon}{|\mathcal{A}|}; \tag{2.2}$$

2) and for all $a \in \mathcal{A}$ with $p(a) = 0$, $N(a|x^n) = 0$,
where $N(a|x^n)$ is the number of occurrences of the symbol a in the sequence x^n.

Definition 10 *A pair of sequences $(x^n, y^n) \in \mathcal{A}^n \times \mathcal{B}^n$ is said to be ϵ-strongly typical with respect to a distribution $p : \mathcal{A} \times \mathcal{B} \to [0, 1]$ if*
1) for all $(a, b) \in \mathcal{A} \times \mathcal{B}$ with $p(a, b) > 0$, we have

$$|\frac{1}{n}N(a, b|x^n, y^n) - p(a, b)| < \frac{\epsilon}{|\mathcal{A}||\mathcal{B}|}; \tag{2.3}$$

2) and for all $(a, b) \in \mathcal{A} \times \mathcal{B}$ with $p(a, b) = 0$, $N(a, b|x^n, y^n) = 0$,
where $N(a, b|x^n, y^n)$ is the number of occurrences of the symbol (a, b) in the sequence (x^n, y^n).

The set of all ϵ-strongly typical sequences $x^n \in \mathcal{A}^n$ with respect to $p : \mathcal{A} \to [0, 1]$ is denoted by $A_\epsilon^{*(n)}(X)$, and the set of all jointly ϵ-strongly typical sequences $(x^n, y^n) \in \mathcal{A}^n \times \mathcal{B}^n$ with respect to $p : \mathcal{A} \times \mathcal{B} \to [0, 1]$ is denoted by $A_\epsilon^{*(n)}(X, Y)$.

Lemma 1 *Let X_i be drawn IID $\sim p(x)$. Then $Pr(A_\epsilon^{*(n)}(X)) \to 1$ as $n \to \infty$.*

Lemma 2 *Let (X_i, Y_i) be drawn IID $\sim p(x, y)$. Then $Pr(A_\epsilon^{*(n)}(X, Y)) \to 1$ as $n \to \infty$.*

Lemma 3 *Let $Y_1, Y_2, ..., Y_n$ be drawn IID $\sim \prod p(y)$. For $x^n \in A_\epsilon^{*(n)}(X)$, the probability that $(x^n, Y^n) \in A_\epsilon^{*(n)}$ is bounded by*

$$2^{-n(I(X;Y)+\epsilon_1)} \leq Pr((x^n, Y^n) \in A_\epsilon^{*(n)}) \leq 2^{-n(I(X;Y)-\epsilon_1)} \tag{2.4}$$

where ϵ_1 goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

Lemma 4 ***(Markov-Lemma)*** *Let (X, Y, Z) form a Markov chain $X \rightarrow Y \rightarrow Z$, i.e., $p(x, y, z) = p(x, y)p(z|y)$. If for a given $(y^n, z^n) \in A_\epsilon^{*(n)}(Y, Z)$, X^n is drawn $\sim \prod_{i=1}^n p(x_i|y_i)$, then $Pr\{(X^n, y^n, z^n) \in A_\epsilon^{*(n)}(X, Y, Z)\} > 1 - \epsilon$ for n sufficiently large.*

Lemma 1 to Lemma 3 can be also generated to stationary and ergodic sources X^n and Y^n correspondingly, and an *extended Markov lemma* is established for stationary ergodic sources in Chapter 3.

2.2 Rate-distortion Theory

Shannon originated the studies on source coding with a fidelity criterion (later coined the term "rate-distortion theory" in [37]) in his early paper [36]. In rate-distortion theory, one is interested in determining the minimum amount of bits with which data generated by a given source can be compressed via a code from a class of codes, subject to a constraint on the distortion in reconstruction of the encoded data. In this section, we present an overview of the history and significant results of rate-distortion theory, including rate-distortion theory emphases on the point-to-point communication system and multiterminal system, respectively, and also the computation algorithm for calculating corresponding information quantities.

2.2.1 Classic Rate-distortion Theory

In the early days when there was less need for communication than today, point-to-point communication systems were the dominant. In a point-to-point system, there is only one source at the encoder and one receiver at the decoder.

The encoder describes the source sequence $X^n = \{X_i\}_{i=1}^n \in \mathcal{X}^n$ by a compact binary representation, and the decoder reconstructs X^n by an estimate $\hat{X}^n \in \hat{\mathcal{X}}^n$ which is not

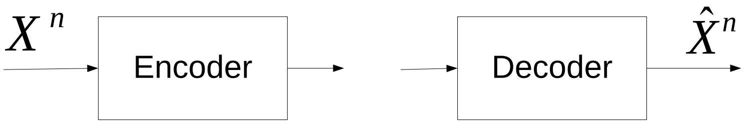

Figure 2.1: Source coding with one encoder and one decoder

identical to X^n from the binary representation, as illustrated in Figure 2.1.

To determine the minimum amount of bits that should be communicated over a noiseless channel, so that the source can be approximately reconstructed at the receiver without exceeding a given distortion, the rate-distortion function is introduced.

The simplest case to which we shall restrict attention for now, is that:

1. X_i is an IID source with distribution $p(x), x \in \mathcal{X}$.

2. The distortion between source sequence $x^n \in \mathcal{X}^n$ and reproduction sequence $\hat{x}^n \in \hat{\mathcal{X}}^n$ is defined by

$$d(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i),$$

where $d(\cdot,\cdot) : \mathcal{X} \times \hat{\mathcal{X}} \to \mathcal{R}^+$ is called a single-letter distortion measure. Examples of popular distortion measures are:

(a) *Hamming (probability of error) distortion.* The Hamming distortion is given by

$$d(x,\hat{x}) = \begin{cases} 0 & \text{if } x = \hat{x} \\ 1 & \text{if } x \neq \hat{x} \end{cases}, \tag{2.5}$$

which results in a probability of error distortion, since $Ed(X,\hat{X}) = Pr(X \neq \hat{X})$. This distortion measure is usually used for discrete alphabets.

(b) *Squared error distortion.* The squared error distortion is defined by

$$d(x,\hat{x}) = (x - \hat{x})^2, \tag{2.6}$$

which is the most popular distortion measure used for continuous alphabets.

3. Formally, we define an order-n rate-distortion code C_n for X^n consisting of an encoding function,

$$f_n : \mathcal{X}^n \to \{0,1\}^*,$$

and a decoding function,

$$g_n : \{0,1\}^* \to \hat{\mathcal{X}}^n,$$

where $\{0,1\}^*$ denotes a prefix set of all binary sequences of finite length. The encoded and reconstructed sequences of X^n are given respectively by $S = f_n(X^n)$ and $\hat{X}^n = g_n(S)$.

The distortion between X^n and $\hat{X}^n$ is given by

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

the average distortion D_x per symbol is then equal to

$$D_x \triangleq \frac{1}{n}\mathbf{E}\left[d(X^n, \hat{X}^n)\right],$$

and the average transmission rate R_x per symbol is defined by

$$R_x \triangleq \frac{1}{n}\mathbf{E}|S|,$$

where $|S|$ denotes the length of the binary sequence S.

Definition 11 *The rate-distortion pair (R, D) is said to be achievable if $\forall \epsilon > 0$, there exists an order-n rate-distortion code (f_n, g_n) for all sufficiently large n such that*

$$R_x \leq R + \epsilon \text{ and } D_x \leq D + \epsilon. \tag{2.7}$$

Definition 12 *The rate-distortion region for a source is the closure of the set of achievable rate-distortion pairs* (R, D)*.*

Definition 13 *The rate-distortion function* $R(D)$ *is the infimum of rates* R *such that* (R, D) *is in the rate distortion region of the source for a given distortion* D*.*

Shannon defined the rate-distortion function $R(D)$ of an IID source as follows.

Definition 14 *The rate-distortion function for an IID source* $\{X_k\}$ *with distribution* $p(x)$ *with respect to the single-letter fidelity criterion generated by* $d(\cdot, \cdot)$ *is defined by*

$$R(D) = \inf_{p(\hat{x}|x): \sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X}) \tag{2.8}$$

2.2.2 Multiterminal Rate-distortion Theory

In practical communications, it is always the case that not all the sources of interest are generated in one place. If a system has more than one source and at least one user, it is called a multiterminal system. Unlike the classic point-to-point system, the source of multiterminal source coding problems is produced at each of several physical distinct places (terminals). Due to the complexity and power constraints, the transmitters are often not allowed to communicate with each other. When the source generated at one terminal is independent of that generated at every other terminals, then the multiterminal system degenerates into many classic point-to-point systems that are operationally independent of each other. A non-trivial case is that a source produced at one terminal is correlated with some other terminals, which often is possible to save a considerable amount of coding rate by designing a more sophisticated multiterminal system than simply estimating each source individually. The discipline that treats such situations is known as multiterminal information theory.

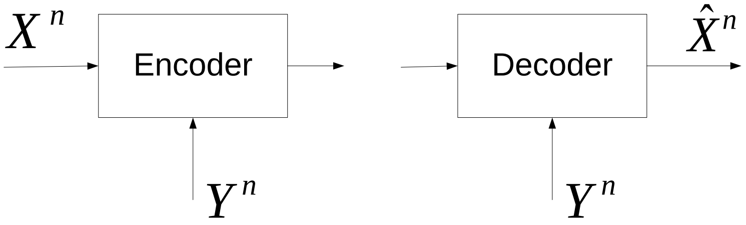

Figure 2.2: Source coding with side information at both encoder and decoder

A seminal research on multiterminal source coding theory was launched by Slepian and Wolf [40] in 1973, which considered separate lossless compression of two correlated sources, and showed that the sum of the coding rates required by the two encoders is no more than when the two sources are connected to both encoders. This work was subsequently extended to the lossy source coding case. In 1976, Wyner and Ziv [44] considered the situation in which one of the sources is fully available at the decoder, and it is used to make an estimation of the other source subject to a given distortion criterion. Unlike Slepian-Wolf coding, there is in general a rate loss[1] [20][2, Sec. 6.1.1] in Wyner-Ziv coding [44] compared to the lossy source coding problem when the encoder also has full information on the other source, which considered lossy source coding with side information[2] at both encoder and decoder.

We illustrate the case of side information available at both encoder and decoder in Fig 2.2. Let $(X^n, Y^n) = \{(X_i, Y_i)\}_{i=1}^{\infty}$ be a sequence of independent copies of an information source vector (X, Y) taking values from $\mathcal{X}$ and $\mathcal{Y}$ respectively. The source X is encoded with the help of side information Y. The decoder, which also access to Y, estimates X within a distortion constraint $D \geq 0$ for some distortion measure $d(\cdot, \cdot)$, then we use

[1] An exception happens when the source and side information are jointly Gaussian and the distortion measure is mean-squared error (MSE) [43].

[2] Besides the message at the encoder, and the received signal at the decoder, if any type of useful information is added to encoder or decoder's information, it will be called side information.

conditional rate-distortion function [20] to characterize the minimum achievable rate per symbol for (X, Y) under the constraint that X is recovered with distortion level no greater than D, which is

$$R_{X|Y}(D) \triangleq \min_{\hat{X}: Ed(X,\hat{X}) \leq D} I(X; \hat{X}|Y) \tag{2.9}$$

where the minimum is taken over all auxiliary random variables $\hat{X} \in \hat{\mathcal{X}}$ jointly distributed with (X, Y) such that $Ed(X, \hat{X}) \leq D$.

Soon after the seminal piece of research by Wyner and Ziv, a number of papers on multiterminal lossy source coding were presented, summarized by Berger [3] in 1977. Some celebrated works by Berger and Tung [4][41], Chang [8], Omura and Housewright [30], Shohara [39], and Sgarro [35] are included therein. Furthermore, some new multiterminal rate-distortion models and problems are proposed which laid foundations that supported new developments on both theoretical and practical fronts, including the multiple description problem [19][18] , the sucessive refinement problem [16], and the CEO problem [6][5][31], applications of which to image, audio and video coding are currently under development.

2.2.3 Iterative Computation Algorithms

Since 1970s, the area of computation of rate-distortion functions was rejuvenated, which deepened the understanding of the development and research insights.

In 1972, Blahut [7] found a numerical algorithm for calculating rate-distortion functions. The algorithm is an elegant iterative technique for numerically obtaining $R(D)$ defined in (2.8) of arbitrary finite input/output alphabet sources. The iterations are between marginal distribution $q_{\hat{X}}$ and transition probability $p_{\hat{X}|X}$ to minimizes the mutual information in (2.8) subject to the distortion constraint D. Specifically, we first rewrite (2.8) as a double minimization, that is

$$
\begin{aligned}
R(D) &= \min_{r(\hat{x})} \min_{p(\hat{x}|x):\sum p(x)p(\hat{x}|x)d(x,\hat{x})\le D} \sum_{x}\sum_{\hat{x}} p(x)p(\hat{x}|x)\log\frac{p(\hat{x}|x)}{r(\hat{x})} \\
&= \min_{p(x)p(\hat{x}|x)\in\mathcal{A}} \min_{p(x)r(\hat{x})\in\mathcal{B}} D(p(x)p(\hat{x}|x)||p(x)r(\hat{x})). \qquad (2.10)
\end{aligned}
$$

Since both of the two alternating sets– one is the set $\mathcal{A}$ of all joint distributions with marginal $p(x)$ satisfying the distortion constraints, and the other one is the set $\mathcal{B}$ of product distributions $p(x)r(\hat{x})$ with arbitrary $r(\hat{x})$ – are convex and the distance measure between $\mathcal{A}$ and $\mathcal{B}$ is the relative entropy, we can use the method of Lagrange multipliers to solve the minimization problem above with a choice of λ and an initial output distribution $r(\hat{x})$ such that $r(\hat{x}) > 0$ for all $\hat{x} \in \hat{\mathcal{X}}$, and obtain

$$
p(\hat{x}|x) = \frac{r(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}} r(\hat{x})e^{-\lambda d(x,\hat{x})}}. \qquad (2.11)
$$

For this $p(\hat{x}|x)$, we then calculate the output distribution $r(\hat{x})$ as

$$
r(\hat{x}) = \sum_{x} p(x)p(\hat{x}|x). \qquad (2.12)
$$

This $r(\hat{x})$ is used as the starting point of the next iteration. Each step in the iteration, minimizing over $p(\hat{x}|x)$ and $r(\hat{x})$ reduces the right-hand-side of (2.10). Thus there is a limit, and the limit has been shown to be $R(D)$ by Csiszar [12] in 1974, which proved the global convergence[3] of Blahut's algorithm, i.e., starting from *any* initial probability distribution $p_{\hat{X}}$ over the reproduction alphabet that has only positive components, the algorithm always terminates at one single optimal point $(D, R(D))$. In the same year, Blahut [7] and Arimoto [1] found an analogous algorithm to compute the capacity of channels independently. More general max-max and min-min alternating optimization algorithms are developed by

[3] An iterative algorithm is called convergent if the corresponding objective function value in each iteration step converges for given initial points.

Csiszar and Tusnady [13] in 1984. The alternating minimization procedure of Csiszar and Tusnady can be specialized to many other situations as well, including the EM algorithm [14] and the algorithm for finding the log-optimal portfolio for a stock market [10].

2.3 Conventional Optimum Fixed-Rate Scalar Quantization Design

Rate-distortion theory can be applied to both discrete and continuous random variables. Since a continuous source can not be reproduced exactly using a finite rate code, to represent such a source by using a finite number of bits, the problem is to find the best possible representation for any given rate, that is, to design optimum quantization in the sense that the expected distortion between the source and output representation is minimized for a given rate.

The simplest quantization is scalar quantization which is the process of using a quantization function to map a scalar (one-dimensional) input value to a scalar output value. An L-level scalar quantizer ϕ is a mapping from the real line $\mathcal{R}$ onto a finite set consisting of L reproduction levels $y_0 < y_1 < \cdots < y_{L-1}$:

$$x \in \mathcal{R} \to \phi(x) \in \mathcal{Y} \overset{\Delta}{=} \{y_0, y_1, \cdots, y_{L-1}\}. \tag{2.13}$$

Let

$$C_i = \{x \in \mathcal{R} : \phi(x) = y_i\}, 0 \leq i \leq L-1 \tag{2.14}$$

Given a real random variable X, the distortion resulting from quantizing X by ϕ is

$$D_\phi = \mathbf{E}[d(X, \phi(X))] \tag{2.15}$$

Fix L, an optimum fixed-rate scalar quantizer ϕ is uniquely determined by $\{(C_i, y_i)\}_{i=0}^{L-1}$ such that D_ϕ is minimized among all fixed-date scalar quantizers with rate $R = \lceil \log L \rceil$.

To design optimum scalar quantizers, Lloyd [24] and Max [28] independently proposed an algorithm which will converge to a local minimum of the distortion, which is called Lloyd-Max algorithm: starting with a set of initial reproduction points $\mathcal{Y}^{(0)}$, we first find the optimal set of classifications $C_i^{(0)}, 0 \leq i \leq L-1$, for each reproduction point in $\mathcal{Y}^{(0)}$, and then find the new optimal reproduction points $\mathcal{Y}^{(1)}$ for every classification set $C_i^{(0)}, 0 \leq i \leq L-1$. Repeating the above iterations, the expected distortion is decreased at each step in the algorithm until it converges to a local minimum of D_ϕ.

2.4 Related Works

When $N = 2$, the causal coding model is the same as the sequential coding[4] of correlated source proposed by Viswanathan and Berger in [42], in which they analytically characterize the achievable rate region and minimum total rate in the usual information theoretic sense. However, how to determine the bits that have to be allocated at different frames to minimize the total rate remains open, and it is pointed out as one of the future works in [42]. Recently, some works of Nan and Prakash [26], [25], and [27] investigate the sequential coding of correlated sources with different configurations of encoding and/or decoding frame delay to see how frame delay will have impact on the performance; its focus is only on the performance characterization for the IID case. The overlap with our work lies in the characterization of achievable rate region and minimum total rate of 3-frame causal coding, which is presented in the work of Ma and Ishwar [26] with an informal sketch proof by following the classic approach, and which is contained as a special case in our Chapter 3. However, when $N > 2$, which is a typical case in MPEG-series and H-series

[4] The name of sequential coding was used in [42] to refer to a special video coding paradigm where the encoder for frame X_k, $k > 1$, can only use the previous frame X_{k-1} as a helper and the corresponding decoder uses only the previous encoded frame S_{k-1} and reconstructed frame $\hat{X}_{k-1}$ as a helper.

video coding, the causal coding model considered here is quite different from sequential coding. In a special case where all frames are identical, which rarely happens in practical video coding, the CVC model is reduced to the successive refinement setting considered in [16]. Notwithstanding, when frames are not identical, CVC is drastically different from successive refinement even though the decoding structure looks similar in both cases.

Chapter 3

Information Theoretic Characterization

3.1 Problem Formulation and Definitions

Recall the CVC model, as shown in Figure 3.1. The encoder of frame $X_k, k = 1, \cdots, N$, can use all previous frames X_j, in addition to previously encoded frames S_j, $j = 1, 2, \cdots, k-1$ as side information, and the corresponding decoder can use only all previously encoded frames. Each video frame X_k itself is a discrete stationary source $X_k = \{X_k(i)\}_{i=1}^{\infty}$ taking values in a finite alphabet $\mathcal{X}_k$—that is, the N frames can be regarded as N correlated sources taking values in the product alphabet $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$.

Formally, we define an order-n causal video code C_n for $X_1, \cdots, X_N$ by using N encoder and decoder pairs as follows[1]:

[1] It is worthwhile to point out that as far as CVC alone is concerned, there is no need to explicitly list previous encoded frames S_k^- as inputs to the encoder for the current frame X_k in both the CVC diagram shown in Figure 1.3 and the formal definition of causal video code given here, and all results and their respective derivations presented in the thesis remain the same. The reason for us to explicitly list S_k^-

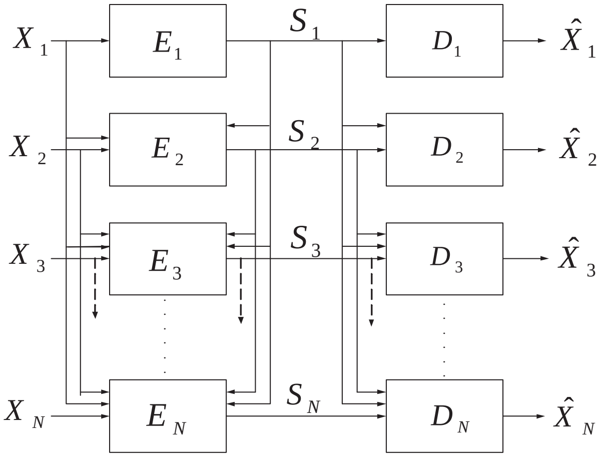

Figure 3.1: Causal video coding model

1. For X_1, an encoder of order n is defined by a function f_1 from $\mathcal{X}_1^n$ to $\{0,1\}^*$, the set of all binary sequences of finite length, satisfying the property that the range of f_1 is a prefix set, and a decoder of order n is defined by a function

$$g_1 : \{0,1\}^* \to \hat{\mathcal{X}}_1^n.$$

The encoded and reconstructed sequences of $X_1(1;n)$ are given respectively by $S_1 = f_1(X_1(1;n))$ and $\hat{X}_1(1;n) = g_1(S_1)$.

as inputs to the encoder for the current frame X_k is two-fold: (1) it makes the subsequent information quantities more transparent and intuitive—connecting those information quantities to the diagram with S_k^- linked to the respective encoder is easier than to that without S_k^- linked to the respective encoder—and (2) more importantly it gives us a simple, unified way to describe PVC in the context of CVC and contrast the two coding paradigms in our forthcoming work on the information theoretic performance comparison of PVC and CVC.

2. For X_k, $k = 2, \cdots, N$, an encoder of order n is defined by a function

$$\begin{aligned} f_k : \quad & \mathcal{X}_1^n \times \cdots \times \mathcal{X}_k^n \times \overbrace{\{0,1\}^* \times \cdots \times \{0,1\}^*}^{k-1 \text{ times}} \\ & \to \{0,1\}^* \end{aligned}$$

satisfying the property that the range of f_k given any $k-1$ binary sequences is a prefix set, and a decoder of order n is defined by a function

$$g_k : \overbrace{\{0,1\}^* \times \cdots \times \{0,1\}^*}^{k \text{ times}} \to \hat{\mathcal{X}}_k^n.$$

The encoded and reconstructed sequences of $X_k(1;n)$ are given respectively by $S_k = f_k(X_k^-(1;n), X_k(1;n), S_k^-)$ and $\hat{X}_k(1;n) = g_k(S_k^-, S_k)$.

For $k = 1, \cdots, N$, the distortion between $X_k(1;n)$ and $\hat{X}_k(1;n)$ is given by

$$d_k(X_k(1;n), \hat{X}_k(1;n)) = \sum_{i=1}^{n} d_k(X_k(i), \hat{X}_k(i))$$

the corresponding average distortion per symbol is then equal to

$$D_{xk} \triangleq \frac{1}{n} E\left[d_k(X_k(1;n), \hat{X}_k(1;n))\right]$$

and the average rate in bits per symbol of the kth encoder is

$$R_{xk} \triangleq \frac{1}{n} E|S_k|$$

where $|S_k|$ denotes the length of the binary sequence S_k. The performance of the order-n causal video code C_n is then measured by the N rate distortion pairs (R_{xk}, D_{xk}), $k = 1, \cdots, N$.

Definition 15 *Let $(R_1, \cdots, R_N)$ be a rate vector and $(D_1, \cdots, D_N)$ a distortion vector. The rate distortion pair vector $(R_1, \cdots, R_N, D_1, \cdots, D_N)$ is said to be achievable by CVC if $\forall \epsilon > 0$, there exists an order-n causal video code $\{(f_k, g_k)\}_{k=1}^N$ for all sufficiently large n such that*

$$R_{xk} \leq R_k + \epsilon \text{ and } D_{xk} \leq D_k + \epsilon \tag{3.1}$$

for $k = 1, \cdots, N$.

Let $\mathcal{R}_c^*$ denote the set of all rate distortion pair vectors $(R_1, \cdots, R_N, D_1, \cdots, D_N)$ achievable by CVC. From the above definition, it follows that $\mathcal{R}_c^*$ is a closed set in the $2N$-dimensional Euclidean space. As in the usual video compression applications, we are interested in the minimum total rate $R_c^*(D_1, \cdots, D_N)$ required to achieve the distortion level $(D_1, \cdots, D_N)$, which is defined by

$$\begin{gathered} R_c^*(D_1, \cdots, D_N) \triangleq \min\{R_1 + R_2 + \cdots + R_N : \\ (R_1, \cdots, R_N, D_1, \cdots, D_N) \in \mathcal{R}_c^*\} . \end{gathered}$$

One of our purposes in this thesis is to numerically compute, analytically characterize, and compare $R_c^*(D_1, \cdots, D_N)$ so that deep insights can be gained regarding how each frame should be encoded in order to have a minimum total rate.

3.2 Achievable Region and Minimum Total Rate

3.2.1 Totally Ergodic Sources

Suppose now that $(X_1, X_2, \cdots, X_N)$ is jointly stationary and totally ergodic across samples (pixels). Define $\mathcal{R}_{c,n}$ to be the region consisting of all rate distortion pair vectors $(R_1, \cdots, R_N, D_1, \cdots, D_N)$ for which there exist auxiliary random variables U_k, $k =$

$1, 2, \cdots, N-1$, and $\hat{X}_N(1; n)$ such that

$$\begin{aligned}
&R_1 \geq \frac{1}{n} I(X_1(1;n); U_1) \\
&R_k \geq \frac{1}{n} I(X_1(1;n), \cdots, X_k(1;n); U_k | U_k^-) \\
&k = 2, 3, \cdots, N-1 \\
&R_N \geq \frac{1}{n} I(X_N(1;n); \hat{X}_N(1;n) | U_N^-)
\end{aligned} \tag{3.2}$$

and the following requirements[2] are satisfied:

(R1) $\hat{X}_1(1;n) = g_1(U_1)$ for some deterministic function g_1,

(R2) $\hat{X}_k(1;n) = g_k(U_k^-, U_k)$ for some deterministic function g_k, $k = 2, \cdots, N-1$,

(R3) for any $1 \leq k \leq N$, $\frac{1}{n} E[d_k(X_k(1;n), \hat{X}_k(1;n))] \leq D_k$, and

(R4) the Markov chain conditions $U_k \to (X_k(1;n), X_k^-(1;n), U_k^-) \to X_k^+(1;n)$, $k = 1, \cdots, N-1$, and $X_N^-(1;n) \to (X_N(1;n), U_N^-) \to \hat{X}_N(1;n)$ are met.

In (3.2) and throughout the rest of the thesis, the notation I stands for mutual information or conditional mutual information (as the case may be) measured in bits, and the notation H stands for entropy or conditional entropy (as the case may be) measured in bits. Although there is no restriction on the size of the alphabet of each U_k in (3.2), one can show, by using the standard cardinality bound argument based on the Caratheodory theorem (see, for example, Appendix A of [42]), that the alphabet size of each U_k in (3.2) can be bounded. Let $\mathcal{R}'_c = \bigcup_{n=1}^{\infty} \mathcal{R}_{c,n}$. Denote its convex hull closure by $co(\mathcal{R}'_c)$. Then we have the following result.

[2] Throughout the thesis, $\hat{X}_k(1;n)$, $k = 1, 2, \cdots, N$, represents a random variable taking values over $\hat{\mathcal{X}}_k^n$, the n-fold product of the reproduction alphabet $\hat{\mathcal{X}}_k$; on the other hand, U_k, $k = 1, 2, \cdots, N-1$, represents a random variable taking values over an arbitrary finite alphabet.

Theorem 1 *For jointly stationary and totally ergodic sources $X_1, \cdots, X_N$, $\mathcal{R}_c^* = co(\mathcal{R}_c')$.*

The positive part of Theorem 1 (i.e., $co(\mathcal{R}_c') \subseteq \mathcal{R}_c^*$) will be proved in Appendix B by adopting a random coding argument similar to that for IID vector sources. Here we present the proof of the converse part (i.e., $\mathcal{R}_c^* \subseteq co(\mathcal{R}_c')$).

Proof of the converse part of Theorem 1: Pick any achievable rate distortion pair vector $(R_1, \cdots, R_N, D_1, \cdots, D_N) \in \mathcal{R}_c^*$. It follows from Definition 15 that for any $\epsilon > 0$, there exists an order-n causal video code $C_n = \{(f_k, g_k)\}_{k=1}^N$ for all sufficiently large n such that (3.1) holds. Let S_k and $\hat{X}_k(1;n)$ be the respective encoded frame of and reconstructed frame for $X_k(1;n)$ given by C_n. Let $U_k = S_k$, $k = 1, 2, \cdots, N-1$. It is easy to see that the Markov conditions $U_k \to (X_k(1;n), X_k^-(1;n), U_k^-) \to X_k^+(1;n)$, $k = 1, \cdots, N-1$, are satisfied. However, since S_N depends in general on $X_N^-(1;n)$ in addition to $X_N(1;n)$ and S_N^-, the random variables $X_N^-(1;n)$, $(X_N(1;n), S_N^-)$, and $\hat{X}_N(1;n)$ do not necessarily form a Markov chain in the indicted order. To overcome this problem, let q denote the conditional probability distribution of $\hat{X}_N(1;n)$ given $(X_N(1;n), S_N^-)$. Define a new random variable $\tilde{X}_N(1;n)$ which is the output of the channel q in response to the input $(X_N(1;n), S_N^-)$. Then it is easy to see that $(X_N(1;n), S_N^-, \tilde{X}_N(1;n))$ and $(X_N(1;n), S_N^-, \hat{X}_N(1;n))$ have the same distribution, and $X_N^-(1;n)$, $(X_N(1;n), S_N^-)$, and $\tilde{X}_N(1;n)$ form a Markov chain. This, together with (3.1), implies the following distortion upper bounds:

$$\frac{1}{n} E[d_k(X_k(1;n), \hat{X}_k(1;n))] \leq D_k + \epsilon \tag{3.3}$$

for any $k = 1, 2, \cdots, N-1$, and

$$\frac{1}{n} E[d_N(X_N(1;n), \tilde{X}_N(1;n))] \leq D_N + \epsilon \; . \tag{3.4}$$

Let us now verify rate lower bounds. In view of (3.1), we have

$$\begin{aligned} n(R_1+\epsilon) &\geq H(S_1) \\ &= I(X_1(1;n);S_1) \\ &= I(X_1(1;n);U_1) \end{aligned} \tag{3.5}$$

and for $k=2,\cdots,N-1$,

$$\begin{aligned} n(R_k+\epsilon) &\geq H(S_k|S_k^-) \\ &\overset{1)}{=} I(X_k^-(1;n),X_k(1;n);S_k|S_k^-) \\ &= I(X_k^-(1;n),X_k(1;n);U_k|U_k^-) \end{aligned} \tag{3.6}$$

where equality 1) is due to the fact that S_k is a function of $(X_k^-(1;n),X_k(1;n),S_k^-)$. For the last frame, we have

$$\begin{aligned} n(R_N+\epsilon) &\geq H(S_N|S_N^-) \\ &= H(S_N,\hat{X}_N(1;n)|S_N^-) \\ &\geq H(\hat{X}_N(1;n)|S_N^-) \\ &\geq I(X_N(1;n);\hat{X}_N(1;n)|S_N^-) \\ &= I(X_N(1;n);\tilde{X}_N(1;n)|U_N^-). \end{aligned} \tag{3.7}$$

With auxiliary random variables U_k, $k=1,\cdots,N-1$, and $\tilde{X}_N(1;n)$ defined above, it now follows from (3.3) to (3.7) and the desired Markov conditions that $(R_1+\epsilon,\cdots,R_N+\epsilon,D_1+\epsilon,\cdots,D_N+\epsilon)\in\mathcal{R}_{c,n}\subseteq\mathcal{R}'_c$. Letting $\epsilon\to 0$ yields $(R_1,\cdots,R_N,D_1,\cdots,D_N)\in co(\mathcal{R}'_c)$, which in turn implies $\mathcal{R}^*_c\subseteq co(\mathcal{R}'_c)$. This completes the proof of the converse part.

To determine $R_c^*(D_1, \cdots, D_N)$ in terms of information quantities, we define for each $n \geq 1$

$$\begin{aligned} & R_{c,n}(D_1, \cdots, D_N) \\ & \quad \stackrel{\Delta}{=} \frac{1}{n} \min_{\{\hat{X}_k(1;n)\}_{k=1}^N} [I(X_1(1;n); \hat{X}_1(1;n)) + \\ & \quad \sum_{t=2}^{N-1} I(X_1(1;n), \cdots, X_t(1;n); \hat{X}_t(1;n) | \hat{X}_t^-(1;n)) + \\ & \quad I(X_N(1;n); \hat{X}_N(1;n) | \hat{X}_N^-(1;n))] \end{aligned} \tag{3.8}$$

where the minimum is taken over all auxiliary random vectors $\hat{X}_k(1;n)$, $k = 1, 2, \cdots, N$, satisfying the following two requirements

(R5) for any $1 \leq j \leq N$, $\frac{1}{n} E d_j(X_j(1;n), \hat{X}_j(1;n)) \leq D_j$, and

(R6) the Markov chains $\hat{X}_k(1;n) \rightarrow (X_k(1;n), X_k^-(1;n), \hat{X}_k^-(1;n)) \rightarrow X_k^+(1;n)$, $k = 1, \cdots, N-1$, and $X_N^-(1;n) \rightarrow (X_N(1;n), \hat{X}_N^-(1;n)) \rightarrow \hat{X}_N(1;n)$ hold.

We further define

$$R_c(D_1, \cdots, D_N) \stackrel{\Delta}{=} \inf\{R_{c,n}(D_1, \cdots, D_N) : n \geq 1\}. \tag{3.9}$$

Then we have the following result.

Theorem 2 *For jointly stationary and totally ergodic sources $X_1, \cdots, X_N$,*

$$R_c^*(D_1, \cdots, D_N) = R_c(D_1, \cdots, D_N)$$

for any distortion level $D_1 > 0, \cdots, D_N > 0$.

To prove Theorem 2, we need the following lemma, which is also interesting on its own right.

Lemma 5 *The function $R_c(D_1, \cdots, D_N)$ is convex and hence continuous over the open region $D_1 > 0, \cdots, D_N > 0$.*

Proof of Lemma 5: Fix $D_1 \geq 0, \cdots, D_N \geq 0$. In view of the definition given in (3.8), it is not hard to show that the sequence $\{nR_{c,n}(D_1, \cdots, D_N)\}$ is sub-additive, that is,

$$\begin{aligned} & (n+m)R_{c,n+m}(D_1, \cdots, D_N) \\ \leq \; & nR_{c,n}(D_1, \cdots, D_N) + mR_{c,m}(D_1, \cdots, D_N) \end{aligned}$$

for any n and m. As such, $R_c(D_1, \cdots, D_N)$ can also be expressed as

$$R_c(D_1, \cdots, D_N) = \lim_{n \to \infty} R_{c,n}(D_1, \cdots, D_N). \tag{3.10}$$

Next we derive an equivalent expression for $R_{c,n}(D_1, \cdots,$
$D_N)$. Define

$$\begin{aligned} \tilde{R}_{c,n}(D_1, \cdots, D_N) \overset{\Delta}{=} & \inf\{R_1 + \cdots + R_N : \\ & (R_1, \cdots, R_N, D_1, \cdots, D_n) \in \mathcal{R}_{c,n}\}. \end{aligned}$$

That is,

$$\begin{aligned} & \tilde{R}_{c,n}(D_1, \cdots, D_N) \\ = \; & \frac{1}{n} \inf[I(X_1(1;n); U_1) + \sum_{t=2}^{N-1} I(X_1(1;n), \cdots, X_t(1;n); U_t | U_t^-) + \\ & I(X_N(1;n); \hat{X}_N(1;n) | U_N^-)] \end{aligned} \tag{3.11}$$

where the infimum is taken over all auxiliary random variables $U_1, \cdots, U_{N-1}$ and $\hat{X}_N(1;n)$ satisfying the requirements (R1) to (R4). By comparing (3.11) with (3.8), it is easy to see that

$$R_{c,n}(D_1, \cdots, D_N) \geq \tilde{R}_{c,n}(D_1, \cdots, D_N). \tag{3.12}$$

On the other hand, pick any auxiliary random variables $U_1, \cdots, U_{N-1}$ and $\hat{X}_N(1;n)$ satisfying the requirements (R1) to (R4). Let $\hat{X}_1(1;n), \hat{X}_2(1;n) \cdots, \hat{X}_{N-1}(1;n)$ be defined as in the requirements (R1) and (R2). Then in view of the Markov conditions in the requirement (R4), we have

$$\begin{aligned}
&I(X_1(1;n);U_1)+ \\
&\sum_{t=2}^{N-1} I(X_1(1;n),\cdots,X_t(1;n);U_t|U_t^-)+ \\
&I(X_N(1;n);\hat{X}_N(1;n)|U_N^-) \\
&= \; I(X_1(1;n),\cdots,X_N(1;n);U_1,\cdots,U_{N-1},\hat{X}_N(1;n)) \\
&\geq \; I(X_1(1;n),\cdots,X_N(1;n);\hat{X}_1(1;n),\cdots,\hat{X}_N(1;n)) \qquad (3.13)
\end{aligned}$$

where the last inequality is due to the fact that $\hat{X}_k(1;n)$ is a function of $U_1, \cdots, U_k$ for any $k = 1, \cdots, N-1$. To continue, we now verify Markov conditions involving $\hat{X}_k(1;n)$. It is not hard to see that the first $N-1$ Markov conditions in the requirement (R4), $U_k \rightarrow (X_k(1;n), X_k^-(1;n), U_k^-) \rightarrow X_k^+(1;n)$, $k = 1, \cdots, N-1$, are equivalent to the following conditions:

(R7) for any $1 \leq k \leq N-1$, $X_k^+(1;n)$ and $(U_1, \cdots, U_k)$ are conditionally independent given $X_k^-(1;n)$ and $X_k(1;n)$.

From this, it follows that for any $1 \leq k \leq N-1$, $X_k^+(1;n)$ and $(\hat{X}_1(1;n), \cdots, \hat{X}_k(1;n))$ are conditionally independent given $X_k^-(1;n)$ and $X_k(1;n)$. Applying the equivalence again, we see that the first $N-1$ Markov conditions involving $\hat{X}_k(1;n)$ in the requirement (R6) are satisfied. Therefore, we have

$$\begin{aligned}
&I(X_1(1;n),\cdots,X_N(1;n);\hat{X}_1(1;n),\cdots,\hat{X}_N(1;n))\\
&\quad= I(X_1(1;n),\cdots,X_N(1;n);\hat{X}_1(1;n))+\\
&\qquad \sum_{k=2}^{N} I(X_1(1;n),\cdots,X_N(1;n);\hat{X}_k(1;n)|\hat{X}_k^-(1;n))\\
&\quad\overset{1)}{=} I(X_1(1;n);\hat{X}_1(1;n))+\\
&\qquad \sum_{k=2}^{N-1} I(X_1(1;n),\cdots,X_k(1;n);\hat{X}_k(1;n)|\hat{X}_k^-(1;n))\\
&\qquad +I(X_1(1;n),\cdots,X_N(1;n);\hat{X}_N(1;n)|\hat{X}_N^-(1;n))\\
&\quad\geq I(X_1(1;n);\hat{X}_1(1;n))+\\
&\qquad \sum_{k=2}^{N-1} I(X_1(1;n),\cdots,X_k(1;n);\hat{X}_k(1;n)|\hat{X}_k^-(1;n))\\
&\qquad +I(X_N(1;n);\hat{X}_N(1;n)|\hat{X}_N^-(1;n))
\end{aligned}\tag{3.14}$$

where the equality 1) follows from the $N-1$ Markov conditions involving $\hat{X}_N^-(1;n)$. Note that the last Markov condition in the requirement (R6) may not be valid for $\hat{X}_N(1;n)$. To overcome this problem, we use the same technique as in the proof of the converse part of Theorem 1 to construct a new random vector $\tilde{X}_N(1;n)$ such that the following hold:

- $(X_N(1;n),\hat{X}_N^-(1;n),\hat{X}_N(1;n))$ and $(X_N(1;n),\hat{X}_N^-(1;n),\tilde{X}_N(1;n))$ have the same distribution, and
- the Markov condition $X_N^-(1;n)\rightarrow(X_N(1;n),\hat{X}_N^-(1;n))\rightarrow\tilde{X}_N(1;n)$ is met.

Therefore, the random variables $\hat{X}_N^-(1;n)$ and $\tilde{X}_N(1;n)$ satisfy the requirements (R5) and (R6). This, together with (3.14), (3.13), and (3.8), implies

$$\begin{aligned}
& I(X_1(1;n);U_1)+ \\
& \sum_{t=2}^{N-1} I(X_1(1;n),\cdots,X_t(1;n);U_t|U_t^-)+ \\
& I(X_N(1;n);\hat{X}_N(1;n)|U_N^-) \\
& \quad \geq \ I(X_1(1;n);\hat{X}_1(1;n)) + \\
& \qquad \sum_{k=2}^{N-1} I(X_1(1;n),\cdots,X_k(1;n);\hat{X}_k(1;n)|\hat{X}_k^-(1;n)) \\
& \qquad + I(X_N(1;n);\tilde{X}_N(1;n)|\hat{X}_N^-(1;n)) \\
& \quad \geq \ nR_{c,n}(D_1,\cdots,D_N). \qquad (3.15)
\end{aligned}$$

Note that (3.15) is valid for any auxiliary random variables $U_1,\cdots,U_{N-1}$ and $\hat{X}_N(1;n)$ satisfying the requirements (R1) to (R4). It then follows from (3.15) and (3.11) that

$$\tilde{R}_{c,n}(D_1,\cdots,D_N) \geq R_{c,n}(D_1,\cdots,D_N)$$

which, together with (3.12), implies that

$$R_{c,n}(D_1,\cdots,D_N) = \tilde{R}_{c,n}(D_1,\cdots,D_N)$$

and (3.11) is an equivalent expression for $R_{c,n}(D_1,\cdots,D_N)$.

In comparison with (3.8), the equivalent expression (3.11) makes it easier to apply the well-known time-sharing argument. By applying the time sharing argument to (3.11), it is now not hard to see that $R_{c,n}(D_1,\cdots,D_N)$ is a convex function of $(D_1,\cdots,D_N)$ for each $n \geq 1$. The convexity of $R_c(D_1,\cdots,D_N)$ as a function of $(D_1,\cdots,D_N)$ then follows from its equivalent expression (3.10) and the convexity of each $R_{c,n}(D_1,\cdots,D_N)$. Since a convex function is continuous over an open region [34], this completes the proof of Lemma 5.

Proof of Theorem 2: In view of the positive part of Theorem 1, it is not hard to see that

$$R_c^*(D_1, \cdots, D_N) \le R_c(D_1, \cdots, D_N)$$

for any $D_1 \ge 0, \cdots, D_N \ge 0$. Therefore, in what follows, it suffices to show

$$R_c^*(D_1, \cdots, D_N) \ge R_c(D_1, \cdots, D_N) \tag{3.16}$$

for any $D_1 > 0, \cdots, D_N > 0$.

Now fix $D_1 > 0, \cdots, D_N > 0$. Pick any rate vector $(R_1, \cdots, R_N)$ such that $(R_1, \cdots, R_N, D_1, \cdots, D_N) \in \mathcal{R}_c^*$. From the proof of the converse part of Theorem 1, it follows that for any $\epsilon > 0$ and sufficiently large n, there exist auxiliary random variables U_k, $k = 1, \cdots, N-1$, and $\hat{X}_N(1;n)$ satisfying the requirements (R1) to (R4) with each D_k replaced by $D_k + \epsilon$ such that

$$\begin{aligned} & n(R_1 + \cdots + R_N + N\epsilon) \\ & \quad \ge \; I(X_1(1;n); U_1) + \sum_{t=2}^{N-1} I(X_1(1;n), \cdots, X_t(1;n); U_t | U_t^-) + \\ & \qquad\quad I(X_N(1;n); \hat{X}_N(1;n) | U_N^-) \end{aligned}$$

which, coupled with the equivalent expression (3.11) for $R_{c,n}(D_1, \cdots, D_N)$, further implies

$$\begin{aligned} & n(R_1 + \cdots + R_N + N\epsilon) \\ & \quad \ge \; nR_{c,n}(D_1 + \epsilon, \cdots, D_N + \epsilon) \\ & \quad \ge \; nR_c(D_1 + \epsilon, \cdots, D_N + \epsilon). \end{aligned} \tag{3.17}$$

In view of Lemma 5, dividing both sides of (3.17) by n and then letting $\epsilon \to 0$ yield

$$R_1 + \cdots + R_N \ge R_c(D_1, \cdots, D_N)$$

from which (3.16) follows. This completes the proof of Theorem 2.

Remark 1 *Theorems 1 and 2 remain valid for general stationary ergodic sources $X_1, \cdots, X_N$. However, the technique adopted in the proof of the classic source coding theorem for a single ergodic source [17], [2] can not be applied here. As such, a new proof technique has to be applied; this will be addressed in the next section.*

3.2.2 General Ergodic Sources

In this section, we extend Theorem 1 and Theorem 2 to general stationary and ergodic sources. Suppose throughout this section that $(X_1, X_2, \cdots, X_N)$ is general stationary and ergodic.

Theorem 3 *For general stationary and* ergodic *sources $X_1, \cdots, X_N$, $\mathcal{R}_c^* = co(\mathcal{R}_c')$.*

The main difficulty ro prove Theorem 3 is that if the vector source is not totally ergodic, X_i and S_i^- may not be order-n ergodic, and as a consequence, S_i may not be found in the proposed coding scheme. This difficulty can be overcome by constructing sliding-block codes instead of traditional block codes. Such strategy has been used by Kieffer in [22] to prove some multi-terminal sliding-block source coding theorems. It was shown in [22] that it is possible to construct the desired sliding-block codes directly without first finding block codes. Furthermore, it follows [21, Theorem 1] that the existence of sliding-block codes imply the existence of block codes. In this section, the same technique as in [22] is applied to prove the sliding-block source coding theorem of CVC for general stationary ergodic sources, and we further verify that Theorem 1 and Theorem 2 remain valid for general stationary ergodic sources $X_1, \cdots, X_N$.

Without losing generality, we will consider the case where $N = 3$. All results and discussions can be easily extended to the case of $N > 3$. If there is no ambiguity, we write $X_1(1;n), X_2(1;n), X_3(1;n)$ as X_1, X_2, X_3 respectively. Using a similar terminology as in

[22], if X, Y are processes defined on the same measurable space, we write $X < Y$ as a shorthand for the property that X is a stationary coding[3] of Y. The stationary code ϕ is called a sliding-block code if there exists for some integer m a map $\phi' : A^{2m+1} \to B$ such that $\phi(x)_i = \phi'(x_{i-m}, \cdots, x_{i+m}), x \in A^\infty, i \in Z$, where Z denotes the set of integers. We call a process X with state A aperiodic if $Pr[X = x] = 0$ for every $x \in A^\infty$.

We write $X < Y(D)$ if X, Y are jointly stationary and

$$Pr[X_0 \neq \phi(Y)_0] < D,$$

for some stationary code ϕ. That is, $X < Y(D)$ means that we can decode Y to obtain an estimate of X to within the distortion level D. Note that, the relation "<" is transitive.

In this proof, we extend the two-step technique in [22] to our case.

Step I: The first step of our method is similar to [22] that we replace each $U_i, i = 1, \cdots, N$, one by one with a sliding-block encoding $\tilde{U}_i$ of $(X_1^i, \tilde{U}_1^{i-1})$, so that the same rate and distortion levels are maintained. ([22, Lemma 1 and Lemma 2] allow us to immediately perform this step.)

Step II: The second step is to sliding-block encode each block code $(\tilde{U}_1, \cdots, \tilde{U}_i), i = 1, \cdots, N$, into a process $\hat{U}_i$ with an arbitrarily small increase in rates. Since our encoders are cascaded, a "random punctuation" construction method specified in [38] can be applied to convert a block code to a lossless sliding-block code so that a small amount of additional distortion is introduced in going from $\hat{U}_i$ back to $\tilde{U}_i$ that maintains the original distortion levels for recovering $X_1, \cdots, X_N$.

Before proving Theorem 3, we need the following theorem, which also provides a sliding-block version of source coding theorem for causal video coding itself.

[3] If A, B are finite sets, a map $\phi : A^\infty \to B^\infty$ is called a stationary code if it is measurable and if $\phi(T_A x) = T_B(\phi(x))$, for all $x \in A^\infty$, where T_A, T_B denote the shifts on A^∞, B^∞, respectively.

Theorem 4 *For jointly stationary and* ergodic *sources* X_1, X_2, X_3*, let* $X_1, X_2, X_3, U_1, U_2, U_3$ *be jointly stationary processes, and the following Markov chains are satisfied:*

(R8) $(X_2, X_3) \rightarrow X_1 \rightarrow U_1$,

(R9) $X_3 \rightarrow (U_1, X_1, X_2) \rightarrow U_2$,

(R10) $(X_1, X_2) \rightarrow (U_1, U_2, X_3) \rightarrow U_3$.

Let R_1, R_2, R_3 *and* D_1, D_2, D_3 *be positive numbers such that*

a) $R_1 > I(X_1; U_1)$*,* $R_2 > I(X_1 X_2; U_2 | U_1)$*,* $R_3 > I(X_3; U_3 | U_1 U_2)$*,*

b) $X_1 < U_1(D_1), X_2 < U_1 U_2(D_2)$*, and* $X_3 < U_1 U_2 U_3(D_3)$*.*

Then there exist processes $\hat{U}_1, \hat{U}_2, \hat{U}_3$ *such that*

c) $\hat{U}_1 < X_1, \hat{U}_2 < X_1 X_2 \hat{U}_1, \hat{U}_3 < X_3 \hat{U}_1 \hat{U}_2$*,*

d) $H(\hat{U}_1) < R_1, H(\hat{U}_2) < R_2, H(\hat{U}_3) < R_3$*,*

e) $X_1 < \hat{U}_1(D_1), X_2 < \hat{U}_1 \hat{U}_2(D_2)$*, and* $X_3 < \hat{U}_1 \hat{U}_2 \hat{U}_3(D_3)$*.*

Proof of Theorem 4:

Case 1) $H(X_1) = 0, H(X_2) = 0, H(X_3) = 0$: take $\hat{U}_1 = X_1, \hat{U}_2 = X_2, \hat{U}_3 = X_3$.

Case 2) $H(X_1) = 0, H(X_2) = 0, H(X_3) \neq 0$: since $X_2 X_3 \rightarrow X_1 \rightarrow U_1$, $X_3 \rightarrow X_1 X_2 U_1 \rightarrow U_2$ and $H(X_1) = 0, H(X_2) = 0$, we have $I(X_3; U_1 U_2) = I(X_3; U_1 U_2 | X_1 X_2) = 0$. Hence

$$\begin{aligned} I(X_3; U_3 | X_1 X_2) &\leq I(X_3; U_1 U_2 U_3) \\ &= I(X_3; U_3 | U_1 U_2) + I(X_3; U_1 U_2) \\ &< R_3. \end{aligned} \tag{3.18}$$

Using [22, Lemma 1 and 2] to replace U_3 by deterministic encoding that $\tilde{U}_3 < X_3$ so that $H(\tilde{U}_3|X_1X_2) < R_3$, and $X_3 < X_1X_2\tilde{U}_3(D_3)$. We then apply Step II to encode $\tilde{U}_3$ into a lossless sliding-block encoding $\hat{U}_3$ so that the desired rate and distortion are obtained. Setting $\hat{U}_1 = X_1$, and $\hat{U}_2 = X_2$, c) to e) hold.

Case 3) $H(X_1) = 0, H(X_2) \neq 0, H(X_3) = 0$: since $X_2X_3 \to X_1 \to U_1$ and $H(X_1) = 0$, we have $I(X_2; U_1) = I(X_2; U_1|X_1) = 0$. Hence

$$\begin{aligned} I(X_2; U_2|X_1) &\leq I(X_2; U_1U_2) \\ &= I(X_2; U_2|U_1) \\ &\overset{1)}{=} I(X_1X_2; U_2|U_1) \\ &< R_2, \end{aligned}$$

where the equality 1) follows from $H(X_1) = 0$. Using [22, Lemma 1 and 2] to replace U_2 by deterministic encodings that $\tilde{U}_2 < X_2$, so that $H(\tilde{U}_2|X_1) < R_2$, and $X_2 < X_1\tilde{U}_2(D_2)$. Applying Step II much as the case 2) to encode $\tilde{U}_2$ into a lossless sliding-block encoding $\hat{U}_2$ with the required rate and distortion level for recovering X_2 maintained, and setting $\hat{U}_1 = X_1, \hat{U}_3 = X_3$, c) to e) hold.

Case 4) $H(X_1) = 0, H(X_2) \neq 0, H(X_3) \neq 0$: since $X_2X_3 \to X_1 \to U_1$ and $H(X_1) = 0$, we have $I(X_2X_3; U_1) = 0$. Hence

$$\begin{aligned} I(X_3; U_3|X_1U_2) &= I(X_3; U_3|U_1U_2) \\ &< R_3 \\ I(X_2; U_2|X_1) &= I(X_2; U_2|U_1) \\ &= I(X_1X_2; U_2|U_1) \\ &< R_2 \end{aligned}$$

Using [22, Lemma 1 and 2] to successively replace U_2, U_3 by deterministic encodings that $\tilde{U}_2 < X_2, \tilde{U}_3 < X_3X_1\tilde{U}_2$ so that

$$\begin{aligned} H(\tilde{U}_2|X_1) &< R_2 \\ H(\tilde{U}_3|\tilde{U}_2X_1) &< R_3 \end{aligned}$$

and $X_2 < X_1\tilde{U}_2(D_2), X_3 < X_1\tilde{U}_2\tilde{U}_3(D_3)$. Applying Step II to cascaded encode a block code $(\tilde{U}_2, \tilde{U}_3)$ into a lossless sliding-block encoding $(\hat{U}_2, \hat{U}_3)$ without changing R_2 and R_3 significantly, and setting $\hat{U}_1 = X_1$, c) to e) hold.

Case 5) $H(X_1) \neq 0, H(X_3) = 0$: in this case, no matter $H(X_2)$ equals to zero or not, we can always distinguish between two subcases: i) $R_2 = 0$, and ii) $R_2 > 0$. In subcase i), repeat the discussion in Case 2) and Case 3) for X_1, it is not hard to find a $\hat{U}_1 < X_1$ so that $H(\hat{U}_1) < R_1$ and $X_1 < \hat{U}_1(D_1)$. Setting $\hat{U}_2 = X_2, \hat{U}_3 = X_3$, c) to e) hold. In subcase ii), we observe that

$$\begin{aligned} I(X_1;U_1) &< R_1 \\ I(X_1X_2;U_2|U_1) &< R_2. \end{aligned}$$

Using [22, Lemma 1 and 2] much as in Case 4), we can successively replace U_1, U_2 by deterministic encodings that $\tilde{U}_1 < X_1, \tilde{U}_2 < X_1X_2\tilde{U}_1$ so that

$$\begin{aligned} H(\tilde{U}_1) &< R_1 \\ H(\tilde{U}_2|\tilde{U}_1) &< R_2 \end{aligned}$$

and $X_1 < \tilde{U}_1(D_1), X_2 < \tilde{U}_1\tilde{U}_2(D_2)$. Applying Step II to cascaded encode a block code $(\tilde{U}_1, \tilde{U}_2)$ into a lossless sliding-block encoding $(\hat{U}_1, \hat{U}_2)$ so that $H(\hat{U}_1) < R_1, H(\hat{U}_2) < R_2$, and setting $\hat{U}_3 = X_3$, c) to e) hold.

Case 6) $H(X_1) \neq 0, H(X_3) \neq 0$: in this case, no matter $H(X_2)$ is zero or not, we can always distinguish between two subcases: i) $R_2 = 0$, and ii) $R_2 > 0$. In subcase i), repeat the discussion in Case 4) for (X_1, X_3), it is not hard to find a $\hat{U}_1 < X_1$ so that $H(\hat{U}_1) < R_1$ and $X_1 < \hat{U}_1(D_1)$, and $\hat{U}_3 < X_3U_1X_2$ so that $H(\hat{U}_3) < R_3$ and $X_3 < \hat{U}_1X_2\hat{U}_3(D_3)$. Setting $\hat{U}_2 = X_2$, c) to e) hold. In subcase ii), the processes U_1, U_2, U_3 are stochastic encoding of the aperiodic processes $X_1, X_1X_2U_1, X_3U_1U_2$, respectively. Therefore, we can use [22, Lemma 1 and 2] to successively replace U_1, U_2, U_3, in that order, by deterministic encodings. Consequently, we may assume that $U_1 < X_1, U_2 < X_1X_2U_1, U_3 < X_3U_1U_2$. The assumptions a) reduce to

a') $R_1 > H(U_1)$, $R_2 > H(U_2|U_1)$, $R_3 > H(U_3|U_1U_2)$.

Applying Step II to cascaded encode (U_1, U_2, U_3) into a lossless sliding-block code $(\hat{U}_1, \hat{U}_2, \hat{U}_3)$ so that $H(\hat{U}_1) < R_1, H(\hat{U}_2) < R_2, H(\hat{U}_3) < R_3$. It is easy to verify that c) to e) hold.

Note that the encoders and decoders in this theorem are stationary codes, however, by [9, Theorem 3.1] they may be replaced by sliding-block codes. This completes the proof of Theorem 4.

Proof of Theorem 3: In view of Theorem 4 and [21, Theorem 1] that sliding-block coders can always imply the existence of block coders, it follows that $(R_1, R_2, R_3, D_1, D_2, D_3) \in \mathcal{R}_{c,n} \subseteq co(\mathcal{R}'_c)$, which in turn implies $\mathcal{R}^*_c \subseteq co(\mathcal{R}'_c)$. This completes the proof of Theorem 3 under the case $N = 3$. Theorem 4 can be easily extended to the case of $N > 3$, and Theorem 3 remains valid.

Having Theorem 3, we see that an argument similar to that used in the proof of Theorem 2 leads to the following result.

Theorem 5 *For jointly stationary and* ergodic *sources* $X_1, \cdots, X_N$,

$$R_c^*(D_1, \cdots, D_N) = \inf\{R_{c,n}(D_1, \cdots, D_N) : n \geq 1\}.$$

For general stationary ergodic sources $X_1, \cdots, X_N$, Theorem 5 is probably the best result one could hope for in terms of analytically characterizing $R_c^*(D_1, \cdots, D_N)$. However, its impact on practical video coding will be limited if the optimization problem involved can not be solved by an effective algorithm. To a large extent, this is also true even if $R_c^*(D_1, \cdots, D_N)$ admits a single-letter characterization, and true for many other multi-user information theoretic problems. In the following section, we will develop an iterative algorithm to compute $R_{c,n}(D_1, \cdots, D_N)$ defined in (3.8), and establish its convergence to the global minimum.

Chapter 4

An Iterative Algorithm

In this chapter, an iterative algorithm is proposed to calculate $R_{c,n}(D_1, \cdots, D_N)$ defined in (3.8), which serves three purposes in this thesis: first, it allows us to do numerical calculations; second, the global convergence of this algorithm provides a completely different approach to establish a single-letter characterization of $R_c^*(D_1, .., D_N)$ when the N sources are IID; and third, it allows us to do comparisons and gain deep insights into $R_c^*(D_1, .., D_N)$.

4.1 Algorithm Description

Without loss of generality, we consider the case of $N = 3$ and denote three sources by $\{X(i)\}_{i=1}^n, \{Y(i)\}_{i=1}^n,$ and $\{Z(i)\}_{i=1}^n$, which in turn will be written as $X^n, Y^n,$ and Z^n respectively to simplify our notation for describing the iterative algorithm.

Let $p_{X^nY^n}$ and $p_{X^nY^nZ^n}$ denote joint distributions of random vectors (X^n, Y^n) and (X^n, Y^n, Z^n), respectively; and let p_{X^n} denote the marginal distribution of X^n. If there is no ambiguity, subscripts in distributions will be omitted. For example, we may write $p(x)$ instead of $p_X(x)$. In order to find the random variables $\hat{X}^n, \hat{Y}^n$ and $\hat{Z}^n$ that achieve

$R_{c,n}(D_1, D_2, D_3)$, we try to find transition probability and probability functions $p_{\hat{X}^n|X^n}$, $p_{\hat{Y}^n|\hat{X}^nY^nX^n}$, $p_{\hat{Z}^n|\hat{X}^n\hat{Y}^nZ^n}$, and $q_{\hat{X}^n\hat{Y}^n\hat{Z}^n}$ that minimize

$$\begin{aligned}
& F_{s,n}(p_{\hat{X}^n|X^n}, p_{\hat{Y}^n|\hat{X}^nY^nX^n}, p_{\hat{Z}^n|\hat{X}^n\hat{Y}^nZ^n}, q_{\hat{X}^n\hat{Y}^n\hat{Z}^n}) \\
\stackrel{\Delta}{=} & \sum_{x^n,y^n,z^n,\hat{x}^n,\hat{y}^n,\hat{z}^n} p(x^n, y^n, z^n)p(\hat{x}^n|x^n) \times \\
& p(\hat{y}^n|\hat{x}^ny^nx^n)p(\hat{z}^n|\hat{x}^n\hat{y}^nz^n) \times \\
& \log[\frac{p(\hat{x}^n|x^n)p(\hat{y}^n|\hat{x}^ny^nx^n)p(\hat{z}^n|\hat{x}^n\hat{y}^nz^n)}{q(\hat{x}^n\hat{y}^n\hat{z}^n)}] + \\
& \alpha \sum_{x^n,\hat{x}^n} p(x^n)p(\hat{x}^n|x^n)d_1(x^n, \hat{x}^n) + \\
& \beta \sum_{x^n,y^n,\hat{x}^n,\hat{y}^n} p(x^n, y^n)p(\hat{x}^n|x^n)p(\hat{y}^n|\hat{x}^ny^nx^n)d_2(y^n, \hat{y}^n) + \\
& \nu \sum_{x^n,y^n,z^n,\hat{x}^n,\hat{y}^n,\hat{z}^n} p(x^n, y^n, z^n)p(\hat{x}^n|x^n) \times \\
& p(\hat{y}^n|\hat{x}^ny^nx^n)p(\hat{z}^n|\hat{x}^n\hat{y}^nz^n)d_3(z^n, \hat{z}^n) \qquad (4.1)
\end{aligned}$$

where $s \stackrel{\Delta}{=} (\alpha, \beta, \nu)$, $\alpha \geq 0, \beta \geq 0, \nu \geq 0$, denotes the standard Lagrange multiplier, and the base of the logarithm is 2. For brevity, we shall denote $(p_{\hat{X}^n|X^n}, p_{\hat{Y}^n|\hat{X}^nY^nX^n}, p_{\hat{Z}^n|\hat{X}^n\hat{Y}^nZ^n})$ by $\mathbf{P}_n$, and $q_{\hat{X}^n\hat{Y}^n\hat{Z}^n} = q_{\hat{X}^n}q_{\hat{Y}^n|\hat{X}^n}q_{\hat{Z}^n|\hat{X}^n\hat{Y}^n}$ by $\mathbf{Q}_n$. Write $F_{s,n}(p_{\hat{X}^n|X^n}, p_{\hat{Y}^n|\hat{X}^nY^nX^n}, p_{\hat{Z}^n|\hat{X}^n\hat{Y}^nZ^n}, q_{\hat{X}^n\hat{Y}^n\hat{Z}^n})$ accordingly as $F_{s,n}(\mathbf{P}_n, \mathbf{Q}_n)$. When there is no ambiguity, the superscript or subscript n will be dropped. The iterative algorithm works as follows.

Step 1: Initialize $i = 0$ and set $\mathbf{Q}^{(0)} \stackrel{\Delta}{=} q^{(0)}_{\hat{X}\hat{Y}\hat{Z}}$ as a joint distribution function over $\hat{\mathcal{X}}$, $\hat{\mathcal{Y}}$, and $\hat{\mathcal{Z}}$, where $q^{(0)}_{\hat{X}\hat{Y}\hat{Z}}(\hat{x}\hat{y}\hat{z}) > 0$ for any $(\hat{x}, \hat{y}, \hat{z}) \in \hat{\mathcal{X}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Z}}$.

Step 2: Fix $\mathbf{Q}^{(i)}$. Find $\mathbf{P}^{(i+1)} \stackrel{\Delta}{=} (p^{(i+1)}_{\hat{X}|X}, p^{(i+1)}_{\hat{Y}|\hat{X}YX}, p^{(i+1)}_{\hat{Z}|\hat{X}\hat{Y}Z})$ such that

$$\mathbf{P}^{(i+1)} \stackrel{\Delta}{=} \arg\min_{\mathbf{P}} F_s(\mathbf{P}, \mathbf{Q}^{(i)}) \qquad (4.2)$$

where the minimum is taken over all transition probability functions $\mathbf{P} = (p_{\hat{X}|X}, p_{\hat{Y}|\hat{X}YX}, p_{\hat{Z}|\hat{X}\hat{Y}Z})$. In view of the nested structure in (4.1), we solve the problem in (4.2) in three stages. First let us find $p^{(i+1)}(\hat{z}|\hat{x}\hat{y}z)$. From (4.1),

$$\begin{aligned}
&\sum_{\hat{z}} p(\hat{z}|\hat{x}\hat{y}z)\log\frac{p(\hat{z}|\hat{x}\hat{y}z)}{q^{(i)}(\hat{z}|\hat{x}\hat{y})2^{-\nu d_3(z,\hat{z})}} \\
&= \sum_{\hat{z}} p(\hat{z}|\hat{x}\hat{y}z)\log\frac{p(\hat{z}|\hat{x}\hat{y}z)}{\Delta^{(i)}(z,\hat{x},\hat{y})\frac{q^{(i)}(\hat{z}|\hat{x}\hat{y})2^{-\nu d_3(z,\hat{z})}}{\Delta^{(i)}(z,\hat{x},\hat{y})}} \\
&\geq \log\frac{1}{\Delta^{(i)}(z,\hat{x},\hat{y})} \qquad (4.3)
\end{aligned}$$

where $\Delta^{(i)}(z,\hat{x},\hat{y}) \triangleq \sum_{\hat{z}} q^{(i)}(\hat{z}|\hat{x}\hat{y})2^{-\nu d_3(z,\hat{z})}$. In the above, the last inequality follows from the log-sum inequality, and becomes an equality if and only if

$$p(\hat{z}|\hat{x}\hat{y}z) = p^{(i+1)}(\hat{z}|\hat{x}\hat{y}z) \triangleq \frac{q^{(i)}(\hat{z}|\hat{x}\hat{y})2^{-\nu d_3(z,\hat{z})}}{\Delta^{(i)}(z,\hat{x},\hat{y})} \qquad (4.4)$$

for any $(\hat{x},\hat{y},z,\hat{z}) \in \hat{\mathcal{X}}\times\hat{\mathcal{Y}}\times\mathcal{Z}\times\hat{\mathcal{Z}}$.

We next find $p^{(i+1)}(\hat{y}|\hat{x}yx)$. In view of (4.1) and (4.3), we have

$$\begin{aligned}
&\sum_{\hat{y}} p(\hat{y}|\hat{x}yx)\log\frac{p(\hat{y}|\hat{x}yx)}{q^{(i)}(\hat{y}|\hat{x})2^{-\beta d_2(y,\hat{y})+\sum_z p(z|yx)\log\Delta^{(i)}(z,\hat{x},\hat{y})}} \\
&= \sum_{\hat{y}} p(\hat{y}|\hat{x}yx)\log\frac{p(\hat{y}|\hat{x}yx)}{\Lambda^{(i)}(x,y,\hat{x})\frac{q^{(i)}(\hat{y}|\hat{x})2^{-\beta d_2(y,\hat{y})+\sum_z p(z|yx)\log\Delta^{(i)}(z,\hat{x},\hat{y})}}{\Lambda^{(i)}(x,y,\hat{x})}} \\
&\geq \log\frac{1}{\Lambda^{(i)}(x,y,\hat{x})} \qquad (4.5)
\end{aligned}$$

where $\Lambda^{(i)}(x,y,\hat{x}) \triangleq \sum_{\hat{y}} q^{(i)}(\hat{y}|\hat{x})2^{-\beta d_2(y,\hat{y})}2^{\sum_z p(z|yx)\log\Delta^{(i)}(z,\hat{x},\hat{y})}$. In the above, the last inequality again follows from the log-sum inequality, and becomes an equality if and only if

$$\begin{aligned}
p(\hat{y}|\hat{x}yx) &= p^{(i+1)}(\hat{y}|\hat{x}yx) \\
&\triangleq \frac{q^{(i)}(\hat{y}|\hat{x})2^{-\beta d_2(y,\hat{y})}2^{\sum_z p(z|yx)\log\Delta^{(i)}(z,\hat{x},\hat{y})}}{\Lambda^{(i)}(x,y,\hat{x})} \qquad (4.6)
\end{aligned}$$

for any $(x,y,\hat{x},\hat{y}) \in \mathcal{X}\times\mathcal{Y}\times\hat{\mathcal{X}}\times\hat{\mathcal{Y}}$.

Finally let us find $p^{(i+1)}(\hat{x}|x)$. Continuing from (4.1) and (4.5), we have

$$\begin{aligned}
&\sum_{\hat{x}} p(\hat{x}|x) \log \frac{p(\hat{x}|x)}{q^{(i)}(\hat{x}) 2^{-\alpha d_1(x,\hat{x}) + \sum_y p(y|x) \log \Lambda^{(i)}(x,y,\hat{x})}} \\
&= \sum_{\hat{x}} p(\hat{x}|x) \log \frac{p(\hat{x}|x)}{\Gamma^{(i)}(x) \frac{q^{(i)}(\hat{x}) 2^{-\alpha d_1(x,\hat{x}) + \sum_y p(y|x) \log \Lambda^{(i)}(x,y,\hat{x})}}{\Gamma^{(i)}(x)}} \\
&\geq \log \frac{1}{\Gamma^{(i)}(x)} \qquad (4.7)
\end{aligned}$$

where $\Gamma^{(i)}(x) \overset{\Delta}{=} \sum_{\hat{x}} q^{(i)}(\hat{x}) 2^{-\alpha d_1(x,\hat{x})} 2^{\sum_y p(y|x) \log \Lambda^{(i)}(x,y,\hat{x})}$. An argument similar to that leading to (4.3) and (4.5) can be used to show that (4.7) becomes an equality if and only if

$$\begin{aligned}
p(\hat{x}|x) &= p^{(i+1)}(\hat{x}|x) \\
&\overset{\Delta}{=} \frac{q^{(i)}(\hat{x}) 2^{-\alpha d_1(x,\hat{x})} 2^{\sum_y p(y|x) \log \Lambda^{(i)}(x,y,\hat{x})}}{\Gamma^{(i)}(x)} \qquad (4.8)
\end{aligned}$$

for any $(x, \hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}$.

Step 3: Fix $\mathbf{P}^{(i+1)}$. Find $\mathbf{Q}^{(i+1)} = q^{(i+1)}$ such that

$$\mathbf{Q}^{(i+1)} \overset{\Delta}{=} \arg\min_{\mathbf{Q}} F_s(\mathbf{P}^{(i+1)}, \mathbf{Q}) \qquad (4.9)$$

where the minimum is taken over all joint distribution functions $\mathbf{Q}$ over $\hat{\mathcal{X}}$, $\hat{\mathcal{Y}}$ and $\hat{\mathcal{Z}}$. In view of (4.1), we see that

$$\begin{aligned}
&\sum_{x,y,z,\hat{x},\hat{y},\hat{z}} p(xyz) p^{(i+1)}(\hat{x}|x) p^{(i+1)}(\hat{y}|\hat{x}yx) p^{(i+1)}(\hat{z}|\hat{x}\hat{y}z) \times \\
&\log \frac{p^{(i+1)}(\hat{x}|x) p^{(i+1)}(\hat{y}|\hat{x}yx) p^{(i+1)}(\hat{z}|\hat{x}\hat{y}z)}{q(\hat{x}\hat{y}\hat{z})} \\
&= I(XYZ; \hat{X}^{(i+1)} \hat{Y}^{(i+1)} \hat{Z}^{(i+1)}) + \sum_{\hat{x},\hat{y},\hat{z}} q^{(i+1)}(\hat{x}\hat{y}\hat{z}) \log \frac{q^{(i+1)}(\hat{x}\hat{y}\hat{z})}{q(\hat{x}\hat{y}\hat{z})} \\
&\geq I(XYZ; \hat{X}^{(i+1)} \hat{Y}^{(i+1)} \hat{Z}^{(i+1)}) \qquad (4.10)
\end{aligned}$$

where $\hat{X}^{(i+1)}\hat{Y}^{(i+1)}\hat{Z}^{(i+1)}$ is the output of the channel $\mathbf{P}^{(i+1)}$ in response to the input XYZ, and $q^{(i+1)}(\hat{x}\hat{y}\hat{z})$ is the distribution of $\hat{X}^{(i+1)}\hat{Y}^{(i+1)}\hat{Z}^{(i+1)}$, i.e.,

$$q^{(i+1)}(\hat{x}\hat{y}\hat{z}) \triangleq \sum_{x,y,z} p(xyz)p^{(i+1)}(\hat{x}|x)p^{(i+1)}(\hat{y}|\hat{x}yx)p^{(i+1)}(\hat{z}|\hat{x}\hat{y}z) \tag{4.11}$$

for any $(\hat{x}, \hat{y}, \hat{z})$. The inequality (4.10) becomes an equality if and only if $q(\hat{x}\hat{y}\hat{z}) = q^{(i+1)}(\hat{x}\hat{y}\hat{z})$ for any $(\hat{x}, \hat{y}, \hat{z}) \in \hat{\mathcal{X}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Z}}$.

Step 4: Repeat Steps 2 and 3 for $i = 1, 2, \cdots$ until $F_s(\mathbf{P}^{(i-1)}, \mathbf{Q}^{(i-1)}) - F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i)})$ is smaller than a prescribed threshold.

4.2 Global Convergence

For any $\mathbf{P}$, let

$$\mathbf{Q}(\mathbf{P}) \triangleq \arg\min_{\mathbf{Q}} F_s(\mathbf{P}, \mathbf{Q}).$$

Similarly, for any $\mathbf{Q}$, let

$$\mathbf{P}(\mathbf{Q}) \triangleq \arg\min_{\mathbf{P}} F_s(\mathbf{P}, \mathbf{Q}).$$

The above iterative algorithm can also be described succinctly by $\mathbf{P}^{(i)} = \mathbf{P}(\mathbf{Q}^{(i-1)})$ and $\mathbf{Q}^{(i)} = \mathbf{Q}(\mathbf{P}^{(i)})$, $i = 1, 2, \cdots$. The following theorem shows that the sequence $\{(\mathbf{P}^{(i)}, \mathbf{Q}^{(i)}) : i \geq 1\}$ converges to a quadruple of distributions that achieves

$$F^*(s) \triangleq \inf F_s(p_{\hat{X}|X}, p_{\hat{Y}|\hat{X}YX}, p_{\hat{Z}|\hat{X}\hat{Y}Z}, q_{\hat{X}\hat{Y}\hat{Z}}) \tag{4.12}$$

where the infimum is taken over all possible $p_{\hat{X}|X}$, $p_{\hat{Y}|\hat{X}YX}$, $p_{\hat{Z}|\hat{X}\hat{Y}Z}$, and $q_{\hat{X}\hat{Y}\hat{Z}}$.

Theorem 6 *For any initial $\boldsymbol{Q}^{(0)}$ satisfying $q^{(0)}_{\hat{X}\hat{Y}\hat{Z}}(\hat{x}, \hat{y}, \hat{z}) > 0$ for any $(\hat{x}, \hat{y}, \hat{z}) \in \hat{\mathcal{X}} \times \hat{\mathcal{Y}} \times \hat{\mathcal{Z}}$, there exists $\boldsymbol{Q}^*$ such that $F_s(\boldsymbol{P}(\boldsymbol{Q}^*), \boldsymbol{Q}^*) = F^*(s)$, and*

$$\boldsymbol{P}^{(i)} \to \boldsymbol{P}(\boldsymbol{Q}^*), \;\; \boldsymbol{Q}^{(i)} \to \boldsymbol{Q}^*, \;\; \text{and } F_s(\boldsymbol{P}^{(i)}, \boldsymbol{Q}^{(i)}) \to F^*(s)$$

as $i \to \infty$.

Proof of Theorem 6: From the description of the iterative algorithm, it follows that

$$F_s(\mathbf{P}^{(1)}, \mathbf{Q}^{(0)}) \geq F_s(\mathbf{P}^{(1)}, \mathbf{Q}^{(1)}) \geq F_s(\mathbf{P}^{(2)}, \mathbf{Q}^{(1)}) \geq \cdots . \tag{4.13}$$

To show the desired convergence, let us first verify that the algorithm has the so-called "five-point property" (as defined in [13]), that is for any $\mathbf{P} = (p_{\hat{X}|X}, p_{\hat{Y}|\hat{X}YX}, p_{\hat{Z}|\hat{X}\hat{Y}Z})$, and the corresponding $\mathbf{Q} = \mathbf{Q}(\mathbf{P})$,

$$F_s(\mathbf{P}, \mathbf{Q}^{(i-1)}) - F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i)}) \geq F_s(\mathbf{P}, \mathbf{Q}^{(i)}) - F_s(\mathbf{P}, \mathbf{Q}). \tag{4.14}$$

To this end, let us calculate both sides of (4.14). In view of Steps 2 and 3, we have

$$\begin{aligned}
& F_s(\mathbf{P}, \mathbf{Q}^{(i-1)}) - F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i)}) \\
\geq \;& F_s(\mathbf{P}, \mathbf{Q}^{(i-1)}) - F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i-1)}) \\
\overset{1)}{=} \;& \sum_{x,\hat{x}} p(x)p(\hat{x}|x)[\log \frac{p(\hat{x}|x)}{p^{(i)}(\hat{x}|x)} + \\
& \sum_{y,\hat{y}} p(y|x)p(\hat{y}|\hat{x}yx)[\log \frac{p(\hat{y}|\hat{x}yx)}{p^{(i)}(\hat{y}|\hat{x}yx)} + \\
& \sum_{z,\hat{z}} p(z|xy)p(\hat{z}|\hat{x}\hat{y}z) \log \frac{p(\hat{z}|\hat{x}\hat{y}z)}{p^{(i)}(\hat{z}|\hat{x}\hat{y}z)}]] \\
\geq \;& \sum_{\hat{x},\hat{y},\hat{z}} q(\hat{x}\hat{y}\hat{z}) \log \frac{q(\hat{x}\hat{y}\hat{z})}{q^{(i)}(\hat{x}\hat{y}\hat{z})},
\end{aligned} \tag{4.15}$$

where the equality 1) follows from the following derivation:

$$F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i-1)}) = -\sum_x p(x) \log \Gamma^{(i-1)}(x) \tag{4.16}$$

and

$$
\begin{aligned}
& F_s(\mathbf{P}, \mathbf{Q}^{(i-1)}) \\
= \ & \sum_{x,y,z,\hat{x}} p(xyz)p(\hat{x}|x)[\log \frac{p(\hat{x}|x)}{q^{(i-1)}(\hat{x})2^{-\alpha d_1(x,\hat{x})}} + \\
& \sum_{\hat{y}} p(\hat{y}|\hat{x}yx)[\log \frac{p(\hat{y}|\hat{x}yx)}{q^{(i-1)}(\hat{y}|\hat{x})2^{-\beta d_2(y,\hat{y})}\Delta^{(i-1)}(z,\hat{x},\hat{y})} + \\
& D(p(\hat{z}|\hat{x}\hat{y}z)||p^{(i)}(\hat{z}|\hat{x}\hat{y}z))]] \\
= \ & \sum_{x,y,\hat{x}} p(xy)p(\hat{x}|x)[\log \frac{p(\hat{x}|x)}{q^{(i-1)}(\hat{x})2^{-\alpha d_1(x,\hat{x})}} + \\
& \sum_{\hat{y}} p(\hat{y}|\hat{x}yx)[\log \frac{p(\hat{y}|\hat{x}yx)}{q^{(i-1)}(\hat{y}|\hat{x})2^{-\beta d_2(y,\hat{y})}} - \sum_{z} p(z|xy)\log \Delta^{(i-1)}(z,\hat{x},\hat{y})]] + \\
& \sum_{x,y,z,\hat{x}} p(xyz)p(\hat{x}|x) \sum_{\hat{y}} p(\hat{y}|\hat{x}yx) D(p(\hat{z}|\hat{x}\hat{y}z)||p^{(i)}(\hat{z}|\hat{x}\hat{y}z)) \\
= \ & \sum_{x,y,\hat{x}} p(xy)p(\hat{x}|x)[\log \frac{p(\hat{x}|x)}{q^{(i-1)}(\hat{x})2^{-\alpha d_1(x,\hat{x})}\Lambda^{(i-1)}(x,y,\hat{x})} + \\
& D(p(\hat{y}|\hat{x}yx)||p^{(i)}(\hat{y}|\hat{x}yx))] + \\
& \sum_{x,y,z,\hat{x}} p(xyz)p(\hat{x}|x) \sum_{\hat{y}} p(\hat{y}|\hat{x}yx) D(p(\hat{z}|\hat{x}\hat{y}z)||p^{(i)}(\hat{z}|\hat{x}\hat{y}z)) \\
= \ & \sum_{x} p(x)[D(p(\hat{x}|x)||p^{(i)}(\hat{x}|x)) - \log \Gamma^{(i-1)}(x)] + \\
& \sum_{x,y,\hat{x}} p(xy)p(\hat{x}|x) D(p(\hat{y}|\hat{x}yx)||p^{(i)}(\hat{y}|\hat{x}yx)) + \\
& \sum_{x,y,z,\hat{x}} p(xyz)p(\hat{x}|x) \sum_{\hat{y}} p(\hat{y}|\hat{x}yx) D(p(\hat{z}|\hat{x}\hat{y}z)||p^{(i)}(\hat{z}|\hat{x}\hat{y}z)) \\
= \ & -\sum_{x} p(x)\log \Gamma^{(i-1)}(x) + \sum_{x} p(x) D(p(\hat{x}|x)||p^{(i)}(\hat{x}|x)) + \\
& \sum_{x,y,\hat{x}} p(xy)p(\hat{x}|x) D(p(\hat{y}|\hat{x}yx)||p^{(i)}(\hat{y}|\hat{x}yx)) + \\
& \sum_{x,y,z,\hat{x}} p(xyz)p(\hat{x}|x) \sum_{\hat{y}} p(\hat{y}|\hat{x}yx) D(p(\hat{z}|\hat{x}\hat{y}z)||p^{(i)}(\hat{z}|\hat{x}\hat{y}z)). \qquad (4.17)
\end{aligned}
$$

Combining (4.16) and (4.17), we immediately have the equality 1) in (4.15).

On the other hand,

$$\begin{aligned} & F_s(\mathbf{P}, \mathbf{Q}^{(i)}) - F_s(\mathbf{P}, \mathbf{Q}) \\ = & \sum_{\hat{x},\hat{y},\hat{z}} q(\hat{x}\hat{y}\hat{z}) \log \frac{q(\hat{x}\hat{y}\hat{z})}{q^{(i)}(\hat{x}\hat{y}\hat{z})}. \end{aligned} \tag{4.18}$$

Combining (4.15) with (4.18) yields the desired “five-point property” in (4.14).

The rest of the proof is similar to that adopted in [12] to show the convergence of the Blahut-Arimoto algorithm [7]. Suppose

$$F_s(\mathbf{P}, \mathbf{Q}) \leq \lim_{i\to\infty} F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i)}) = \lim_{i\to\infty} F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i-1)}) \tag{4.19}$$

for some $\mathbf{P}$ and $\mathbf{Q} = \mathbf{Q}(\mathbf{P})$. From (4.14), it then follows that for any $N > M \geq 0$,

$$\begin{aligned} 0 & \leq \sum_{i=M+1}^{N} [F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i)}) - F_s(\mathbf{P}, \mathbf{Q})] \\ & \leq \sum_{i=M+1}^{N} [F_s(\mathbf{P}, \mathbf{Q}^{(i-1)}) - F_s(\mathbf{P}, \mathbf{Q}^{(i)})] \\ & \leq \sum_{i=M+1}^{N} \sum_{\hat{x},\hat{y},\hat{z}} q(\hat{x}\hat{y}\hat{z}) \log \frac{q^{(i)}(\hat{x}\hat{y}\hat{z})}{q^{(i-1)}(\hat{x}\hat{y}\hat{z})} \\ & = \sum_{\hat{x},\hat{y},\hat{z}} q(\hat{x}\hat{y}\hat{z}) \log \frac{q^{(N)}(\hat{x}\hat{y}\hat{z})}{q^{(M)}(\hat{x}\hat{y}\hat{z})} \\ & = D(\mathbf{Q}||\mathbf{Q}^{(M)}) - D(\mathbf{Q}||\mathbf{Q}^{(N)}) \end{aligned} \tag{4.20}$$

which, together with $D(\mathbf{Q}||\mathbf{Q}^{(0)}) < \infty$, implies

$$\sum_{i=1}^{\infty} [F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i)}) - F_s(\mathbf{P}, \mathbf{Q})] < \infty \tag{4.21}$$

and hence

$$\lim_{i\to\infty} [F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i)}) - F_s(\mathbf{P}, \mathbf{Q})] = 0. \tag{4.22}$$

Note that (4.22) is valid for any $\mathbf{P}$ and $\mathbf{Q} = \mathbf{Q}(\mathbf{P})$ satisfying (4.19). From this, we have

$$\lim_{i\to\infty} F_s(\mathbf{P}^{(i)}, \mathbf{Q}^{(i)}) = F^*(s) = \inf F_s(\mathbf{P}, \mathbf{Q}). \tag{4.23}$$

To prove the convergence of $(\mathbf{P}^{(i)}, \mathbf{Q}^{(i)})$, pick a convergent subsequence of $\mathbf{Q}^{(i)}$, say $\mathbf{Q}^{(i_j)} \to \mathbf{Q}^*$. Then $\mathbf{P}^{(i_j+1)} = \mathbf{P}(\mathbf{Q}^{(i_j)}) \to \mathbf{P}^* = \mathbf{P}(\mathbf{Q}^*)$ and

$$F_s(\mathbf{P}^{(i_j+1)}, \mathbf{Q}^{(i_j)}) \to F_s(\mathbf{P}^*, \mathbf{Q}^*). \tag{4.24}$$

In view of (4.23), we have $F_s(\mathbf{P}^*, \mathbf{Q}^*) = F^*(s)$; thus $\mathbf{Q}^* = \mathbf{Q}(\mathbf{P}^*)$, and hence (4.20) applies to $\mathbf{Q}^*$ and $\mathbf{P}^*$. In particular, $D(\mathbf{Q}^*||\mathbf{Q}^{(i)})$ is a nonincreasing sequence. Since $\mathbf{Q}^{(i_j)} \to \mathbf{Q}^*$ implies $D(\mathbf{Q}^*||\mathbf{Q}^{(i_j)}) \to 0$, this means $D(\mathbf{Q}^*||\mathbf{Q}^{(i)}) \to 0$. Hence $\mathbf{Q}^{(i)} \to \mathbf{Q}^*$ and $\mathbf{P}^{(i)} \to \mathbf{P}^*$ as $i \to \infty$. This completes the proof of Theorem 6.

Remark 2 *The above iterative algorithm can be easily extended to the case of $N > 3$, and Theorem 6 remains valid. By setting $\nu = 0$, it also reduces to the case of $N = 2$.*

Remark 3 *The iterative algorithm can be further extended to work for coupled distortion measures (as defined in [42]) $d'_k : \mathcal{X}_k \times \hat{\mathcal{X}}_k \times \hat{\mathcal{X}}_{k-1} \times \cdots \times \hat{\mathcal{X}}_1 \to [0, \infty)$, $k = 2, \cdots, N$, where the distortion $d'_k(X_k, \hat{X}_k|\hat{X}_k^-)$ depends not only on $(X_k, \hat{X}_k)$ but also on $(\hat{X}_1, \cdots, \hat{X}_{k-1})$. The global convergence as expressed in Theorem 6 is still guaranteed.*

Remark 4 *Although $R_{c,n}(D_1, \cdots, D_N)$ as a function of $D_1, \cdots, D_N$ is convex as shown in the proof of Lemma 5, both the optimization problems* (3.8) *and* (4.12) *are actually a non-convex optimization problem. It is therefore kind of surprising to see the global convergence of our proposed iterative algorithm. As shown in the proof of Theorem 6, the key for the global convergence is the five-point property* (4.14)*.*

Remark 5 *There are many other ways (including, for example, the greedy alternative algorithm [56]) to derive iterative procedures. However, it is not clear whether their global* *convergence can be guaranteed. Having algorithms with global convergence is important to not only numerical computation itself, but also single-letter characterization of performance. One of the purposes of this thesis is indeed to demonstrate for the first time that single-letter characterization of performance can also be established in a computational way via algorithms with global convergence, as shown in the next section.*

We conclude this section by presenting an alternative expression for $R_{c,n}(D_1, \cdots, D_N)$. Once again, we illustrate this by considering the case of $N = 3$. In view of the definitions (3.8) and (4.12), it is not hard to show (for example, by using the technique demonstrated in the proof of Property 1 in [52]) that for any $s = (\alpha, \beta, \nu)$, $\alpha \geq 0, \beta \geq 0, \nu \geq 0$,

$$\frac{F^*(s)}{n} = \inf\{R_{c,n}(D_1, D_2, D_3) + \alpha D_1 + \beta D_2 + \nu D_3 : \; D_1 \geq 0, D_2 \geq 0, D_3 \geq 0\}. \tag{4.25}$$

In other words, $F^*(s)/n$ as a function of s is the conjugate of $R_{c,n}(D_1, D_2, D_3)$. Since $R_{c,n}(D_1, D_2, D_3)$ is convex and lower semi-continuous over the whole region $D_1 \geq 0, D_2 \geq 0, D_3 \geq 0$, it follows from [34, Theorem 12.2, pp. 104] that for any $D_1 \geq 0, D_2 \geq 0, D_3 \geq 0$,

$$R_{c,n}(D_1, D_2, D_3) = \sup\{F^*(s)/n - \alpha D_1 - \beta D_2 - \nu D_3 : \; s = (\alpha, \beta, \nu) \text{ and } \alpha \geq 0, \beta \geq 0, \nu \geq 0\}. \tag{4.26}$$

In the next section, (4.26) will be used in the process of establishing a single-letter characterization for $R_c^*(D_1, \cdots, D_N)$ when the vector source $(X_1, \cdots, X_N)$ is IID.

4.3 Single-letter Characterization: IID Causal Case

Suppose now that the vector source $(X_1, \cdots, X_N)$ is IID. In this section, we will use our iterative algorithm proposed in Section 4.1 and its global convergence to establish a single-letter characterization for $R_c^*(D_1, \cdots, D_N)$.

Theorem 7 *If* $(X_1, \cdots, X_N)$ *is IID, then*

$$R_c^*(D_1, \cdots, D_N) = R_{c,1}(D_1, \cdots, D_N)$$

for any $D_1 \geq 0, D_2 \geq 0, \cdots, D_N \geq 0$.

Proof: We first show that for any $D_1 \geq 0, D_2 \geq 0, \cdots, D_N \geq 0$,

$$R_{c,n}(D_1, \cdots, D_N) = R_{c,1}(D_1, \cdots, D_N) \tag{4.27}$$

for any $n > 1$. Without loss of generality, we demonstrate (4.27) in the case of $N = 3$ by using our iterative algorithm in Section 4.1. Denote three sources by $X, Y,$ and Z. Since the vector source (X, Y, Z) is IID, we have $p_{X^n Y^n Z^n} = \prod_{i=1}^n p_{X(i)Y(i)Z(i)}$. In view of (4.26), we have

$$\begin{aligned} R_{c,n}(D_1, D_2, D_3) = \sup\{F_n^*(s)/n - \alpha D_1 - \beta D_2 - \\ \nu D_3 : \ s = (\alpha, \beta, \nu) \text{ and } \alpha \geq 0, \beta \geq 0, \nu \geq 0\} \end{aligned} \tag{4.28}$$

for any $n \geq 1$, where $F_n^*(s)$ is defined in (4.12). Here and throughout the rest of this proof, the subscript or superscript n dropped for convenience for notation in Section 4.1 is brought back to distinguish between the cases of $n = 1$ and $n > 1$. Therefore, it suffices to show that

$$F_n^*(s) = nF_1^*(s) \tag{4.29}$$

for any $s = (\alpha, \beta, \nu)$, $\alpha \geq 0, \beta \geq 0, \nu \geq 0$. To this end, we will run the iterative algorithm in both cases of $n = 1$ and $n > 1$ to calculate $F_1^*(s)$ and $F_n^*(s)$. Pick any initial positive distribution $\mathbf{Q}_1^{(0)}$, and run the iterative algorithm in the case of $n = 1$. We then get a sequence $\{(\mathbf{P}_1^{(i)}, \mathbf{Q}_1^{(i)}) : i \geq 1\}$ which, according to Theorem 6, satisfies

$$\lim_{i \to \infty} F_{s,1}(\mathbf{P}_1^{(i)}, \mathbf{Q}_1^{(i)}) = F_1^*(s). \tag{4.30}$$

Now let $\mathbf{Q}_n^{(0)}$ be the n-fold product distribution of $\mathbf{Q}_1^{(0)}$. Clearly, $\mathbf{Q}_n^{(0)}$ is also positive. Use $\mathbf{Q}_n^{(0)}$ as an initial distribution and run the iterative algorithm in the case of $n > 1$. Then we get a sequence $\{(\mathbf{P}_n^{(i)}, \mathbf{Q}_n^{(i)}) : i \geq 1\}$ which, according to Theorem 6 again, satisfies

$$\lim_{i\to\infty} F_{s,n}(\mathbf{P}_n^{(i)}, \mathbf{Q}_n^{(i)}) = F_n^*(s). \tag{4.31}$$

Since $p_{X^nY^nZ^n}$ is the n-fold product of $p_{X(1)Y(1)Z(1)}$ and $\mathbf{Q}_n^{(0)}$ is the n-fold product of $\mathbf{Q}_1^{(0)}$, careful examination on (4.4), (4.6), (4.8), and (4.11) reveals that for any $i \geq 1$, $\mathbf{P}_n^{(i)}$ is the n-fold product of $\mathbf{P}_1^{(i)}$, and $\mathbf{Q}_n^{(i)}$ is the n-fold product of $\mathbf{Q}_1^{(i)}$. (To see this is the case, let us look at (4.4) for example. Let us temporarily drop the subscripts indicating random variables in all notation. When $p(x^ny^nz^n) = \prod_{j=1}^n p(x_jy_jz_j)$ and $q^{(i)}(\hat{x}^n\hat{y}^n\hat{z}^n) = \prod_{j=1}^n q^{(i)}(\hat{x}_j\hat{y}_j\hat{z}_j)$, it can be verified that in (4.4),

$$q^{(i)}(\hat{z}^n|\hat{x}^n\hat{y}^n) = \prod_{j=1}^n q^{(i)}(\hat{z}_j|\hat{x}_j\hat{y}_j)$$

and

$$\Delta^{(i)}(z^n, \hat{x}^n, \hat{y}^n) = \prod_{j=1}^n \Delta^{(i)}(z_j, \hat{x}_j, \hat{y}_j).$$

Since

$$d_3(z^n, \hat{z}^n) = \sum_{j=1}^n d_3(z_j, \hat{z}_j)$$

it follows from (4.4) that

$$p^{(i+1)}(\hat{z}^n|\hat{x}^n\hat{y}^nz^n) = \prod_{j=1}^n p^{(i+1)}(\hat{z}_j|\hat{x}_j\hat{y}_jz_j).$$

Similar argument can be applied to (4.6), (4.8), and (4.11).) Therefore, for any $i \geq 1$,

$$F_{s,n}(\mathbf{P}_n^{(i)}, \mathbf{Q}_n^{(i)}) = nF_{s,1}(\mathbf{P}_1^{(i)}, \mathbf{Q}_1^{(i)})$$

which, coupled with (4.30) and (4.31), implies (4.29) and hence (4.27).

Combining (4.27) with (3.9) yields

$$R_c(D_1, \cdots, D_N) = R_{c,1}(D_1, \cdots, D_N)$$

for any $D_1 \geq 0, \cdots, D_N \geq 0$. This, together with Theorem 2, implies

$$R_c^*(D_1, \cdots, D_N) = R_{c,1}(D_1, \cdots, D_N) \tag{4.32}$$

for any $D_1 > 0, \cdots, D_N > 0$. Since by their definitions, both functions $R_c^*(D_1, \cdots, D_N)$ and $R_{c,1}(D_1, \cdots, D_N)$ are right continuous in the sense that for any $D_1 \geq 0, \cdots, D_N \geq 0$,

$$\lim_{\epsilon \downarrow 0} R_c^*(D_1 + \epsilon, \cdots, D_N + \epsilon) = R_c^*(D_1, \cdots, D_N)$$

and

$$\lim_{\epsilon \downarrow 0} R_{c,1}(D_1 + \epsilon, \cdots, D_N + \epsilon) = R_{c,1}(D_1, \cdots, D_N)$$

it follows that (4.32) remains valid for boundary points where some D_i may be 0. This completes the proof of Theorem 7.

Theorem 7 can also be proved by using the classical auxiliary random variable converse and positive proof (hereafter referred to as "the classic approach"). Indeed, one can establish the following single-letter characterization for the achievable region $\mathcal{R}_c^*$, the proof of which is given in Appendix A.

Theorem 8 *If $(X_1, \cdots, X_N)$ is an IID vector source, then*[1] $\mathcal{R}_c^* = co(\mathcal{R}_{c,1})$.

Remark 6 *It is instructive to compare the computational approach to single-letter characterization (as illustrated in the proofs of Theorems 2, 6, and 7) with the classic approach.*

[1] *Since the alphabet size of each U_k in (3.2) can be bounded, $\mathcal{R}_{c,n}$, $n \geq 1$, is actually convex and closed. As such, $co(\mathcal{R}_{c,1}) = \mathcal{R}_{c,1}$. We leave $co(\mathcal{R}_{c,1})$ in the statement of Theorem 8 just for the sake of consistency with the norm in the literature [11].*

In the computational approach, the converse is first established for multiple letters (blocks); its proof is often straightforward and the required Markov chain conditions are satisfied automatically as shown in the proof of Theorem 2. The key is then to have an algorithm with global convergence for computing all block terms and later show that all these block terms are the same. On the other hand, in the classic approach, the converse proof is quite involved; coming up with auxiliary random variables with right Markov chain conditions is always challenging and sometimes seems impossible. Since single-letter characterization has to be computed any way, the computational approach is preferred whenever it is possible.

Remark 7 *When $N = 2$, Theorems 8 and 7 reduce to Theorems 1 and 3 in [42], respectively. However, the proofs in [42] are incomplete due to the invalid claim of the Markov condition made in the proofs therein; as such formulas therein can not be extended to the case of $N > 2$. Theorems 8 and 7 in a slightly different, but equivalent form were also reported in [26], [25], and [27] by following the classic approach. The difference lies in the extra Markov chain condition for the reconstruction $\hat{X}_N(1)$ shown as Condition (R4). For example, in the specific formulas shown in [26, Theorem 1] in the case of $N = 3$, the Markov chain condition $\hat{X}_3(1) \to (X_3(1), U_1, U_2) \to (X_1(1), X_2(1))$ is not required.*

4.4 Comparisons

To gain deep insights into CVC, in this section, we use our iterative algorithm proposed in Section 4.1 to compare: 1) CVC with greedy coding; and 2) $R_c^*(D_1, \cdots, D_N)$ among different values of N.

4.4.1 Causal vs. Greedy

All MPEG-series and H-series video coding standards [32], [49] proposed so far fall into PVC, where at the encoder for each frame X_k, only previous encoded frames are used as a helper. By using a technique called soft decision quantization [49], [47], [48], it has been demonstrated in a series of papers [49], [50], [46] that the greedy coding method[2] offers significant gains (ranging from 10% to 30% rate reduction at the same quality) over the respective reference codecs[3] of these standards. As such, it is instructive to compare the performance of causal coding characterized by $R_c^*(D_1, \cdots, D_N)$ with the performance of greedy coding characterized by the total rate $R_g(D_1, \cdots, D_N)$ offered by the greedy coding method. In this section, we present specific examples to numerically compare $R_c^*(D_1, \cdots, D_N)$ with $R_g(D_1, \cdots, D_N)$.

Example 1: Suppose that $\mathcal{X}_i = \hat{\mathcal{X}}_i = \{0, 1, 2, 3\}$, $i = 1, 2, 3$, and the Hamming distortion measure is used. In this example, we consider a Markov chain: $X_1 \to X_2 \to X_3$. The transition probability $p_{X_2|X_1}$ is given by

$$\left(\begin{array}{cccccc} & X_2 & 0 & 1 & 2 & 3 \\ X_1 & & & & & \\ 0 & & 0 & 1 & 0 & 0 \\ 1 & & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 2 & & 0 & 0 & 0 & 1 \\ 3 & & 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{array}\right)$$

[2] The greedy coding method is a special form of PVC; based on all previous encoded frames, it encodes each current frame in a local optimum manner so as to achieve the best rate distortion trade-off for the current frame only.

[3] Both the greedy coding method and reference codecs are special forms of PVC. At this point, the best rate distortion performance of PVC is still unknown in general.

and the other transition probability $p_{X_3|X_2}$ is given by

$$\left(\begin{array}{cccccc} & X_3 & 0 & 1 & 2 & 3 \\ X_2 & & & & & \\ 0 & & 0 & 1 & 0 & 0 \\ 1 & & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 2 & & 0 & 0 & 0 & 1 \\ 3 & & 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{array}\right).$$

Figure 4.1 shows the rate-distortion curves of $R_c^*(D_1, D_2, D_3)$ and $R_g(D_1, D_2, D_3)$ versus D_3 when X_1 is uniformly distributed, $D_1 = 0.5488$, and $D_2 = 0.3927$. As shown in Figure 4.1,

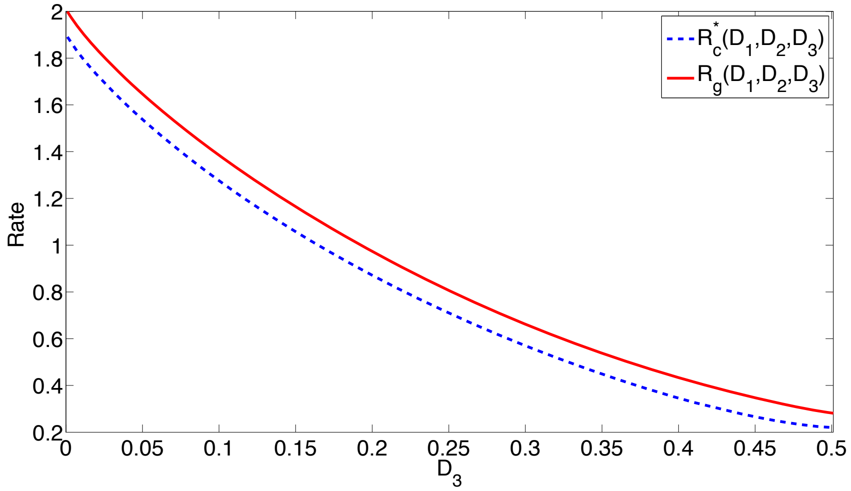

Figure 4.1: Comparison of $R_c^*(D_1, D_2, D_3)$ and $R_g(D_1, D_2, D_3)$ versus D_3 for fixed $D_1 = 0.5488$ and $D_2 = 0.3927$ in Example 1.

when $D_3 = 0.4768$, $R_c^*(D_1, D_2, D_3) = 0.2354$, which is more than 31 percent less than $R_g(D_1, D_2, D_3) = 0.3086$.

Let us now look at another example in which X_1, X_2, and X_3 do not form a Markov chain.

Example 2: Suppose that $\mathcal{X}_i = \hat{\mathcal{X}}_i = \{0, 1, 2, 3\}$, $i = 1, 2, 3$, and the Hamming distortion measure is used. In this example, X_1, X_2, and X_3 do not form a Markov chain, but $X_2 \rightarrow X_1 \rightarrow X_3$ does form a Markov chain in the indicated order. The transition probability $p_{X_2|X_1}$ is given by

$$\left(\begin{array}{cccccc} & X_2 & 0 & 1 & 2 & 3 \\ X_1 & & & & & \\ 0 & & 0 & 1 & 0 & 0 \\ 1 & & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 2 & & 0 & 0 & 0 & 1 \\ 3 & & 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{array}\right)$$

and the other transition probability $p_{X_3|X_1}$ is given by

$$\left(\begin{array}{cccccc} & X_3 & 0 & 1 & 2 & 3 \\ X_1 & & & & & \\ 0 & & 0 & 1 & 0 & 0 \\ 1 & & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 2 & & 0 & 0 & 0 & 1 \\ 3 & & 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{array}\right).$$

Figure 4.2 shows the rate-distortion curves of $R_c^*(D_1, D_2, D_3)$ and $R_g(D_1, D_2, D_3)$ versus D_3 when X_1 is uniformly distributed, $D_1 = 0.5488$, and $D_2 = 0.3927$. As shown in Figure 4.2, when $D_3 = 0.3927$, $R_c^*(D_1, D_2, D_3) = 0.2210$, which is 34.8 percent less than $R_g(D_1, D_2, D_3) = 0.2979$.

The above two examples are of course toy examples. However, if the performance improvement is indicative of the performance of CVC for real video data, it is definitely worthwhile to make the CVC idea materialize in video codecs.

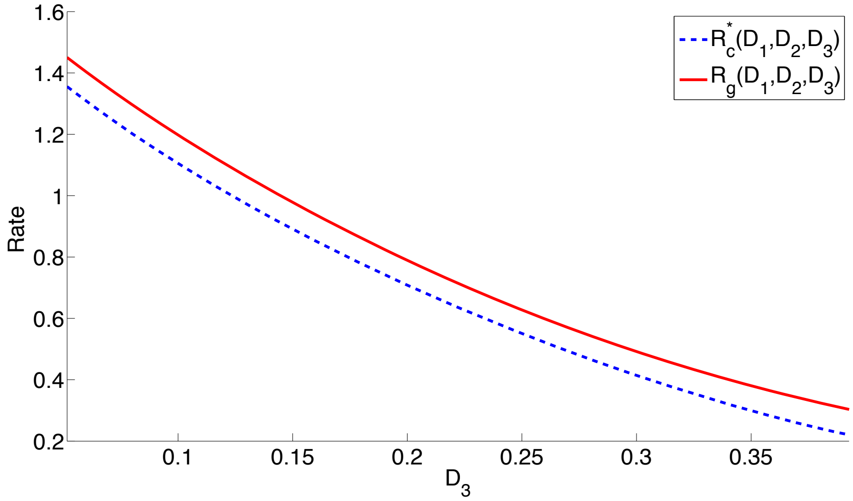

Figure 4.2: Comparison of $R_c^*(D_1, D_2, D_3)$ and $R_g(D_1, D_2, D_3)$ versus D_3 for fixed $D_1 = 0.5488$ and $D_2 = 0.3927$ in Example 2.

4.4.2 More and Less Coding Theorem

To gain deep insights into CVC, in this section, we use our iterative algorithm proposed in Section 4.1 to compare $R_c^*(D_1, \cdots, D_N)$ among different values of N. To be specific, whenever we need to bring out the dependence of $R_c^*(D_1, \cdots, D_N)$ and $R_{c,n}(D_1, \cdots, D_N)$ on the sources $X_1, \cdots, X_N$, we will write $R_c^*(D_1, \cdots, D_N)$ as $R_c^{X_1 \cdots X_N}(D_1, \cdots, D_N)$, and $R_{c,n}(D_1, \cdots, D_N)$ as $R_{c,n}^{X_1 \cdots X_N}(D_1, \cdots, D_N)$. In particular, we will compare $R_c^{X_1 \cdots X_N}(D_1, \cdots, D_N)$ with $R_c^{X_2 \cdots X_N}(D_2, \cdots, D_N)$.

Without loss of generality again, we will consider the case of $N = 3$. All results and discussions in this section can be easily extended to the case of $N > 3$. We first have the following result.

Theorem 9 *Suppose that (X_1, X_2, X_3) is jointly stationary and ergodic, and X_1, X_2, and*

X_3 *form a Markov chain in the indicated order. Then for any* $D_1, D_2, D_3 \geq 0$,

$$R_c^{X_1X_2X_3}(D_1, D_2, D_3) \geq R_c^{X_2X_3}(D_2, D_3). \tag{4.33}$$

Proof: We distinguish between two cases: (1) $D_1D_2D_3 > 0$, and (2) $D_1D_2D_3 = 0$. In Case (1), it follows from Theorem 2 and (3.9) that it suffices to show

$$R_{c,n}^{X_1X_2X_3}(D_1, D_2, D_3) \geq R_{c,n}^{X_2X_3}(D_2, D_3) \tag{4.34}$$

for any $n \geq 1$ and $D_1 > 0, D_2 > 0, D_3 > 0$. To this end, pick any auxiliary random variables $\hat{X}_i(1;n)$, $i = 1, 2, 3$, satisfying the requirements (R5) and (R6) with $N = 3$. It is not hard to verify that

$$\begin{aligned}
&I(X_1(1;n); \hat{X}_1(1;n)) + I(X_1(1;n)X_2(1;n); \hat{X}_2(1;n)|\hat{X}_1(1;n)) \\
&\quad + I(X_3(1;n); \hat{X}_3(1;n)|\hat{X}_1(1;n)\hat{X}_2(1;n)) \\
&\quad = \; I(X_1(1;n)X_2(1;n)X_3(1;n); \hat{X}_1(1;n)\hat{X}_2(1;n)\hat{X}_3(1;n)) \\
&\quad \geq \; I(X_2(1;n)X_3(1;n); \hat{X}_2(1;n)\hat{X}_3(1;n)) \\
&\quad = \; I(X_2(1;n)X_3(1;n); \hat{X}_2(1;n)) + I(X_2(1;n)X_3(1;n); \hat{X}_3(1;n)|\hat{X}_2(1;n)) \\
&\quad \overset{1)}{=} \; I(X_2(1;n); \hat{X}_2(1;n)) + I(X_2(1;n)X_3(1;n); \hat{X}_3(1;n)|\hat{X}_2(1;n)) \\
&\quad \geq \; I(X_2(1;n); \hat{X}_2(1;n)) + I(X_3(1;n); \hat{X}_3(1;n)|\hat{X}_2(1;n))
\end{aligned} \tag{4.35}$$

where the equality 1) follows from the fact that the requirement (R6) plus the Markov condition $X_1 \rightarrow X_2 \rightarrow X_3$ implies that the Markov condition $\hat{X}_2(1;n) \rightarrow X_2(1;n) \rightarrow X_3(1;n)$ is satisfied. In (4.35), the Markov condition $X_2(1;n) \rightarrow \hat{X}_2(1;n)X_3(1;n) \rightarrow \hat{X}_3(1;n)$ may not be valid. However, to overcome this problem, we can use the the same technique as in the proof of the converse part of Theorem 1 and also in the proof of Lemma 5 to construct a new random vector $\tilde{X}_3(1;n)$ such that the following hold:

- $(X_3(1;n), \hat{X}_2(1;n), \hat{X}_3(1;n))$ and $(X_3(1;n), \hat{X}_2(1;n), \tilde{X}_3(1;n))$ have the same distribution, and

- the Markov condition $X_2(1;n) \to (X_3(1;n), \hat{X}_2(1;n)) \to \tilde{X}_3(1;n)$ is met.

Therefore, the random variables $\hat{X}_2(1;n)$ and $\tilde{X}_3(1;n)$ satisfy the requirements (R5) and (R6) with $N = 2$ with respect to $X_2(1;n)$ and $X_3(1;n)$. This, together with (4.35) and (3.8), implies

$$\begin{aligned} & I(X_1(1;n); \hat{X}_1(1;n)) + I(X_1(1;n)X_2(1;n); \hat{X}_2(1;n)|\hat{X}_1(1;n)) \\ & \quad + I(X_3(1;n); \hat{X}_3(1;n)|\hat{X}_1(1;n)\hat{X}_2(1;n)) \\ & \quad \geq \; I(X_2(1;n); \hat{X}_2(1;n)) + I(X_3(1;n); \tilde{X}_3(1;n)|\hat{X}_2(1;n)) \\ & \quad \geq \; nR_{c,n}^{X_2X_3}(D_2, D_3). \end{aligned} \tag{4.36}$$

Since (4.36) is valid for any auxiliary random variables $\hat{X}_i(1;n)$, $i = 1,2,3$, satisfying the requirements (R5) and (R6) with $N = 3$, (4.34) then follows from the definition (3.8). This completes the proof of (4.33) in Case (1).

To prove (4.33) in Case (2), note that both $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ are right continuous in the sense that for any $D_1 \geq 0, D_2 \geq 0, D_3 \geq 0$,

$$\lim_{\epsilon \downarrow 0} R_c^{X_1X_2X_3}(D_1 + \epsilon, D_2 + \epsilon, D_3 + \epsilon) = R_c^{X_1X_2X_3}(D_1, D_2, D_3)$$

and

$$\lim_{\epsilon \downarrow 0} R_c^{X_2X_3}(D_2 + \epsilon, D_3 + \epsilon) = R_c^{X_2X_3}(D_2, D_3).$$

The validity of (4.33) in Case (2) then follows from its validity in Case (1). This completes the proof of Theorem 9.

Theorem 9 is what one would expect and consistent with our intuition. Let us now look at the case where X_1, X_2, and X_3 do not form a Markov chain, and (X_1, X_2, X_3) is an IID vector source. Define for any i

$$D_{i,max} \overset{\Delta}{=} \min\{D_i : R_{X_i}(D_i) = 0\} \tag{4.37}$$

where $R_X(D)$, for any source X, is the classical rate distortion function of X. Assume that $D_{1,max} > 0$. In view of Theorem 7 and the proof of Lemma 5, both $R_c^{X_1X_2X_3}(D_{1,max}, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ are convex as functions of D_2 and D_3 over the region $\{(D_2, D_3) : D_2 \geq 0, D_3 \geq 0\}$. As such, they are subdifferentiable at any point (D_2, D_3) with $D_2 > 0$ and $D_3 > 0$. (See [34, Chapter 23] for discussions on the subdifferential and subgradients of a convex function.) From Section 4.1, they can also be computed via our iterative algorithm through their respective conjugates. Since (X_1, X_2, X_3) is an IID vector source, in view of Theorem 7, we will drop the subscript or superscript n for all notation in Section 4.1 with understanding of $n = 1$ throughout the rest of this section. Once again, to bring out the dependence of $F^*(s)$ on the source (X_1, X_2, X_3), we will write $F^*(s)$ for X_1, X_2, X_3 as $F^{X_1X_2X_3}(s)$, $F^*(s)$ for $(X_1X_2), X_3$ as $F^{(X_1X_2)X_3}(s)$—the notation (X_1X_2) means that (X_1X_2) is regarded as a super source (see Figure 4.3)—and $F^*(s)$ for X_2, X_3 as $F^{X_2X_3}(s)$. This convention will apply to other notation in Section 4.1 as well. In particular,

$$F^{X_2X_3}(s) = \inf\{R_c^{X_2X_3}(D_2, D_3) + \beta D_2 + \nu D_3 : D_2 \geq 0, D_3 \geq 0\} \tag{4.38}$$

for any $s = (\beta, \nu)$.

Condition A: A point (D_2, D_3) with $D_2 > 0$ and $D_3 > 0$ is said to satisfy Condition A if $R_c^{X_1X_2X_3}(D_{1,max}, D_2, D_3)$ as a function of D_2 and D_3 has a negative subgradient $-s = (-\beta, -\nu)$, $\beta > 0, \nu > 0$, at (D_2, D_3) such that there is a distribution $\mathbf{Q} = \{q(\hat{x}_2, \hat{x}_3) = q(\hat{x}_2)q(\hat{x}_3|\hat{x}_2) : \hat{x}_2 \in \hat{\mathcal{X}}_2, \hat{x}_3 \in \hat{\mathcal{X}}_3\}$ satisfying the following requirements:

(R11) $F_s^{X_2X_3}(\mathbf{P(Q)}, \mathbf{Q}) = F^{X_2X_3}(s)$.

(R12) Define (as in Step 2 of the iterative algorithm)

$$p_{(X_1X_2)X_3}(\hat{x}_3|\hat{x}_2x_3) \triangleq \frac{q(\hat{x}_3|\hat{x}_2)2^{-\nu d(x_3,\hat{x}_3)}}{\Delta(\hat{x}_2,x_3)} \tag{4.39}$$

and

$$p_{(X_1X_2)X_3}(\hat{x}_2|x_1x_2) \triangleq \frac{q(\hat{x}_2)2^{-\beta d(x_2,\hat{x}_2)}2^{\sum_{x_3} p(x_3|x_1x_2)\log\Delta(\hat{x}_2,x_3)}}{\Gamma(x_1,x_2)} \tag{4.40}$$

where

$$\Delta(\hat{x}_2,x_3) \triangleq \sum_{\hat{x}_3\in\hat{\mathcal{X}}_3} q(\hat{x}_3|\hat{x}_2)2^{-\nu d(x_3,\hat{x}_3)} \tag{4.41}$$

and

$$\Gamma(x_1,x_2) \triangleq \sum_{\hat{x}_2\in\hat{\mathcal{X}}_2} q(\hat{x}_2)2^{-\beta d(x_2,\hat{x}_2)}2^{\sum_{x_3} p(x_3|x_1x_2)\log\Delta(\hat{x}_2,x_3)}. \tag{4.42}$$

Denote the two conditional distributions $p_{(X_1X_2)X_3}(\hat{x}_3|\hat{x}_2x_3)$ and $p_{(X_1X_2)X_3}(\hat{x}_2|x_1x_2)$ by $\mathbf{P}_{(X_1X_2)X_3}\mathbf{(Q)}$. Then either

$$F^{(X_1X_2)X_3}(s) < F_s^{(X_1X_2)X_3}(\mathbf{P}_{(X_1X_2)X_3}\mathbf{(Q)},\mathbf{Q})$$

or $p_{(X_1X_2)X_3}(\cdot|x_1x_2)$ depends on x_1, i.e., there exist $\hat{x}_2$, x_2, x_1, and $x_1' \in \mathcal{X}_1$ with $x_1' \neq x_1$ such that

$$p_{(X_1X_2)X_3}(\hat{x}_2|x_1x_2) \neq p_{(X_1X_2)X_3}(\hat{x}_2|x_1'x_2).$$

We are now ready to state a somewhat surprising result dubbed the more and less coding theorem.

Theorem 10 (More and less coding theorem) *Suppose that (X_1,X_2,X_3) is an IID vector source with $D_{1,max} > 0$, and X_1, X_2, and X_3 do not form a Markov chain. Then for any point (D_2,D_3), $D_2 > 0, D_3 > 0$, satisfying Condition A, there is a critical value $D_1^* < D_{1,max}$ such that for any $D_1 > D_1^*$,*

$$R_c^{X_1X_2X_3}(D_1,D_2,D_3) < R_c^{X_2X_3}(D_2,D_3) \tag{4.43}$$

and for any $D_1 < D_1^*$,

$$R_c^{X_1X_2X_3}(D_1, D_2, D_3) > R_c^{X_2X_3}(D_2, D_3). \tag{4.44}$$

Remark 8 *In Theorem 10, if $D_1^* > 0$, then at D_1^*,*

$$R_c^{X_1X_2X_3}(D_1^*, D_2, D_3) = R_c^{X_2X_3}(D_2, D_3).$$

Proof of Theorem 10: Since $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ as a function of D_1 is continuous over $D_1 > 0$ and non-increasing, it suffices to show that

$$R_c^{X_1X_2X_3}(D_{1,max}, D_2, D_3) < R_c^{X_2X_3}(D_2, D_3) \tag{4.45}$$

for any point (D_2, D_3), $D_2 > 0, D_3 > 0$, satisfying Condition A. To this end, we consider a new two-layer causal coding model shown in Figure 4.3, where X_1 and X_2 together are regarded as one super source. Let $R_c^{(X_1X_2)X_3}(D_2, D_3)$ denote its minimum total rate function. Since at $D_{1,max}$, $R_{X_1}(D_{1,max}) = 0$, a random variable $\hat{X}_1(1)$ independent of $X_1(1)$, $X_2(1)$, and $X_3(1)$ can be constructed in such a way that $Ed_1(X_1(1), \hat{X}_1(1)) = D_{1,max}$. Therefore, it is easy to see that

$$R_c^{X_1X_2X_3}(D_{1,max}, D_2, D_3) \leq R_c^{(X_1X_2)X_3}(D_2, D_3) \tag{4.46}$$

for any $D_2 \geq 0$ and $D_3 \geq 0$. On the other hand, in view of the definition of causal vide codes, it is not hard to see that any causal code for encoding X_1, X_2, and X_3 with respective distortions $D_1, D_2,$ and D_3 can also be used for encoding (X_1X_2) and X_3 in Fig 4.3 with distortions $D_2,$ and D_3 without changing the total rate. Thus

$$R_c^{X_1X_2X_3}(D_1, D_2, D_3) \geq R_c^{(X_1X_2)X_3}(D_2, D_3)$$

for any $D_1, D_2, D_3 \geq 0$. This, coupled with (4.46), implies

$$R_c^{X_1X_2X_3}(D_{1,max}, D_2, D_3) = R_c^{(X_1X_2)X_3}(D_2, D_3) \tag{4.47}$$

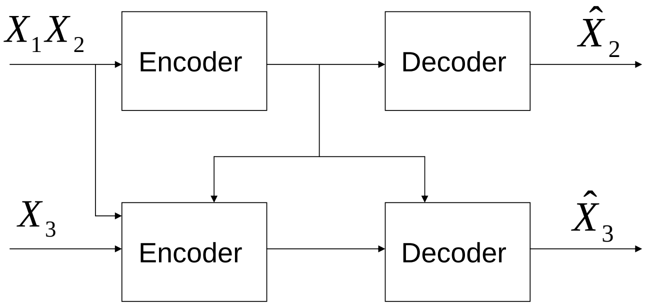

Figure 4.3: One special case of two-layer causal coding.

for any $D_2 \geq 0$ and $D_3 \geq 0$.

To continue, we are now led to show

$$R_c^{(X_1X_2)X_3}(D_2, D_3) < R_c^{X_2X_3}(D_2, D_3) \tag{4.48}$$

for any point (D_2, D_3), $D_2 > 0, D_3 > 0$, satisfying Condition A. First note that from the definition of causal video codes,

$$R_c^{(X_1X_2)X_3}(D_2, D_3) \leq R_c^{X_2X_3}(D_2, D_3) \tag{4.49}$$

for any $D_2 \geq 0$ and $D_3 \geq 0$. Fix now any point (D_2, D_3), $D_2 > 0, D_3 > 0$, satisfying Condition A. We prove (4.48) by contradiction. Suppose that

$$R_c^{(X_1X_2)X_3}(D_2, D_3) = R_c^{X_2X_3}(D_2, D_3) \tag{4.50}$$

at the point (D_2, D_3). Let $-s = (-\beta, -\nu)$ be the negative subgradient of $R_c^{X_1X_2X_3}(D_{1,max}, D_2, D_3)$ at the point (D_2, D_3) in Condition A. From (4.47), $-s$ is also a negative subgradient of $R_c^{(X_1X_2)X_3}(D_2, D_3)$ at the point (D_2, D_3). This implies that for any $D_2' \geq 0$ and $D_3' \geq 0$,

$$R_c^{(X_1X_2)X_3}(D_2', D_3') \geq R_c^{(X_1X_2)X_3}(D_2, D_3) - \beta(D_2' - D_2) - \nu(D_3' - D_3)$$

which, coupled with (4.50) and (4.49), in turn implies

$$R_c^{X_2X_3}(D_2', D_3') \geq R_c^{X_2X_3}(D_2, D_3) - \beta(D_2' - D_2) - \nu(D_3' - D_3)$$

for any $D_2' \geq 0$ and $D_3' \geq 0$. In other words, under the assumption (4.50), $-s$ is also a negative subgradient of $R_c^{X_2X_3}(D_2, D_3)$ at the point (D_2, D_3). In view of (4.25), (4.26), and (4.38), it then follows that

$$R_c^{(X_1X_2)X_3}(D_2, D_3) = F^{(X_1X_2)X_3}(s) - \beta D_2 - \nu D_3 \tag{4.51}$$

and

$$R_c^{X_2X_3}(D_2, D_3) = F^{X_2X_3}(s) - \beta D_2 - \nu D_3. \tag{4.52}$$

In view of the requirement (R11) in Condition A, we have

$$F^{X_2X_3}(s) = F_s^{X_2X_3}(\mathbf{P(Q)}, \mathbf{Q}). \tag{4.53}$$

From Step 2 of the iterative algorithm, it follows that

$$\begin{aligned} F^{(X_1X_2)X_3}(s) &\leq F_s^{(X_1X_2)X_3}(\mathbf{P}_{(X_1X_2)X_3}\mathbf{(Q)}, \mathbf{Q}) && (4.54) \\ &\leq F_s^{X_2X_3}(\mathbf{P(Q)}, \mathbf{Q}) && (4.55) \end{aligned}$$

where the inequality in (4.55) is strict when $\mathbf{P}_{(X_1X_2)X_3}\mathbf{(Q)}$ depends on X_1. Therefore, according to the requirement (R12) in Condition A, no matter which choice in the requirement (R12) is valid, we always have

$$F^{(X_1X_2)X_3}(s) < F_s^{X_2X_3}(\mathbf{P(Q)}, \mathbf{Q})$$

which, together with (4.51) to (4.53), implies that

$$R_c^{(X_1X_2)X_3}(D_2, D_3) < R_c^{X_2X_3}(D_2, D_3).$$

This contradicts the assumption (4.50), hence completing the proof of (4.48) and (4.45).

Define

$$D_1^* = \min\{D_1 : R_c^{X_1X_2X_3}(D_1, D_2, D_3) < R_c^{X_2X_3}(D_2, D_3)\}.$$

Then from (4.45), it is easy to see that D_1^* is the desired critical value. This completes the proof of Theorem 10.

Remark 9 *Theorem 10, in particular,* (4.43) *is really counter intuitive. It says that whenever the conditions specified in Theorem 10 are met, the more source frames need to be encoded and transmitted, the less amount of data after encoding has to be actually sent! If the cost of data transmission is proportional to the transmitted data volume, this translates literally into a scenario where the more frames you download, the less you would pay. To help the reader better understand this phenomenon, let us examine where the gain of* $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ *over* $R_c^{X_2X_3}(D_2, D_3)$ *comes from whenever the conditions specified in Theorem 10 are met. The availability of* X_1 *to the encoder of* X_2 *does not really help the encoder of* X_2 *and its corresponding decoder achieve a better rate distortion trade-off* (R_2, D_2)*. Likewise, the availability of* X_1 *and* X_2 *to the encoder of* X_3 *does not really help the encoder of* X_3 *and its corresponding decoder achieve a better rate distortion trade-off* (R_3, D_3) *either. What really matters is that the availability of* X_1 *to the encoder of* X_2 *will help the encoder of* X_2 *choose better side information* $\hat{X}_2$ *for the encoder and decoder of* X_3*. If the rate reduction of the encoder of* X_3 *arising from this better* $\hat{X}_2$ *along with* $\hat{X}_1$ *is more than the overhead associated with the rate* R_1 *and the selection of this better* $\hat{X}_2$*, then the total rate* $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ *is smaller. (Here the overhead associated with the rate* R_1 *and the selection of this better* $\hat{X}_2$ *is meant to be the difference between the sum of*

R_1 and R_2 in $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and the rate R_2 in $R_c^{X_2X_3}(D_2, D_3)$. Depending on how helpful $\hat{X}_1$ is, the rate R_2 in $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ can be more or less than the rate R_2 in $R_c^{X_2X_3}(D_2, D_3)$.) This is further confirmed in Examples 1 and 2 at the end of this section.

Condition A is generally met at points (D_2, D_3), $D_2 > 0, D_3 > 0$, for which positive bit rates are needed at both the decoder for X_2 and the decoder for X_3 in order for them to produce the respective reproductions with the desired distortions D_2 and D_3. Such distortion points will be called points with positive rates. By using the technique demonstrated in the proof of Property 1 in [52], it can be shown that $R_c^{X_1X_2X_3}(D_{1,max}, D_2, D_3)$ has a negative subgradient at any point (D_2, D_3), $D_2 > 0, D_3 > 0$, with positive rates. In addition, the distribution $\mathbf{P}_{(X_1X_2)X_3}\mathbf{(Q)}$, if optimal, generally depends on X_1 (except for some corner cases) when X_1, X_2, and X_3 do not form a Markov chain. We illustrate this in the following theorem in the binary case.

Theorem 11 *Assume that $\mathcal{X}_i = \hat{\mathcal{X}}_i = \{0, 1\}$, $i = 1, 2, 3$, and the Hamming distortion measure is used. Let (X_1, X_2, X_3) be an IID vector source with $I(X_2(1); X_3(1)) > 0$. Suppose that X_1, X_2, and X_3 do not form a Markov chain. Then for $s = (\beta, \nu)$ with $\beta > 0$ and $\nu > 0$, if $(\boldsymbol{P}, \boldsymbol{Q})$ ($\boldsymbol{P} = (p_{\hat{X}_3|\hat{X}_2X_3}, p_{\hat{X}_2|X_1X_2})$ and $\boldsymbol{Q} = (q_{\hat{X}_2\hat{X}_3})$) achieves $F^{(X_1X_2)X_3}(s)$, i.e.,*

$$F_s^{(X_1X_2)X_3}(\boldsymbol{P}, \boldsymbol{Q}) = F^{(X_1X_2)X_3}(s) \tag{4.56}$$

then $p_{\hat{X}_2|X_1X_2}(\cdot|x_1x_2)$ depends on x_1, i.e., there exists x_2 such that the condition distributions $p_{\hat{X}_2|X_1X_2}(\cdot|x_1 = 0, x_2)$ and $p_{\hat{X}_2|X_1X_2}(\cdot|x_1 = 1, x_2)$ are different.

Proof of Theorem 11: Fix $s = (\beta, \nu)$ with $\beta > 0$ and $\nu > 0$. We first derive some bounds on $F^{(X_1X_2)X_3}(s)$. It is not hard to verify that

$$\begin{aligned} F^{(X_1X_2)X_3}(s) &\leq F^{X_2X_3}(s) \\ &= \inf\{R_c^{X_2X_3}(D_2, D_3) + \beta D_2 + \nu D_3 : D_2 \geq 0, D_3 \geq 0\} \\ &\overset{1)}{\leq} \inf\{R_{X_2}(D_2) + \beta D_2 : D_2 \geq 0\} \\ &\quad + \inf\{R_{X_3}(D_3) + \nu D_3 : D_3 \geq 0\} \\ &= R_{X_2}(D_2(\beta)) + \beta D_2(\beta) \\ &\quad + R_{X_3}(D_3(\nu)) + \nu D_3(\nu) \end{aligned} \tag{4.57}$$

where $0 < D_2(\beta) < D_{2,max}$ is the unique value of D_2 at which the derivative of $R_{X_2}(D_2)$ is equal to β, and $0 < D_3(\nu) < D_{3,max}$ is the unique value of D_3 at which the derivative of $R_{X_3}(D_3)$ is equal to ν. In the above, the inequality 1) is due to the fact that

$$R_c^{X_2X_3}(D_2, D_3) \leq R_{X_2}(D_2) + R_{X_3}(D_3) \tag{4.58}$$

for any $D_2 \geq 0, D_3 \geq 0$. Under the condition that $I(X_2(1); X_3(1)) > 0$, the inequality (4.58) is strict at $(D_2(\beta), D_3(\nu))$. Therefore,

$$\begin{aligned} F^{(X_1X_2)X_3}(s) &\leq R_c^{X_2X_3}(D_2(\beta), D_3(\nu)) + \beta D_2(\beta) + \nu D_3(\nu) \\ &< R_{X_2}(D_2(\beta)) + \beta D_2(\beta) \\ &\quad + R_{X_3}(D_3(\nu)) + \nu D_3(\nu). \end{aligned} \tag{4.59}$$

In view of (4.56), it follows from the iterative algorithm that

$$\mathbf{P} = \mathbf{P}_{(X_1X_2)X_3}(\mathbf{Q}) \tag{4.60}$$

and

$$\mathbf{Q} = \mathbf{Q}_{(X_1X_2)X_3}(\mathbf{P}) \tag{4.61}$$

where $(X_1X_2)X_3$ appears as subscripts to indicate that the operations $\mathbf{P(Q)}$ and $\mathbf{Q(P)}$ defined in Section 4.1 are for the sources (X_1X_2) and X_3. Let $(\hat{X}_2(1), \hat{X}_3(1))$ be the output of the channel $\mathbf{P}$ in response to the input $((X_1(1)X_2(1)), X_3(1))$. Then the joint distribution of $(\hat{X}_2(1), \hat{X}_3(1))$ is $\mathbf{Q}$, and (4.56) implies

$$\begin{aligned} F^{(X_1X_2)X_3}(s) &= I(X_1(1)X_2(1); \hat{X}_2(1)) + I(X_3(1); \hat{X}_3(1)|\hat{X}_2(1)) \\ &\quad + \beta E[d_2(X_2(1), \hat{X}_2(1))] + \nu E[d_3(X_3(1), \hat{X}_3(1))]. \end{aligned} \tag{4.62}$$

Putting (4.62) and (4.59) together, we can conclude that $H(\hat{X}_2) > 0$ and hence $q_{\hat{X}_2}(\hat{x}_2) > 0$ for any $\hat{x}_2$. Otherwise, from (4.62) we would have that

$$\begin{aligned} &F^{(X_1X_2)X_3}(s) \\ &\quad \geq \beta D_{2,max} + I(X_3(1); \hat{X}_3(1)) + \nu E[d_3(X_3(1), \hat{X}_3(1))] \\ &\quad \geq \beta D_{2,max} + R_{X_3}(D_3(\nu)) + \nu D_3(\nu) \\ &\quad > R_{X_2}(D_2(\beta)) + \beta D_2(\beta) \\ &\qquad + R_{X_3}(D_3(\nu)) + \nu D_3(\nu) \end{aligned} \tag{4.63}$$

which contradicts (4.59).

We now prove Theorem 11 by contradiction. Suppose that $p_{\hat{X}_2|X_1X_2}(\cdot|x_1x_2)$ does not depend on x_1. Then for any x_2 and $\hat{x}_2$,

$$p_{\hat{X}_2|X_1X_2}(\hat{x}_2|x_1 = 0, x_2) = p_{\hat{X}_2|X_1X_2}(\hat{x}_2|x_1 = 1, x_2) \tag{4.64}$$

which, together with (4.60), (4.39) to (4.42), and the fact that $q_{\hat{X}_2}(\hat{x}_2) > 0$, implies

$$\begin{aligned} &\sum_{\hat{x}_2^*} q_{\hat{X}_2}(\hat{x}_2^*) 2^{-\beta d(x_2, \hat{x}_2^*)} 2^{\sum_{x_3} p(x_3|x_1=0,x_2) \log \frac{\Delta(\hat{x}_2^*, x_3)}{\Delta(\hat{x}_2, x_3)}} = \\ &\sum_{\hat{x}_2^*} q_{\hat{X}_2}(\hat{x}_2^*) 2^{-\beta d(x_2, \hat{x}_2^*)} 2^{\sum_{x_3} p(x_3|x_1=1,x_2) \log \frac{\Delta(\hat{x}_2^*, x_3)}{\Delta(\hat{x}_2, x_3)}} \end{aligned} \tag{4.65}$$

Simplifying (4.65) yields

$$\sum_{x_3} p(x_3|x_1 = 0, x_2) \log \frac{\Delta(\hat{x}_2', x_3)}{\Delta(\hat{x}_2, x_3)} = \sum_{x_3} p(x_3|x_1 = 1, x_2) \log \frac{\Delta(\hat{x}_2', x_3)}{\Delta(\hat{x}_2, x_3)} \tag{4.66}$$

where $\hat{x}_2' = 1 - \hat{x}_2$.

To continue, we now consider specific values of x_2 and $\hat{x}_2$. Let us first look at the case of $x_2 = 0$ and $\hat{x}_2 = 0$. It follows from (4.66) that

$$\sum_{x_3} [p(x_3|x_1 = 0, x_2 = 0) - p(x_3|x_1 = 1, x_2 = 0)] \log \frac{\Delta(\hat{x}_2^* = 1, x_3)}{\Delta(\hat{x}_2 = 0, x_3)} = 0 \tag{4.67}$$

which implies

$$\begin{aligned} &[p(x_3 = 0|x_1 = 0, x_2 = 0) - p(x_3 = 0|x_1 = 1, x_2 = 0)] \log \frac{b + (1-b)2^{-\nu}}{a + (1-a)2^{-\nu}} + \\ &[p(x_3 = 1|x_1 = 0, x_2 = 0) - p(x_3 = 1|x_1 = 1, x_2 = 0)] \log \frac{b2^{-\nu} + (1-b)}{a2^{-\nu} + (1-a)} = 0 \end{aligned} \tag{4.68}$$

where $a = q_{\hat{X}_2\hat{X}_3}(\hat{x}_3 = 0|\hat{x}_2 = 0)$, and $b = q_{\hat{X}_2\hat{X}_3}(\hat{x}_3 = 0|\hat{x}_2 = 1)$. Further simplifying (4.68) yields

$$\begin{aligned} &[p(x_3 = 0|x_1 = 0, x_2 = 0) - p(x_3 = 0|x_1 = 1, x_2 = 0)] \times \\ &\log \frac{(b + (1-b)2^{-\nu})(a2^{-\nu} + (1-a))}{(a + (1-a)2^{-\nu})(b2^{-\nu} + (1-b))} = 0. \end{aligned} \tag{4.69}$$

Since $\nu > 0$, it can be verified that $\log \frac{(b+(1-b)2^{-\nu})(a2^{-\nu}+(1-a))}{(a+(1-a)2^{-\nu})(b2^{-\nu}+(1-b))}$ is equal to 0 if and only if $a = b$.

Next we show that $a \neq b$. To this end, first note that $a = b$ is equivalent to saying that $\mathbf{Q} = (q_{\hat{X}_2\hat{X}_3})$ is a product distribution, i.e.,

$$q_{\hat{X}_2\hat{X}_3} = q_{\hat{X}_2} \times q_{\hat{X}_3} \tag{4.70}$$

By plugging (4.70) into (4.60), it follows from the Step 2 of the iterative algorithm that in $\mathbf{P} = (p_{\hat{X}_3|\hat{X}_2X_3}, p_{\hat{X}_2|X_1X_2})$, $p_{\hat{X}_3|\hat{X}_2X_3}(\cdot|\hat{x}_2, x_3)$ does not depend on $\hat{x}_2$ and $p_{\hat{X}_2|X_1X_2}(\cdot|x_1, x_2)$

does not depend on x_1, i.e.,

$$\begin{aligned} p_{\hat{X}_3|\hat{X}_2 X_3}(\hat{x}_3|\hat{x}_2, x_3) &= p(\hat{x}_3|x_3) \stackrel{\Delta}{=} \frac{q_{\hat{X}_3}(\hat{x}_3) 2^{-\nu d(x_3,\hat{x}_3)}}{\Delta(x_3)} \\ p_{\hat{X}_2|X_1 X_2}(\hat{x}_2|x_1, x_2) &= p(\hat{x}_2|x_2) \stackrel{\Delta}{=} \frac{q_{\hat{X}_2}(\hat{x}_2) 2^{-\beta d(x_2,\hat{x}_2)}}{\Gamma(x_2)} \end{aligned} \tag{4.71}$$

where $\Delta(x_3)$ and $\Gamma(x_2)$ are the normalization factors so that the respective terms are indeed distributions. It is easy to see that (4.70) and (4.71) imply

$$I(\hat{X}_2(1); \hat{X}_3(1)) = 0 \tag{4.72}$$

$$I(X_1(1)X_2(1); \hat{X}_2(1)) = I(X_2(1); \hat{X}_2(1)) \tag{4.73}$$

and

$$H(\hat{X}_3(1)|X_3(1)\hat{X}_2(1)) = H(\hat{X}_3(1)|X_3(1)). \tag{4.74}$$

Combining (4.72) to (4.74) with (4.62) yields

$$\begin{aligned} F^{(X_1 X_2) X_3}(s) &= I(X_2(1); \hat{X}_2(1)) + \beta E[d_2(X_2(1), \hat{X}_2(1))] \\ &\quad + I(X_3(1); \hat{X}_3(1)) + \nu E[d_3(X_3(1), \hat{X}_3(1))] \\ &\geq R_{X_2}(D_2(\beta)) + \beta D_2(\beta) \\ &\quad + R_{X_3}(D_3(\nu)) + \nu D_3(\nu) \end{aligned}$$

which contradicts (4.59). Therefore, $a \neq b$.

Go back to (4.69). Since $a \neq b$, (4.69) is equivalent to

$$p(x_3|x_1 = 0, x_2 = 0) = p(x_3|x_1 = 1, x_2 = 0) \text{ for any } x_3 \in \{0, 1\}. \tag{4.75}$$

Repeat the above argument for the case of $x_2 = 1$ and $\hat{x}_2 = 0$. We then have accordingly

$$p(x_3|x_1 = 0, x_2 = 1) = p(x_3|x_1 = 1, x_2 = 1) \text{ for any } x_3 \in \{0, 1\}. \tag{4.76}$$

Putting (4.75) and (4.76) together, we have shown that (4.64) implies that X_1, X_2, and X_3 form a Markov chain, which contradicts our assumption. This completes the proof of Theorem 11.

Remark 10 *From Theorem 11, it follows that for any sources X_1, X_2, and X_3 satisfying the conditions of Theorem 11, Condition A is met at any point (D_2, D_3), $D_2 > 0, D_3 > 0$, at which $R_c^{X_1X_2X_3}(D_{1,max}, D_2, D_3)$ has a negative subgradient.*

We conclude this section with examples illustrating Theorem 10.

Example 3: Suppose that $\mathcal{X}_i = \hat{\mathcal{X}}_i = \{0, 1\}$, $i = 1, 2, 3$, and that the Hamming distortion measure is used. Let $p_{X_1}(0) = 1/3$, $p_{X_2|X_1}(0|1) = p_{X_2|X_1}(1|0) = 3/5$, and

$$p_{X_3|X_1X_2} = \left(\begin{array}{cccccc} & X_1X_2 & 00 & 01 & 10 & 11 \\ X_3 & & & & & \\ 0 & & 0.97 & 0.03 & 0.03 & 0.97 \\ 1 & & 0.03 & 0.97 & 0.97 & 0.03 \end{array} \right).$$

It is easy to see that X_1, X_2 and X_3 do not form a Markov chain. We consider the following three cases:

Case 1: $D_1 = 0.31 < D_{1,max}$, and $D_3 = 0.15$;

Case 2: $D_2 = 0.20 < D_{2,max}$, and $D_3 = 0.15$; and

Case 3: $D_2 = 0.22 < D_{2,max}$, and $D_3 = 0.23$.

For Case 1, Figure 4.4 shows the rate-distortion curves of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_2. Over the interval of D_2 shown in Figure 4.4, it is clear that $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ is always strictly less than $R_c^{X_2X_3}(D_2, D_3)$.

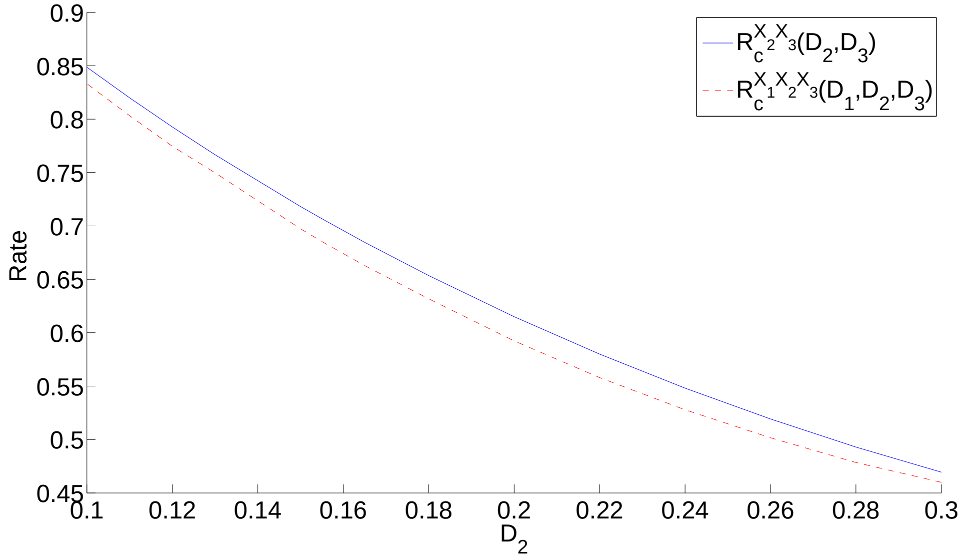

Figure 4.4: Comparison of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_2 for fixed $D_1 = 0.31$ and $D_3 = 0.15$.

For Case 2, Figure 4.5 shows $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1 with fixed $D_2 < D_{2,max}$ and D_3. It is observed that the critical point at which $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ meets $R_c^{X_2X_3}(D_2, D_3)$ is the intersection of the two curves. Denote this critical point by D_1^*. Then it is clear that when $D_1 > D_1^*$, $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ is indeed strictly less than $R_c^{X_2X_3}(D_2, D_3)$. Table 4.1 shows the rate allocation across different encoders in both cases of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ for several sample values of D_1, where R_i, $i = 1, 2, 3$, represents the rate allocated to the encoder of X_i in both cases, and $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ are denoted as $R_c^{X_1X_2X_3}$ and $R_c^{X_2X_3}$, respectively to save space. It is clear from Table 4.1 that the allocated rates confirm the explanation mentioned in Remark 9.

When we assign different values to $D_2 < D_{2,max}$ and D_3, we observe the same phe-

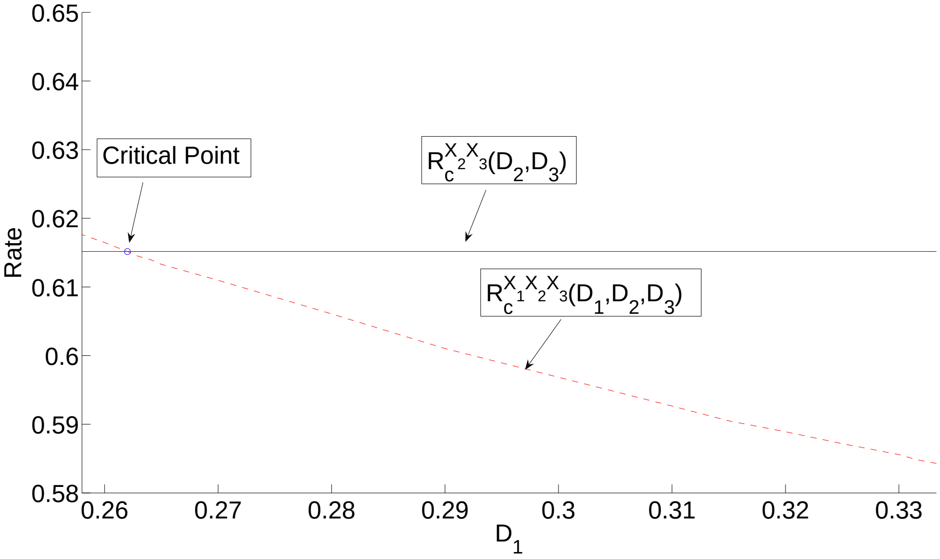

Figure 4.5: Comparison of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1 for fixed $D_2 = 0.20$ and $D_3 = 0.15$ in Example 3.

nomenon, as shown again in Figure 4.6 and Table 4.2 for Case 3.

Let us now look at another example with a different joint distribution.

Example 4: Suppose that $\mathcal{X}_i = \hat{\mathcal{X}}_i = \{0, 1\}$, $i = 1, 2, 3$, and that the Hamming distortion measure is used. Let $p_{X_1}(0) = 1/10$, $p_{X_2|X_1}(0|1) = p_{X_2|X_1}(1|0) = 1/10$, and

$$p_{X_3|X_1X_2} = \left(\begin{array}{cccccc} & X_1X_2 & 00 & 01 & 10 & 11 \\ X_3 & & & & & \\ 0 & & 0.80 & 0.05 & 0.1 & 0.92 \\ 1 & & 0.20 & 0.95 & 0.9 & 0.08 \end{array} \right).$$

Once again, X_1, X_2 and X_3 do not form a Markov chain. Fix $D_2 = 0.0988 < D_{2,max}$ and $D_3 = 0.0911$. Figure 4.7 shows the two rate distortion curves $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1, and Table 4.3 lists their respective rate allocations for several

Table 4.1: Rate allocation of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1 for fixed $D_2 = 0.20$ and $D_3 = 0.15$ in Example 3.

$D_1 = 0.3300$	$R_1 = 0.0035$	$R_2 = 0.3546$	$R_3 = 0.2274$	$R_c^{X_1X_2X_3} = 0.5854$
$D_1 = 0.3000$	$R_1 = 0.0369$	$R_2 = 0.3855$	$R_3 = 0.1737$	$R_c^{X_1X_2X_3} = 0.5961$
$D_1 = 0.2700$	$R_1 = 0.0769$	$R_2 = 0.4224$	$R_3 = 0.1111$	$R_c^{X_1X_2X_3} = 0.6103$
$D_1 = 0.2620$ (critical point)	$R_1 = 0.0886$	$R_2 = 0.4327$	$R_3 = 0.0939$	$R_c^{X_1X_2X_3} = 0.6150$
$D_1 = 0.2600$	$R_1 = 0.0916$	$R_2 = 0.4425$	$R_3 = 0.0821$	$R_c^{X_1X_2X_3} = 0.6160$

N/A	N/A	$R_2 = 0.2748$	$R_3 = 0.3402$	$R_c^{X_2X_3} = 0.6150$

Table 4.2: Rate allocation of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1 for fixed $D_2 = 0.22$ and $D_3 = 0.23$ in Example 3.

$D_1 = 0.3222$	$R_1 = 0.0115$	$R_2 = 0.3138$	$R_3 = 0.0602$	$R_c^{X_1X_2X_3} = 0.3855$
$D_1 = 0.3093$	$R_1 = 0.0260$	$R_2 = 0.3170$	$R_3 = 0.0486$	$R_c^{X_1X_2X_3} = 0.3915$
$D_1 = 0.2944$	$R_1 = 0.0440$	$R_2 = 0.3203$	$R_3 = 0.0359$	$R_c^{X_1X_2X_3} = 0.4002$
$D_1 = 0.2771$ (critical point)	$R_1 = 0.0668$	$R_2 = 0.3237$	$R_3 = 0.0212$	$R_c^{X_1X_2X_3} = 0.4119$
$D_1 = 0.2730$	$R_1 = 0.0726$	$R_2 = 0.3243$	$R_3 = 0.0180$	$R_c^{X_1X_2X_3} = 0.4149$

N/A	N/A	$R_2 = 0.2366$	$R_3 = 0.1753$	$R_c^{X_2X_3} = 0.4119$

sample values of D_1. The same phenomenon is revealed as in Example 3 .

For all cases shown in Examples 1 and 2, in comparison with $R_c^{X_2X_3}(D_2, D_3)$, when we include X_1 in the encoding and transmission, we not only get the reconstruction of X_1 (with distortion $\geq D_1^*$) free at the receiver end, but are also able to reduce the total number of bits to be transmitted. In other words, we can achieve a double gain.

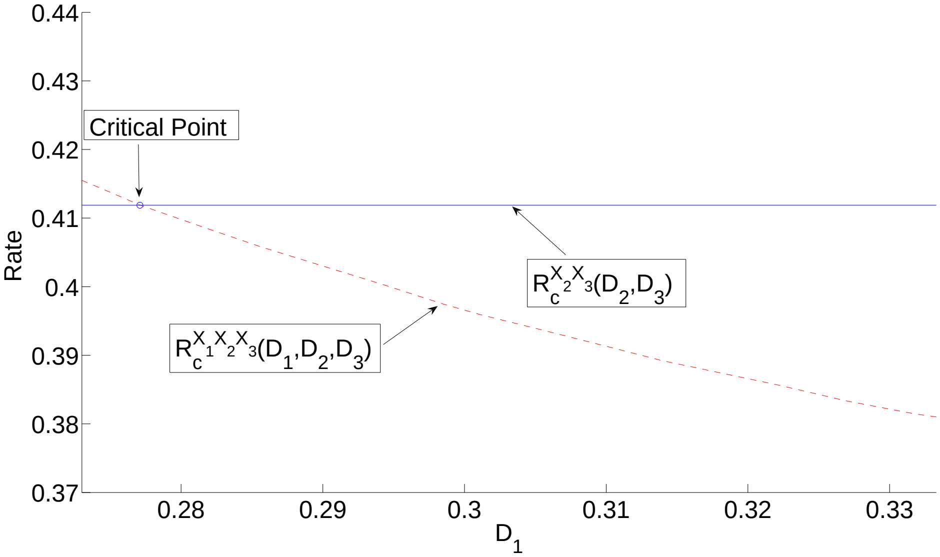

Figure 4.6: Comparison of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1 for fixed $D_2 = 0.22$ and $D_3 = 0.23$ in Example 3.

Table 4.3: Rate allocation of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1 for fixed $D_2 = 0.0988$ and $D_3 = 0.0911$ in Example 4.

$D_1 = 0.1000$	$R_1 = 1.3265 \times 10^{-7}$	$R_2 = 0.2412$	$R_3 = 0.1311$	$R_c^{X_1X_2X_3} = 0.3722$
$D_1 = 0.0977$	$R_1 = 0.0074$	$R_2 = 0.2378$	$R_3 = 0.1307$	$R_c^{X_1X_2X_3} = 0.3758$
$D_1 = 0.0922$	$R_1 = 0.0252$	$R_2 = 0.2281$	$R_3 = 0.1303$	$R_c^{X_1X_2X_3} = 0.3836$
$D_1 = 0.0872$ (critical point)	$R_1 = 0.0421$	$R_2 = 0.2191$	$R_3 = 0.1303$	$R_c^{X_1X_2X_3} = 0.3914$
$D_1 = 0.0849$	$R_1 = 0.0495$	$R_2 = 0.2150$	$R_3 = 0.1303$	$R_c^{X_1X_2X_3} = 0.3949$
N/A	N/A	$R_2 = 0.2150$	$R_3 = 0.1764$	$R_c^{X_2X_3} = 0.3914$

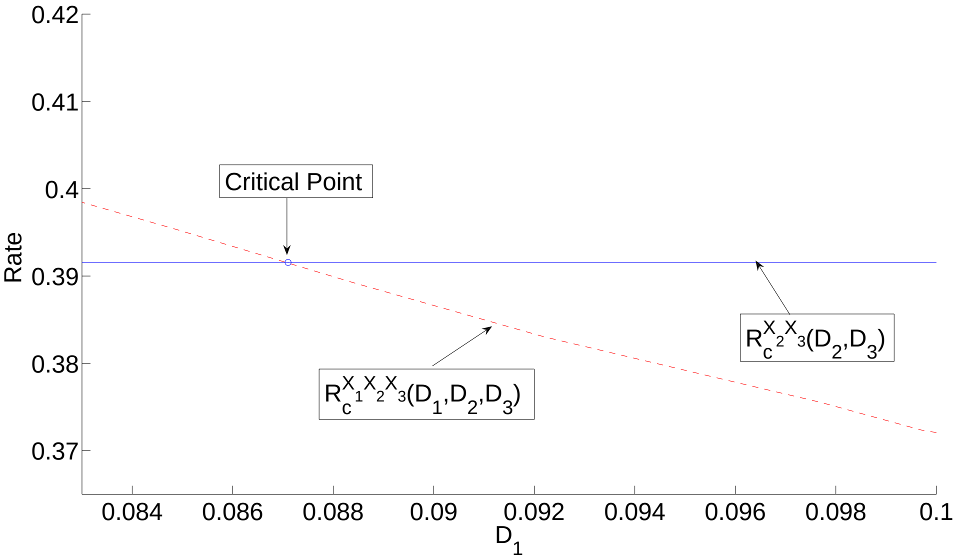

Figure 4.7: Comparison of $R_c^{X_1X_2X_3}(D_1, D_2, D_3)$ and $R_c^{X_2X_3}(D_2, D_3)$ versus D_1 for fixed $D_2 = 0.0988$ and $D_3 = 0.0911$ in Example 4.

Chapter 5

Information Theoretic Performance Comparison of Causal Video Coding and Predictive Video Coding

Recall Figure 1.3 for the paradigm of CVC and PVC. From their respective definitions, it follows that CVC includes PVC as a special case where the original video frames $X_1, \cdots, X_{k-1}$ are discarded at the encoder. In other words, these original video frames can be regarded as the side information available only to the encoder of X_k. The loss of access to original video frames in PVC has important implications. For example, despite the fact that PVC is widely used in practice, the single-letter characterization of the minimum total rate $R_p^*(D_1, \cdots, D_N)$ of PVC required to achieve a given distortion level $D_1, \cdots, D_N > 0$ in the usual information theoretic sense (see [54]), if any, is unknown in general, let alone any algorithm to actually compute it. These are in contrast to CVC, of which not only the total rate distortion function $R_c^*(D_1, \cdots, D_N)$ has a single-letter characterization, but also it is computable. The fact that the total rate distortion function $R_p^*(D_1, \cdots, D_N)$ of PVC is not computable makes the comparison between CVC and PVC difficult and technically challenging. In this chapter, our purpose is to address this issue and provide some results comparing the rate distortion performance of PVC against that of CVC. Note that such comparison is of particular interest in the practice of video compression: if $R_c^*(D_1, \cdots, D_N)$ is strictly smaller than $R_p^*(D_1, \cdots, D_N)$, then it implies a possible paradigm shift from PVC to CVC in the design of future video coding systems and standards.

Before presenting our findings, let us review some relevant results on source coding with encoder only side information. As shown in [2, Ch. 6, Case 2, p. 180], in the case where there are only one encoder and one decoder, the best rate distortion performance achievable asymptotically does not improve with encoder only side information. Conveniently, one might extrapolate that the encoder only side information in CVC does not help improve the total rate distortion function either, or in other words, $R_c^*(D_1, \cdots, D_N)$ is equal to $R_p^*(D_1, \cdots, D_N)$.

In this chapter, we show that such extrapolation is in general incorrect by identifying cases where $R_c^*(D_1, \cdots, D_N)$ is strictly less than $R_p^*(D_1, \cdots, D_N)$. Specifically, by fixing $N = 3$, we first show that for general stationary ergodic sources X_1, X_2, X_3 which do not form a (first-orer) Markov chain, the minimum total rate $R_{c,n}(D_1, D_2, D_3)$ of nth order causal video codes is always strictly less than the minimum total rate $R_{p,n}(D_1, D_2, D_3)$ of nth order predictive video codes for any finite $n > 0$ under mild conditions on source frames and distortions. We next establish a single-letter characterization of $R_p^*(D_1, D_2, D_3)$ for an IID vector source (X_1, X_2, X_3) where X_1 and X_2 are independent. Finally we provide a specific condition under which $R_c^*(D_1, D_2, D_3) < R_p^*(D_1, D_2, D_3)$ for binary sources, which implies that CVC is an attractive framework for future video coding systems and standards.

5.1 Information Theoretic Characterization for Predictive Video Coding

Before characterizing the rate-distortion performance for PVC, we first define an order-n predictive video code C_n for $X_1, \cdots, X_N$ by using a sequence of function pairs $(f_1, g_1), \cdots, (f_N, g_N)$ such that

$$
\begin{aligned}
&f_1 : \mathcal{X}_1^n \to \{0,1\}^*, \quad g_1 : \{0,1\}^* \to \hat{\mathcal{X}}_1^n, \\
&f_i : \mathcal{X}_j^n \times \overbrace{\{0,1\}^* \times \cdots \times \{0,1\}^*}^{i-1 \text{ times}} \to \{0,1\}^*, \text{ and} \\
&g_i : \overbrace{\{0,1\}^* \times \cdots \times \{0,1\}^*}^{i \text{ times}} \to \hat{\mathcal{X}}_i^n, \quad i = 2, \cdots, N
\end{aligned}
$$

where $\mathcal{X}_k^n$ and $\hat{\mathcal{X}}_k^n, k = 1, \cdots, N$, denote the n-fold product of $\mathcal{X}_k$ and $\hat{\mathcal{X}}_k$, respectively, and $\{0,1\}^*$ denotes the set of all finite binary strings. For an order-n predictive video code C_n, $S_1 = f_1(X_1(1;n))$, $S_i = f_i(X_i(1;n), S_i^-)$, $\hat{X}_1(1;n) = g_1(S_1)$, and $\hat{X}_i(1;n) = g_i(S_i^-, S_i)$.

To evaluate the rate distortion performance of C_n, we define the following quantities.

$$
D_{C_n,i} \triangleq \frac{1}{n}\mathbf{E}\sum_{j=1}^{n} d_i(X_i(j), \hat{X}_i(j)) \tag{5.1}
$$

$$
R_{C_n,i} \triangleq \frac{1}{n}\mathbf{E}|S_i|, \quad 1 \le i \le N \tag{5.2}
$$

where $|B|$ denotes the number of symbols in a string B.

Fix a vector source $(X_1, \cdots, X_N)$. Let $(R_1, \cdots, R_N)$ be a rate vector and $(D_1, \cdots, D_N)$ be a distortion vector. The rate distortion pair vector $(R_1, \cdots, R_N, D_1, \cdots, D_N)$ is said to be achievable by PVC if for any $\epsilon > 0$, there exists an order-n predictive video code C_n for all sufficiently large n such that

$$
D_{C_n,i} \le D_i + \epsilon, \text{ and } R_{C_n,i} \le R_i + \epsilon \tag{5.3}
$$

for $i = 1, \cdots, N$.

Let $\mathcal{R}_p^*$ denote the set of all rate distortion pair vectors $(R_1, \cdots, R_N, D_1, \cdots, D_N)$ achievable by predictive source coding. As in the practice of video compression, we are interested in the minimum total rate $R_p^*(D_1, \cdots, D_N)$ to achieve the distortion level $(D_1, \cdots, D_N)$ required by PVC. Specifically, $R_p^*(D_1, \cdots, D_N)$ is defined by

$$R_p^*(D_1, \cdots, D_N) \triangleq \min\Big\{\sum_{i=1}^{N} R_i\colon\ (R_1, \cdots, R_N) \in \mathcal{R}_p(D_1, \cdots, D_N)\Big\}. \tag{5.4}$$

In the following, we characterize the rate-distortion performance of PVC for general stationary and ergodic sources. To that end, let us define $\mathcal{R}_{p,n}$ to be the region consisting of all rate distortion pair vectors $(R_1, \cdots, R_N, D_1, \cdots, D_N)$ for which there exist auxiliary random variables U_k, $k = 1, 2, \cdots, N-1$, and $\hat{X}_N(1;n)$ such that

$$\begin{aligned} & R_1 \geq \frac{1}{n} I(X_1(1;n); U_1) \\ & R_k \geq \frac{1}{n} I(X_k(1;n); U_k | U_k^-) \\ & k = 2, 3, \cdots, N-1 \\ & R_N \geq \frac{1}{n} I(X_N(1;n); \hat{X}_N(1;n) | U_N^-) \end{aligned} \tag{5.5}$$

and the following requirements are satisfied:

(R13) $\hat{X}_1(1;n) = g_1(U_1)$ for some deterministic function g_1,

(R14) $\hat{X}_k(1;n) = g_k(U_k^-, U_k)$ for some deterministic function g_k, $k = 2, \cdots, N-1$,

(R15) for any $1 \leq k \leq N$, $\frac{1}{n} E[d_k(X_k(1;n), \hat{X}_k(1;n))] \leq D_k$, and

(R16) the Markov chain conditions $U_k \to (X_k(1;n), U_k^-) \to X_k^+(1;n)$, $k = 1, \cdots, N-1$, $U_j \to (X_j(1;n), U_j^-) \to (X_j^-(1;n), X_j^+(1;n))$, $j = 2, \cdots, N-1$, and $X_N^-(1;n) \to (X_N(1;n), U_N^-) \to \hat{X}_N(1;n)$ are met.

Let $\mathcal{R}'_p = \bigcup_{n=1}^{\infty} \mathcal{R}_{p,n}$. Denote its convex hull closure by $co(\mathcal{R}'_p)$. In parallel to Theorem 3, we have the following result for PVC. The proof of Theorem 12 is similar to that of Theorem 3 and is thus not reproduced here.

Theorem 12 *For general stationary and* ergodic *sources* $X_1, \cdots, X_N$, $\mathcal{R}^*_p = co(\mathcal{R}'_p)$.

To determine $R^*_p(D_1, \cdots, D_N)$ in terms of information quantities, we define

$$\begin{aligned} & R_{p,n}(D_1, \cdots, D_N) \\ = \; & \frac{1}{n} \inf[I(X_1(1;n); U_1) + \\ & \sum_{t=2}^{N-1} I(X_t(1;n); U_t | U_t^-) + \\ & I(X_N(1;n); \hat{X}_N(1;n) | U_N^-)] \end{aligned} \tag{5.6}$$

where the infimum is taken over all auxiliary random variables $U_1, \cdots, U_{N-1}$ and $\hat{X}_N(1;n)$ satisfying the requirements (R13) to (R16).

In parallel to Theorem 5, we have the following result.

Theorem 13 *For jointly stationary and* ergodic *sources* $X_1, \cdots, X_N$,

$$R^*_p(D_1, \cdots, D_N) = \inf\{R_{p,n}(D_1, \cdots, D_N) : n \geq 1\}$$

for any distortion level $D_1 > 0, \cdots, D_N > 0$.

To ease the subsequent discussions, we need the following lemma in parallel to Lemma 1.

Lemma 6 *The function* $R^*_p(D_1, \cdots, D_N)$ *is convex and hence continuous over the open region* $D_1 > 0, \cdots, D_N > 0$.

The proofs of Theorem 13 and Lemma 6 are similar to that of Theorem 5 and Lemma 5, respectively, and are thus omitted here.

5.2 Predictive vs. Causal

In this section, we compare CVC against PVC. Specifically, since it follows from their definitions that

$$R_c^*(D_1, \cdots, D_N) \le R_p^*(D_1, \cdots, D_N) \tag{5.7}$$

for any video source $X_1, \cdots, X_N$, our comparison is focused on identifying cases of interest in practice for which the inequality in Equation (5.7) becomes, if possible, equality or strict inequality.

Without losing generality, we consider only the case where $N = 3$, and write the three jointly stationary and ergodic sources $X_1(1;n)$, $X_2(1;n)$, $X_3(1;n)$ as X^n, Y^n, and Z^n respectively. Let $p_{X^nY^n}$ and $p_{X^nY^nZ^n}$ denote the joint distributions of random vectors (X^n, Y^n) and (X^n, Y^n, Z^n), respectively; and let p_{X^n} denote the marginal distribution of X^n. If there is no ambiguity, subscripts in distributions will be omitted. For example, we may write $p(x)$ instead of $p_X(x)$.

In the following discussions, $R_{p,n}(D_1, D_2, D_3)$ is defined in (5.6) when $N = 3$ and $R_{c,n}(D_1, D_2, D_3)$ can be rewritten as below[1]:

$$\min_{U_1, U_2, \hat{Z}^n} \Big[I(X^n; U_1) + I(X^nY^n; U_2|U_1) + I(Z^n; \hat{Z}^n|U_1U_2) \Big] \tag{5.8}$$

where the minimization is taken over all random variables $U_1 \in \mathcal{U}_1$, $U_2 \in \mathcal{U}_2$ and $\hat{Z}^n \in \hat{\mathcal{Z}}^n$ satisfying the following conditions: i) there exist a function g_1 such that $\hat{X}^n = g_1(U_1)$ and $\frac{1}{n}\mathbf{E}d_1(X^n, \hat{X}^n) \le D_1$; ii) there exist a function g_2 such that $\hat{Y}^n = g_2(U_1, U_2)$ and $\frac{1}{n}\mathbf{E}d_2(Y^n, \hat{Y}^n) \le D_2$; iii) $(Y^n, Z^n) \to X^n \to U_1$ is a Markov chain; iv) $Z^n \to (X^n, Y^n, U_1) \to U_2$ is a Markov chain; and v) $(X^n, Y^n) \to (Z^n, U_1, U_2) \to \hat{Z}^n$ is a Markov

[1] It follows from the proof of Theorem 2 that (5.8) is equal to the right-hand-side of (3.8) when $N = 3$.

chain. Define

$$F_{c,n}(s) \overset{\Delta}{=} \inf F_{c,n}(p_{U_1|X^n}, p_{U_2|U_1X^nY^n}, p_{\hat{Z}^n|U_1U_2Z^n}, Q_{U_1U_2\hat{Z}^n}) \tag{5.9}$$

for any $s = (\alpha, \beta, \nu)$, where $\alpha, \beta, \nu \geq 0$, denotes the standard Lagrange multipliers, and

$$\begin{aligned} & F_{c,n}(p_{U_1|X^n}, p_{U_2|U_1X^nY^n}, p_{\hat{Z}^n|U_1U_2Z^n}, Q_{U_1U_2\hat{Z}^n}) \\ \overset{\Delta}{=} & \sum_{x^n,u_1} P(x^n)P(u_1|x^n)\Bigg[\log \frac{P(u_1|x^n)}{Q(u_1)2^{-\alpha d_1(x^n,g_1(u_1))}} + \\ & \sum_{y^n,u_2} P(y^n|x^n)P(u_2|x^ny^nu_1)\Bigg[\log \frac{P(u_2|x^ny^nu_1)}{Q(u_2|u_1)2^{-\beta d_2(y^n,g_2(u_1,u_2))}} + \\ & \sum_{z^n,\hat{z}^n} P(z^n|x^ny^n)P(\hat{z}^n|z^nu_1^2)\log \frac{P(\hat{z}^n|z^nu_1^2)}{Q(\hat{z}^n|u_1^2)2^{-\nu d_3(z^n,\hat{z}^n)}}\Bigg]\Bigg]. \end{aligned} \tag{5.10}$$

It follows (4.25) and (4.26) that

$$\begin{aligned} \frac{F_{c,n}(s)}{n} = & \inf\{R_{c,n}(D_1, D_2, D_3) + \alpha D_1 + \beta D_2 + \nu D_3 : \\ & D_1 \geq 0, D_2 \geq 0, D_3 \geq 0\} \end{aligned} \tag{5.11}$$

and

$$\begin{aligned} R_{c,n}(D_1, D_2, D_3) = \sup\{ & F_{c,n}(s)/n - \alpha D_1 - \beta D_2 - \\ & \nu D_3 : s = (\alpha, \beta, \nu) \text{ and } \alpha \geq 0, \beta \geq 0, \nu \geq 0\}. \end{aligned} \tag{5.12}$$

For brevity, let us denote $(p_{U_1|X^n}, p_{U_2|U_1X^nY^n}, p_{\hat{Z}^n|U_1U_2Z^n})$ in (5.10) by $\mathbf{P}_{c,n}$, and $Q_{U_1U_2\hat{Z}^n}$ in (5.10) by $Q_{c,n}$. The algorithm computes $R_{c,n}(D_1, D_2, D_3)$ or equivalently $F_{c,n}(s)$ by finding transition probability and probability functions $\mathbf{P}_{c,n}$ and $Q_{c,n}$ iteratively until convergence, that is, for any $\mathbf{P}_{c,n}$, find

$$Q_{c,n}(\mathbf{P}_{c,n}) \overset{\Delta}{=} \arg\min_{Q_{c,n}} F_{c,n}(\mathbf{P}_{c,n}, Q_{c,n}),$$

and for any $Q_{c,n}$, find

$$\mathbf{P}_{c,n}(Q_{c,n}) \stackrel{\Delta}{=} \arg\min_{\mathbf{P}_{c,n}} F_{c,n}(\mathbf{P}_{c,n}, Q_{c,n}).$$

It was shown in Section 4.1 that the iterative algorithm converges globally.

We then define

$$F_{p,n}(s) \stackrel{\Delta}{=} \inf F_{p,n}(p_{U_1|X^n}, p_{U_2|U_1Y^n}, p_{\hat{Z}^n|U_1U_2Z^n}, Q_{U_1U_2\hat{Z}^n}) \tag{5.13}$$

for any $s = (\alpha, \beta, \nu)$ with $\alpha, \beta, \nu \geq 0$, and

$$\begin{aligned} & F_{p,n}(p_{U_1|X^n}, p_{U_2|U_1Y^n}, p_{\hat{Z}^n|U_1U_2Z^n}, Q_{U_1U_2\hat{Z}^n}) \\ \stackrel{\Delta}{=} & \sum_{x^n,u_1} P(x^n)P(u_1|x^n)\Bigg[\log \frac{P(u_1|x^n)}{Q(u_1)2^{-\alpha d_1(x^n,g_1(u_1))}} + \\ & \sum_{y^n,u_2} P(y^n|x^n)P(u_2|y^n u_1)\Bigg[\log \frac{P(u_2|y^n u_1)}{Q(u_2|u_1)2^{-\beta d_2(y^n,g_2(u_1,u_2))}} + \\ & \sum_{z^n,\hat{z}^n} P(z^n|x^n y^n)P(\hat{z}^n|z^n u_1^2) \log \frac{P(\hat{z}^n|z^n u_1^2)}{Q(\hat{z}^n|u_1^2)2^{-\nu d_3(z^n,\hat{z}^n)}}\Bigg]\Bigg]. \end{aligned} \tag{5.14}$$

In order to find the random variables $U_1, U_2, \hat{Z}^n$ that achieve $R_{p,n}(D_1, D_2, D_3)$, we try to find the transition probability and probability functions $p_{U_1|X^n}, p_{U_2|U_1Y^n}, p_{\hat{Z}^n|U_1U_2Z^n}$, and $Q_{U_1U_2\hat{Z}^n}$ that minimize (5.14). We denote $(p_{U_1|X^n}, p_{U_2|U_1Y^n}, p_{\hat{Z}^n|U_1U_2Z^n})$ and $Q_{U_1U_2\hat{Z}^n}$ in (5.14) by $\mathbf{P}_{p,n}$ and $Q_{p,n}$ respectively. For any $\mathbf{P}_{p,n}$, let

$$Q_{p,n}(\mathbf{P}_{p,n}) \stackrel{\Delta}{=} \arg\min_{Q_{p,n}} F_{p,n}(\mathbf{P}_{p,n}, Q_{p,n}).$$

Similarly, for any $Q_{p,n}$, let

$$\mathbf{P}_{p,n}(Q_{p,n}) \stackrel{\Delta}{=} \arg\min_{\mathbf{P}_{p,n}} F_{p,n}(\mathbf{P}_{p,n}, Q_{p,n}).$$

However, the problem of computing $R_{p,n}(D_1, D_2, D_3)$ to achieve (5.13) is very challenging and still open in general.

In view of (5.6) and (5.13), it is not hard to show the conjugate of $R_{p,n}(D_1, D_2, D_3)$ is

$$\begin{aligned}\frac{F_{p,n}(s)}{n} &= \inf\{R_{p,n}(D_1, D_2, D_3) + \alpha D_1 + \beta D_2 + \nu D_3 : \\ &\qquad D_1 \geq 0, D_2 \geq 0, D_3 \geq 0\}, \end{aligned} \tag{5.15}$$

for any $s = (\alpha, \beta, \nu), \alpha \geq 0, \beta \geq 0, \nu \geq 0$. In view of Lemma 6, $R_{p,n}(D_1, D_2, D_3)$ is convex and lower semi-continuous over the whole region $D_1 \geq 0, D_2 \geq 0, D_3 \geq 0$, it follows from [34, Theorem 12.2, pp. 104] that for any $D_1 \geq 0, D_2 \geq 0, D_3 \geq 0$, we have an alternative expression for $R_{p,n}(D_1, D_2, D_3)$ similar to that in (5.12),

$$\begin{aligned} R_{p,n}(D_1, D_2, D_3) = \sup\{F_{p,n}(s)/n - \alpha D_1 - \beta D_2 - \\ \nu D_3 : s = (\alpha, \beta, \nu) \text{ and } \alpha \geq 0, \beta \geq 0, \nu \geq 0\}. \end{aligned} \tag{5.16}$$

5.2.1 Markov Case

In this subsection, we consider the case under which the inequality in Equation (5.7) becomes equality.

Theorem 14 *If the jointly stationary and* ergodic *sources X_1, X_2, X_3 form a (first-order) Markov chain in the indicated order, then*

$$R_c^*(D_1, D_2, D_3) = R_p^*(D_1, D_2, D_3) \tag{5.17}$$

for any $D_1, D_2, D_3 \geq 0$.

Proof of Theorem 14: We discuss the following two cases: (1) $D_1D_2D_3 > 0$, and (2) $D_1D_2D_3 = 0$. In case (1), it follows from the Theorem 13 and Theorem 5 that it suffices to show $R_{p,n}(D_1, D_2, D_3) = R_{c,n}(D_1, D_2, D_3)$ for any $n \geq 1$ and $D_1D_2D_3 > 0$, where $R_{p,n}(D_1, D_2, D_3)$ is defined in (5.6) when $N = 3$ and $R_{c,n}(D_1, D_2, D_3)$ can be rewritten as (5.8) satisfying condition i) to v) below it.

In view of the definition of causal video codes and predictive video codes, it is not hard to see that

$$R_c^*(D_1, D_2, D_3) \leq R_p^*(D_1, D_2, D_3) \tag{5.18}$$

for any $D_1 D_2 D_3 \geq 0$. Therefore, in what follows, it suffices to show

$$R_c^*(D_1, D_2, D_3) \geq R_p^*(D_1, D_2, D_3) \tag{5.19}$$

for any $D_1 D_2 D_3 \geq 0$, which, together with Theorem 13 and Theorem 5, implies that case 1) suffices to prove that

$$R_{c,n}(D_1, D_2, D_3) \geq R_{p,n}(D_1, D_2, D_3) \tag{5.20}$$

for any $D_1 D_2 D_3 > 0$ and $n \geq 1$.

To this end, pick U_1, U_2, and $\hat{X}_3(1;n)$ satisfying the requirements (R5) and (R6). It is not hard to verify that the requirement (R6) plus the Markov condition $X_1 \to X_2 \to X_3$ implies that the Markov condition $U_2 \to (X_2(1;n), U_1) \to X_3(1;n)$ is satisfied. On the other hand, one can verify that if the Markov condition $X_1 \to X_2 \to X_3$ holds, $X_1(1;n)$ can be removed from the Markov chain $U_2 \to (X_2(1;n), U_1) \to X_1(1;n) X_3(1;n)$ in the requirement (R10) without changing the minimum total rate of predictive video coding with respect to the same distortion constraints. Therefore, the random variables $U_1, U_2, \hat{X}_3(1;n)$ also satisfy the requirements (R13) to (R15). This, together with (5.6), implies

$$\begin{aligned}
& nR_{c,n}(D_1, D_2, D_3) \\
= \; & I(X_1(1;n); U_1) + I(X_1(1;n) X_2(1;n); U_2 | U_1) + I(X_3(1;n); \hat{X}_3(1;n) | U_1 U_2) \\
= \; & I(X_1(1;n); U_1) + I(X_2(1;n); U_2 | U_1) + I(X_3(1;n); \hat{X}_3(1;n) | U_1 U_2) \\
\geq \; & nR_{p,n}(D_1, D_2, D_3)
\end{aligned} \tag{5.21}$$

Since (5.21) is valid for any auxiliary random variables $U_1, U_2, \hat{X}_3(1;n)$ satisfying the requirements (R5) and (R6), (5.19) then follows from the definition (5.6). This completes the proof of this Theorem in Case (1).

To prove (5.19) in Case (2), note that both $R_p^*(D_1, D_2, D_3)$ and $R_c^*(D_1, D_2, D_3)$ are right continuous in the sense that for any $D_1 D_2 D_3 \geq 0$,

$$\lim_{\epsilon \downarrow 0} R_p^*(D_1 + \epsilon, D_2 + \epsilon, D_3 + \epsilon) = R_p^*(D_1, D_2, D_3)$$

and

$$\lim_{\epsilon \downarrow 0} R_c^*(D_1 + \epsilon, D_2 + \epsilon, D_3 + \epsilon) = R_c^*(D_1, D_2, D_3).$$

The validity of (5.19) in Case (2) then follows from its validity in Case (1). In view of (5.18) and (5.19), this completes the proof of Theorem 14.

Apart from the classic information theoretic approach, Theorem 14 can also be proved by computational approach, that is, using our proposed iterative algorithm with global convergence, and an alternative proof is provided as below. The key strategy of the computational approach is to show that if sources form a (first-order) Markov chain in the indicated order, then any causal code, which can be solved by our iterative algorithm, is easy to be verified as a predictive code. The computational approach uses our iterative algorithm in a novel way, not for computing, but for comparing. The computational approach sometimes shows its superiority to the classic approach. One example is to show the strict inequality in Theorem 15 using our computational approach, however, the solution is still unknown by using the classic approach.

Next, we provide an alternative proof of Theorem 14. The proof presented here is to compare $R_{c,n}(D_1, D_2, D_3)$ with $R_{p,n}(D_1, D_2, D_3)$ through a novel application of the iterative computation algorithm proposed in [54]. For convenience, in the following we shall write $X_1(1;n), X_2(1;n), X_3(1;n)$ as X^n, Y^n, and Z^n, respectively. Note that $X^n \rightarrow Y^n \rightarrow Z^n$ is a Markov chain by assumption .

An alternative proof of Theorem 14: In view of (5.7), we see that in order to prove (5.17), it suffices to show that

$$R_c^*(D_1, D_2, D_3) \geq R_p^*(D_1, D_2, D_3) \tag{5.22}$$

for any $D_1, D_2, D_3 \geq 0$, which, together with Theorem 2 and Theorem 13, implies that it suffices to prove that

$$R_{c,n}(D_1, D_2, D_3) \geq R_{p,n}(D_1, D_2, D_3) \tag{5.23}$$

for any $n \geq 1$.

Let us first consider the case where $D_1 D_2 D_3 > 0$. As mentioned above, our strategy to prove (5.23) is as follows. Since $R_{c,n}(D_1, D_2, D_3)$ or equivalently $F_{c,n}(s)$ in (5.9) is computable by using the iterative algorithm proposed in Section 4.1, where $s = (\alpha, \beta, \gamma)$ denotes the Lagrange multipliers. Note that $D_1 D_2 D_3 > 0$ implies $\alpha\beta\gamma < \infty$. This allows us to verify that if $X^n \to Y^n \to Z^n$ is a Markov chain, the solution to $F_{c,n}(s)$ with $\alpha\beta\gamma < \infty$ is a predictive code. Consequently, $R_{c,n}(D_1, D_2, D_3) \geq R_{p,n}(D_1, D_2, D_3)$.

Along this line, let $\mathbf{P}_{c,n} = (p_{U_1|X^n}, p_{U_2|U_1 X^n Y^n}, p_{\hat{Z}^n|U_1 U_2 Z^n})$ denote a vector of three transition probability functions in $\mathcal{P}_{\mathcal{U}_1|\mathcal{X}^n} \times \mathcal{P}_{\mathcal{U}_2|\mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{U}_1} \times \mathcal{P}_{\hat{\mathcal{Z}}^n|\mathcal{Z}^n \times \mathcal{U}_1 \times \mathcal{U}_2}$. Let $Q_{c,n} = Q_{U_1 U_2 \hat{Z}^n}$ denote a function in $\mathcal{P}_{\mathcal{U}_1 \times \mathcal{U}_2 \times \hat{\mathcal{Z}}^n}$ derived from $\mathbf{P}_{c,n}$ as in Step 3 of the iterative algorithm, that is,

$$Q(u_1 u_2 \hat{z}^n) = \sum_{x^n, y^n, z^n} p(x^n) p(y^n|x^n) p(z^n|y^n) P(u_1|x^n) P(u_2|x^n y^n u_1) P(\hat{z}^n|u_1^2 z^n). \tag{5.24}$$

Suppose that $(\mathbf{P}_{c,n}, Q_{c,n})$ is a stationary point, i.e.,

$$F_{c,n}(\mathbf{P}_{c,n}, Q_{c,n}) = F_{c,n}(s). \tag{5.25}$$

Note that the existence of such stationary point was guaranteed by Theorem 6. It follows immediately that

$$\mathbf{P}_{c,n} = \mathbf{P}_{c,n}(Q_{c,n}), \tag{5.26}$$

which implies that

$$\begin{aligned} P(\hat{z}^n|z^n u_1^2) &= \frac{Q(\hat{z}^n|u_1^2)2^{-\nu d_3(z^n,\hat{z}^n)}}{\Delta(z^n,u_1,u_2)}, & (5.27)\\ P(u_2|x^n y^n u_1) &= \frac{Q(u_2|u_1)2^{-\beta d_2(y^n,g_2(u_1,u_2))}}{\Lambda(y^n,u_1)} \times 2^{\sum_{z^n} p(z^n|y^n)\log\Delta(z^n,u_1,u_2)}, & (5.28)\\ P(u_1|x^n) &= \frac{Q(u_1)2^{-\alpha d_1(x^n,g_1(u_1))}}{\Gamma(x^n)} \times 2^{\sum_{y^n} p(y^n|x^n)\log\Lambda(y^n,u_1)}. & (5.29) \end{aligned}$$

where

$$\begin{aligned} \Delta(z^n,u_1,u_2) &\triangleq \sum_{\hat{z}^n} Q(\hat{z}^n|u_1^2)2^{-\nu d_3(z^n,\hat{z}^n)}, & (5.30)\\ \Lambda(y^n,u_1) &\triangleq \sum_{u_2} Q(u_2|u_1)2^{-\beta d_2(y^n,g_2(u_1,u_2))} \times 2^{\sum_{z^n} p(z^n|y^n)\log\Delta(z^n,u_1,u_2)}, & (5.31)\\ \Gamma(x^n) &\triangleq \sum_{u_1} Q(u_1)2^{-\alpha d_1(x^n,g_1(u_1))} \times 2^{\sum_{y^n} p(y^n|x^n)\log\Lambda(y^n,u_1)}. & (5.32) \end{aligned}$$

It is observed that the right-hand-side of (5.28) does not depend on x^n, i.e., for any u_2, u_1, y^n, x^n and $\tilde{x}^n \in \mathcal{X}^n$ with $\tilde{x}^n \neq x^n$, we have $P(u_2|x^n y^n u_1) = P(u_2|\tilde{x}^n y^n u_1)$, which implies the validity of an extra Markov chain $X^n \to (Y^n, U_1) \to U_2$ besides the requirements (R5) and (R6) for causal codes. Thus, it is not hard to see that if $X^n \to Y^n \to Z^n$, then any causal codes $U_1, U_2, \hat{Z}^n$ for encoding X^n, Y^n, Z^n with respective distortion $D_1, D_2, D_3 > 0$ also satisfy the requirements (R13) to (R16) with respect to X^n, Y^n, Z^n and D_1, D_2, D_3 respectively. Therefore

$$R_{p,n}(D_1, D_2, D_3) \leq R_{c,n}(D_1, D_2, D_3). \tag{5.33}$$

for any $D_1, D_2, D_3 > 0$.

Let us then consider the cases where some distortion levels are 0.

Case 1) all D_1, D_2, D_3 are 0: in this case, CVC does not offer any performance gain over PVC, i.e., (5.17) is satisfied automatically.

Case 2) any two of D_1, D_2, D_3 are 0: in this case, it is equivalent to encode a single source given a positive distortion level in both causal coding and predictive coding settings, which is not hard to verify the validity of (5.17).

Case 3) $D_3 = 0, D_1 \neq 0, D_2 \neq 0$: along the same line of the case $D_1 D_2 D_3 > 0$, we consider the computation of $R_{c,n}(D_1, D_2, 0)$, i.e.,

$$\min_{U_1, U_2} \Big[I(X^n; U_1) + I(X^n Y^n; U_2 | U_1) + H(Z^n | U_1 U_2) \Big], \tag{5.34}$$

where the minimization is taken over all random variables U_1 from a finite alphabet $\mathcal{U}_1$ and U_2 from $\mathcal{U}_2$ satisfying the following conditions: i) there exists a function g_1 such that $\hat{X}^n = g_1(U_1)$ and $\frac{1}{n}\mathbf{E} d_1(X^n, \hat{X}^n) \leq D_1$; ii) there exists a function g_2 such that $\hat{Y}^n = g_2(U_1, U_2)$ and $\frac{1}{n}\mathbf{E} d_2(Y^n, \hat{Y}^n) \leq D_2$; iii) $U_1 \to X^n \to Y^n Z^n$ is a Markov chain; and (iv) $U_2 \to X^n Y^n U_1 \to Z^n$ is a Markov chain.

In order to calculate (5.34), we equivalently calculate $F_{c,n}(s)$ with only two Lagrange multipliers $\alpha\beta < \infty$ by iteratively updating $p_{U_1|X^n}, p_{U_2|U_1 X^n Y^n}$, and $Q_{Z^n U_1 U_2}$. Since $X^n \to Y^n \to Z^n$, (5.24), along with (5.27) to (5.32) reduce to:

$$Q(u_1 u_2 z^n) = \sum_{x^n, y^n} p(x^n) p(y^n | x^n) p(z^n | y^n) P(u_1 | x^n) P(u_2 | x^n y^n u_1), \tag{5.35}$$

$$P(u_2 | x^n y^n u_1) = \frac{Q(u_2 | u_1) 2^{-\beta d_2(y^n, g_2(u_1, u_2))}}{\Lambda(y^n, u_1)} \times 2^{\sum_{z^n} p(z^n | y^n) \log \Delta(z^n, u_1, u_2)}, \tag{5.36}$$

$$P(u_1 | x^n) = \frac{Q(u_1) 2^{-\alpha d_1(x^n, g_1(u_1))}}{\Gamma(x^n)} \times 2^{\sum_{y^n} p(y^n | x^n) \log \Lambda(y^n, u_1)}, \tag{5.37}$$

where

$$\Delta(z^n, u_1, u_2) \triangleq Q(z^n | u_1^2), \tag{5.38}$$

$$\Lambda(y^n, u_1) \triangleq \sum_{u_2} Q(u_2 | u_1) 2^{-\beta d_2(y^n, g_2(u_1, u_2))} \times 2^{\sum_{z^n} p(z^n | y^n) \log \Delta(z^n, u_1, u_2)}, \tag{5.39}$$

$$\Gamma(x^n) \triangleq \sum_{u_1} Q(u_1) 2^{-\alpha d_1(x^n, g_1(u_1))} \times 2^{\sum_{y^n} p(y^n | x^n) \log \Lambda(y^n, u_1)}. \tag{5.40}$$

Similar to the discussion of case $D_1D_2D_3 > 0$ by verifying the validity of the Markov chain $X^n \to (Y^n, U_1) \to U_2$ via checking (5.36), we conclude that if $X^n \to Y^n \to Z^n$, then any causal codes U_1, U_2 for encoding X^n, Y^n, Z^n with respective distortion $D_1 > 0, D_2 > 0$ and $D_3 = 0$ also satisfy the requirements (R13) to (R16) with respect to X^n, Y^n, Z^n and $D_1 > 0, D_2 > 0, D_3 = 0$ respectively. Thus (5.33) also holds in this case.

Case 4) $D_1 = 0, D_2 \neq 0, D_3 \neq 0$: in this case, it is equivalent to compare CVC and PVC when $N = 2$, and (5.17) follows immediately.

Case 5) $D_2 = 0, D_1 \neq 0, D_3 \neq 0$: in this case, $\hat{X}_2$ is identical to X_2, and it is easy to see the equivalence between CVC and PVC.

Then following from Theorem 2 and Theorem 13, (5.33) implies (5.22), which, together with the definition of causal video codes and predictive video codes that

$$R_c^*(D_1, D_2, D_3) \leq R_p^*(D_1, D_2, D_3),$$

for any $D_1, D_2, D_3 \geq 0$, implies (5.17) valid for any $D_1D_2D_3 > 0$ and Case 3) as well. Putting all cases together, we have shown that if $X^n \to Y^n \to Z^n$, then (5.17) holds for any $D_1, D_2, D_3 \geq 0$, which completes our proof.

Theorem 14 implies that all the information theoretic results and the computation algorithm on CVC can be directly applied to PVC when sources form a (first-order) Markov chain. However, when sources do not form a (first-order) Markov chain, the problems of single-letter characterizing, computing, and comparing $R_p^*(D_1, D_2, D_3)$ are still open. In the rest of the section, we look into the case when sources do not form a (first-order) Markov chain, and investigate the single-letter characterization and comparison of $R_p^*(D_1, D_2, D_3)$ for the first time.

5.2.2 Non-Markov Case

In this subsection, we consider the case under which $R_{c,n}(D_1, D_2, D_3)$ is strictly less than $R_{p,n}(D_1, D_2, D_3)$ for any finite $n \geq 1$ and $D_1 > 0, D_2 > 0, D_3 \geq 0$. Again, we denote the three jointly stationary and ergodic sources $X_1(1;n), X_2(1;n), X_3(1;n)$ as X^n, Y^n, and Z^n respectively for brevity. In the following discussions, we distinguish between two cases: 1)$D_1D_2D_3 > 0$, and 2)$D_1D_2 > 0, D_3 = 0$.

Since $R_{p,n}(D_1, D_2, D_3)$ and $R_{c,n}(D_1, D_2, D_3)$ are convex as functions of D_1, D_2, D_3 over the region $\{(D_1, D_2, D_3) : D_1 \geq 0, D_2 \geq 0, D_3 \geq 0\}$. As such, in Case 1), they are subdifferentiable at any point (D_1, D_2, D_3) with $D_1 > 0$, $D_2 > 0$ and $D_3 > 0$. (See [34, Chapter 23] for discussions on the subdifferential and subgradients of a convex function.) From (5.12) and (5.16), the comparison between $R_{c,n}(D_1, D_2, D_3)$ and $R_{p,n}(D_1, D_2, D_3)$ can also be made through comparing their respective conjugates.

The discussion of Case 2) is more involved due to the difficulty to deal with $\nu \to \infty$ in the conjugates of $R_{c,n}(D_1, D_2, D_3)$ and $R_{p,n}(D_1, D_2, D_3)$ (as shown in (5.11) and (5.15) respectively). In view of (5.8), we rewrite $R_{c,n}(D_1, D_2, 0)$ as

$$\min_{U_1,U_2} \Big[I(X^n; U_1) + I(X^nY^n; U_2|U_1) + H(Z^n|U_1U_2) \Big], \tag{5.41}$$

where the minimization is taken over all random variables U_1 from a finite alphabet $\mathcal{U}_1$ and U_2 from $\mathcal{U}_2$ satisfying the following conditions: i) there exists a function g_1 such that $\hat{X}^n = g_1(U_1)$ and $\frac{1}{n}\mathbf{E}d_1(X^n, \hat{X}^n) \leq D_1$; ii) there exists a function g_2 such that $\hat{Y}^n = g_2(U_1, U_2)$ and $\frac{1}{n}\mathbf{E}d_2(Y^n, \hat{Y}^n) \leq D_2$; iii) $U_1 \to X^n \to Y^nZ^n$ is a Markov chain; and (iv) $U_2 \to X^nY^nU_1 \to Z^n$ is a Markov chain. Define

$$F_{c,n}(s) \stackrel{\Delta}{=} \inf F_{c,n}(p_{U_1|X^n}, p_{U_2|U_1X^nY^n}, Q_{U_1U_2Z^n}) \tag{5.42}$$

for any $s = (\alpha, \beta)$, where $\alpha, \beta \geq 0$, denotes the standard Lagrange multipliers, and

$$\begin{aligned}
& F_{c,n}(p_{U_1|X^n}, p_{U_2|U_1X^nY^n}, Q_{U_1U_2Z^n}) \\
\triangleq\ & \sum_{x^n,u_1} P(x^n)P(u_1|x^n)\Bigg[\log \frac{P(u_1|x^n)}{Q(u_1)2^{-\alpha d_1(x^n,g_1(u_1))}} + \\
& \sum_{y^n,u_2} P(y^n|x^n)P(u_2|x^ny^nu_1)\Bigg[\log \frac{P(u_2|x^ny^nu_1)}{Q(u_2|u_1)2^{-\beta d_2(y^n,g_2(u_1,u_2))}} + \\
& \sum_{z^n} P(z^n|x^ny^n)\log \frac{1}{P(z^n|u_1^2)}\Bigg]\Bigg]. \qquad (5.43)
\end{aligned}$$

It follows that (5.11) and (5.12) can be rewritten as:

$$\begin{aligned}
\frac{F_{c,n}(s)}{n} &= \inf\{R_{c,n}(D_1, D_2, 0) + \alpha D_1 + \beta D_2 : \\
&\qquad D_1 \geq 0, D_2 \geq 0\} \qquad (5.44)
\end{aligned}$$

and

$$\begin{aligned}
R_{c,n}(D_1, D_2, 0) = \sup\{&F_{c,n}(s)/n - \alpha D_1 - \beta D_2 : \\
& s = (\alpha, \beta) \text{ and } \alpha \geq 0, \beta \geq 0\}. \qquad (5.45)
\end{aligned}$$

For brevity, we denote $(p_{U_1|X^n}, p_{U_2|U_1X^nY^n})$ in (5.43) by $\mathbf{P}_{c,n}$, and $Q_{U_1U_2Z^n}$ in (5.43) by $Q_{c,n}$.

Similarly, $R_{p,n}(D_1, D_2, 0)$ can be written as

$$R_{p,n}(D_1, D_2, 0) = \min_{U_1,U_2}\Big[I(X^n; U_1) + I(Y^n; U_2|U_1) + H(Z^n|U_1U_2)\Big], \qquad (5.46)$$

where the minimization is taken over all random variables U_1 from a finite alphabet $\mathcal{U}_1$ and U_2 from $\mathcal{U}_2$ satisfying the following conditions: i) there exists a function g_1 such that $\hat{X}^n = g_1(U_1)$ and $\frac{1}{n}\mathbf{E}d_1(X^n, \hat{X}^n) \leq D_1$; ii) there exists a function g_2 such that $\hat{Y}^n = g_2(U_1, U_2)$ and $\frac{1}{n}\mathbf{E}d_2(Y^n, \hat{Y}^n) \leq D_2$; iii) $U_1 \to X^n \to Y^nZ^n$ is a Markov chain; and (iv)

$U_2 \to Y^n U_1 \to X^n Z^n$ is a Markov chain. Define

$$F_{p,n}(s) \triangleq \inf F_{p,n}(p_{U_1|X^n}, p_{U_2|U_1Y^n}, Q_{U_1U_2Z^n}) \tag{5.47}$$

for any $s = (\alpha, \beta)$ with $\alpha, \beta \geq 0$, and

$$\begin{aligned} & F_{p,n}(p_{U_1|X^n}, p_{U_2|U_1Y^n}, Q_{U_1U_2Z^n}) \\ \triangleq\; & \sum_{x^n, u_1} P(x^n)P(u_1|x^n)\Bigg[\log \frac{P(u_1|x^n)}{Q(u_1)2^{-\alpha d_1(x^n, g_1(u_1))}} + \\ & \sum_{y^n, u_2} P(y^n|x^n)P(u_2|y^n u_1)\Bigg[\log \frac{P(u_2|y^n u_1)}{Q(u_2|u_1)2^{-\beta d_2(y^n, g_2(u_1,u_2))}} + \\ & \sum_{z^n} P(z^n|x^n y^n) \log \frac{1}{P(z^n|u_1^2)}\Bigg]\Bigg]. \end{aligned} \tag{5.48}$$

We denote $(p_{U_1|X^n}, p_{U_2|U_1Y^n})$ and $Q_{U_1U_2Z^n}$ in (5.48) as $\mathbf{P}_{p,n}$ and $Q_{p,n}$ accordingly, and repeat the same argument in Case 1), we have

$$\begin{aligned} \frac{F_{p,n}(s)}{n} &= \inf\{R_{p,n}(D_1, D_2, 0) + \alpha D_1 + \beta D_2 : \\ &\qquad D_1 \geq 0, D_2 \geq 0\}, \end{aligned} \tag{5.49}$$

and

$$\begin{aligned} R_{p,n}(D_1, D_2, 0) = \sup\{&F_{p,n}(s)/n - \alpha D_1 - \beta D_2 : \\ &s = (\alpha, \beta) \text{ and } \alpha \geq 0, \beta \geq 0\}. \end{aligned} \tag{5.50}$$

Accordingly, $R_{p,n}(D_1, D_2, 0)$ and $R_{c,n}(D_1, D_2, 0)$ are subdifferentiable at any point (D_1, D_2) with $D_1 > 0$, and $D_2 > 0$, and the comparison between $R_{c,n}(D_1, D_2, 0)$ and $R_{p,n}(D_1, D_2, 0)$ can also be made through comparing their respective conjugates.

Combine Case 1) and 2), specifically, we state the following condition to facilitate our discussion.

Condition B: A point (D_1, D_2, D_3) with $D_1 > 0, D_2 > 0,$ and $D_3 \geq 0$ is said to satisfy Condition A if $R_{c,n}(D_1, D_2, D_3)$ as a function of D_1, D_2, D_3 has a negative subgradient $-s = (-\alpha, -\beta, -\nu), \alpha > 0, \beta > 0, \nu > 0,$ at $(D_1, D_2, D_3), D_1D_2D_3 > 0$ (or $-s = (-\alpha, -\beta), \alpha > 0, \beta > 0$ at $(D_1, D_2), D_1D_2 > 0$ under the case of $D_3 = 0$) such that there is a distribution $Q = Q(u_1u_2\hat{z}^n) = Q(u_1)Q(u_2|u_1)Q(\hat{z}^n|u_1u_2)$, where $u_1 \in \mathcal{U}_1, u_2 \in \mathcal{U}_2, \hat{z}^n \in \hat{\mathcal{Z}}^n$ (or $Q = Q(u_1u_2z^n) = Q(u_1)Q(u_2|u_1)Q(z^n|u_1u_2)$, where $u_1 \in \mathcal{U}_1, u_2 \in \mathcal{U}_2, z^n \in \mathcal{Z}^n$) satisfying the following requirements:

(B1) $F_{p,n}(\mathbf{P}_{p,n}(Q), Q) = F_{p,n}(s)$.

(B2) Either

$$F_{c,n}(s) < F_{c,n}(\mathbf{P}_{c,n}(Q), Q)$$

or $p_{U_2|U_1X^nY^n}(\cdot|u_1x^ny^n)$ depends on x^n, i.e., there exist $u_2 \in \mathcal{U}_2$, $u_1 \in \mathcal{U}_1$, $y^n \in \mathcal{Y}^n$, $x^n \in \mathcal{X}^n$, and $\tilde{x}^n \in \mathcal{X}^n$ with $\tilde{x}^n \neq x^n$ such that

$$p_{U_2|U_1X^nY^n}(u_2|u_1x^ny^n) \neq p_{U_2|U_1X^nY^n}(u_2|u_1\tilde{x}^ny^n).$$

The following theorem summarizes the main result.

Theorem 15 *Suppose that X_1, X_2, X_3 are jointly stationary and* ergodic *sources, and $X_1, X_2,$ and X_3 do not form a (first-order) Markov chain, then for any point $(D_1, D_2, D_3), D_1 > 0, D_2 > 0,$ and $D_3 \geq 0$, satisfying Condition B,*

$$R_{c,n}(D_1, D_2, D_3) < R_{p,n}(D_1, D_2, D_3),$$

for any $n \geq 1$.

Proof of Theorem 15 : For brevity, we rewrite three sources $X_1(1;n), X_2(1;n), X_3(1;n)$ as $X^n, Y^n,$ and Z^n respectively in the proof. First note that from the definition of causal video codes and predictive video codes,

$$R_{c,n}(D_1, D_2, D_3) \leq R_{p,n}(D_1, D_2, D_3) \tag{5.51}$$

for any $D_1, D_2, D_3 \geq 0$. Fix now any point $(D_1, D_2, D_3), D_1, D_2, D_3 > 0$, satisfying Condition B. We prove Theorem 15 by contradiction. Suppose that

$$R_{c,n}(D_1, D_2, D_3) = R_{p,n}(D_1, D_2, D_3) \tag{5.52}$$

at point (D_1, D_2, D_3). Let $-s = (-\alpha, -\beta, -\nu)$ be the negative subgradient of $R_{c,n}(D_1, D_2, D_3)$ at the point (D_1, D_2, D_3) in Condition B. This implies that for any $D_1' \geq 0, D_2' \geq 0,$ and $D_3' \geq 0$,

$$R_{c,n}(D_1', D_2', D_3') \geq R_{c,n}(D_1, D_2, D_3) - \alpha(D_1' - D_1) - \beta(D_2' - D_2) - \nu(D_3' - D_3) \tag{5.53}$$

which, coupled with (5.51) and (5.52), in turn implies

$$R_{p,n}(D_1', D_2', D_3') \geq R_{p,n}(D_1, D_2, D_3) - \alpha(D_1' - D_1) - \beta(D_2' - D_2) - \nu(D_3' - D_3) \tag{5.54}$$

for any $D_1' \geq 0, D_2' \geq 0,$ and $D_3' \geq 0$. In other words, under the assumption (5.52), $-s$ is also a negative subgradient of $R_p^*(D_1, D_2, D_3)$ at the point (D_1, D_2, D_3). In view of (5.11), (5.12), (5.15) and (5.16), it then follows that

$$R_{c,n}(D_1, D_2, D_3) = \frac{1}{n} F_{c,n}(s) - \alpha D_1 - \beta D_2 - \nu D_3, \tag{5.55}$$

and

$$R_{p,n}(D_1, D_2, D_3) = \frac{1}{n} F_{p,n}(s) - \alpha D_1 - \beta D_2 - \nu D_3. \tag{5.56}$$

Repeat the above arguments for any point $(D_1, D_2, D_3), D_1 D_2 > 0,$ and $D_3 = 0$, satisfying Condition A. In view of (5.44), (5.45), (5.49) and (5.50), we then have accordingly

$$R_{c,n}(D_1, D_2, 0) = \frac{1}{n} F_{c,n}(s) - \alpha D_1 - \beta D_2, \tag{5.57}$$

and

$$R_{p,n}(D_1, D_2, 0) = \frac{1}{n} F_{p,n}(s) - \alpha D_1 - \beta D_2. \tag{5.58}$$

In view of the requirement (R17) in Condition B, we have

$$F_{p,n}(s) = F_{p,n}(\mathbf{P}_{p,n}(Q), Q). \tag{5.59}$$

From Step 2 of the iterative algorithm presented in Section 4.1, it follows that

$$\begin{aligned} F_{c,n}(s) &\le F_{c,n}(\mathbf{P}_{c,n}(Q), Q) && (5.60) \\ &\le F_{p,n}(\mathbf{P}_{p,n}(Q), Q) && (5.61) \end{aligned}$$

where the inequality in (5.61) is strict when $p_{U_2|U_1 X^n Y^n}$ depends on X^n. Therefore, according to the requirement (R18) in Condition B, no matter which choice in the requirement (R18) is valid, we always have

$$F_{c,n}(s) < F_{p,n}(\mathbf{P}_{p,n}(Q), Q), \tag{5.62}$$

which, together with (5.55) to (5.59), implies that for any $D_1 D_2 > 0, \text{and } D_3 \ge 0$,

$$R_{c,n}(D_1, D_2, D_3) < R_{p,n}(D_1, D_2, D_3). \tag{5.63}$$

This contradicts the assumption (5.52), hence complete the proof of this theorem.

To help the reader better understand where the gain of $R_{c,n}(D_1, D_2, D_3)$ over $R_{p,n}(D_1, D_2, D_3)$ comes from, we give a high level explanation as follows. The availability of X_1 to the encoder of X_2 does not really help the encoder of X_2 and its corresponding decoder achieve a better rate distortion trade-off (R_2, D_2). Likewise, the availability of X_1 and X_2 to the encoder of X_3 does not really help the encoder of X_3 and its corresponding decoder achieve a better rate distortion trade-off (R_3, D_3) either. What really matters is that the availability of X_1 to the encoder of X_2 will help the encoder of X_2 choose better side information $\hat{X}_2$ for the encoder and decoder of X_3. If the rate reduction of the encoder of X_3 arising from this better $\hat{X}_2$ is more than the overhead associated with the selection of this better $\hat{X}_2$, then the total rate $R_{c,n}(D_1, D_2, D_3)$ is smaller than $R_{p,n}(D_1, D_2, D_3)$.

Remark 11 *When sources X_1, X_2, X_3 do not form a (first-order) Markov chain in the indicated order, the single-letter characterization of $R^*_p(D_1, D_2, D_3)$ is still unknown. Therefore, the strict inequality $R^*_c(D_1, D_2, D_3) < R^*_p(D_1, D_2, D_3)$ can not be inferred from Theorem 15. However, Theorem 15 indeed implies that if $R^*_p(D_1, D_2, D_3)$ admits a single-letter characterization, the strict inequality $R^*_c(D_1, D_2, D_3) < R^*_p(D_1, D_2, D_3)$ holds. In the following case study, we derive for the first time that the single-letter characterization of $R^*_p(D_1, D_2, D_3)$ can be established for an IID vector source $\mathbf{X} = (X_1, X_2, X_3)$ where X_1 and X_2 are independent, and illustrate Condition B specifically in the binary case.*

5.2.3 Case Study

In this subsection, we first derive the single-letter expression of $R^*_p(D_1, D_2, D_3)$ in the case of $N = 3$ for an IID vector source $\mathbf{X} = (X_1, X_2, X_3)$ where X_1 and X_2 are independent.

Theorem 16 *Let $\mathbf{X} = (X_1, X_2, X_3)$ be an IID vector video source such that X_1 and X_2 are independent. Then*

$$\begin{aligned} R^*_p(D_1, D_2, D_3) = \min_{U, \hat{X}_2, \hat{X}_3} \Big[I(X_1; U) + \\ I(X_2; \hat{X}_2 | U) + I(X_3; \hat{X}_3 | U \hat{X}_2) \Big], \end{aligned} \tag{5.64}$$

where the minimization is taken over all random variables U from a finite alphabet $\mathcal{U}$ and $\hat{X}_2, \hat{X}_3$ from $\hat{\mathcal{X}}_2$, $\hat{\mathcal{X}}_3$ satisfying the following conditions: i) there exists a function g such that $\hat{X}_1 = g(U)$ and $\mathbf{E}d_1(X_1, \hat{X}_1) \leq D_1$; ii) $\mathbf{E}d_2(X_2, \hat{X}_2) \leq D_2$; iii) $\mathbf{E}d_3(X_3, \hat{X}_3) \leq D_3$; iv)

$U \to X_1 \to X_2X_3$ is a Markov chain; v) $\hat{X}_3 \to X_3U\hat{X}_2 \to X_1X_2$ is a Markov chain; and vi) $\hat{X}_2 \to X_2U \to X_1X_3$ is a Markov chain.

Proof of Theorem 16: First, we prove the converse. For brevity, we write $X_1X_2X_3$ as XYZ, $X_1(1;n)$ as X^n, $X_2(1;n)$ as Y^n, and $X_3(1;n)$ as Z^n. Let $S_1S_2S_3$ be the output of a length n predictive video code achieving rates (R_1, R_2, R_3) at distortion levels (D_1, D_2, D_3). Then

$$\begin{aligned}
& n(R_1 + R_2 + R_3) \geq H(S_1S_2S_3) \\
\geq\ & H(S_1S_2\hat{Z}^n) \\
=\ & I(X^nY^nZ^n; S_1S_2\hat{Z}^n) \\
\overset{1)}{=}\ & I(X^n; S_1) + I(Y^n; S_2|S_1) + I(X^nY^nZ^n; \hat{Z}^n|S_1S_2) \\
\overset{2)}{=}\ & H(X^n) - H(X^n|S_1) + H(Y^n) - H(Y^n|S_1S_2) \\
& +I(X^nY^nZ^n; \hat{Z}^n|S_1S_2) \\
=\ & \sum_{i=1}^{n} \Big[H(X_i) - H(X_i|X_i^- S_1) + H(Y_i) \\
& -H(Y_i|Y_i^- S_1S_2) + I(X_iY_iZ_i; \hat{Z}^n|X_i^-Y_i^-Z_i^-S_1S_2) \Big] \\
\geq\ & \sum_{i=1}^{n} \Big[I(X_i; X_i^- S_1) + H(Y_i) - H(Y_i|Y_i^- S_1S_2) \\
& +I(Z_i; \hat{Z}_i|X_i^-Y_i^-Z_i^-S_1S_2) \Big]
\end{aligned} \tag{5.65}$$

where $\hat{Z}^n$ denote the reconstruction of Z^n. In the above, the equality 1) follows from the fact that S_1 is a function of X^n and S_2 is a function of (Y^n, S_1); the equality 2) is due to that X^n and Y^n are independent; and the last inequality is due to the fact that mutual information is always non-negative. On the right-hand-side of (5.65),

$$\begin{aligned}
& H(Y_i) - H(Y_i|Y_i^- S_1 S_2) \\
= \; & H(Y_i|X_i^- S_1) - H(Y_i|Y_i^- S_2 X_i^- S_1) \\
= \; & I(Y_i; Y_i^- S_2|X_i^- S_1) \qquad (5.66)
\end{aligned}$$

where the first equality follows from the independence between X^n and Y^n; and

$$\begin{aligned}
& I(Z_i; \hat{Z}_i|X_i^- Y_i^- Z_i^- S_1 S_2) \\
\geq \; & H(Z_i|X_i^- Y_i^- Z_i^- S_1 S_2) - H(Z_i|\hat{Z}_i X_i^- Y_i^- S_1 S_2) \\
= \; & H(Z_i|X_i^- Y_i^- S_1 S_2) - H(Z_i|\hat{Z}_i X_i^- Y_i^- S_1 S_2) \\
= \; & I(Z_i; \hat{Z}_i|X_i^- Y_i^- S_1 S_2) \qquad (5.67)
\end{aligned}$$

where the first equality follows from verifying $I(Z_i; Z_i^-|X_i^- Y_i^- S_1 S_2) = 0$. Putting (5.66) and (5.67) back into (5.65), and letting $U_i \stackrel{\Delta}{=} X_i^- S_1$ and $V_i \stackrel{\Delta}{=} Y_i^- S_2$, we have

$$\begin{aligned}
n(R_1 + R_2 + R_3) \; \geq \; & \sum_{i=1}^{n} [I(X_i; U_i) + I(Y_i; V_i|U_i) + \\
& I(Z_i; \hat{Z}_i|U_i V_i)]. \qquad (5.68)
\end{aligned}$$

It is easy to verify that $U_i \to X_i \to Y_i Z_i$ is a Markov chain. To see that $V_i \to Y_i U_i \to X_i Z_i$ is also a Markov chain, we first verify that $I(V_i; X_i|Y_i U_i) = 0$ by checking

$$\begin{aligned}
H(V_i|U_i Y_i X_i) \; = \; & H(X_i U_i Y_i V_i) - H(X_i U_i Y_i) \\
\stackrel{3)}{=} \; & H(X_i U_i) + H(V_i Y_i|X_i U_i) - \\
& H(Y_i|U_i) - H(X_i U_i) \\
\stackrel{4)}{=} \; & H(V_i Y_i|U_i) - H(Y_i|U_i) \\
= \; & H(V_i|Y_i U_i). \qquad (5.69)
\end{aligned}$$

In the above, the equality 3) is due to the independence between X^n and Y^n; and the equality 4) follows from verifying $I(V_iY_i; X_i|U_i) = 0$. Similarly, we can show that $I(V_i; Z_i|Y_iU_iX_i) = 0$.

To continue, we introduce a timesharing random variable I that is independent of XYZ and uniformly distributed over $\{1, 2, \cdots, n\}$. It follows from (5.68) that

$$\begin{aligned} R_1 + R_2 + R_3 &\geq I(X; U_I|I) + I(Y; V_I|U_I I) \\ &\quad +I(Z; \hat{Z}_I|U_I V_I I) \\ &\geq I(X; U) + I(Y; V|U) + \\ &\quad I(Z; \hat{Z}|UV), \end{aligned} \tag{5.70}$$

where $U \triangleq (U_I, I)$, $V \triangleq (V_I, I)$, and we abuse the notation to use X, Y, Z, and $\hat{Z}$ as random variables.

To complete the proof of the converse, we note that the reconstruction $\hat{Y}$ of Y is a function of (U, V). Thus

$$\begin{aligned} I(Y; V|U) &= I(Y; V\hat{Y}|U) \\ &= I(Y; \hat{Y}|U) + I(Y; V|U\hat{Y}) \end{aligned} \tag{5.71}$$

and

$$\begin{aligned} I(Z; \hat{Z}|UV) &= H(Z|UV) - H(Z|UV\hat{Z}) \\ &\geq H(Z|UV\hat{Y}) - H(Z|U\hat{Y}\hat{Z}) \\ &= I(Z; \hat{Z}|U\hat{Y}) - I(Z; V|U\hat{Y}). \end{aligned} \tag{5.72}$$

Putting (5.71) and (5.72) back into (5.70), and invoking the inequality $I(Y; V|U\hat{Y}) - I(Z; V|U\hat{Y}) \geq 0$, we obtain

$$\begin{aligned} R_1 + R_2 + R_3 &\geq I(X; U) + I(Y; \hat{Y}|U) + \\ &\quad I(Z; \hat{Z}|U\hat{Y}). \end{aligned} \tag{5.73}$$

Note that $\hat{Y} \to YU \to XZ$ is a Markov chain following from

$$I(\hat{Y}; XZ|YU) \leq I(V; XZ|YU) = 0, \tag{5.74}$$

and the non-negativity of mutual information. Finally observe that one can force $\hat{Z} \to ZU\hat{Y} \to XY$ to be a Markov chain without affecting (5.73) and thus satisfies Condition v) below (5.64). This completes the proof of the converse of Theorem 16.

The direct part of Theorem 16 can be shown by using the standard random coding argument, which is omitted in this thesis. This completes the proof of Theorem 16.

Theorem 15 and Theorem 16 together imply that for an IID vector video source (X_1, X_2, X_3) such that X_1 and X_2 are independent. If X_1, X_2, and X_3 do not form a (first-order) Markov chain, then under some mild conditions on D_1, D_2, and D_3 ($D_1, D_2 > 0, D_3 \geq 0$ and satisfying Condition B),

$$R_c^*(D_1, D_2, D_3) < R_p^*(D_1, D_2, D_3).$$

Condition B is generally met at points (D_1, D_2, D_3), $D_1 > 0, D_2 > 0, D_3 \geq 0$, for which positive bit rates are needed at all the decoder for X_1, X_2, and X_3 in order for them to produce the respective reproductions with the desired distortions D_1, D_2, and D_3. Such distortion points will be called points with positive rates. By using the technique demonstrated in the proof of Property 1 in [52], it can be shown that $R_c^*(D_1, D_2, D_3)$ has a negative subgradient at any point (D_1, D_2, D_3), $D_1 > 0, D_2 > 0, D_3 > 0$ (or at any point (D_1, D_2), $D_1 > 0, D_2 > 0,$ when $D_3 = 0$), with positive rates. In addition, the distribution $\mathbf{P}_c(Q)$, if optimal, generally depends on X_1 (except for some corner cases) when X_1, X_2, and X_3 do not form a Markov chain. In the following, we present a simple example to illustrate Theorem 15.

Example 5: Let $\mathbf{X} = (X_1, X_2, X_3)$ denote a memoryless video source with $\mathcal{X}_i = \mathcal{X} \overset{\Delta}{=} \{0, 1\}, i = 1, 2, 3$. Suppose that $\hat{\mathcal{X}_i} = \hat{\mathcal{X}} \overset{\Delta}{=} \{0, 1\}$, and Hamming distortion measure, denoted by d, is used for all three sources. Further suppose that X_1 and X_2 are independent, and $X_3(j) = X_1(j) + X_2(j)$ for $j \geq 1$, where '+' denotes modulo-2 addition. Let $p_1 \stackrel{\Delta}{=} P(X_1 = 1)$ and $p_2 \stackrel{\Delta}{=} P(X_2 = 1)$ such that $0 < p_1 \leq 0.5$ and $0 < p_2 \leq 0.5$. In this example, we would like to show

$$R_c^*(D_1, D_2, 0) < R_p^*(D_1, D_2, 0) \tag{5.75}$$

when $D_1 > 0$ and $0 < D_2 < p_2$.

The main difficulty of deriving (5.75) lies in the fact that there is no known algorithm to compute $R_p^*(D_1, D_2, 0)$ even though it has a single-letter expression. To circumvent this problem, we instead look at the computation of $R_c^*(D_1, D_2, 0)$. More specifically, we consider the computation of[2]

$$\min_{U, \hat{X}_2} \Big[I(X_1; U) + I(X_1 X_2; \hat{X}_2 | U) + H(X_3 | U \hat{X}_2) \Big], \tag{5.76}$$

where the minimization is taken over all random variables U from a finite alphabet $\mathcal{U}$ and $\hat{X}_2$ from $\hat{\mathcal{X}}$ satisfying the following conditions: i) there exists a function g such that $\hat{X}_1 = g(U)$ and $\mathbf{E}d_1(X_1, \hat{X}_1) \leq D_1$; ii) $\mathbf{E}d_2(X_2, \hat{X}_2) \leq D_2$; iii) $U \to X_1 \to X_2 X_3$ is a Markov chain; and iv) $\hat{X}_2 \to X_1 X_2 U \to X_3$ is a Markov chain.

In order to calculate (5.76), we define a function $F_{\alpha,\beta}$ by

$$\begin{aligned} & F_{\alpha,\beta}(P_{U|X_1}, P_{\hat{X}_2|X_1^2 U}, Q_{U\hat{X}_2X_3}) \stackrel{\Delta}{=} \\ & \sum_{x_1,u} P(x_1) P(u|x_1) \Bigg[\log \frac{P(u|x_1)}{Q(u) 2^{-\alpha d(x_1, g(u))}} + \\ & \sum_{x_2, \hat{x}_2} P(x_2) P(\hat{x}_2 | x_1^2 u) \log \frac{P(\hat{x}_2 | x_1^2 u)}{Q(\hat{x}_2 x_3 | u) 2^{-\beta d(x_2, \hat{x}_2)}} \Bigg], \end{aligned} \tag{5.77}$$

[2] It follows from the proof of Theorem 2 that (5.76) is equal to the right-hand-side of (3.8) for this particular IID vector source (X_1, X_2, X_3) when $D_3 = 0$.

where $x_3 = x_1 + x_2$, and α and β are Lagrange multipliers. In the above, for brevity we use the convention that $x_i \in \mathcal{X}$ and $\hat{x}_i \in \hat{\mathcal{X}}$ are values taken by X_i and $\hat{X}_i$, respectively. Thus, the summation $\sum_{x_i,\hat{x}_i}$ is over all $x_i \in \mathcal{X}$ and $\hat{x}_i \in \hat{\mathcal{X}}$, $P(x_i)$ means $P_{X_i}(x_i)$, $P(\hat{x}_i|x_i)$ means $P_{\hat{X}_i|X_i}(\hat{x}_i|x_i)$, and so on. Such a convention will be applied whenever it does not cause confusion. Further define

$$F^*_{\alpha,\beta} \triangleq \min_{\mathbf{P},Q} F_{\alpha,\beta}(\mathbf{P}, Q),$$

where the minimization is take over all $(P_{U|X_1}, P_{\hat{X}_2|X_1^2U}) \in \mathcal{P}_{\mathcal{U}|\mathcal{X}} \times \mathcal{P}_{\hat{\mathcal{X}}|\mathcal{X}^2\times\mathcal{U}}$ denoted by $\mathbf{P}$, and $Q_{U\hat{X}_2X_3} \in \mathcal{P}_{\mathcal{U}\times\hat{\mathcal{X}}\times\mathcal{X}}$ denoted by Q.

The definition of $F_{\alpha,\beta}(\mathbf{P}, Q)$ suggests that in order to find a solution to $F^*_{\alpha,\beta}$, one can alternatively compute $\mathbf{P}$ and Q. Indeed, we specialize our alternating minimization algorithm (hereafter "Algorithm A") to compute $F^*_{\alpha,\beta}$ as follows.

Step 1: Initialize $i = 0$ and $Q^{(0)} \in \mathcal{P}_{\hat{\mathcal{X}}\times\mathcal{X}\times\mathcal{U}}$.

Step 2: Fix $Q^{(i)}$. Find $\mathbf{P}^{(i+1)} \in \mathcal{P}_{\mathcal{U}|\mathcal{X}} \times \mathcal{P}_{\hat{\mathcal{X}}|\mathcal{X}^2\times\mathcal{U}}$ such that

$$\mathbf{P}^{(i+1)} \triangleq \operatorname*{arg\,min}_{\mathbf{P}\in\mathcal{P}_{\mathcal{U}|\mathcal{X}}\times\mathcal{P}_{\hat{\mathcal{X}}|\mathcal{X}^2\times\mathcal{U}}} F_{\alpha,\beta}(\mathbf{P}, Q^{(i)}). \tag{5.78}$$

Specifically, $\mathbf{P}^{(i+1)}$ can be further derived as follows: for any $(x_1, x_2, u, \hat{x}_2) \in \mathcal{X}^2 \times \mathcal{U} \times \hat{\mathcal{X}}$,

$$P^{(i+1)}(\hat{x}_2|x_1^2u) = \frac{Q^{(i)}(\hat{x}_2x_3|u)2^{-\beta d(x_2,\hat{x}_2)}}{\Delta^{(i)}_{x_1^2u}}$$

where $\Delta^{(i)}_{x_1^2u} \triangleq \sum_{\hat{x}_2} Q^{(i)}(\hat{x}_2x_3|u)2^{-\beta d(x_2,\hat{x}_2)}$, and $x_3 = x_1 + x_2$; and

$$P^{(i+1)}(u|x_1) = \frac{Q^{(i)}(u)2^{-\alpha d(x_1,g(u))+\sum_{x_2} p(x_2)\log\Delta^{(i)}_{x_1^2u}}}{\Gamma^{(i)}_{x_1}}$$

where $\Gamma^{(i)}_{x_1} \triangleq \sum_u Q^{(i)}(u)2^{-\alpha d(x_1,g(u))+\sum_{x_2} p(x_2)\log\Delta^{(i)}_{x_1^2u}}$.

Step 3: Fix $\mathbf{P}^{(i+1)}$. Find $Q^{(i+1)}$ such that

$$Q^{(i+1)} \triangleq \arg\min_{Q \in \mathcal{P}_{\hat{\mathcal{X}} \times \mathcal{X} \times \mathcal{U}}} F_{\alpha,\beta}(\mathbf{P}^{(i+1)}, Q). \tag{5.79}$$

(5.79) is solved by the following equation: for any $(u, x_3, \hat{x}_2) \in \mathcal{U} \times \mathcal{X} \times \hat{\mathcal{X}}$,

$$Q^{(i+1)}(\hat{x}_2 x_3 u) = \sum_{\substack{x_1, x_2 : \\ x_1 + x_2 = x_3}} p(x_1) p(x_2) P^{(i+1)}(u|x_1) P^{(i+1)}(\hat{x}_2|x_1^2 u).$$

Step 4: Increase i by 1.

Step 5: Repeat Steps 2–4 until a stationary point is reached.

Algorithm A generates a sequence of distribution pairs $(\mathbf{P}^{(1)}, Q^{(1)}), (\mathbf{P}^{(2)}, Q^{(2)}), \cdots$ such that

$$F_{\alpha,\beta}(\mathbf{P}^{(1)}, Q^{(0)}) \geq F_{\alpha,\beta}(\mathbf{P}^{(1)}, Q^{(1)}) \geq F_{\alpha,\beta}(\mathbf{P}^{(2)}, Q^{(1)}) \cdots .$$

By using an argument similar to that used to prove the convergence of the Blahut-Arimoto algorithm in [12], [13], it was shown in Section 4.2 that $(P^{(1)}, Q^{(1)}), (P^{(2)}, Q^{(2)}), \cdots$ indeed converges to a solution to $F^*_{\alpha,\beta}$.

Equipped with Algorithm A, which was proven that it converges to a solution to $F^*_{\alpha,\beta}$, we are now ready to show (5.75). Our strategy is to show that for any finite $0 < \alpha, \beta < \infty$, any $P \in \mathcal{P}_{\hat{\mathcal{X}}|\mathcal{X}^2 \times \mathcal{U}}$ that gives rise to a Markov chain $\hat{X}_2 \to X_2 U \to X_1 X_3$ cannot be a stationary point in computing $F^*_{\alpha,\beta}$ by using Algorithm A.

Let $\mathbf{P} = (P_{U|X_1}, P_{\hat{X}_2|X_1^2 U})$ denote a vector of two transition probability functions in $\mathcal{P}_{\mathcal{U}|\mathcal{X}} \times \mathcal{P}_{\hat{\mathcal{X}}|\mathcal{X}^2 \times \mathcal{U}}$ such that

$$P_{\hat{X}_2|X_1^2 U}(\hat{x}_2|0 x_2 u) = P_{\hat{X}_2|X_1^2 U}(\hat{x}_2|1 x_2 u) \tag{5.80}$$

for all $(x_2, u, \hat{x}_2) \in \mathcal{X} \times \mathcal{U} \times \hat{\mathcal{X}}$. Let Q denote a function in $\mathcal{P}_{\hat{\mathcal{X}} \times \mathcal{X} \times \mathcal{U}}$ derived from $\mathbf{P}$ as in Step 3 of Algorithm A, that is,

$$Q(\hat{x}_2 x_3 u) = \sum_{x_1, x_2 : x_1 + x_2 = x_3} p(x_1) p(x_2) P(u|x_1) P(\hat{x}_2 | x_1^2 u). \tag{5.81}$$

Suppose that $(\mathbf{P}, Q)$ is a stationary point. We would like to derive a contradiction from the set of equations (5.80) and (5.81). To that end, let us initialize $Q^{(0)}$ as Q, and compute $P^{(1)}_{\hat{X}_2|X_1^2 U}$ by using Algorithm A as

$$P^{(1)}(\hat{x}_2 | x_1^2 u) = \frac{Q^{(0)}(\hat{x}_2 x_3 | u) 2^{-\beta d(x_2, \hat{x}_2)}}{\sum_{\hat{x}_2} Q^{(0)}(\hat{x}_2 x_3 | u) 2^{-\beta d(x_2, \hat{x}_2)}}. \tag{5.82}$$

According to the assumed fact that $(\mathbf{P}, Q)$ is a stationary point, we have that $\mathbf{P}^{(1)} = \mathbf{P}$.

For any u, define $q_u^{(0)}(\hat{x}_2 | x_3) \triangleq Q^{(0)}(\hat{x}_2 x_3 | u) / Q^{(0)}(x_3 | u)$, $\bar{q}_u^{(0)} \triangleq 0.5(q_u^{(0)}(\cdot|0) + q_u^{(0)}(\cdot|1))$, and $\delta_u^{(0)} \triangleq 0.5(q_u^{(0)}(0|0) - q_u^{(0)}(0|1))$. Then

$$\begin{aligned}
P^{(1)}(\hat{x}_2 | 00u) &= \frac{q_u^{(0)}(\hat{x}_2|0) 2^{-\beta d(0,\hat{x}_2)}}{q_u^{(0)}(0|0) + q_u^{(0)}(1|0) 2^{-\beta}}, && (5.83)\\
P^{(1)}(\hat{x}_2 | 10u) &= \frac{q_u^{(0)}(\hat{x}_2|1) 2^{-\beta d(0,\hat{x}_2)}}{q_u^{(0)}(0|1) + q_u^{(0)}(1|1) 2^{-\beta}}, && (5.84)\\
P^{(1)}(\hat{x}_2 | 01u) &= \frac{q_u^{(0)}(\hat{x}_2|1) 2^{-\beta d(1,\hat{x}_2)}}{q_u^{(0)}(0|1) 2^{-\beta} + q_u^{(0)}(1|1)}, \text{ and} && (5.85)\\
P^{(1)}(\hat{x}_2 | 11u) &= \frac{q_u^{(0)}(\hat{x}_2|0) 2^{-\beta d(1,\hat{x}_2)}}{q_u^{(0)}(0|0) 2^{-\beta} + q_u^{(0)}(1|0)}. && (5.86)
\end{aligned}$$

Calculate

$$\begin{aligned}
&|P^{(1)}_{\hat{X}_2|X_1^2 U}(\hat{x}_2 | 00u) - P^{(1)}_{\hat{X}_2|X_1^2 U}(\hat{x}_2 | 10u)| \\
= &\left| \frac{\delta_u^{(0)} 2^{1-\beta}}{(\bar{q}_u^{(0)}(0) + \bar{q}_u^{(0)}(1) 2^{-\beta})^2 - (\delta_u^{(0)}(1 - 2^{-\beta}))^2} \right|. && (5.87)
\end{aligned}$$

Since $P^{(1)}_{\hat{X}_2|X_1^2U} = P_{\hat{X}_2|X_1^2U}$ which satisfies (5.80), (5.87) implies that

$$\delta_u^{(0)} = 0 \text{ for } \beta < \infty. \tag{5.88}$$

Let $p_u \overset{\Delta}{=} \frac{p_1 P_{U|X_1}(u|1)}{(1-p_1)P_{U|X_1}(u|0)+p_1 P_{U|X_1}(u|1)}$. It follows from the definition of $\delta_u^{(0)}$ above that

$$\delta_u^{(0)} = \frac{(1-2p_u)p_2(1-p_2)(P(0|00u)-P(0|01u))}{2(p_u * p_2)(1-p_u * p_2)}, \tag{5.89}$$

where $p_u * p_2 \overset{\Delta}{=} (1-p_u)p_2 + p_u(1-p_2)$. Equation (5.89), together (5.88), implies that one of the following equalities must hold

$$p_u = 0.5; \tag{5.90}$$

$$P(0|00u) = P(0|01u). \tag{5.91}$$

In the following we argue that (5.90) cannot hold for any u with $Q(u) > 0$. Suppose to the contrary there exists u with $p_u = 0.5$. Then

$$p_1 P_{U|X_1}(u|1) = (1-p_1)P_{U|X_1}(u|0), \tag{5.92}$$

which, together with Step 2 in Algorithm A, implies that the following equality must hold:

$$\frac{p_1 2^{-\alpha d(1,g(u))}}{\Gamma_1} = \frac{(1-p_1)2^{-\alpha d(0,g(u))}}{\Gamma_0}, \tag{5.93}$$

where $\Gamma_{x_1} \overset{\Delta}{=} \sum_u Q(u) 2^{-\alpha d(x_1,g(u)) + \sum_{x_2} p(x_2)\log \Delta_{x_2 u}}$, and $\Delta_{x_2 u} \overset{\Delta}{=} \sum_{\hat{x}_2} Q(\hat{x}_2|u) 2^{-\beta d(x_2,\hat{x}_2)}$ due to (5.88). Without losing generality, we assume that $g(u) = 0$, and (5.93) reduces to

$$\begin{aligned} & p_1 \sum_u Q(u) 2^{-\alpha(1+d(0,g(u)))+\Lambda_u} \\ = \; & (1-p_1) \sum_u Q(u) 2^{-\alpha d(1,g(u))+\Lambda_u}, \end{aligned} \tag{5.94}$$

where $\Lambda_u \triangleq \sum_{x_2} p(x_2) \log \Delta_{x_2 u}$. (5.94) can be rewritten as

$$\begin{aligned}&\sum_u Q(u) 2^{\Lambda_u} [p_1 2^{-\alpha(1+d(0,g(u)))} - \\ &(1-p_1) 2^{-\alpha d(1,g(u))}] = 0.\end{aligned} \tag{5.95}$$

Observe that $\forall u$ with $Q(u) > 0$, we have $2^{\Lambda_u} > 0$, and

$$p_1 2^{-\alpha(1+d(0,g(u)))} \leq (1-p_1) 2^{-\alpha d(1,g(u))}$$

due to that $d(1, g(u)) \leq 1 + d(0, g(u))$ for any u, and $p_1 \leq 0.5 \leq 1 - p_1$. Therefore, (5.95) holds if and only if

$$p_1 2^{-\alpha(1+d(0,g(u)))} - (1-p_1) 2^{-\alpha d(1,g(u))} = 0. \tag{5.96}$$

which implies that for $\alpha < \infty$,

$$p_1 = \begin{cases} 0.5 & \text{if } g(u) = 0 \\ \frac{1}{1+2^{-2\alpha}} & \text{if } g(u) = 1 \end{cases}. \tag{5.97}$$

Since $\frac{1}{1+2^{-2\alpha}} \neq 0.5$ whenever $\alpha > 0$, the equality (5.96) cannot hold for every u. Consequently, the equality in (5.95) cannot hold, which implies $p_u \neq 0.5$ for any u with $Q(u) > 0$.

Since (5.90) does not hold, it follows from our argument above that (5.91) must hold for all u with $Q(u) > 0$, which, together with (5.80), implies that given U, $\hat{X}_2$ is independent of X_2. Since X_2 and U are independent, one readily sees that in this case the distortion constraint $D_2 < p_2$ cannot be achieved. Thus (5.91) cannot hold either.

The fact that neither (5.90) nor (5.91) holds contradicts directly (5.80). We thus conclude that $(\mathbf{P}, Q)$ is not a stationary point in Algorithm A, which, together with the convergence property of Algorithm A, leads to the desired inequality (5.75).

Chapter 6

On Optimum Fixed-Rate Causal Scalar Quantization Design for Causal Video Coding

Following the line of modeling each frame as a stationary source, in this chapter, we look at how to design specific codes for CVC. As a starting point, we shall focus on scalar quantization. To this end, we first put forth a concept called causal scalar quantization for the CVC shown in Figure 1.3 and investigate how to design optimal fixed-rate causal scalar quantizers (CSQ). By extending the classic Lloyd-Max algorithm for a single source to this multiple sources case, we then propose an algorithm for designing optimum fixed-rate CSQ to minimize the total distortion among all sources. The proposed algorithm converges in the sense that the total distortion cost is monotonically decreasing until a stationary point is reached. Simulation results show that in comparison with fixed-rate predictive scalar quantization, fixed-rate causal scalar quantization offers as large as 16% quality improvement (distortion reduction). Since PVC is what all previous and current video coding standards fall into, the rate-distortion performance gain of fixed-rate CSQ for CVC over fixed-rate PSQ for PVC would be instructive to practice.

6.1 Formal Definition of Causal Scalar Quantizers(CSQ)

In this thesis, the design of CVC is considered from an information theoretic perspective by modeling each frame $X_k, k = 1, \cdots, N$, as a stationary information source $X_k = \{X_k(i)\}_{i=1}^{\infty}$ taking values in the real line $\mathcal{R}$.

As a starting point, in this thesis, we consider the simplest block codes: one-dimensional block codes, which are also called scalar quantizers. Formally we define a causal scalar quantizer (CSQ) by using N encoder and decoder pairs $(\psi_k, g_k), k = 1, \cdots, N$, such that

$$\begin{aligned} &\psi_1 : \mathcal{R} \to \mathcal{U}_1,\ g_1 : \mathcal{U}_1 \to \mathcal{R}, \\ &\psi_i : \Pi_{j=1}^{i} \mathcal{R} \times \Pi_{j=1}^{i-1} \mathcal{U}_j \to \mathcal{U}_i, \\ &g_i : \Pi_{j=1}^{i} \mathcal{U}_i \to \mathcal{R},\ \text{for } i = 2, \cdots, N. \end{aligned}$$

The encoding output of X_1 is given by $U_1 = \psi_1(X_1)$, where $U_1 \in \mathcal{U}_1 \overset{\Delta}{=} \{1, 2, \cdots, L_1\}$, and the reconstruction of X_1 is defined as $\hat{X}_1 = g_1(U_1)$ taking values from the reproduction alphabet $\mathcal{A}_1 \overset{\Delta}{=} \{a_1, \cdots, a_{L_1}\} \subset \mathcal{R}$. In addition, the encoding outputs of $X_i, i = 2, \cdots, N$, are given by $U_i = \psi_i(X_i^-, U_i^-, X_i)$, where $U_i \in \mathcal{U}_i \overset{\Delta}{=} \{1, 2, \cdots, L_i\}$, and conditional on each U_i^-, the reconstruction of X_i is defined as $\hat{X}_{U_i^-,i} = g_i(U_i^-, U_i)$ drawn from the reproduction alphabet $\mathcal{A}_{\mathcal{U}_i^-,i} \overset{\Delta}{=} \{a_{U_i^-,1}, \cdots, a_{U_i^-,L_i}\} \subset \mathcal{R}$. In the above, $g_k, k = 1, \cdots, N$ is called a decoder, and CSQ reduces to predictive scalar quantizers (PSQ) for PVC if ψ_i becomes

$$\psi_i : \mathcal{R} \times \Pi_{j=1}^{i-1} \mathcal{U}_j \to \mathcal{U}_i,\ \text{for } i = 2, \cdots, N.$$

If we use CSQ (PSQ, respectively) to encode N length-n stationary sources $X_1(1; n), \cdots, X_N(1; n)$, the corresponding encoding output is denoted by $U_1(1; n) = \{\psi_1(X_1(i))\}_{i=1}^{n}, U_k(1; n) =$

$\{\psi_k(X_k^-(i), U_k^-(i), X_k(i))\}_{i=1}^n$,$k = 2, \cdots, N$, ($U_k(1;n) = \{\psi_k(U_k^-(i), X_k(i))\}_{i=1}^n$ in PSQ case, respectively) drawn from $\mathcal{U}_j = \{1, \cdots, L_j\}, j = 1, \cdots, N$, respectively. Specifically, for every individual sequence $\mathbf{x}_k = \{x_k(i)\}_{i=1}^n \in \mathcal{R}^n$, the encoder finds an index sequence $\mathbf{u}_k = \{u_k(i)\}_{i=1}^n \in \mathcal{U}_k^n$ on observing $\mathbf{x}_1^k$ and $\mathbf{u}_1^{k-1}$ (on observing $\mathbf{x}_k$ and $\mathbf{u}_1^{k-1}$ in PSQ case, respectively), and then encodes $\mathbf{x}_k$ into a binary codeword associated with $\mathbf{u}_k$ via some lossless codeword compression method. Let $\mathcal{U}_k^n$ denote the set of all sequences of length n from $\mathcal{U}_k$. A function $l : \mathcal{U}_k^n \to \{1, 2, \cdots\}$ is called a lossless codeword length function if for any $n \geq 1$, and we have $\sum_{\mathbf{u}_k \in \mathcal{U}_k^n} 2^{-l(\mathbf{u}_k)} \leq 1$. It is easy to see that for any lossless codeword length function l, there exists a prefix code $\phi_k : \mathcal{U}_k^n \to \{0, 1\}^*$ such that for any $\mathbf{u}_k \in \mathcal{U}_k^n$, $l(\mathbf{u}_k)$ is the length of $\phi_k(\mathbf{u}_k)$. In other words, ϕ_k denotes a mapping from the range of ψ_k^n (denoted by $\mathbf{u}_k = \psi_k^n(\mathbf{x}_k)$) in $\mathcal{U}_k^n$ to a prefix subset of $\{0, 1\}^*$ of finite binary strings. After receiving the binary codeword, the decoder first recovers $\mathbf{u}_k$ and then outputs $g_k(\mathbf{u}_k^-, \mathbf{u}_k) = \{g_k(u_k^-(i), u_k(i))\}_{i=1}^n$ as a reproduction of $\mathbf{x}_k$. To convert $\mathbf{u}_k$ into a binary sequence, one may encode each CSQ index $u_k(i) \stackrel{\Delta}{=} \psi_k(x_k^-(i), u_k^-(i), x_k(i))$ (PSQ index $u_k(i) \stackrel{\Delta}{=} \psi_k(u_k^-(i), x_k(i))$, respectively) into a binary sequence of length $\lceil \log_2 L_k \rceil$ with L_k being fixed; in this case, the corresponding block codes (or scalar quantizers) are called fixed-rate CSQ (fixed-rate PSQ, respectively). The performance of fixed-rate CSQ (PSQ, respectively) is measured by its average total rate $R = \sum_{k=1}^N \lceil \log_2 L_k \rceil$ among all sources in bits per symbol and the resulting average total distortion per symbol

$$\begin{aligned} D &= \frac{1}{n} \sum_{k=1}^N \mathbf{E} d_k(X_k(1;n), g_k(U_k^-(1;n), U_k(1;n))) \\ &= \sum_{k=1}^N \mathbf{E} d_k(X_k(1), g_k(U_k^-(1), U_k(1))), \end{aligned} \tag{6.1}$$

where the second equality follows from the stationarity of sources. On the other hand, one may apply a universal lossless codeword compression algorithms such as Lempel-Ziv coding [57], and Grammar-based coding [23] to encode the sequence $\mathbf{u}_k$; in this case, the corresponding block codes are called variable-rate CSQ (variable-rate PSQ, respectively). The performance of variable-rate CSQ (PSQ, respectively) is measured by its average total rate $R = \sum_{k=1}^{N} R_k = \sum_{k=1}^{N} \mathbf{E}\frac{l(U_k(1;n))}{n}$ among all sources in bits per symbol and the resulting average total distortion D defined in (6.1). Of course, the rate R_k for encoding $X_k(1;n)$ depends on the lossless codeword length function l. So if l is selected to be some universal lossless codeword length function, the remaining problem for designing variable-rate CSQ (PSQ, respectively) is how to jointly optimize D, R, and actually $U_k(1;n)$, which may be regarded as variable-rate *soft-decision* [51] CSQ (PSQ, respectively). It is not hard to see that fixed-rate CSQ (PSQ, respectively) is a special case of variable-rate CSQ (PSQ, respectively), and better compression performance can be achieved by using variable-rate CSQ (PSQ, respectively).

Remark 12 *It is worthwhile to point out the difference between CVC and the causal source coding setup in [29], in which a source encoder is causal in the sense that the encoder for a source frame $X = \{X(i)\}_{i=1}^{\infty}$ can only view all past source pixels up to the current pixel position, but not allowed to access to future ones. On the other hand, as shown in Figure 1.3, CVC is causal in a frame-temporal sense that the encoder for X_k can only view all previous source frames $X_1, \cdots, X_{k-1}$ up to the current frame without enlisting the help from future frames in the encoder order. In other words, in [29], the number of frames N is always equal to 1, and the code is causal in terms of the pixel position i only; while in CVC, the code is causal in terms of the frame number index $k \in \{1, 2, \cdots, N\}$ for $N \geq 1$. As such, [29] only handles a frame-averaged expected distortion criterion as opposed to frame-specific individual distortion constraints treated in CVC.*

In this chapter, we use mean-squared error distortion criterion to evaluate D, and only focus on designing the optimum fixed-rate CSQ such that D in (6.1) is minimized among all fixed-rate CSQ with L_i being fixed, i.e., the average total compression rate $R = \sum_{i=1}^{N} \lceil \log_2 L_i \rceil$ being fixed. Therefore, to design the optimal fixed-rate CSQ (PSQ, respectively), one has to solve the following problems.

Q1: For each U_i^- value, how to design the decoder $g_i, i = 1, \cdots, N$, (or equivalently, the reproduction alphabet $\mathcal{A}_{\mathcal{U}_i^-,i}$) in an optimum way?

Q2: For each (U_i^-, X_i^-) pair in CSQ case (each U_i^- in PSQ case, respectively), how to map each source symbol $X_i \in \mathcal{R}$ to one of L_i reproduction symbols from $\mathcal{A}_{\mathcal{U}_i^-,i} \in \mathcal{R}$?

Note that a real number x_i will be quantized to different output levels conditional on different (x_i^-, u_i^-) pairs (conditional on different u_i^- values in PSQ case, respectively). Thus, the information of (x_i^-, u_i^-) (u_i^- in PSQ case, respectively) are conveyed in the choice of different quantization output of x_i. In this chapter, our purpose is to design an algorithm to simultaneously solve Problem Q1 and Q2.

6.2 An iterative algorithm to design optimum fixed-rate CSQ

In this section, we extend the well-known Lloyd-Max algorithm [24][28] for designing an optimal fixed-rate scalar quantizer for a single source X to the multiple sources case. The modified Lloyd-Max algorithm allows us to design the optimum fixed-rate CSQ for CVC. The proposed algorithm converges in the sense that the total distortion (6.1) is monotonically decreasing until a stationary point is reached. In addition, it allows us to do comparisons between the rate distortion performance of fixed-rate CSQ and that of fixed-rate PSQ.

Without loss of generality, we consider the case $N = 3$. Three memoryless video sources $X, Y,$ and Z taking values from $\mathcal{R}$ are drawn from three random variables with the joint distributions $p_{XYZ}(xyz)$. For simplicity, we write the quantization index vector (U_1, U_2, U_3) in response to (X, Y, Z) as $\mathbf{U} \overset{\Delta}{=} (I, J, K)$, and their realizations $i \in \{1, \cdots, L_1\}, j \in \{1, \cdots, L_2\}$, and $k \in \{1, \cdots, L_3\}$. In addition, we denote the codebook $(\hat{X}_I, \hat{Y}_{I,J}, \hat{Z}_{I,J,K})$ as $\mathbf{B}$, where $\hat{X}_I \overset{\Delta}{=} g_1(I), \hat{Y}_{I,J} \overset{\Delta}{=} g_2(I, J), \hat{Z}_{I,J,K} \overset{\Delta}{=} g_3(I, J, K)$, and their realizations $\hat{x}_i, \hat{y}_{i,j}$, and $\hat{z}_{i,j,k} \in \mathcal{R}$. In order to design the optimal fixed-rate CSQ to achieve the minimum total distortion defined in (6.1), we try to find a $(\mathbf{B}^*, \mathbf{U}^*)$ pair that minimizes

$$D(\mathbf{B}, \mathbf{U}) = \mathbf{E}\Big[(X - \hat{X}_I)^2 + (Y - \hat{Y}_{I,J})^2 + (Z - \hat{Z}_{I,J,K})^2\Big]$$

over all possible $(\mathbf{B}, \mathbf{U})$ pairs. The iterative algorithm works as follows.

Algorithm B: A Modified Lloyd-Max Algorithm

Step 1: Select an initial quantization index vector $\mathbf{U}^{(0)} \overset{\Delta}{=} (I^{(0)}, J^{(0)}, K^{(0)})$ and set the initial codebook $\mathbf{B}^{(0)} \overset{\Delta}{=} (\hat{X}^{(0)}_{I^{(0)}} \hat{Y}^{(0)}_{I^{(0)}, J^{(0)}}, \hat{Z}^{(0)}_{I^{(0)}, J^{(0)}, K^{(0)}})$ such that $\hat{x}^{(0)}_i = \mathbf{E}\left[X \mid I^{(0)} = i\right], \hat{y}^{(0)}_{i,j} = \mathbf{E}\left[Y \mid I^{(0)} = i, J^{(0)} = j\right]$, and $\hat{z}^{(0)}_{i,j,k} = \mathbf{E}\left[Z \mid I^{(0)} = i, J^{(0)} = j, K^{(0)} = k\right]$.

Step 2: Fix $\mathbf{B}^{(t)} \overset{\Delta}{=} (\hat{X}^{(t)}_{I^{(t)}}, \hat{Y}^{(t)}_{I^{(t)}, J^{(t)}}, \hat{Z}^{(t)}_{I^{(t)}, J^{(t)}, K^{(t)}})$. Given $\mathbf{U}^{(t)}$, update $\mathbf{U}^{(t+1)} = (I^{(t+1)}, J^{(t+1)}, K^{(t+1)})$ that minimizes

$$\mathbf{E}\left[(X - \hat{X}^{(t)}_I)^2 + (Y - \hat{Y}^{(t)}_{I,J})^2 + (Z - \hat{Z}^{(t)}_{I,J,K})^2\right]$$

over all possible quantization index vectors $\mathbf{U} = (I, J, K)$.

Step 3: Fix $\mathbf{U}^{(t+1)} = (I^{(t+1)}, J^{(t+1)}, K^{(t+1)})$. Find $\mathbf{B}^{(t+1)} \overset{\Delta}{=} (\hat{X}^{(t+1)}_{I^{(t+1)}}, \hat{Y}^{(t+1)}_{I^{(t+1)}, J^{(t+1)}}, \hat{Z}^{(t+1)}_{I^{(t+1)}, J^{(t+1)}, K^{(t+1)}})$ to minimize

$$\begin{aligned}&\mathbf{E}[(X - \hat{X}_{I^{(t+1)}})^2 + (Y - \hat{Y}_{I^{(t+1)}, J^{(t+1)}})^2 + \\ &(Z - \hat{Z}_{I^{(t+1)}, J^{(t+1)}, K^{(t+1)}})^2]\end{aligned}$$

over all codebook sets $\mathbf{B}$.

Step 4: Increase t by 1. Record $D(\mathbf{B}^{(t)}, \mathbf{U}^{(t)})$ by $D^{(t)}$.

Step 5: Repeat steps $2 - 4$ until $D^{(t)} - D^{(t+1)}$ is smaller than a prescribed threshold.

In step 3, $\mathbf{B}^{(t+1)}$ can be updated as

$$\begin{aligned} \hat{x}_i^{(t+1)} &= \mathbf{E}_X\left[X \mid I^{(t+1)} = i\right], \\ \hat{y}_{i,j}^{(t+1)} &= \mathbf{E}_Y\left[Y \mid I^{(t+1)} = i, J^{(t+1)} = j\right], \\ \hat{z}_{i,j,k}^{(t+1)} &= \mathbf{E}_Z\left[Z \mid I^{(t+1)} = i, J^{(t+1)} = j, K^{(t+1)} = k\right], \end{aligned}$$

where $1 \leq i \leq L_1$, $1 \leq j \leq L_2$ and $1 \leq k \leq L_3$, $\mathbf{E}_V[\cdot]$ denotes the expectation with respect to the random variable V, and $\mathbf{E}_V\left[\cdot|A = a\right]$ denotes the conditional expectation with respect to V given the event $A = a$. If there is no ambiguity, we sometimes omit the random variables in the condition. For example, we may write $\mathbf{E}_Y\left[Y \mid x\right]$ instead of $\mathbf{E}_Y\left[Y \mid X = x\right]$.

In step 2, fix $\mathbf{B}^{(t)}$, and given $\mathbf{U}^{(t)}$, we update $\mathbf{U}^{(t+1)}$ as follows:

$$k^{(t+1)} = \underset{k \in \{1,\cdots,L_3\}}{\arg\min}\,(z - \hat{z}_{i^{(t)},j^{(t)},k})^2; \tag{6.2}$$

$$j^{(t+1)} = \underset{j \in \{1,\cdots,L_2\}}{\arg\min}\,[(y - \hat{y}_{i^{(t)},j})^2 + \Delta^{(t)}(x, y, i^{(t)}, j)], \tag{6.3}$$

where $\Delta^{(t)}(x, y, i^{(t)}, j) \triangleq \mathbf{E}_Z\left[(Z - \hat{Z}_{i^{(t)},j,K^{(t+1)}})^2 \mid x, y\right]$, in which $K^{(t+1)}$ is the updated index in response to Z; and

$$i^{(t+1)} = \underset{i \in \{1,\cdots,L_1\}}{\arg\min}\,[(x - \hat{x}_i)^2 + \Lambda^{(t)}(x, i)], \tag{6.4}$$

where

$$\begin{aligned} \Lambda^{(t)}(x, i) \triangleq & \mathbf{E}_Y\Big[(Y - \hat{Y}_{i,J^{(t+1)}})^2 + \\ & \mathbf{E}_Z[(Z - \hat{Z}_{i,J^{(t+1)},K^{(t+1)}})^2 \mid x, Y] \mid x\Big], \end{aligned}$$

and $J^{(t+1)}$ is the updated index in response to Y. In the above, (6.2) is from the classic nearest neighbor rule. In (6.3) and (6.4), think of $(y - \hat{y}_{i^{(t)},j})^2 + \Delta^{(t)}(x, y, i^{(t)}, j)$ and $(x - \hat{x}_i)^2 + \Lambda^{(t)}(x, i)$ as a new distortion measure respectively, and the nearest neighbor rule can be extended to our present case. Specifically, we utilize the Monte-Carlo method to handle the conditional expectation terms $\Delta^{(t)}(x, y, i^{(t)}, j)$ in (6.3) and $\Lambda^{(t)}(x, i)$ in (6.4). It can be verified that the algorithm does converge in the sense that for $t \geq 1$, $D(\mathbf{B}^{(t+1)}, \mathbf{U}^{(t+1)}) \leq D(\mathbf{B}^{(t)}, \mathbf{U}^{(t)})$, and hence, the above algorithm will produce a sequence of codebooks and index vectors with monotonically decreasing values of the object function $D(\mathbf{B}^{(t)}, \mathbf{U}^{(t)})$ which converges as $t \to \infty$.

Remark 13 *Note that the optimum fixed-rate scalar quantization design method for a single source [28] can not be applied directly to this multiple sources case, since the quantization of source X_k in CVC depends on not only the current source but also all previous sources and all previous quantization outputs.*

Remark 14 *Compared with the computation algorithm proposed in Section 4.1 to calculate $R_c^*(D_1, \cdots, D_N)$ via finding the optimal transitional probability and probability functions, Algorithm B finds the optimum mappings from real-value source symbols to finite sets of output levels in a deterministic way.*

The advantage of fixed-rate CSQ lies in its low implementation complexity, low time delay and immunity to error propagation for transmission over noisy channel. Many other advanced lossy compression methods can be built on fixed-rate CSQ, such as variable-rate CSQ, and vector causal quantization which extends the code design from one-dimensional block codes to n-dimensional block codes for any given dimension $n \geq 1$.

6.3 Simulation

In this section, we use specific example to numerically compare the minimum total distortion of fixed-rate CSQ with fixed-rate PSQ by running Algorithm B. In the implementation, we generate three length-n sequences $\{X(s), Y(s), X(s)\}_{s=1}^{n}$ from the joint probability density function p_{XYZ} with $n = 30,000$, and $\{I(s), J(s), K(s)\}_{s=1}^{n}$ represent the quantization index in response to $\{X(s), Y(s), X(s)\}_{s=1}^{n}$. Since PSQ is a special case of CSQ, we shall assign the same initialization $\mathbf{U}^{(0)}$ and $\mathbf{B}^{(0)}$ to both fixed-rate CSQ and fixed-rate PSQ cases. All updates in Algorithm B are according to the empirical distributions.

Example 6: Suppose (X, Y, Z) is jointly memoryless Gaussian sources. We consider two different covariance matrices of XYZ as follows:

$$\text{Case 1:} \begin{pmatrix} 1 & 0.1 & 1.1 \\ 0.1 & 1 & 1.1 \\ 1.1 & 1.1 & 2.21 \end{pmatrix}; \quad \text{Case 2:} \begin{pmatrix} 1 & 0.4 & 1.4 \\ 0.4 & 1 & 1.4 \\ 1.4 & 1.4 & 2.8 \end{pmatrix}.$$

In both of the two cases, we fix $L_1 = 5, L_3 = 3$, and select L_2 from 5 to 9. Table 6.3 shows the minimum total distortion D versus L_2 of optimum fixed-rate CSQ and that of optimum fixed-rate PSQ respectively for above two cases in Example 6.

L_2	5	6	7	8	9
Case 1:					
D of fixed-rate CSQ	0.2371	0.2048	0.1866	0.1722	0.1613
D of fixed-rate PSQ	0.2523	0.2202	0.2060	0.1931	0.1815
Case 2:					
D of fixed-rate CSQ	0.2228	0.1980	0.1736	0.1642	0.1554
D of fixed-rate PSQ	0.2433	0.2163	0.1956	0.1903	0.1807

Table 6.1: Comparison of D versus L_2 between fixed-rate CSQ and fixed-rate PSQ for fixed $L_1 = 5$ and $L_3 = 3$

As shown in Table 6.3, the quality improvement (distortion reduction) of optimum fixed-rate CSQ over optimum fixed-rate PSQ is as large as 12.5 percent and 16.3 percent for Case 1 and Case 2 respectively.

Chapter 7

Conclusion and Future Research

7.1 Conclusion

In this thesis, we have investigated CVC for encoding source frames $X_1, \cdots, X_N$ from an information theoretic point of view.

An iterative algorithm has been proposed to numerically compute the minimum total rate $R_c^*(D_1, \cdots, D_N)$ achievable asymptotically by CVC for jointly stationary and ergodic sources at distortion levels $D_1, \cdots, D_N \geq 0$, and analytically characterize $R_c^*(D_1, \cdots, D_N)$ for IID sources $(X_1, \cdots, X_N)$. The algorithm has been shown to converge globally. With the help of the algorithm, we have further established a somewhat surprising more and less coding theorem—under some conditions on source frames and distortion, the more frames need to be coded and transmitted, the less amount of data after encoding has to be sent! If the cost of data transmission is proportional to the transmitted data volume, this translates literally into a scenario where the more frames you download, the less you would pay. Numerical comparisons between CVC and greedy coding have shown that CVC offers significant performance gains over greedy coding. Along the way, we have advocated that whenever possible, the computational approach as illustrated in the thesis is a preferred approach to multi-user problems in information theory. In addition, we have also established an extended Markov lemma for correlated ergodic sources, which will be useful to other multi-user problems in information theory as well.

An interesting comparison has been made between CVC and PVC from an information theoretic point of view, where all MPEG-series and H-series video coding standards proposed so far are without exception defined within the paradigm of PVC. We first show that for general stationary ergodic sources $X_1, X_2, \cdots, X_N$, $R_p^*(D_1, \cdots, D_N)$ is equal to the infimum of the nth order total rate distortion function $R_{p,n}(D_1, \cdots, D_N)$ over all n, where $R_{p,n}(D_1, \cdots, D_N)$ itself is given by the minimum of an information quantity over a set of auxiliary random variables. We then prove that if the general stationary ergodic sources $X_1, \cdots, X_N$ form a (first-order) Markov chain, we have $R_p^*(D_1, \cdots, D_N) = R_c^*(D_1, \cdots, D_N)$. However, this is not true in general if $X_1, \cdots, X_N$ do not form a (first-order) Markov chain. Specifically, we demonstrate that in the case of $N = 3$, if X_1, X_2, X_3 do not form a (first-order) Markov chain, then under some conditions on source frames and distortion, $R_{c,n}(D_1, D_2, D_3)$ is strictly less than $R_{p,n}(D_1, D_2, D_3)$ in general for any finite $n > 0$. We then derive a single-letter characterization of $R_p^*(D_1, D_2, D_3)$ for an IID vector source (X_1, X_2, X_3) where X_1 and X_2 are independent, and demonstrate the existence of such X_1, X_2, X_3 for which $R_p^*(D_1, D_2, D_3) > R_c^*(D_1, D_2, D_3)$ under some conditions on source frames and distortion. We also present a simple example to illustrate it.

At the end, we consider the code design problem of CVC, and propose an iterative algorithm for designing optimum fixed-rate CSQ as a starting point. With the help of this algorithm, we observe as large as 16% quality improvement (distortion reduction) of fixed-rate CSQ for CVC over fixed-rate PSQ for PVC.

7.2 Direction for Future Research

If the information theoretic analysis as demonstrated in this thesis is indicative of the real performance of CVC for real video data, then the more and less coding theorem, the significant performance gain of CVC over greedy coding, the strict rate reduction from PVC to CVC, plus the quality improvement of fixed-rate CSQ over fixed-rate PSQ really point out a bright future for CVC. In this section, we offer some interesting directions for future research that emphases on both information theoretic perspective and practical implementation perspective of CVC.

1. In Chapter 4, we propose an iterative algorithm to calculate $R_{c,n}(D_1, \cdots, D_N)$ for N sources $X_1, \cdots, X_N$ with finite alphabet. For sources with continuous alphabet, we can either extend our algorithm from discrete sources to continuous case with global convergence, or obtain close-forms of $R_c^*(D_1, \cdots, D_N)$ in some special cases.

   In [45], we have found a way to extend our algorithm to continuous sources. Instead of using continuous reproduction alphabets, the extended algorithm in [45] utilizes finite reproduction alphabets and iteratively updates them along with transitional probabilities from the continuous source to reproduction letters, thus overcoming the computation complexity problem encountered when applying out algorithm proposed in Section 4.1 for discrete sources to continuous sources. The proposed algorithm converges in the sense that the rate-distortion cost is monotonically decreasing until a stationary point is reached, however, the global convergence is still unknown. Another line is to obtain the close-form of $R_c^*(D_1, \cdots, D_N)$, and in some special cases, close-forms would be solvable. One natural assumption is that N sources are jointly Gaussian distributed. In [26], Nan and Prakash obtained the close-form for Gauss-Markov sources when $N = 3$ and $X_1 \to X_2 \to X_3$. Then what about the jointly

Gaussian case without the Gauss-Markov sources assumption?

2. From an information theoretic point of view, what is the rate-distortion performance of universal CVC? In the studies of CVC in this thesis, they are all based on a key premise: that the source statistics is fully known. In some real-life situations, such a premise is not always hold.

   Universal source coding is an important research area in information theory in which the main concern is the source coding performance and realization when source statistics is unknown or insufficient. For lossless source coding, it has been proved that the entropy rate can be achieved without using the source statistics. Algorithms such as dynamic arithmetic coding, Lempel-Ziv coding and grammar-based coding are developed. For lossy source coding, it has also been proved rate-distortion function can be achieved without the source statistics explicitly. For multi-user universal source coding, the situation is more involved. It is known that Slepian-Wolf universal coding is impossible as well as in many other multi-user source coding models such as Wyner-Ziv coding. It is very likely that the universality of CVC does not exist neither. The reason is that the code statistics strongly depends on the source statistics, without knowing exact source statistics, the coder cannot construct the optimal code, therefore the problem is no longer the optimality for a particular source, rather it becomes a problem of selection of a coding strategy which is optimal in certain sense for a class of sources. To this end, three immediate questions are under our consideration: **Q1)** How to specify the class of sources with practical implication and meaningful solution? **Q2)** How to select a way to characterize the rate-distortion performance of such universal CVC? and **Q3)** Since code construction algorithm design is the key to link the theory to applications, how to design an algorithm for computing and gaining deep insights?

3. To make the idea of CVC materialize in real video codecs, future research efforts should be towards designing effective CVC algorithms. For example, in view of the optimum fixed-rate CSQ design from information theoretic point of view in Chapter 6, one could explore to further improve the RD performance of current video coding standard on how quantization should be performed conditionally given previous frames and coded frames in real video codecs.

Appendix A

In this appendix, we prove Theorem 8. As usual, we divide the proof of Theorem 8 into its converse part and its positive part.

Proof of the converse part: Pick any achievable rate distortion pair vector

$$(R_1, \cdots, R_N, D_1, \cdots, D_N) \in \mathcal{R}_c^*.$$

For any $\epsilon > 0$, there exists an order-n causal video code $C_n = \{(f_k, g_k)\}_{k=1}^N$ for all sufficiently large n such that (3.1) holds. Let S_k and $\hat{X}_k(1;n)$ be the respective encoded frame of and reconstructed frame for $X_k(1;n)$ given by C_n. It follows from the definition of causal video codes that the Markov conditions $S_k \rightarrow (X_k(1;n), X_k^-(1;n), S_k^-) \rightarrow X_k^+(1;n)$, $k = 1, \cdots, N-1$, are satisfied, and

$$\begin{aligned} & \frac{1}{n} E[d_k(X_k(1;n), \hat{X}_k(1;n))] \\ = & \ \frac{1}{n} \sum_{i=1}^{n} E[d_k(X_k(i), \hat{X}_k(i))] \\ \leq & \ D_k + \epsilon \end{aligned} \tag{A.1}$$

for $k = 1, 2, \cdots, N$.

Define auxiliary random variables

$$U_k(i) \stackrel{\Delta}{=} (X_k(i-), S_k)$$

for any $1 \leq i \leq n$ and $1 \leq k \leq N-1$, where $X_k(i-) = \{X_k(j)\}_{j=1}^{i-1}$. Since $(X_1, X_2, \cdots, X_N)$ is an IID vector source, it is not hard to verify that the Markov chain $U_k(i) \rightarrow (X_k(i), X_k^-(i), U_k^-(i)) \rightarrow X_k^+(i)$ is valid for any $1 \leq i \leq n$ and $1 \leq k \leq N-1$. In view of (3.1), and the assumption that $(X_1, X_2, \cdots, X_N)$ is an IID vector source, we have

$$\begin{aligned} n(R_1+\epsilon) &\geq H(S_1) \\ &= I(X_1(1;n); S_1) \\ &= H(X_1(1;n)) - H(X_1(1;n)|S_1) \\ &= \sum_{i=1}^{n}[H(X_1(i)) - H(X_1(i)|X_1(i-)S_1)] \\ &= \sum_{i=1}^{n} I(X_1(i); U_1(i)) \end{aligned} \tag{A.2}$$

and for $k = 2, \cdots, N-1$,

$$\begin{aligned} n(R_k+\epsilon) &\geq H(S_k|S_k^-) \\ &= I(X_k^-(1;n), X_k(1;n); S_k|S_k^-) \\ &= H(X_k^-(1;n), X_k(1;n)|S_k^-) - \\ &\quad H(X_k^-(1;n), X_k(1;n)|S_k^-, S_k) \\ &= \sum_{i=1}^{n}[H(X_k^-(i), X_k(i)|X_k^-(i-), X_k(i-), S_k^-) - \\ &\quad H(X_k^-(i), X_k(i)|X_k^-(i-), X_k(i-), S_k^-, S_k)] \\ &\overset{1)}{=} \sum_{i=1}^{n}[H(X_k^-(i), X_k(i)|X_k^-(i-), S_k^-) - \\ &\quad H(X_k^-(i), X_k(i)|X_k^-(i-), X_k(i-), S_k^-, S_k)] \\ &= \sum_{i=1}^{n} I(X_k^-(i), X_k(i); U_k(i)|U_k^-(i)) \end{aligned} \tag{A.3}$$

where the equality 1) is due to the Markov chain $(X_k^-(i), X_k(i)) \rightarrow (X_k^-(i-), S_k^-) \rightarrow X_k(i-)$.

For the last frame, we have

$$\begin{aligned}
n(R_N+\epsilon) \geq H(S_N|S_N^-)& \\
= \;& H(S_N, \hat{X}_N(1;n)|S_N^-) \\
\geq \;& I(X_N^-(1;n)X_N(1;n); \hat{X}_N(1;n)|S_N^-) \\
= \;& \sum_{i=1}^{n}[H(X_N^-(i), X_N(i)|X_N^-(i-), X_N(i-), S_N^-) - \\
& H(X_N^-(i), X_N(i)|X_N^-(i-), X_N(i-), S_N^-, \hat{X}_N(1;n))] \\
\overset{2)}{=} \;& \sum_{i=1}^{n}[H(X_N^-(i), X_N(i)|U_N^-(i)) - \\
& H(X_N^-(i), X_N(i)|U_N^-(i), X_N(i-), \hat{X}_N(1;n))] \\
\geq \;& \sum_{i=1}^{n}[H(X_N^-(i), X_N(i)|U_N^-(i)) - \\
& H(X_N^-(i), X_N(i)|U_N^-(i), \hat{X}_N(i))] \\
= \;& \sum_{i=1}^{n} I(X_N^-(i), X_N(i); \hat{X}_N(i)|U_N^-(i)) \\
\geq \;& \sum_{i=1}^{n} I(X_N(i); \hat{X}_N(i)|U_N^-(i))
\end{aligned} \tag{A.4}$$

where the equality 2) is due to the Markov chain $(X_N^-(i), X_N(i)) \to (X_N^-(i-), S_N^-) \to X_N(i-)$.

To continue, we introduce a timesharing random variable J that is uniformly distributed over $\{1, 2, \cdots, n\}$, and independent of $X_k, k = 1, 2, \cdots, N$, and hence of all random variables appearing in (A.1) to (A.4). Define $U_k \overset{\Delta}{=} (U_k(J), J)$, for $k = 1, 2, \cdots, N-1$. Then it is not hard to verify that the Markov chain $U_k \to (X_k(J), X_k^-(J), U_k^-) \to X_k^+(J)$ is valid for $k = 1, 2, \cdots, N-1$, and (A.2), (A.3), (A.4), and (A.1) can be rewritten respectively as

$$R_1 + \epsilon \geq I(X_1(J); U_1(J)|J) = I(X_1(J); U_1) \tag{A.5}$$

$$R_k + \epsilon \geq I(X_k^-(J), X_k(J); U_k|U_k^-) \tag{A.6}$$

$$R_N + \epsilon \geq I(X_N(J); \hat{X}_N(J)|U_N^-) \tag{A.7}$$

and

$$D_k + \epsilon \geq E[d_k(X_k(J), \hat{X}_k(J))]. \tag{A.8}$$

Note that $(X_1(J), \cdots, X_N(J))$ and $(X_1(1), \cdots, X_N(1))$ have the same distribution, and $\hat{X}_k(J)$, $k = 1, \cdots, N-1$, is a function of (U_k, U_k^-). Therefore, in comparison with the requirements (R1) to (R4) in the definition (3.2), the only thing missing is that the Markov chain $X_N^-(J) \to (X_N(J), U_N^-) \to \hat{X}_N(J)$ may not be valid. To overcome this problem, we can use the same technique as in the proof of the converse part of Theorem 1 and also in the proof of Lemma 5 to construct a new random vector $\tilde{X}_N(J)$ such that the following hold:

- $(X_N(J), U_N^-, \hat{X}_N(J)$ and $(X_N(J), U_N^-, \tilde{X}_N(J)$ have the same distribution, and
- the Markov condition $X_N^-(J) \to (X_N(J), U_N^-) \to \tilde{X}_N(J)$ is met.

This, together with (A.5) to (A.8) and the definition (3.2), implies that

$$(R_1 + \epsilon, \cdots, R_N + \epsilon, D_1 + \epsilon, \cdots, D_N + \epsilon) \in \mathcal{R}_{c,1}. \tag{A.9}$$

Letting $\epsilon \to 0$ yields

$$(R_1, \cdots, R_N, D_1, \cdots, D_N) \in co(\mathcal{R}_{c,1})$$

and hence $\mathcal{R}_c^* \subseteq co(\mathcal{R}_{c,1})$. This completes the proof of the converse part of Theorem 8.

The positive part of Theorem 8, $co(\mathcal{R}_{c,1}) \subseteq \mathcal{R}_c^*$, can be proved by using the standard random coding argument in multi-user information theory [11], [3]. For the sake of completeness, we present a sketch of proof below.

Proof sketch of the positive part: For convenience, we shall use bold letters to denote vectors throughout the rest of this section. For example, $\mathbf{X}_k = X_k(1;n)$. Since $\mathcal{R}_{c,1}$ is convex and $\mathcal{R}_c^*$ is closed, it suffices to show that $\mathcal{R}_{c,1} \subseteq \mathcal{R}_c^*$.

Pick any rate distortion pair vector

$$(R_1, \cdots, R_N, D_1, \cdots, D_N) \in \mathcal{R}_{c,1}.$$

We shall show that it is achievable. Let U_k, $k = 1, 2, \cdots, N-1$, and $\hat{X}_N$ be the auxiliary random variables in (3.2) (for the definition of $\mathcal{R}_{c,1}$) satisfying the requirements (R1) to (R4) with functions g_k, $k = 1, \cdots, N-1$. Denote the alphabets of U_k, $k = 1, 2, \cdots, N-1$, by $\mathcal{U}_k$, respectively. For any $\epsilon > 0$, define

$$M_k \stackrel{\Delta}{=} \lfloor 2^{n(R_k+\epsilon)} \rfloor, 1 \leq k \leq N.$$

Let $A_\epsilon^n(X_1, U_1)$ be the set of ϵ-strongly jointly typical sequences of length n with respect to the joint distribution of $(X_1(1), U_1)$. Similarly, for any $k = 2, \cdots, N-1$, let $A_\epsilon^n(X_k^-, X_k, U_k^-, U_k)$ be the set of ϵ-strongly jointly typical sequences of length n with respect to the joint distribution of $(X_k^-(1), X_k(1), U_k^-, U_k)$, and let $A_\epsilon^n(X_N, U_N^-, \hat{X}_N)$ be the set of ϵ-strongly jointly typical sequences of length n with respect to the joint distribution of $(X_N(1), U_N^-, \hat{X}_N)$. Similar notation will be used for other sets of strongly typical sequences with respect to other joint distributions. (For the definition of strong typicality, please refer to, for example, [11, Page 326].) In what follows, the values of ϵ in different strongly typical sets should be understood as $\sqrt[t]{\epsilon}$ multiplied by different constants for different t. We are now ready to describe random codebooks and how encoders/decoders work.

Generation of codebooks:

1) Generate independently M_1 codewords $\mathbf{U}_1^1, \mathbf{U}_1^2, \cdots, \mathbf{U}_1^{M_1}$ (the set of which is denoted by $\mathcal{C}_1$), where each codeword $\mathbf{U}_1^l, l \in \{1, 2, \cdots, M_1\}$ is drawn according to the n-fold product distribution of p_{U_1}.

2) For $1 < k < N$, for every combination $(\mathbf{U}_1^{i_1}, \mathbf{U}_{2|i_1}^{i_2}, \cdots, \mathbf{U}_{k-1|i_1 \cdots i_{k-2}}^{i_{k-1}})$, where $i_j \in \{1, 2, \cdots, M_j\}$ for $j = 1, 2, \cdots, k-1$, generate independently M_k codewords $\mathbf{U}_{k|i_1 i_2 \cdots i_{k-1}}^1, \cdots,$

$\mathbf{U}^{M_k}_{k|i_1 i_2 \cdots i_{k-1}}$ (the set of which is denoted by $\mathcal{C}_{k|i_1 i_2 \cdots i_{k-1}}$), where each $\mathbf{U}^{l}_{k|i_1 i_2 \cdots i_{k-1}}, l \in \{1, 2, \cdots, M_k\}$ is drawn according to the n-fold product conditional distribution of $p_{U_k|U_k^-}$ conditionally given $(\mathbf{U}_1^{i_1}, \mathbf{U}_{2|i_1}^{i_2}, \cdots, \mathbf{U}^{i_{k-1}}_{k-1|i_1 \cdots i_{k-2}})$.

3) For every combination $(\mathbf{U}_1^{i_1}, \mathbf{U}_{2|i_1}^{i_2}, \cdots, \mathbf{U}^{i_{N-1}}_{N-1|i_1 \cdots i_{N-2}})$, where $i_j \in \{1, 2, \cdots, M_j\}$ for $j = 1, 2, \cdots, N-1$, generate independently M_N codewords $\hat{\mathbf{X}}^1_{N|i_1 \cdots i_{N-1}}, \cdots, \hat{\mathbf{X}}^{M_N}_{N|i_1 \cdots i_{N-1}}$ (the set of which is denoted by $\mathcal{C}_{N|i_1 \cdots i_{N-1}}$), where each $\hat{\mathbf{X}}^l_{N|i_1 \cdots i_{N-1}}, l \in \{1, 2, \cdots, M_N\}$ is drawn according to the n-fold product conditional distribution of $p_{\hat{X}_N|U_N^-}$ conditionally given $(\mathbf{U}_1^{i_1}, \mathbf{U}_{2|i_1}^{i_2}, \cdots, \mathbf{U}^{i_{N-1}}_{N-1|i_1 \cdots i_{N-2}})$.

Encoding:

1) Given a sequence $\mathbf{X}_1$, encode $\mathbf{X}_1$ into the index, say s_1, of the first codeword in $\mathcal{C}_1$ such that $(\mathbf{X}_1, \mathbf{U}_1^{s_1}) \in A^n_\epsilon(X_1, U_1)$ if such a codeword exists. Otherwise, set $s_1 = 1$. Denote the resulting codeword $\mathbf{U}_1^{s_1}$ by $\mathbf{C}_1$.

2) For $1 < k < N$, with the knowledge of all historical codewords $\mathbf{U}_1^{s_1}, \mathbf{U}_{2|s_1}^{s_2}, \cdots, \mathbf{U}^{s_{k-1}}_{k-1|s_1 \cdots s_{k-2}}$, denoted by $\mathbf{C}_k^-$, the encoder for $\mathbf{X}_k$ finds the index, say s_k, of the first codeword in $\mathcal{C}_{k|s_1 s_2 \cdots s_{k-1}}$ such that $(\mathbf{X}_k^-, \mathbf{X}_k, \mathbf{C}_k^-, \mathbf{U}^{s_k}_{k|s_1 s_2 \cdots s_{k-1}}) \in A^n_\epsilon(X_k^-, X_k, U_k^-, U_k)$ if such a codeword exist, and set $s_k = 1$ otherwise. Denote the resulting codeword $\mathbf{U}^{s_k}_{k|s_1 s_2 \cdots s_{k-1}}$ by $\mathbf{C}_k$.

3) With the knowledge of all historical codewords $\mathbf{U}_1^{s_1}, \mathbf{U}_{2|s_1}^{s_2}, \cdots, \mathbf{U}^{s_{N-1}}_{N-1|s_1 \cdots s_{N-2}}$, denoted by $\mathbf{C}_N^-$, the encoder for $\mathbf{X}_N$ finds the index, say s_N, of the first codeword in $\mathcal{C}_{N|s_1 s_2 \cdots s_{N-1}}$ such that $(\mathbf{X}_N, \mathbf{C}_N^-, \hat{\mathbf{X}}^{s_k}_{N|s_1 s_2 \cdots s_{N-1}}) \in A^n_\epsilon(X_N, U_N^-, \hat{X}_N)$ if such a codeword exist, and set $s_N = 1$ otherwise. Denote the resulting codeword $\hat{\mathbf{X}}^{s_N}_{N|s_1 s_2 \cdots s_{N-1}}$ by $\mathbf{C}_N$.

Decoding:

1) The decoder for $\mathbf{X}_1$ first reproduces the codeword $\mathbf{C}_1$ from s_1, and then calculates $\hat{\mathbf{X}}_1$ by applying the function g_1 to each component of $\mathbf{C}_1$.

2) Upon receiving s_k, $1 < k \leq N$, the decoder for $\mathbf{X}_k$ reproduces the codeword $\mathbf{C}_k$ from $\mathcal{C}_{k|s_1 s_2 \cdots s_{k-1}}$, and then calculates $\hat{\mathbf{X}}_k$ by applying the function g_k to each component of $(\mathbf{C}_k^-, \mathbf{C}_k)$.

3) Upon receiving s_N, the decoder for $\mathbf{X}_N$ reproduces the codeword $\mathbf{C}_N$ from $\mathcal{C}_{N|s_1 s_2 \cdots s_{N-1}}$, and then outputs $\hat{\mathbf{X}}_N^{s_N} = \mathbf{C}_N$.

Analysis of bit rates, typicality, and distortions:

1) From the construction of encoders, the bit rate in bits per symbol for each X_k is upper bounded by $R_k + \epsilon$.

2) In view of the law of large numbers, standard probability bounds associated with typicality (see, for example, [11, Lemma 10.6.2, Chapter 10]), and the Markov lemma [11, Lemma 15.8.1, Chapter 15], [3], it follows that with probability approaching 1 as $n \to \infty$, $\mathbf{X}_1, \cdots, \mathbf{X}_N, \mathbf{C}_1, \cdots, \mathbf{C}_{N-1}$ are strongly typical, and $\mathbf{X}_N$ and $\mathbf{C}_N$ are strongly typical.

3) In view of Requirements (R1) to (R3) in the definition (3.2) and of the above two paragraphs, it follows that the distortion per symbol between each $\mathbf{X}_k$ and $\hat{\mathbf{X}}_k$, $k = 1, \cdots, N$, is upper bounded by $D_k + O(\epsilon)$ with probability approaching 1 as $n \to \infty$.

Existence of a deterministic causal video code with desired performance:

In the above analysis, all probabilities are with respect to both the random sources $X_1, \cdots, X_N$, and the random codebooks. By the well-known Markov inequality, it follows that there exists a deterministic causal video code (i.e., a deterministic codebook) for which the distortion per symbol between each $\mathbf{X}_k$ and $\hat{\mathbf{X}}_k$, $k = 1, \cdots, N$, is upper bounded by $D_k + O(\epsilon)$ with probability approaching 1 as $n \to \infty$[1]. Therefore, for this

[1] This step is necessary since we have multiple distortion inequalities to satisfy, in which case declaring the existence of a deterministic code immediately from several inequalities with average performance over the codebook ensemble would fail.

deterministic causal video code, the average distortion per symbol between each $\mathbf{X}_k$ and $\hat{\mathbf{X}}_k$, $k = 1, \cdots, N$, is upper bounded by $D_k + O(\epsilon)$. Note that all rates are fixed.

Putting all pieces together, we have shown that

$$(R_1 + \epsilon, \cdots, R_N + \epsilon, D_1 + O(\epsilon), \cdots, D_N + O(\epsilon)) \in \mathcal{R}_c^*.$$

Letting $\epsilon \to 0$ yields

$$(R_1, \cdots, R_N, D_1, \cdots, D_N) \in \mathcal{R}_c^*.$$

This completes the proof of the positive part of Theorem 8.

Appendix B

In this Appendix, we prove the positive part (i.e., $co(\mathcal{R}'_c) \subseteq \mathcal{R}^*_c$) of Theorem 1. Since $\mathcal{R}'_c = \bigcup_{m=1}^{\infty} \mathcal{R}_{c,m}$, each $\mathcal{R}_{c,m}$ is convex, and $\mathcal{R}^*_c$ is closed, it suffices to show that for each $m \geq 1$, $\mathcal{R}_{c,m} \subseteq \mathcal{R}^*_c$.

Proof of $\mathcal{R}_{c,1} \subseteq \mathcal{R}^*_c$: Unless otherwise specified, notation below is the same as in the proof of the positive part in Appendix A. Indeed, our proof is similar to the random coding argument made for the IID case in Appendix A. However, since the vector source $(X_1, \cdots, X_N)$ now is not IID, but stationary and totally ergodic, the Markov lemma in its simple form as expressed in [11, Lemma 15.8.1, Chapter 15] is not valid any more. To overcome this difficulty, we will modify the concept of typical sequences and make it even stronger. With $A^n_\epsilon(X_1, U_1)$ and $A^n_\epsilon(X^-_k, X_k, U^-_k, U_k)$, $k = 2, \cdots, N-1$, defined as in Appendix A, we define for each sequence $\mathbf{x}_1 \in \mathcal{X}^n_1$ and $\mathbf{u}_1 \in \mathcal{U}^n_1$, where for any alphabet $\mathcal{U}$, $\mathcal{U}^n$ denotes the set of all sequences of length n from $\mathcal{U}$,

$$\begin{aligned} A^n_\epsilon(X^+_1 | \mathbf{x}_1, \mathbf{u}_1) \stackrel{\Delta}{=} & \{\mathbf{x}^+_1 = (\mathbf{x}_2, \cdots, \mathbf{x}_N) : \\ & (\mathbf{x}_1, \mathbf{x}^+_1, \mathbf{u}_1) \in A^n_\epsilon(X_1, X^+_1, U_1)\} \end{aligned} \tag{B.1}$$

and similarly, for each $\mathbf{x}_k^-, \mathbf{x}_k, \mathbf{u}_k^-, \mathbf{u}_k$,

$$\begin{aligned} A_\epsilon^n(X_k^+|\mathbf{x}_k^-, \mathbf{x}_k, \mathbf{u}_k^-, \mathbf{u}_k) \overset{\Delta}{=} &\{\mathbf{x}_k^+ : \\ &(\mathbf{x}_k^-, \mathbf{x}_k, \mathbf{x}_k^+, \mathbf{u}_k^-, \mathbf{u}_k) \in A_\epsilon^n(X_k^-, X_k, X_k^+, U_k^-, U_k)\}. \end{aligned} \tag{B.2}$$

We then define our modified joint typical sets as follows

$$\begin{aligned} \hat{A}_\epsilon^n(X_1, U_1) \overset{\Delta}{=} &\{(\mathbf{x}_1, \mathbf{u}_1) : (\mathbf{x}_1, \mathbf{u}_1) \in A_\epsilon^n(X_1, U_1) \\ &\& \Pr\left\{\mathbf{X}_1^+ \in A_\epsilon^n(X_1^+|\mathbf{x}_1, \mathbf{u}_1)\middle|\mathbf{X}_1 = \mathbf{x}_1\right\} > 1 - \epsilon\right\} \end{aligned} \tag{B.3}$$

and for $1 < k < N$,

$$\begin{aligned} \hat{A}_\epsilon^n(X_k^-, X_k, U_k^-, U_k) \overset{\Delta}{=} &\big\{(\mathbf{x}_k^-, \mathbf{x}_k, \mathbf{u}_k^-, \mathbf{u}_k) : \\ &(\mathbf{x}_k^-, \mathbf{x}_k, \mathbf{u}_k^-, \mathbf{u}_k) \in A_\epsilon^n(X_k^-, X_k, U_k^-, U_k) \\ &\& \Pr\Big\{\mathbf{X}_k^+ \in A_\epsilon^n(X_k^+|\mathbf{x}_k^-, \mathbf{x}_k, \mathbf{u}_k^-, \mathbf{u}_k)\Big|\mathbf{X}_k = \mathbf{x}_k, \\ &\mathbf{X}_k^- = \mathbf{x}_k^-\Big\} > 1 - \epsilon\big\}. \end{aligned} \tag{B.4}$$

To get our random CVC scheme in this case, we simply modify the encoding procedure of the random coding scheme constructed in Appendix A by replacing $A_\epsilon^n(X_1, U_1)$ and $A_\epsilon^n(X_k^-, X_k, U_k^-, U_k)$ with $\hat{A}_\epsilon^n(X_1, U_1)$ and $\hat{A}_\epsilon^n(X_k^-, X_k, U_k^-, U_k)$, respectively; the rest of the random coding scheme remains the same. Since the rate of the encoder for each X_k is fixed, the bit rate in bits per symbol for each X_k is upper bounded by $R_k + \epsilon$. To get the desired upper bounds on distortions, we need to analyze the joint typicality of the source sequences and the respective transmitted codeword sequences. At this point, we invoke the following result, which will be proved at the end of this Appendix.

Lemma 7 (Extended Markov Lemma) *Suppose that* $X_1, X_2, \cdots, X_N$ *are jointly stationary and ergodic. Let* U_k*,* $k = 1, 2, \cdots, N-1$*, and* $\hat{X}_N$ *be the auxiliary random variables in* (3.2) *(for the definition of* $\mathcal{R}_{c,1}$*) satisfying the requirements (R1) to (R4). Let* $\{U_1(i)\}_{i=1}^{\infty}$ *be the output process of the memoryless channel given by* $p_{U_1|X_1(1)}$ *in response to the input* X_1*. For any* $1 < k < N$*, let* $\{U_k(i)\}_{i=1}^{\infty}$ *be the output process of the memoryless channel given by* $p_{U_k|X_k^-(1)X_k(1)U_k^-}$ *in response to the inputs* (X_k^-, X_k) *and* $\{\{U_j(i)\}_{i=1}^{\infty} : j = 1, \cdots, k-1\}$*. Let* $\{\hat{X}_N(i)\}_{i=1}^{\infty}$ *be the output process of the memoryless channel given by* $p_{\hat{X}_N|X_N(1)U_N^-}$ *in response to the inputs* X_N *and* $\{\{U_j(i)\}_{i=1}^{\infty} : j = 1, \cdots, N-1\}$*. Then the following properties hold:*

(P1) *The probability* $\Pr\{(\boldsymbol{X}_1, \boldsymbol{U}_1) \in \hat{A}_\epsilon^n(X_1, U_1)\}$*, where* $\boldsymbol{X}_1 = X_1(1;n)$ *and* $\boldsymbol{U}_1 = U_1(1;n)$*, goes to* 1 *as* $n \to \infty$*.*

(P2) *For any* $1 < k < N$ *and sufficiently large* n*,*

$$\begin{aligned}\Pr\big\{&(\boldsymbol{x}_k^-, \boldsymbol{X}_k, \boldsymbol{u}_k^-, \boldsymbol{U}_k) \in \\ &\hat{A}_{\sqrt{2\epsilon}}^n(X_k^-, X_k, U_k^-, U_k)\Big| \boldsymbol{X}_k^- = \boldsymbol{x}_k^-, \\ &\boldsymbol{U}_k^- = \boldsymbol{u}_k^-\big\} \geq 1 - 2\epsilon - \sqrt{2\epsilon}\end{aligned} \tag{B.5}$$

for any $(\boldsymbol{x}_k^-, \boldsymbol{u}_k^-) \in \hat{A}_\epsilon^n(X_k^-, U_k^-)$*.*

(P3) *For sufficiently large* n*,*

$$\begin{aligned}\Pr\Big\{&(\boldsymbol{x}_N^-, \boldsymbol{X}_N, \boldsymbol{u}_N^-, \hat{\boldsymbol{X}}_N) \in \\ &A_{2\epsilon}^n(X_N^-, X_N, U_N^-, \hat{X}_N)\Big| \boldsymbol{X}_N^- = \boldsymbol{x}_N^-, \\ &\boldsymbol{U}_N^- = \boldsymbol{u}_N^-\big\} \geq 1 - 2\epsilon\end{aligned} \tag{B.6}$$

for any $(\boldsymbol{x}_N^-, \boldsymbol{u}_N^-) \in \hat{A}_\epsilon^n(X_N^-, U_N^-)$*.*

Lemma 7 can be regarded as an extended Markov lemma in the ergodic case. In view of Lemma 7, it is not hard to see that with high probability, which approaches 1 as $\epsilon \to 0$, $\mathbf{X}_1, \cdots, \mathbf{X}_N, \mathbf{C}_1, \cdots, \mathbf{C}_{N-1}$ are strongly typical, and $\mathbf{X}_N$ and $\mathbf{C}_N$ are strongly typical. The rest of the proof is identical to the case considered in Appendix A. This completes the proof of $\mathcal{R}_{c,1} \subseteq \mathcal{R}_c^*$.

Proof of $\mathcal{R}_{c,m} \subseteq \mathcal{R}_c^*$: We consider a block of m symbols as a super symbol and regard $(X_1, \cdots, X_N)$ as a vector source over $\mathcal{X}_1^m \times \cdots \times \mathcal{X}_N^m$. Since $(X_1, \cdots, X_N)$ is totally ergodic, it is also ergodic when regarded as a vector source over $\mathcal{X}_1^m \times \cdots \times \mathcal{X}_N^m$. Repeating the above argument for super symbols, i.e., for alphabets $\mathcal{X}_1^m, \cdots \mathcal{X}_N^m, \hat{\mathcal{X}}_1^m, \cdots, \hat{\mathcal{X}}_N^m$, we then have $\mathcal{R}_{c,m} \subseteq \mathcal{R}_c^*$ for any $m \geq 1$. This completes the proof of the positive part of Theorem 1.

We now prove Lemma 7.

Proof of Lemma 7: By construction, it is easy to see that $\{\{U_j(i)\}_{i=1}^{\infty} : j = 1, \cdots, N-1\}$ and $\{\hat{X}_N(i)\}_{i=1}^{\infty}$ are the output of a memoryless channel in response to the input $(X_1, \cdots, X_N)$. Since $X_1, X_2, \cdots, X_N$ are joint stationary and ergodic, it follows from [2, Theorem 7.2.1, Page 272] that the $2N$ processes $\{\{(X_j(i), U_j(i))\}_{i=1}^{\infty} : j = 1, \cdots, N-1\}$ and $\{(X_N(i), \hat{X}_N(i))\}_{i=1}^{\infty}$ are jointly stationary and ergodic as well. By the ergodic theorem, we then have

$$\lim_{n \to \infty} \Pr\{(\mathbf{X}_1, \mathbf{X}_1^+, \mathbf{U}_1) \in A_\epsilon^n(X_1, X_1^+, U_1)\} = 1. \tag{B.7}$$

Let

$$a_n \overset{\Delta}{=} \Pr\{(\mathbf{X}_1, \mathbf{X}_1^+, \mathbf{U}_1) \notin A_\epsilon^n(X_1, X_1^+, U_1)\}.$$

Rewrite a_n as

$$a_n = E\left[\Pr\{\mathbf{X}_1^+ \notin A_\epsilon^n(X_1^+ | \mathbf{X}_1, \mathbf{U}_1) | \mathbf{X}_1\}\right]. \tag{B.8}$$

Applying the Markov inequality to (B.8), we get

$$\begin{aligned}\Pr\Big\{(\mathbf{X}_1,\mathbf{U}_1)\in\Big\{(\mathbf{x}_1,\mathbf{u}_1):\Pr\{\mathbf{X}_1^+\notin\\ A_\epsilon^n(X_1^+|\mathbf{x}_1,\mathbf{u}_1)\Big|\mathbf{x}_1\}<\sqrt{a_n}\Big\}\Big\}\geq 1-\sqrt{a_n}.\end{aligned}\tag{B.9}$$

Since $a_n\to 0$ as $n\to\infty$, combining (B.9) with (B.7) yields Property P1 in Lemma 7.

To prove Property P2 in Lemma 7, note that given any $(\mathbf{x}_k^-,\mathbf{x}_k,\mathbf{x}_k^+,\mathbf{u}_k^-)$, $\{U_k(i)\}_{i=1}^n$ is a conditionally independent sequence. It is not hard to see that

$$\begin{aligned}\lim_{n\to\infty}\Pr\big\{(\mathbf{x}_k^-,\mathbf{x}_k,\mathbf{x}_k^+,\mathbf{u}_k^-,\mathbf{U}_k)\in\\ A_{2\epsilon}^n(X_k^-,X_k,X_k^+,U_k^-,U_k)\Big|\mathbf{x}_k^-,\mathbf{x}_k,\mathbf{x}_k^+,\mathbf{u}_k^-\Big\}=1\end{aligned}\tag{B.10}$$

as long as $(\mathbf{x}_k^-,\mathbf{x}_k,\mathbf{x}_k^+,\mathbf{u}_k^-)\in A_\epsilon^n(X_k^-,X_k,X_k^+,U_k^-)$. Furthermore, the convergence in (B.10) is uniform. This, coupled with the definition of $\hat{A}_\epsilon^n(X_k^-,U_k^-)$, implies that for sufficiently large n and for any $(\mathbf{x}_k^-,\mathbf{u}_k^-)\in\hat{A}_\epsilon^n(X_k^-,U_k^-)$,

$$\begin{aligned}&\Pr\big\{(\mathbf{x}_k^-,\mathbf{X}_k,\mathbf{X}_k^+,\mathbf{u}_k^-,\mathbf{U}_k)\in\\ &\quad A_{2\epsilon}^n(X_k^-,X_k,X_k^+,U_k^-,U_k)\Big|\mathbf{x}_k^-,\mathbf{u}_k^-\Big\}\\ &\quad >1-2\epsilon.\end{aligned}\tag{B.11}$$

Applying the Markov inequality to (B.11), we get

$$\begin{aligned}&\Pr\Big\{(\mathbf{X}_k,\mathbf{U}_k)\in\Big\{(\mathbf{x}_k,\mathbf{u}_k):\Pr\{\mathbf{X}_k^+\notin\\ &\quad A_{2\epsilon}^n(X_k^+|\mathbf{x}_k^-,\mathbf{x}_k,\mathbf{u}_k^-,\mathbf{u}_k)|\mathbf{x}_k^-,\mathbf{x}_k\}<\sqrt{2\epsilon}\Big\}\Big|\mathbf{x}_k^-,\mathbf{u}_k^-\Big\}\\ &\quad >1-\sqrt{2\epsilon}\end{aligned}\tag{B.12}$$

which in turn implies

$$\begin{aligned}&\Pr\Big\{(\mathbf{X}_k,\mathbf{U}_k)\in\Big\{(\mathbf{x}_k,\mathbf{u}_k):\Pr\{\mathbf{X}_k^+\in\\&\quad A^n_{\sqrt{2\epsilon}}(X_k^+|\mathbf{x}_k^-,\mathbf{x}_k,\mathbf{u}_k^-,\mathbf{u}_k)|\mathbf{x}_k^-,\mathbf{x}_k\}>1-\sqrt{2\epsilon}\Big\}\,\Big|\,\mathbf{x}_k^-,\\&\quad\mathbf{u}_k^-\Big\}>1-\sqrt{2\epsilon}\end{aligned}\tag{B.13}$$

whenever $2\epsilon<1$. Combining (B.13) with (B.11) yields (B.5).

A similar argument can be used to prove Property (P3). The completes the proof of Lemma 7.

Bibliography

[1] S. Arimoto. An algorithm for calculating the capacity of an arbitrary discrete memoryless channel. *IEEE Trans. Inform. Theory*, pages 14–20, 1972.

[2] T. Berger. *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[3] T. Berger. Multiterminal source coding. In *The Inform. Theory Approach to Communications*, pages 171–231. New York: Springer-Verlag, Wien, 1977.

[4] T. Berger and S. Y. Tung. Encoding of correlated analog sources. pages 7–10, Piscataway, NJ, 1975. Proc. 1975 IEEE-USSR Joint Work Inform. Theory.

[5] T. Berger and H. Viswanathan. The quadratic Gaussian CEO problem. *IEEE Trans. Inform. Theory*, 43:1549–1561, 1997.

[6] T. Berger, Z. Zhang, and H. Viswanathan. The CEO problem. *IEEE Trans. Inform. Theory*, 42:887–903, 1996.

[7] R. E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inform. Theory*, pages 460–473, 1972.

[8] M. U. Chang. *Rate-distortion with a fully informed decoder and a partially informed encoder.* PhD thesis, Cornell Univ., Ithaca, NY, 1978.

[9] C. L. Chen, W. W. Peterson, and Jr. E. J. Weldon. Some results on quasi-cyclic codes. *Inform. Contr.*, 15:407 – 423, 1969.

[10] T. M. Cover. An algorithm for maximizing expected log investment return. *IEEE Trans. Inform. Theory*, pages 369–373, 1984.

[11] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, Hoboken, NJ, second edition, 2006.

[12] I. Csiszar. On the computation of rate distortion functions. *IEEE Trans. Inform. Theory*, 20:122–124, 1974.

[13] I. Csiszar and G. Tusnady. Information geometry and alternating minimization procedures. *Statistics and Decisions*, pages 205–237, 1984.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Stat. Soc.*, pages 1–38, 1977.

[15] W. Effelsberg and R. Steinmetz. *Video Compression Techniques.* Dpunkt.Verlag, 1998.

[16] W. H. R. Equitz and T. Cover. Successive refinement of information. *IEEE Trans. Inform. Theory*, 37:269–275, 1991.

[17] R. G. Gallager. *Information Theory and Reliable Communication.* Wiley, New York, NY, 1968.

[18] A. E. Gamal and T. M. Cover. Achievable rates for multiple descriptions. *IEEE Trans. Inform. Theory*, pages 851 – 857, 1982.

[19] A. Gersho and A. D. Wyner. The multiple descriptions problem. Presented by A. D. Wyner at the IEEE Inform. Theory Work., 1979. Seven Springs Conf. Ctr., Mt. Kisco, NY.

[20] R. M. Gray. A new class of lower bounds to information rates of staionary sources via conditional rate-distortion function. *IEEE Trans. Inform. Theory*, pages 480–489, 1973.

[21] J. C. Kieffer. Extension of source coding theorems for block codes to sliding-block codes. *IEEE Trans. Inform. Theory*, (6), 1980.

[22] J. C. Kieffer. A metod for proving multiterminal source coding theorems. *IEEE Trans. Inform. Theory*, (5), 1981.

[23] J. C. Kieffer and E.-h. Yang. Grammar-based codes: a new class of universal lossless source codes. *IEEE Trans. Inform. Theory*, 46:737–754, 2000.

[24] S. P. Lloyd. Least squares quantization in PCM. Bell Laboratories Technical Note, 1957.

[25] N. Ma and P. Ishwar. The value of frame-delays in the sequential coding of correlated sources. pages 1496–1500, Nice, France, 2007. Proc. of the 2007 IEEE Intern. Symp. Inform. Theory.

[26] N. Ma and P. Ishwar. On delayed sequential coding of correlated sources. *IEEE Trans. Inform. Theory*, 57:3763–3782 , 2011.

[27] N. Ma, Y. Wang, and P. Ishwar. Delayed sequential coding of correlated sources. pages 214 – 222, San Diego, California, 2007. Proc. of the 2007 Information Theory and Applications Workshop.

[28] J. Max. Quantizing for minimum distortion. *IRE Trans. Inform. Theory*, pages 7–12, 1960.

[29] D. L. Neuhoff and R. K. Gilbert. Causal source codes. *IEEE Trans. Inform. Theory*, 25(5):701–713, 1982.

[30] J. K. Omura and K. B. Housewright. Source coding studies for information networks. Chicago, Ill., 1977. Proc. 1977 Int. Conf. Communications.

[31] Y. Oohama. The rate distortion function for the quadratic Gaussian CEO problem. *IEEE Trans. Inform. Theory*, 44:1057–1070, 1998.

[32] I. E. G. Richardson. *H.264 and MPEG-4 video compression.* NY: Willey, New York, 2003.

[33] M. J. Riely and Richardson I. e.g. *Digital Video Communications.* Artech House, Boston, 1997.

[34] R. T. Rockafellar. *Convex Analysis.* New Jersey: Princeton University Press, 1970.

[35] A. Sgarro. Source coding with side information at several decoders. *IEEE Trans. Inform. Theory*, pages 179–182, 1979.

[36] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 1948.

[37] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE National Convention Record, Part 4*, pages 142–163. 1959.

[38] P. C. Shields and D. L. Neuhoff. Block and sliding-block source coding. *IEEE Trans. Inform. Theory*, pages 211–215, 1977.

[39] A. Shohara. *Source coding theorems for information networks.* PhD thesis, Univ. Calif. Los Angeles, 1974. Tech. Rep. UCLA-ENG-7445.

[40] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inform. Theory*, pages 471–480, 1973.

[41] S. Y. Tung. *Multiterminal rate-distortion theory.* PhD thesis, Cornell Univ., Ithaca, NY, 1977.

[42] H. Viswanathan and T. Berger. Sequential coding of correlated sources. *IEEE Trans. Inform. Theory*, 46:236–246, 2000.

[43] A. D. Wyner. The rate-distortion function for source coding with side information at the decoder-II: general sources. *Information and Control*, 38:60–80, 1978.

[44] A. D. Wyner and J. Ziv. The rate-distortion function for source coding with side-information at the receiver. *IEEE Trans. Inform. Theory*, pages 1–11, 1976.

[45] E.-h. Yang, C. Sun, and L. Zheng. An improved iterative algorithm for calculating the rate distortion performance of causal video coding for continuous sources and its application to real video data. Nagoya, Japan, 2010. Proc. of the 2010 Picture Coding Symposium.

[46] E.-h. Yang and L. Wang. Full rate distortion optimization of MPEG-2 video coding. pages 605–608, Cairo, Egypt, 2009. Proc. of the 2009 IEEE Intern. Conf. Image Process.

[47] E.-h. Yang and L. Wang. Joint optimization of run-length coding, Huffman coding and quantization table with complete baseline JPEG decoder compatibility. *IEEE Trans. Image Process.*, 18:63–74, 2009.

[48] E.-h. Yang and L. Wang. Method, system, and computer program product for optimization of data compression with cost function, Aug. 2009. U.S. Patent No. 7,570,827.

[49] E.-h. Yang and X. Yu. Rate distortion optimization for H.264 inter-frame video coding: A general framework and algorithms. *IEEE Trans. on Image Processing*, 16(7):1774–1784, 2007.

[50] E.-h. Yang and X. Yu. Soft decision quantization for H.264 with main profile compatibility. *IEEE Trans. Circuits Syst. Video Technol.*, 19:122–127, 2009.

[51] E.-h. Yang and Z. Zhang. Variable-rate trellis source encoding. *IEEE Trans. Inform. Theory*, 45(2), 1999.

[52] E.-h. Yang and Zhen Zhang. On the redundancy of lossy source coding with abstract alphabets. *IEEE Trans. Inform. Theory*, 44:1092–1110, 1999.

[53] E.-h. Yang, L. Zheng, D.-k. He, and Z. Zhang. On the rate distortion theory for causal video coding. pages 385 – 391, San Diego, California, U.S.A., 2009. Proc. of the 2009 Information Theory and Applications Workshop.

[54] E.-h. Yang, L. Zheng, D.-k. He, and Z. Zhang. Rate distortion theory for causal video coding: characterization, computation algorithm, and comparison. *IEEE Trans. Inform. Theory*, 57:5258 – 5280, Aug. 2011.

[55] E.-h. Yang, L. Zheng, Z. Zhang, and D.-k. He. A computation approach to the minimum total rate problem of causal video coding. pages 2141–2145, Seoul, Korea, 2009. Proc. of the 2009 IEEE Intern. Symp. Inform. Theory.

[56] R. W. Yeung and T. Berger. Multi-way alternating minimization. Whistler, Canada, 1995. Proc. of 1995 IEEE Intern. Symp. on Inform. Theory.

[57] J. Ziv and A. Lempel. Compression of individual sequences via variable rate coding. *IEEE Trans. Inform. Theory*, 24:530–536, 1978.