UNIVERSITY OF CYPRUS

Nonanticipative Information Theory

by

Christos K. Kourtellaris

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering
Department of Electrical and Computer Engineering

March 2014

Christos K. Kourtellaris

Christos K. Kourtellaris

©Christos K. Kourtellaris, 2014

Declaration of Authorship

The present doctoral dissertation was submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy of the University of Cyprus. It is a product of original work of my own, unless otherwise mentioned through references, notes, or any other statements.

Name: ______________________________

Signed: ______________________________

Christos K. Kourtellaris

ΠΕΡΙΛΗΨΗ

Η κλασσική θεωρία πληροφορίας χρησιμοποιεί την αμοιβαία πληροφορία για να ορίσει τη χωρητικότητα των καναλιών και τη συμπίεση των πηγών πληροφορίας. Για κανάλια και πηγές χωρίς μνήμη και ανατροφοδότηση, το μέτρο αυτό μπορεί να χρησιμοποιηθεί με επιτυχία για να υπολογιστεί η λειτουργική χωρητικότητα των καναλιών και η συμπίεση των πηγών. Για κανάλια με μνήμη ανατροφοδότηση και για συμπίεση πηγών σε πραγματικό χρόνο, χωρίς καθυστέρηση η αμοιβαία πληροφορία δεν αποτελεί κατάλληλο μέτρο πληροφορίας.

Η κατευθυνόμενη πληροφορία μπορεί να χρησιμοποιηθεί για να υπολογιστή τόσο η χωρητικότητα καναλιών με μνήμη και ανατροφοδότηση, για χωρητικότητα δικτύων, ακόμα και για ανάλυση βιολογικών συστημάτων.

Η συγκεκριμένη διατριβή ερευνά, μέσω της κατευθυνόμενης πληροφορίας, την χωρητικότητα καναλιών με μνήμη και ανατροφοδότηση την συμπίεση πηγών σε πραγματικό χρόνο, την κοινή κωδικοποίηση πηγής και καναλιού πραγματικού χρόνου, όσο και την χρησιμοποίηση της πληροφορίας για στοχαστικό έλεγχο συστημάτων. Η παρουσίαση γίνεται σε ένα ενοποιημένο πλαίσιο κατάλληλο για την ανάλυση τέτοιων προβλημάτων χρησιμοποιώντας αρχές και έννοιες της στοχαστικής θεωρίας ελέγχου, του δυναμικού προγραμματισμού και λογισμού των μεταβολών.

Christos K. Kourtellaris

Abstract

Traditional information theoretic measures for capacity and lossy compression are defined via mutual information. For memoryless communication channels and sources these measures have been successfully applied to compute the operation capacity of channels and lossy compression of sources, respectively. For channels with memory and nonanticipative (causal) feedback and nonanticipative lossy compression of sources with memory the valid information measure is the directed information defined via nonanticipative conditional distributions. Directed information is also extensively utilized in networks, communication for real-time stochastic control applications, and in biological system analysis.

This thesis investigates via directed information, capacity of channels with memory and feedback, lossy nonanticipative data compression, Joint Source Channel Coding based on nonanticipative transmission, and communication for real-time stochastic control.

The thesis presents a unifying framework to analyze such extremum problems. It utilizes concepts from stochastic control theory, dynamic programming, and calculus of variations to address extremum problems of capacity of channels with memory and feedback, extremum problems of nonanticipative rate distortion function of sources with memory, extremum problems of Joint Source Channel Coding based on nonanticipative transmission.

Christos K. Kourtellaris

Acknowledgements

Christos K. Kourtellaris

It gives me great pleasure in expressing my gratitude to all those people who have supported me and had their contributions in making this thesis possible.

I express my profound sense of reverence to my supervisor and promoter Prof. Charalambos D. Charalambous. He patiently provided the vision, encouragement and advise necessary for me to proceed through the doctoral program and complete my dissertation. His in-depth knowledge on a broad spectrum of information theory and control theory has been extremely beneficial for me. He has given me enough freedom during my research, and he has always been supportive to my decisions.

This research was mainly supported by Control for Coordination of Distributed Systems (C4C) project which was financially supported by the European Commission project (Program EU.ICT, Objective ICT-2007.3.7, Networked Embedded and Control Systems, Project 223844). In particular, I would like to thank the project coordinator, Professor Jan H. van Schuppen, for all the fruitful discussions, the encouraging and the constructive feedback.

Special thanks to my committee, Professor Christoforos Hadjicostis, Dr. Ioannis Krikidis and Dr. Themistoklis Charalambous, for their support, guidance and helpful suggestions. Their guidance has served me well and I owe them my heartfelt appreciation.

I thank Photis, Ioannis and Ioanna, for being very good friends and brilliant collaborators. Their critical remarks and suggestions have always been very helpful in improving my skills and for strengthening our manuscripts. Moreover, I would like to thank all my friends and especially Chloe for their endless support during the last six years.

Above and beyond all, my heartfelt gratitude to my parents and my brother, for their much needed support, patience, understanding, and encouragement in every possible way.

Contents

Christos K. Kourtellaris

Christos K. Kourtellaris

Christos K. Kourtellaris

List of Figures

Christos K. Kourtellaris

List of Tables

Christos K. Kourtellaris

Abbreviations

RDF	**R**ate **D**istortion **F**unction
OPTA	**O**ptimal **P**erformance **T**heoretically **A**ttainable
SbS	**S**ymbol **b**y **S**ymbol
DMC	**D**iscrete **M**emoryless **C**hannel
PMS	**P**osterior **M**atching **S**cheme
IID	**I**ndependent **I**dentical **D**istributed
JSCC	**J**oint **S**ource **C**hannel **C**oding
JSCM	**J**oint **S**ource **C**hannel **M**atching
AWGN	**A**dditive **W**hite **G**aussian **N**oise
SRDF	**S**equential **R**ate **D**istortion **F**unction
LbL	**L**etter **by** **L**etter
MC	**M**arkov **C**hain
BSMS	**B**inary **S**ymmetric **M**arkov **S**ource
RL	**R**ate **L**oss
RV	**R**andom **V**ariable
BSSC	**B**inary **S**tate **S**ymmetric **C**hannel
POST	**P**revious **O**utput **ST**ate
BSC	**B**inary **S**ymmetric **C**hannel

Christos K. Kourtellaris

Symbols

$\mathbb{N}^n$	$\triangleq$	$\{0,1,2,\ldots,n\}$
$\mathbb{N}$	$\triangleq$	$\{0,1,2,\ldots\}$
X^n	$\triangleq$	$\{X_0,X_1,\ldots,X_n\}$
X_i^n	$\triangleq$	$\{X_i,X_{i+1},\ldots,X_n\}$
$\mathbb{E}\{.\}$	:	Expected value
$\otimes$	:	Convolution
$(\Omega,\mathbb{F},\mathbb{P})$	:	Probability space
Ω	:	Sample space of the probability space
$\mathbb{F}$	:	σ-algebra of events
$\mathbb{P}$	:	Probability measure
$\mathscr{M}_1(\mathscr{X})$	:	The space of probability measure on $(\mathscr{X},\mathscr{B}(\mathscr{X}))$
$\sigma\{X\}$	:	σ-algebra generated by RV X
$\mathscr{X}$	:	Source alphabet
$\mathscr{Y}$	:	Reproduction alphabet
$\mathscr{A}$	:	Channel input alphabet
$\mathscr{B}$	:	Channel output alphabet
$D(\cdot \Vert \cdot)$	:	Kullback-Leibler divergence
$I(\cdot;\cdot)$	:	Mutual Information
$\mathbb{I}_{\cdot;\cdot}(\cdot;\cdot)$	:	Mutual Information Functional
$I(\cdot \rightarrow \cdot)$	:	Forward Directed Information
$I_{\cdot\rightarrow\cdot}(\cdot \rightarrow \cdot)$	:	Forward Directed Information Functional
$R(D)$	:	Rate Distortion Function
$Q(D)$	:	Fidelity set of reproduction conditional distributions

Christos K Kourtellaris

$R^{na}(D)$	:	Nonanticipative Rate Distortion Function
$Q^{na}(D)$	:	Nonanticipative Fidelity set of reproduction conditional distributions
$R^{na,ff}(D)$	:	Nonanticipative Rate Distortion Function with feedforward side information
C	:	Channel Capacity
$C(P)$	:	Capacity with Transmission Cost Constraint
$\mathscr{P}_{0,n}(P)$	:	Transmission cost constraint

Christos K. Kourtellaris

To my parents

Christos K. Kourtellaris

Chapter 1

Introduction

Christos K. Kourtellaris

Communication systems have rapidly changed since Claude Shannon's seminal paper [65] gave birth to the field of Information Theory. The point to point communication diagram is illustrated in Figure. 1.0.1. It consists of an information source described via a probability distribution, a noisy channel described via a conditional probability distribution, a transmitter and a receiver. While the initial task of point to point reliable transmission, has evolved with time to include sources with memory, channels with memory and feedback, control communication schemes and networks, the communication problem remains the same. Send the minimum amount of data from the transmitter to the receiver in order to reconstruct the initial message, with or without distortion, with arbitrarily small probability of error.

Traditional information theory considers transmission schemes where the transmitter requires the whole symbol sequence to construct infinitely large blockcodes, while the receiver requires the blockcodes at the channel output to reconstruct the channel source sequence. This is a significant drawback when dealing with channels with feedback, real time communication schemes and communication for control applications, where delays must be taken into consideration.

The main goal of this thesis is to provide a comprehensive analysis of *nonanticipative* reliable communication. The key mathematical tool when dealing with such problems is nonanticipative conditional distributions which are used to define nonanticipative or causal, information measures. In this thesis, we describe how this approach affects the traditional aspects of the lossy compression of general sources with memory, the capacity for general channels with memory and feedback, and joint source-channel coding. These concepts are embedded

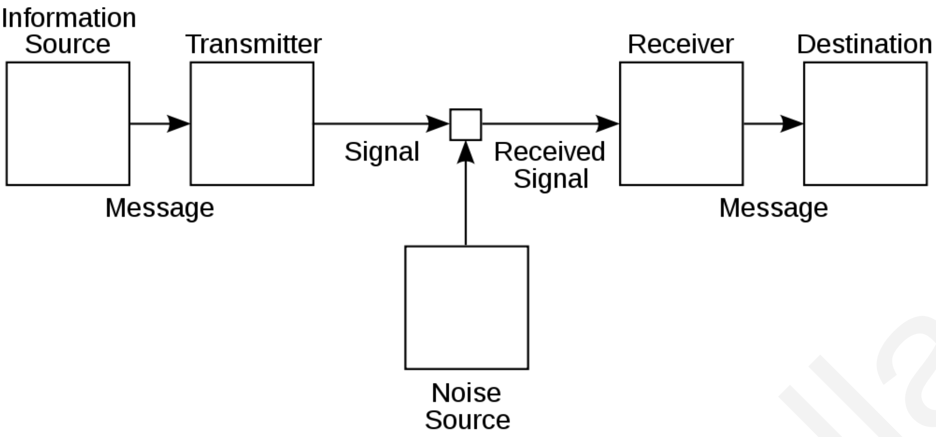

FIGURE 1.0.1: Shannon's communication diagram [65]

into specific examples which are finally merged together to show that nonanticipative transmission is indeed optimal and nothing can be gained in terms of performance by encoding messages into long codewords. Moreover, we expand these concepts for lossy compression of sources with feedforward information at the decoder.

1.1 Outline

1.1.1 Chapter 2: Nonanticipative Rate Distortion Function

In lossy compression problems for general sources with memory, the reproduction symbol at time instant n, Y_n, depends on past source symbols $\{X^{n-1} \overset{\triangle}{=} X_0,\ldots,X_{n-1}\}$, the present source symbol $\{X_n\}$, and future source symbols $\{X_{n+1},X_{n+2},\ldots\}$. This means that the optimal reproduction distribution, $P_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^{n} P_{Y_i|X^n,Y^{i-1}}(dy_i|x^n,y^{i-1})$, and has limitations in terms of real-time applications.

The first limitation is the computational complexity of obtaining an exact expression for the optimal reproduction distribution of the classical Rate Distortion Function (RDF), which gives the Optimal Performance Theoretically Attainable (OPTA) by noncausal codes. For general sources with memory, the exact expression is known only for Gaussian sources.

The second limitation is that the reconstruction distribution cannot be decomposed into a convolution of causal conditional distributions. This directly implies that, in general, the classical RDF cannot be used in nonanticipative joint source channel coding and in probabilistic matching of the source to the channel.

Christos K. Kourtellaris

In Chapter 2, we overcome this limitation by introducing an analytical framework via the nonanticipative RDF for general sources with memory, which give optimal nonanticipative reproduction distributions. The analysis includes discussion on noncausal codes, causal codes and sequential codes, relation of the nonanticipative RDF with Gorbunov and Pinsker nonanticipatory $\varepsilon-$entropy [34], and noisy coding theorems via joint source channel matching, which are further elaborated on Chapter 4.
The contributions are the following:

- Theoretical framework of the nonanticipative RDF.
- Closed form expression for the optimal reconstruction distribution, and solution of the nonanticipative RDF for stationary source-reproduction sequence.
- Bounds on the OPTA by noncausal and causal codes.
- Examples to illustrate the calculation of the nonanticipative RDF and the optimal reconstruction distribution for a binary symmetric Markov source.

1.1.2 Chapter 3: Structural Properties of Encoders for Channels with Memory and Feedback

Capacity of channels with feedback and associated coding theorems are often classified into Discrete Memoryless Channels (DMC) and channels with memory, with or without feedback [41, 65]. In chapter 3 we generalize current and past research in the area of capacity of channels with memory and feedback, and indicate the necessity of considering nonanticipative kernels in capacity optimization problems. We derive structural properties of capacity achieving encoders and channel input distribution for channels with memory and feedback, and structural properties of encoders that maximize directed information measure from the source to the channel output. Moreover, we apply dynamic programming recursions to compute the optimal conditional distributions. Finally, we generalize the Posterior Matching Scheme proposed in (PMS) [70] for channels with memory and feedback.

We derive a closed form expression for the capacity and the optimal channel input distributions, for a unit memory channel (the binary state symmetric channel). This analysis includes both the feedback and no feedback case, as well as, constraint and unconstraint capacity. The final expression of the capacity can be interpreted as the optimal time sharing

Christos K. Kourtellaris

among the two states of the channel.
The contributions are the following:

- Structural encoder properties which maximize directed information from the source to the channel output.
- Structural properties of capacity achieving distribution.
- Dynamic programming recursions to aid the computation of the optimal distributions.
- Optimal form of the channel input distribution that achieves the capacity for the unit memory channel with feedback.
- Capacity and optimal input distribution for the binary state symmetric channel.
- PMS for designing encoders, to achieve the information capacity for channels with memory and feedback.

1.1.3 Chapter 4: Nonanticipative Joint Source Channel Coding for Real-Time Transmission

Coding over infinitely large blocklengths, although optimal under certain conditions, it is not claimed to be the only optimal choice. Two classical memoryless examples, the Independent and Identical Distributed (IID) Bernoulli source with a single letter Hamming distortion criterion transmitted via a binary symmetric channel, and the Gaussian source with a mean square error distortion criterion transmitted via a Gaussian channel, indicate that real-time transmission performs optimally. For the case of the Bernoulli source this is achieved by the absence of an encoder-decoder scheme, hence both cost and complexity are reduced to an absolute minimum.

In Chapter 4, we merge Chapter 2 and Chapter 3 and we introduce the concept of nonanticipative transmission and minimum excess distortion, to show achievability of SbS codes with memory without anticipation via a noisy channel. Subsequently, we show that Joint Source Channel Matching (JSCM) of a binary symmetric Markov source with a single letter Hamming distortion and a binary state symmetric channel subject to a cost constraint is feasible. We additionally show than even in the unmatched case, where the capacity is greater than the nonanticipative rate distortion function, that uncoded schemes performs reliably in terms of average and excess distortion probability.

Christos K. Kourtellaris

The contributions are the following:

- Develop theoretical framework for nonanticipative and SbS transmission for general sources with memory and general channels with memory and feedback.
- Provide noisy coding theorems showing achievability of the nonanticipative code.
- Show that JSCM for a binary symmetric Markov source with single letter Hamming distortion via a binary state symmetric channel subject to a cost constraint.
- Provide unmatched SbS transmission.

1.1.4 Chapter 5: Nonanticipative Rate Distortion Function with Feedforward Information

In Chapter 5, we investigate the role of the nonanticipative feedforward side information, where the decoder has access to the previously transmitted symbols. We begin the analysis by introducing the concepts of feedforward compression and the nonanticipative RDF with feedforward information. We identify and compare the RDF when the decoder and the encoder has the same available information [84], the feedforward RDF [79], and the nonanticipative RDF with feedforward information. Here we prove that the first two measures are equivalent. Then, we elaborate on the nonanticipative RDF with feedforward side information, where we provide a closed form expression for the optimal reproduction distribution.

We continue our analysis focusing on Markov sources with certain distortion criteria, and we show that the feedforward RDF and nonanticipative RDF with feedforward information are equivalent. Finally, we solve examples for Markov sources via the proposed methodology, by calculating directly the optimal reproduction distribution, and the solution of the nonanticipative RDF.
The contributions are the following:

- Formulate the nonanticipative RDF with feedforward information at the decoder, characterize the optimal reproduction distribution and provide a closed form expression for the nonanticipative RDF with feedforward side information.
- Show equivalence between the mutual information with causal conditioning and the feedforward RDF.

Christos K. Kourtellaris

- Prove equivalence between feedforward RDF and nonanticipative RDF with feedforward information, for Markov sources under certain distortion measures.

- Provide a lower bound for the classical rate distortion problem and describe the Rate Loss of causal codes with respect to noncausal codes.

- Calculate the RDF for Markov sources with feedforward information.

Christos K. Kourtellaris

Chapter 2

Nonanticipative Rate Distortion Function

2.1 Introduction

In lossy compression source coding with fidelity constraint [4, 38], the sequence of real-valued symbols $X^n \triangleq \{X_0, X_1, \ldots, X_n\}$, $\mathscr{X}_{0,n} \triangleq \times_{i=0}^{n} \mathscr{X}_i$, $n \in \mathbb{N}$, $X_i \in \mathscr{X}_i$, generated by a source distribution P_{X^n}, is transformed by the encoder into a sequence of symbols, the compressed representation $Z^k \triangleq \{Z_0, Z_1, \ldots, Z_k\}$ (taking values in a finite alphabet set), which is then transmitted over a noiseless channel. The decoder at the channel output upon observing the compressed representation symbols produces the reproduction sequence $Y^n \triangleq \{Y_0, Y_1, \ldots, Y_n\} \in \mathscr{Y}_{0,n}$.

Such a compression system is called causal [56] if the reproduction symbol Y_n, depends on the present and past source symbols $\{X_0, \ldots, X_n\}$ but not on the future source symbols $\{X_{n+1}, X_{n+2}, \ldots\}$. Thus, in a causal source code the cascade of the encoder-decoder, called the reproduction coder, is a family of measurable functions $\{f_n : \ n = 0, 1, \ldots\}$, such that $Y_n \triangleq f_n(X_0, \ldots, X_n)$, while the compressed representation itself may be noncausal and have variable rate [56]. Consequently, the decoder can generate the reproductions with arbitrary delay.

Zero-delay source coding is a sub-class of causal coding, with the additional constraint that the compressed representation symbol Z_n, depends on the past and present source symbols $X^n \triangleq \{X_0, X_1, \ldots, X_n\}$, while the reproduction at the decoder Y_n of the present source symbol X_n, depends only on the compressed representation $Z^n \triangleq \{Z_0,\ Z_1,\ \ldots,\ Z_n\ \}$ received so far.

Christos K. Kourtellaris

Thus, a zero-delay coding system consists of a family of encoding-decoding measurable functions $\{h_i, f_i\} : i = 0, 1, \ldots$, such that $Z_i = h_i(\{X_j : \ j = 0, 1, \ldots, i\})$ and $Y_i = f_i(\{Z_j : \ j = 0, 1, \ldots, i\})$, $\forall i \geq 0$ [2, 25, 27, 48, 75].

On the other hand, the most efficient zero-delay coding systems in information theory is that of uncoded transmission, obtained by Joint Source Channel Coding (JSCC) based on Symbol-by-Symbol (SbS) transmission [31], also called source-channel matching. Two such fascinating examples are a) the Independent Identically Distributed (IID) binary source with Hamming distortion transmitted uncoded over a symmetric channel, and b) the IID Gaussian source with average squared-error distortion transmitted over an Additive White Gaussian Noise (AWGN) channel, with the encoder and decoder scaling their inputs. These examples demonstrate the potential of the joint source-channel coding system operating with zero-delay coding in complexity, when compared to the asymptotic performance of optimally separating the encoder/decoder to the source and channel encoders/decoders which require long processing delays.

In general, very little is known about the Optimal Performance Theoretically Attainable (OPTA) by causal, zero-delay codes, and based on SbS transmission. Often, bounds are introduced to quantify the rate loss due to causality and zero-delay of the coding systems compared to that of the noncausal coding systems.

Clearly, in many delay sensitive applications of lossy compression, limited end-to-end decoding delay is often desirable, while for real-time systems, such as, communication for control over finite rate channels [9, 11, 26, 55, 76], and in general, for systems involving feedback [77], causal and more importantly, zero-delay coding is preferable to noncausal coding.

Before we discuss our results and related literature, we identify some limitations of the classical information RDF with respect to its computation, and its applications to source-channel matching based on SbS transmission. These limitations, together with our interest to develop bounding techniques for noncausal and causal codes, motivated us to consider the nonanticipative RDF.

Recall that the classical information RDF. Given a source probability distribution $P_{X^n}(dx^n)$ and a reproduction probability distribution $P_{Y^n|X^n}(dy^n|x^n)$ the joint probability distribution $P_{Y^n,X^n}(dy^n, dx^n)$ of (Y^n, X^n), its Y^n marginal $P_{Y^n}(dy^n)$, and product measure $P_{X^n}(dx^n) \times$

Christos K. Kourtellaris

$P_{Y^n}(dy^n)$ are uniquely defined. Let

$$d_{0,n} : \mathscr{X}_{0,n} \times \mathscr{Y}_{0,n} \mapsto [0,\infty) \tag{2.1.1}$$

The single letter distortion function is defined by

$$d_{0,n}(x^n,y^n) \triangleq \sum_{i=0}^{n} \rho(x_i,y_i) \tag{2.1.2}$$

With respect to the distortion function, the fidelity set of reproduction conditional distributions is defined by[1]

$$Q_{0,n}(D) \triangleq \left\{ P_{Y^n|X^n} : \frac{1}{n+1} \int_{\mathscr{X}_{0,n}\times\mathscr{Y}_{0,n}} d_{0,n}(x^n,y^n)(P_{Y^n|X^n} \otimes P_{X^n})(dx^n,dy^n) \leq D \right\} \tag{2.1.3}$$

The finite-time information RDF is defined by

$$R_{0,n}(D) \triangleq \inf_{P_{Y^n|X^n} \in Q_{0,n}(D)} I(X^n;Y^n) \tag{2.1.4}$$

where

$$\begin{aligned} I(X^n;Y^n) &\triangleq \mathbb{D}(Q_{X^n,Y^n}||P_{X^n} \times Q_{Y^n}) \\ &= \int \log \frac{Q_{Y^n|X^n}(dy^n|x^n)}{Q_{Y^n}(dy^n)} Q_{Y^n|X^n}(dy^n|x^n) P_{X^n}(dx^n) \equiv \mathbb{I}_{X^n;Y^n}(P_{X^n}, Q_{Y^n|X^n}) \end{aligned}$$

in which $\mathbb{D}(.||.)$ is the Kullback$-$Leibler divergence, defined by

$$\mathbb{D}(P||Q) = \begin{cases} \int \log\left(\frac{P(dx)}{Q(dx)}\right) P(dx) & \text{if } P << Q \text{ and } \log\left(\frac{P(dx)}{Q(dx)}\right) \in L^1(P) \\ 0 & \text{otherwise} \end{cases} \tag{2.1.5}$$

The functional $\mathbb{I}_{X^n;Y^n}(P_{X^n}, Q_{Y^n|X^n})$ is used to denote the functional dependence of the mutual information on the source and reconstruction distributions.

The information RDF is defined by

$$R(D) \triangleq \lim_{n\longrightarrow\infty} \frac{1}{n+1} R_{0,n}(D) \tag{2.1.6}$$

[1] $\otimes$ denotes convolution of distributions.

Christos K. Kourtellaris

provided the limit exists and the infimum in $R_{0,n}(D)$ exists (it is finite) [20, 63]. Under general conditions [4, 38], (i.e., jointly stationary ergodic processes) it is already known that if the infimum over $Q_{0,n}(D)$ exists, then the limit $R(D) = \lim_{n \longrightarrow \infty} \frac{1}{n+1} R_{0,n}(D)$ exists, and $R(D)$ is the OPTA by noncausal codes.

Moreover, it is also known that if the optimal conditional distribution achieving the infimum in (2.1.4) exists, then it is given by the implicit expression

$$P^*_{Y^n|X^n}(dy^n|x^n) = \frac{e^{sd_{0,n}(x^n,y^n)} P^*_{Y^n}(dy^n)}{\int_{\mathscr{Y}_{0,n}} e^{sd_{0,n}(x^n,y^n)} P^*_{Y^n}(dy^n)}, \; s \leq 0 \tag{2.1.7}$$

where $s \in (-\infty, 0]$ is the Lagrange multiplier associated with the fidelity constraint $Q_{0,n}(D)$, and

$$P^*_{Y^n}(dy^n) = \int_{\mathscr{X}_{0,n}} P^*_{Y^n|X^n}(dy^n|x^n) P_{X^n}(dx^n)$$

Although, from the point of view of establishing a noiseless coding theorem giving an operational meaning to $R(D)$ as the OPTA by noncausal codes is by now standard, $R(D)$ has certain limitations.

The first limitation of the classical information RDF is the computational complexity of obtaining the exact expression of $R_{0,n}(D)$ and $P^*_{Y^n|X^n}(dy^n|x^n)$, for finite n, and $R(D)$, in the limit $n \longrightarrow \infty$, even for stationary sources. The exact expression of $R(D)$ is only known for a small class of sources, which are either memoryless or Gaussian, often with respect to a single-letter distortion function. For example, for finite alphabet sources, the exact computation of $R(D)$ is based on its single-letter expression. Indeed, for the Binary Symmetric Markov Source BSMS(p), the complete characterization of the OPTA by noncausal codes is currently unknown; more precisely, it is only known for a certain distortion region $0 \leq D \leq D_c$[2], $D_c = \frac{1}{2}\Big(1 - \sqrt{1-(\frac{q}{p})^2}\Big)$, $p = 1-q$, $q \leq \frac{1}{2}$, and only bounds are available [5, 37, 42].

The second limitation of the classical information RDF is the noncausality or anticipative form of the optimal reproduction distribution (2.1.7), which implies that for any time n, the reproduction at time $i \leq n$ of $x_i \in \mathscr{X}_i$ by $y_i \in \mathscr{Y}_i$ has the form $f_i(x^i, x_{i+1}, \ldots, x_n)$, $\forall i \leq n$, and hence it depends on the past and future source symbols (i.e., its is noncausal). The noncausality of the optimal reproduction distribution (2.1.7) follows directly by Bayes' rule,

[2] This is the region for which the exact value of $R(D)$ is known [5, 37, 42].

Christos K. Kourtellaris

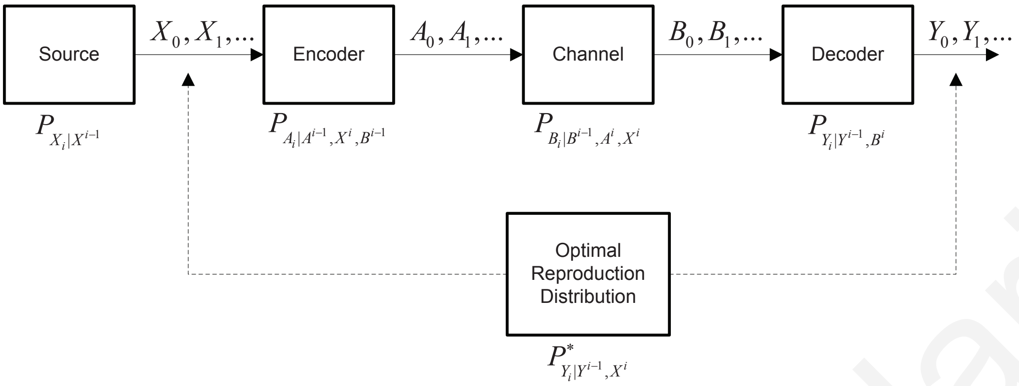

FIGURE 2.1.1: Block diagram of nonanticipative information transmission.

which yields

$$P^*_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^n P^*_{Y_i|Y^{i-1},X^n}(dy_i|y^{i-1},x^n) \tag{2.1.8}$$

$$\neq \otimes_{i=0}^n P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) \tag{2.1.9}$$

Therefore, probabilistically, the optimal reproduction distribution (2.1.7) of the classical information RDF cannot be decomposed into a convolution of causal conditional distribution (i.e., it is anticipative). The anticipative form of the optimal reproduction distribution (2.1.7) (or failure of (2.1.9) to hold) implies that, in general, the classical information RDF cannot be used in JSCC based on nonanticipative transmission (see Figure 2.1.1), and in probabilistic matching on the channel [10, 31, 44, 45]. An exception is the class of independent sources. Indeed, a necessary condition for probabilistic matching of the source to the channel via nonanticipative transmission as illustrated in Figure. 2.1.1, is causal conditioning of the optimal reproduction distribution, that is, (2.1.8) should be equal to (2.1.9), or equivalently, the following causality constraint expressed in terms of Markov chains (MC) should hold.

$$X_{i+1}^n \leftrightarrow (X^i, Y^{i-1}) \leftrightarrow Y_i,\ i = 0, 1, \ldots, n-1,\ \forall\, n \in \mathbb{N} \tag{2.1.10}$$

Among all the classes of sources, the only subclass for which the optimal reproduction distribution (2.1.7) of the classical information RDF is nonanticipative (i.e., causal with respect to future source symbols), and hence satisfies the necessary conditions for probabilistic matching of the source to the channel via nonanticipative transmission, is the independent source $\{X_n :\ n = 0, 1, \ldots\}$ with single letter distortion. In this case we have

$$P^*_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^n P^*_{Y_i|X_i}(dy_i|x_i) \tag{2.1.11}$$

Christos K. Kourtellaris

and hence $P^*_{Y_i|X_i}(dy_i|x_i)$ satisfies the necessary condition for probabilistic matching of the source and the channel (memoryless) via nonanticipative transmission. Alternatively stated, given any source (with or without memory), a necessary condition for probabilistic matching of the source to a noisy channel via nonanticipative transmission (as shown in Fig. 2.1.1) is the realization of the optimal reproduction distribution by an encoder-channel-decoder which process, at each time instant symbols causally. Moreover, such realization of the optimal reproduction distribution is a necessary condition for JSCC via nonanticipative or uncoded transmission [31, 44, 45]. This nonanticipative nature of the reproduction distribution is fundamental in the two examples of nonanticipative transmission mentioned earlier (see also [31]), e.g., the binary IID source with a Hamming distortion, and the IID Gaussian source with mean-square distortion. In fact, by recalling the necessary and sufficient conditions for source-channel matching based on nonanticipative transmission of memoryless sources and channels given in [31, Lemma 2, ii)] it requires (by adopting our notation) that the distortion satisfies

$$d(x_i,y_i) = -c_2 \log P_{X_i|Y_i}(dx_i|y_i) + d_0(x_i) \tag{2.1.12}$$

where $c_2 > 0$ and $d_0(\cdot)$ is an arbitrary function. It is easy to verify that (2.1.12) is just a restatement of (2.1.7), for memoryless sources (i.e., $P_{X^n}(x^n) = \otimes_{i=0}^n P_{X_i|X^{i-1}}(dx_i|x^{i-1}) = \otimes_{i=0}^n P_{X_i}(dx_i)$), where

$$P^*_{Y_i|X_i}(dy_i|x_i) = \frac{e^{sd(x_i,y_i)}P_{Y_i}(dy_i)}{\underbrace{\int_{\mathcal{Y}_i} e^{sd(x_i,y_i)}P_{Y_i}(dy_i)}_{g(x_i)}} \tag{2.1.13}$$

By simple manipulations (i.e., $P^*_{Y_i|X_i}(dy_i|x_i) = \frac{P^*_{X_i|Y_i}(dx_i|y_i)P^*_{Y_i}(dy_i)}{P_{X_i}(dx_i)}$) from (2.1.13) we have

$$d(x_i,y_i) = \frac{1}{s}\log P_{X_i|Y_i}(dx_i|y_i) + \frac{1}{s}\log\Big(\frac{g(x_i)}{P_{X_i}(dx_i)}\Big), \ s<0$$

That is, $c_2 = -\frac{1}{s}$, $d_0(x_i) = -\frac{1}{s}\log\big(\frac{g(x_i)}{P_{X_i}(dx_i)}\big)$, $\forall i$.

The main feature of the optimal reproduction distribution of the information nonanticipative RDF is that at each time instant n, it is described by the conditional distribution $P_{Y_n|Y^{n-1},X^n}(dy_n|y^{n-1},x^n)$, hence it is causal with respect to the past and present source symbols and past reproduction symbols (X^n,Y^{n-1}) for $n=0,1,\ldots$.

Christos K. Kourtellaris

The information nonanticipative RDF is defined as follows. Given a source distribution $P_{X^n}(dx^n)$, a sequence of reproduction distributions $\{P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) : \ i=0,1,\ldots,n\}$, and a measurable distortion function

$$d_{0,n}(.,.) : \mathcal{X}_{0,n} \times \mathcal{Y}_{0,n} \longmapsto [0,\infty), \ d_{0,n} = \sum_{i=0}^{n} \rho_{0,i}(T^i x^n, T^i y^n) \tag{2.1.14}$$

where $T^i x^n$ is for each time instant i, a causal mapping of x^n, i.e., $T^i x^n$ is measurable function of x^i, and similarly for $T^i y^n$, and an average fidelity set

$$\begin{aligned} Q^{na}_{0,n}(D) \triangleq \Big\{ & \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \triangleq \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) : \\ & \frac{1}{n+1} \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} d_{0,n}(x^n,y^n)(\overrightarrow{P}_{Y^n|X^n} \otimes P_{X^n})(dx^n,dy^n) \leq D \Big\} \end{aligned} \tag{2.1.15}$$

the finite-time information nonanticipative RDF is defined by

$$\begin{aligned} R^{na}_{0,n}(D) &\triangleq \inf_{\overrightarrow{P}_{Y^n|X^n}(\cdot|x^n) \in Q^{na}_{0,n}(D)} \int_{\mathcal{X}_{0,n} \times \mathcal{Y}_{0,n}} \log \Big(\frac{\overrightarrow{P}_{Y^n|X^n}(dy^n|x^n)}{P_{Y^n}(dy^n)} \Big) (\overrightarrow{P}_{Y^n|X^n} \otimes P_{X^n})(dx^n,dy^n) \\ &= \inf_{\overrightarrow{P}_{Y^n|X^n}(\cdot|x^n) \in Q^{na}_{0,n}(D)} \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n}) \end{aligned} \tag{2.1.16}$$

Here, $\mathbb{I}_{X^n \rightarrow Y^n}(\cdot,\cdot)$ is used to denote the functional dependence of $R^{na}_{0,n}(D)$ on the two distributions $\{P_{X^n}, \overrightarrow{P}_{Y^n|X^n}\}$. Whenever, the infimum in (2.1.16) does not exist the value of $R^{na}_{0,n}(D)$ is set to $+\infty$.

The information nonanticipative RDF rate is defined by

$$R^{na}(D) \triangleq \lim_{n \longrightarrow \infty} \frac{1}{n+1} R^{na}_{0,n}(D) \tag{2.1.17}$$

provided the limit exists; if the infimum in (2.1.16) does not exist we set $R^{na}(D) = +\infty$.

In this chapter, we consider the information nonanticipative RDF and we describe its applications in

1. Joint source-channel coding via nonanticipative transmission.
2. Bounding the OPTA by noncausal and causal codes [56] for general stationary sources.
3. Computing the OPTA by sequential codes [50, 75] for two dimensional sources.

Christos K. Kourtellaris

4. Show equivalence of the information nonanticipative RDF to Gorbunov and Pinsker nonanticipatory $\varepsilon-$entropy and message generation rates [34–36], which corresponds to Shannon information RDF with an additional causality constraint imposed on the optimal reproduction distribution.

5. Demonstrate that Gorbunov and Pinsker definition of nonanticipatory $\varepsilon-$entropy is limited by its own definition and hence, it cannot be extended to feedback control applications while the information nonanticipative RDF is easily generalized to such applications.

6. Derive the expression of the optimal reproduction distribution of the information nonanticipative RDF and characterize some of its properties.

7. Compute the nonanticipative RDF of the BSMS(p).

2.2 Definition of Lossy Compression Codes

In this section we introduce the precise definitions for the different classes of codes (noncausal, causal, sequential) and their associated information theoretic definitions (classical information RDF, causal information RDF based on entropy rate of reproduction symbols, sequential information RDF), in order to put into context the various applications of the nonanticipaive RDF.

2.2.1 Noncausal Codes and Classical RDF

Assume a random sequence $X^\infty \stackrel{\triangle}{=} \{X_i : i \in \mathbb{N}\}$ taking values in an arbitrary alphabet $\mathcal{X}_{0,\infty}$ and a compression scheme consisting of an encoder and a decoder. The encoder upon observing source sequences $\{X_i : i \in \mathbb{N}\} \in \mathcal{X}_{0,\infty}$ generates a message W, the compressed representation $\{Z_i : i \in \mathbb{N}\}$, taking values in a finite alphabet, which is then transmitted over a noiseless channel of rate R bits per source symbol to the decoder. At the channel output the decoder upon observing W obtains an estimate $\{Y_i : i \in \mathbb{N}\} \in \mathcal{Y}_{0,\infty}$ called the reproduction of the source sequence $\{X_i : i \in \mathbb{N}\}$. The cascade system consisting of the encoder and the decoder is often called the reproduction coder. The reproduction coder is a family of measurable functions $\{f_i : i \in \mathbb{N}\}$ such that $Y_i = f_i(\{X_j\}_{j=0}^{\infty})$ is the reproduction of the source

Christos K. Kourtellaris

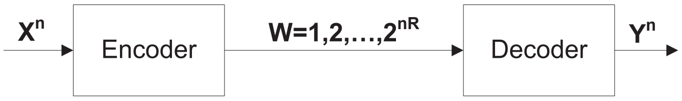

FIGURE 2.2.2: The source coding model

output. This coding scheme consists of the following encoding and decoding mappings.

Definition 2.1. (Noncausal codes)
A noncausal $(n,2^{nR})$ source code of block length n and normalized rate R consists the following encoding and decoding mappings.

$$\text{Encoding mapping:} \quad e_n : \mathscr{X}^n \to \mathscr{W} \triangleq \{1,2,..,2^{nR}\} \ \text{ and } W = e_n(X^n)$$
$$\text{Decoding mapping:} \quad g_i : \mathscr{W} \to \mathscr{Y}_i \ \text{ and } Y_i = g_i(W),\ i\ \in \mathbb{N}^n$$

Note that the ith component of Y^n is $Y_i = g_i(W) = g_i \circ e_n(X^n)$ and $Y^n = g^n(W)$. The distortion associated with the $(n,2^{nR})$ code is defined by

$$D = \frac{1}{n+1}E\Big\{d_{0,n}(X^n,g^n(e_n(X^n)))\Big\} = \frac{1}{n}\int_{\mathscr{X}_{0,n}} d_{0,n}(X^n,g^n(e_n(X^n)))P_{X^n}(dx^n)$$

where the expectation is with respect to the distribution $P_{X^n}(.)$ induced by X^n. The objective is to minimize the rate R subject to the fidelity constraint defined by $\frac{1}{n+1}E\{d_{0,n}(x^n,y^n)\} \leq D,\ D > 0$. The operational definition of a code $(n,2^{nR})$ is defined as follows.

Definition 2.2. (Achievable Rate)
A rate distortion pair (R,D) is called achievable if $\forall \varepsilon > 0$ and sufficiently large n there exists a sequence $(n,2^{nR})$ of noncausal code such that

$$\frac{1}{n+1}E\big\{d_{0,n}(X^n,Y^n)\big\} \leq D+\varepsilon$$

The rate distortion region for a source is the closure of the set of achievable rate distortion pairs (R,D). The classical RDF $R(D)$ is the infimum of rates R such that (R,D) is in the rate distortion region of the source for a given distortion D.

Suppose for a given $n \in \mathbb{N}$, a set of 2^{nR} reproduction sequences $\{y^n_{(i)} \in \mathscr{Y}_{0,n} :\ i = 1,\ldots,2^{nR}\}$ are chosen, and the source sequences $\{x^n_{(i)} \in \mathscr{X}_{0,n} :\ i = 1,\ldots,2^{nR}\}$ are mapped into this set of reproduction sequences. Thus, a noncausal code $(n,2^{nR})$ is specified, in which $\{y^n_{(i)} \in$

Christos K. Kourtellaris

$\mathscr{Y}_{0,n}: \ i = 1,\ldots,2^{nR}\}$ are all the possible codewords. Each codeword in such an encoding can be represented by a sequence of length nR binary bits. These bits are transmitted over a noiseless channel reliably such that the distortion between the reproduction sequence (one of the codewords) and the source sequence is the distortion defined for the encoding.

Given a source distribution $P_{X^n}(dx^n)$ and the encoding specified by a given noncausal code, the joint distribution $P_{X^n,Y^n}(dx^n,dy^n)$ of a joint ensemble $\{(X^n,Y^n) : n \in \mathbb{N}\}$ is defined as follows. The conditional distribution of a reproduction sequence Y^n given a source sequence X^n is given by

$$P_{Y^n|X^n}(dy^n|x^n) = \delta_{e_n(x^n)}(dy^n)$$

where $\delta_z(.)$ denotes the delta measure concentrated at point z. That is, $P_{Y^n|X^n}(.|x^n)$ becomes a point mass measure if y^n is the codeword into which the source sequence x^n is mapped into. The joint distribution is

$$P_{X^n,Y^n}(dx^n,dy^n) = P_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) = \delta_{e_n(x^n)}(dy^n) \otimes P_{X^n}(dx^n)$$

The analysis of the achievable rate is done by utilizing a test-channel, and then studying the behaviour of a randomly selected set of chosen codewords. Specifically, for a given test channel $P_{Y^n|X^n}(dy^n|x^n)$, and a source distribution $P_{X^n}(dx^n)$, an ensemble of source codes is generated by selecting sets of 2^{nR} sequence $y^n_{(1)}, y^n_{(2)}, \ldots, y^n_{(2^{nR})}$ drawn independently according to the distribution $P_{Y^n}(dy^n) = \int_{\mathscr{X}_{0,n}} P_{Y^n|X^n}(y^n|x^n) \otimes P_{X^n}(x^n)$. The probability measure on this ensemble is denoted by P_c. For a given set of codewords $y^n_{(1)}, y^n_{(2)}, \ldots, y^n_{(2^{nR})}$ in the ensemble each source sequence x^n is mapped into that codeword, $y^n_{(j)}$, which minimizes $d_{0,n}(x^n, y^n_{(j)})$, $j \in \{1,2,\ldots,2^{nR}\}$. The selection of codeword is arbitrary if the minimum is not unique.

Next we define the classical information RDF, a functional of the source distribution and reconstruction conditional distribution, which gives the OPTA by noncausal codes (see (2.1.1), (2.1.2)).

Definition 2.3. (Classical Information RDF) Let $Q_{0,n}(D)$ (assuming is non-empty) denote the average distortion or fidelity constraint defined by

$$Q_{0,n}(D) \triangleq \Big\{ P_{Y^n|X^n}(dy^n|x^n) : \frac{1}{n+1} \int_{\mathscr{X}_{0,n},\mathscr{Y}_{0,n}} d_{0,n}(x^n,y^n) P_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) \leq D \Big\} \tag{2.2.18}$$

Christos K. Kourtellaris

where $D \geq 0$. Define

$$R_{0,n}(D) = \inf_{P_{Y^n|X^n} \in \mathcal{Q}_{0,n}(D)} \mathbb{I}_{X^n;Y^n}(P_{X^n}, P_{Y^n|X^n}) \tag{2.2.19}$$

The classical information RDF is defined by

$$R(D) = \lim_{n \longrightarrow \infty} \frac{1}{n+1} R_{0,n}(D) \tag{2.2.20}$$

It is well-known that for stationary ergodic sources, finite alphabet spaces, and single letter distortion functions $d_{0,n}(x^n, y^n) = \sum_{i=0}^{n} \rho(x_i, y_i)$, that the information RDF, $R(D)$, is the OPTA by noncausal codes at distortion D. For memoryless sources (e.g., $\{X_i : i \in \mathbb{N}\}$, IID), R(D) is given by the single letter expression

$$R(D) = \inf_{P_{Y|X} : \int \rho(x,y) P_{Y|X}(dy|x) \otimes P_X(dx) \leq D} \mathbb{I}_{X;Y}(P_X, P_{Y|X}) \tag{2.2.21}$$

For information stable sources and distortion stable [41], $R(D)$ is also the optimal theoretically attainable rate at distortion D. Further, it is shown in [20, 63] that for Polish spaces $(\mathscr{X}_{0,n}, \mathscr{Y}_{0,n})$ in which $\mathscr{Y}_{0,n}$ is compact, and $d(x^n, .)$ continuous on $\mathscr{Y}_{0,n}$, that the infimum is attained at

$$P^*_{Y^n|X^n}(dy^n|x^n) = \frac{e^{s\rho(T^i x^n, T^i y^n)} P_{Y^n}(dy^n)}{\int_{\mathscr{Y}_{0,n}} e^{s\rho(T^i x^n, T^i y^n)} P_{Y^n}(dy^n)} \tag{2.2.22}$$

where $s \in (-\infty, 0]$ is the Lagrange multiplier associated with the distortion constraint. The condition on the compactness $\mathscr{Y}_{0,n}$ in [20] is removed in [63]. The generation of ensemble of codes is done via any test-channel distribution which achieves the RDF (2.2.21) for a given distortion D.

It is clear that in general, the optimal conditional distribution $P^*_{Y^n|X^n}(dy^n|x^n)$ which gives the infimum in (2.2.21) is noncausal, since by Bayes rule

$$P_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},X^n}(dy_i|y^{i-1}, x^n)$$

This conditional distribution cannot be applied in causal compression [56], since it violates the definition of causality that requires conditional independence on future source symbols, given the past and present source symbols. Therefore, in order to characterize the causal information RDF and to establish coding theorems, an information measure needs to be

Christos K. Kourtellaris

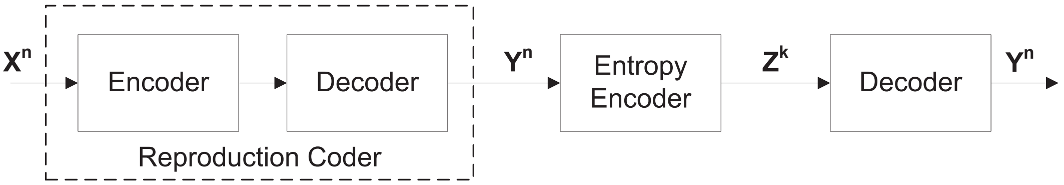

FIGURE 2.2.3: The causal source coding model

introduced, in order to impose a causality or nonanticipation constraint on the admissible set of reproduction conditional distributions. This is elaborated in the next subsection.

2.2.2 Causal Codes and Causal Information RDF

Following Neuhoff and Gilbert [56], the cascade of an encoder and decoder is called the reproduction coder, and a coder is called causal if its reproduction coder is causal. The precise definition is the following.

Definition 2.4. (Causal reproduction coder)
A reproduction coder $f_i : \mathscr{X}_{0,n} \mapsto \mathscr{Y}_i, \ \ i \in \mathbb{N}^n$ is called causal, if the mapping $x^n \mapsto f_i(x^n)$ is measurable, $\forall\, i \in \mathbb{N}^n$, $n \in \mathbb{N}$, and

$$f_i(x^n) = f_i(\hat{x}^n), \,\forall\, x^n \quad \text{such that} \quad x^i = \hat{x}^i \ \ \forall\, i \leq n$$

A source code is causal if the induced reproduction coder is causal. For a given $i \in \mathbb{N}$ the set of such reproduction codes f_i is denoted by $\mathbb{F}_i$, and $\mathbb{F}_{0,n} \overset{\triangle}{=} \times_{i=0}^{n} \mathbb{F}_i = \{f_i \in \mathbb{F}_i : i = 0, 1, \ldots, n\}$.

Therefore for causal codes, the induced reproduction coder, which is the cascade of the encoder and the decoder, must satisfy the causality constraint of Definition 2.4. Thus, causal codes are a subset of noncausal codes. In [56], Neuhoff and Gilbert have also shown that for IID sources one may design the reproduction coder first, followed by a lossless code as shown in figure Fig.2.2.3. Causal codes are dealt with the entropy rate of Y^n, while coding theorems are generalized in [83], in the presence of side information. However, no closed form expression is given for the reconstruction distribution, as in the OPTA of noncausal codes.

The probabilistic equivalent to Definition 2.4 is the following. Since for any $i \in \mathbb{N}^n$ the reconstruction symbol, Y_i, is allowed to depend on past and present source symbols $\{X_j : j = 0, 1, \ldots, i\}$ but not on the future ones $\{X_j : j = i+1, i+2, \ldots, n\}$, then the following

Christos K. Kourtellaris

Markov chain must hold

$$X_{i+1}^n \leftrightarrow (X^i, Y^{i-1}) \leftrightarrow Y_i, \quad \forall\, i\, \in \mathbb{N}^n \tag{2.2.23}$$

Given a source distribution $P_{X^n}(.)$ and a causal reproduction coder $\{f_i : i=0,1,\ldots,n\} \in \mathbb{F}_{0,n}$, the joint distribution $P_{X^n,Y^n}(.,.)$ is specified uniquely as follows.

$$\begin{aligned} P_{X^n,Y^n}(dx^n, dy^n) &= P_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) \\ &= \otimes_{i=0}^n P_{Y_i|Y^{i-1},X^n}(dy_i|y^{i-1},x^n) \otimes P_{X^n}(dx^n) \\ &\overset{(\alpha)}{=} \otimes_{i=0}^n P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) \otimes P_{X^n}(dx^n) \\ &= \otimes_{i=0}^n \delta_{f_i(x^i,y^{i-1})}(dy_i) \otimes P_{X^n}(dx^n) \end{aligned}$$

where the equality in (α) is due to the causality of the reproduction coder which satisfies Markov chain (2.2.23). Next, we provide the definition of causal codes and the operational definition of causal codes for noiseless channels.

The coding scheme for causal codes consists of the following encoding and decoding mappings, as well as the causal reproduction coder.

Definition 2.5. (Causal Codes)
A $(n, 2^{nR})$ causal source code of block length n, and rate R consists the following encoding mappings.

$$\begin{aligned} &\text{Encoding mapping:} \quad e_n : \mathscr{X}^n \rightarrow \mathscr{W} \triangleq \{1,2,..,2^{nR}\} \;\; \text{and } W = e_n(X^n) \\ &\text{Decoding mapping:} \quad g_i : \mathscr{W} \rightarrow \mathscr{Y} \;\; \text{and } Y_i = g_i(W),\; i\, \in \mathbb{N}^n \end{aligned}$$

such that the sequence of reproduction coders $\{f_i = g_i \circ e_n\}_{i=0}^n$ are causal.

Next, we give the operational definition of causal codes for which an information theoretic measure and a coding theorem are derived in [56].

Definition 2.6. (Operation of causal codes)
Let $Q_{0,n}^f(D)$ (assuming is non-empty) denote the average distortion or fidelity constraint defined by

$$Q_{0,n}^f(D) \triangleq \Big\{(f_0, f_1, \ldots, f_n) \in \mathbb{F}_{0,n} : \frac{1}{n+1}\mathbb{E}\Big\{d_{0,n}(x^n, y^n)\Big\} \leq D\Big\} \tag{2.2.24}$$

Christos K. Kourtellaris

where $D \geq 0$. Define

$$R^{c,o}_{0,n}(D) \stackrel{\triangle}{=} \inf_{(f_0,f_1,\ldots,f_n)\in \mathcal{Q}^f_{0,n}(D)} H(Y_0,Y_1,\ldots,Y_n) \tag{2.2.25}$$

The classical information causal RDF is defined by

$$R^{c,o}(D) = \lim_{n\longrightarrow\infty} \frac{1}{n+1} R^{c,o}_{0,n}(D) \tag{2.2.26}$$

provided the limit exists and the infimum of (2.2.25) is finite. If the infimum in (2.2.25) is not finite we set $R^{c,o}_{0,n}(D) = +\infty$.

By Definition 2.6, the classical (noncausal) RDF does not account for (2.2.23), hence in general the optimal reconstruction distribution does not satisfy causality, therefore a new RDF needs to be defined and its operational meaning established.

Before we define the information definition of causal codes we recall Neuhoff and Gilbert [56] operational and information definition of causal codes based on entropy. Consider a causal source code with an induced reproduction coder $\{f_i : i \in \mathbb{N}\} \in \mathbb{F}_{0,\infty}$ applied to a source $\{X_i : i \in \mathbb{N}\}$ and define the average distortion by

$$\bar{d}(\mathbf{x},\mathbf{y}) \stackrel{\triangle}{=} \limsup_{n\longrightarrow\infty} \frac{1}{n+1} \mathbb{E}\{d_{0,n}(X^n,Y^n)\}, \quad d_{0,n}(x^n,y^n) \stackrel{\triangle}{=} \sum_{i=0}^{n} \rho(x_i,y_i)$$

The average operational rate of the reproduction coder is defined by

$$r \stackrel{\triangle}{=} \limsup_{n\longrightarrow\infty} \frac{1}{n+1} \mathbb{E}\Big\{\ell_n(X^\infty)\Big\}$$

where $\ell_n(X^\infty)$ is the total number of bits received by the decoder at the time it reproduces the output sequence $\{Y_j : j \in \mathbb{N}\}$ when the source is $X^\infty \stackrel{\triangle}{=} \{X_i : i \in \mathbb{N}\}$. That is, if $Z_1, Z_2, \ldots,$ is the sequence of bits produced by the encoder in response to $\{X_i : i \in \mathbb{N}\}$, and if Y_n is produced by the decoder after receiving Z_l but before Z_{l+1}, then $\ell_n(X^\infty) = l$ (here it is assumed that the decoder has already produced Y^{n-1}).

Definition 2.7. (Causal Achievable Rate)

A rate distortion pair (R,D) is called achievable if $\forall \varepsilon > 0$ and sufficiently large n there exists a sequence $(n, 2^{nR})$ of causal codes such that

$$\frac{1}{n+1} E\big\{d_{0,n}(X^n,Y^n)\big\} \leq D + \varepsilon$$

Christos K Kourtellaris

The causal rate distortion region for a source is the closure of the set of causal achievable rate distortion pairs (R,D). The operational causal RDF is the infimum of rates R such that (R,D) is in the rate distortion region of a source for a given distortion D.

In [56] the OPTA by causal codes for a source $\{X_i : i \in \mathbb{N}\}$ denoted by $r^c(D)$, is shown to be the infimum of the average rates of all causal codes subject to an average distortion constraint $\bar{d}(\mathbf{x},\mathbf{y}) \leq D$, defined by

$$r^c(D) = \inf_{\{f_n(.):\ Y_n = f_n(X^\infty), f_n(.)\ \text{causal},\ n\in\mathbb{N}\},\ \bar{d}(\mathbf{x},\mathbf{y})\leq D} \limsup_{n\longrightarrow\infty} \frac{1}{n+1} H(Y^n) \tag{2.2.27}$$

Furthermore, it is shown that $r^c(D)$ is determined by properties of the reproduction coders. The general, coding system has an equivalent representation as shown in Figure 2.2.3.

Next, we define the test-channel conditional distribution of causal reproduction coder, and establish a lower bound to $r^c(D)$ via a variation of directed information. By definition, a causal reproduction coder utilizes a test-channel of the form

$$Q_{Y^n|X^n}(dy^n|x^n) = \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \overset{\triangle}{=} \otimes_{i=1}^{n} P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i),\ n \in \mathbb{N}$$

Similarly to the classical case, given a source $P_{X^n}(.)$ and a sequence of causal reproduction distributions $P_{Y_i|Y^{i-1},X^i} : i \in \mathbb{N}$ the joint distribution $P_{X^n,Y^n}(.,.)$ and the marginals are obtained as follows.

$$\begin{aligned}
P_{X^n,Y^n}(dx^n,dy^n) &= \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) \otimes P_{X_i|X^{i-1}}(dx_i|x^{i-1}) \\
&= \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) \\
P_{Y^n}(dy^n) &= P_{X^n,Y^n}(\mathscr{X}_{0,n}, dy^n) \\
P_{Y^n}(dy^n) &= \int_{\mathscr{X}_{0,n}} \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n)
\end{aligned}$$

The sequence uniquely defines the convolution measure $\overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^{n} P_{Y_i|X^i,Y^{i-1}}$ $(dy_i|x^i,y^{i-1})$ and vice-versa. The ensemble of causal codes, should be drawn independently according to the distribution $P_{Y^n}(dy^n)$. Therefore, the new information measure that should be used instead of mutual information, is

$$\mathbb{D}(\overrightarrow{P}_{Y^n|X^n} \otimes P_{X^n} || P_{Y^n} \times P_{X^n}) \equiv \mathbb{I}_{X^n\rightarrow Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n})$$

Christos K. Kourtellaris

The notation $\mathbb{I}_{X^n\rightarrow Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n})$ is used to define the directed information evaluated at $P_{X_i|X^{i-1},Y^{i-1}} = P_{X_i|X^{i-1}}$.

At this stage, it is informative to establish the relation between $r^c(D)$ and the information measure $\mathbb{I}_{X^n\rightarrow Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n})$. Let $Z^{(n)}(X^\infty)$ denote the first $\ell_n(X^\infty)$ bits produced by decoder, e.g., $Z^{(n)}(X^\infty)$ is the sequence of bits received by the decoder at the time it produces Y_n (assuming the decoder already produced Y^{n-1}). Then,

$$\begin{aligned}
\frac{1}{n+1}\mathbb{E}\{\ell_n(X^\infty)\} &\overset{(\alpha_1)}{\geq} \frac{1}{n+1}H(Z^{(n)}) \\
&\overset{(\alpha_2)}{\geq} \frac{1}{n+1}H(Y^n) \\
&\overset{(\alpha_3)}{\geq} \frac{1}{n+1}\sum_{i=0}^{n}\Big(H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1},X^i)\Big) \\
&= \frac{1}{n+1}\sum_{i=0}^{n} I(X^i;Y_i|Y^{i-1}) \\
&= \frac{1}{n+1}\mathbb{I}_{X^n\rightarrow Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n})
\end{aligned}$$

where (α_1) holds since the average length of the sequence of R.V.s is not less than the entropy of the sequence of the R.V.s, (α_2) holds since the entropy of a function of a sequence of R.V.s is not greater than the entropy of a sequence R.V, and (α_3) holds since the entropy is positive for finite alphabets. Note that inequality (α_3) is actually an equality because Y^{i-1} is causally dependent on $\{X_j : j = 0, 1, \dots, j-1\}$. The last equality holds since by (2.2.23), $P_{X_i|X^{i-1},Y^{i-1}} = P_{X_i|X^{i-1}}$ -a.s for all causal reproduction coders. Hence, taking the infimum on both sides

$$\begin{aligned}
r^c(D) &\geq \inf_{\{f_n(.):\, Y_n = f_n(X^\infty), f_n(.) \text{ causal},\, n\in\mathbb{N}\},\, \bar{d}(\mathbf{x},\mathbf{y})\leq D} \limsup_{n\longrightarrow\infty} \frac{1}{n+1}\mathbb{I}_{X^n\rightarrow Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n}) \\
&\geq \limsup_{n\longrightarrow\infty} \frac{1}{n+1} \inf_{\overrightarrow{P}_{Y^n|X^n}\in Q^{na}_{0,n}(D)} \mathbb{I}_{X^n\rightarrow Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n}) \qquad (2.2.28)
\end{aligned}$$

The bound on (2.2.28), obtained by randomizing reproduction coders is the quantity we obtained from (2.1.13), the nonanticipative information RDF. Moreover, since causal reproduction coders impose the additional causality constraint on the test channel of classical

Christos K. Kourtellaris

RDF, then the following relations hold.

$$
\begin{aligned}
r^c(D) &\geq \limsup_{n\longrightarrow\infty} \inf_{P_{Y^n|X^n}\in Q_{0,n}(D), P_{Y^n|X^n}(dy^n|x^n)=\overrightarrow{P}_{Y^n|X^n}(dy^n|x^n)} \frac{1}{n+1}\mathbb{I}_{X^n;Y^n}(P_{X^n},P_{Y^n|X^n}) \\
&= \limsup_{n\longrightarrow\infty} \inf_{\overrightarrow{P}_{Y^n|X^n}\in Q^{na}_{0,n}(D)} \frac{1}{n+1}\mathbb{I}_{X^n\rightarrow Y^n}(P_{X^n},\overrightarrow{P}_{Y^n|X^n}) \\
&\geq \limsup_{n\longrightarrow\infty} \inf_{P_{Y^n|X^n}\in Q_{0,n}(D)} \frac{1}{n+1}\mathbb{I}_{X^n;Y^n}(P_{X^n},Q_{Y^n|X^n})
\end{aligned}
$$

The last inequality holds because the infimum is over a larger set since $Q^{na}_{o,n}(D)\subseteq Q_{o,n}(D)$. Therefore, we can investigate the rate loss of causal codes with respect to the noncausal codes by the equation $RL \stackrel{\triangle}{=} R(D)-r^c(D)\leq R(D)-R^{na}(D)$.

2.2.3 Sequential Codes and Sequential RDF

Clearly, causal coding regards only the information structure of the decoder, since no constraints are imposed on how the compressed representation $\{Z_0,Z_1,Z_2,\ldots\}$ is created from $\{X_0,X_1,X_2,\ldots\}$ by the encoder and interpreted by the decoder. Since no restrictions are imposed on how the index W, or its equivalent binary representation $\{Z_0,Z_1,Z_2,\ldots\}$, is generated, causality does not imply zero-delay between the output of the decoder and the output of the source. This is obvious by the encoding mapping since the encoder takes n time units, until the index W is produced. Thus, the time ordering of the random variables for general causal codes is $X_1,X_2,..,X_n,Y_1,Y_2,..,Y_n$.

A further restriction on causal codes is to require zero-delay between the time X_n enters the decoder and Y_n is produced by the decoder. A causal code is said to have zero delay or it is sequential, if each compressed representation symbol Z_n depends on the past and present source sequence $X_0,X_1,\ldots,X_n$, and the reproduction at the decoder Y_n of the present source symbol X_n, depends on the compressed representation $Z_0,Z_1,\ldots,Z_n$ received so far. Thus, a zero delay or sequential code consists of a family of encoder-decoder measurable function $\{(h_i,f_i):i=0,1,\ldots\}$ such that $Z_i=h_i(\{X_j\}_{j=0}^i)$ and $Y_i=f_i(\{Z_j\}_{j=0}^i), i=0,1,\ldots$. Such codes are desirable for delay-sensitive applications and for communication for control application. Delayless codes are a subset of the family of causal source codes.

Coding theorems for Sequential Rate Distortion Function (SRDF) are discussed and derived in [75] utilizing a two-parameter random processes. Specifically, consider a two dimensional source $X^{T,N}\stackrel{\triangle}{=}\{X_{t,n}:t=0,1,\ldots,T,\ n=0,1,\ldots,N\}$ where t represents the time index and

Christos K. Kourtellaris

n represents the spatial index, such as in video coding applications. The natural time order is with respect to the time index, hence for a fixed t, one observes the spatial process $\{X_{t,0}, X_{t,1}, \ldots, X_{t,N}\}$. In this formulation, the notation used for sequential codes is the following. For each $t \in \{0,1,\ldots,T\}$, the time index alphabet is denoted by $\mathscr{X}_t^N \triangleq \otimes_{n=0}^N \mathscr{X}_{t,n}$, for each $n \in \{0,1,\ldots,N\}$ the spatial index alphabet is denoted by $\mathscr{X}_n^T \triangleq \otimes_{t=0}^T \mathscr{X}_{t,n}$, and the joint time and spatial alphabet is denoted by $\mathscr{X}^{T,N} \triangleq \otimes_{t=0}^T \otimes_{n=0}^N \mathscr{X}_{t,n}$. Thus, for a fixed $t \in \{0,1,\ldots,T\}$, $x_t^N \in \mathscr{X}_t^N$, for a fixed $n \in \{0,1,\ldots,N\}$, $x_n^T \in \mathscr{X}_n^T$, and $x^{T,N} = \{x_{i,j} : i = 0,1,\ldots,T,\ j = 0,1,\ldots,N\} \in \mathscr{X}^{T,N}$.

The precise definition of a sequential quantizer is the two-dimension generalization of the causal reproduction coder defined as follows.

Definition 2.8. (Sequential reproduction coder)
A reproduction coder $\{f_t^N : \mathscr{X}^{T,N} \mapsto \mathscr{Y}_t^N : t = 0,1,\ldots,T\}$ is called sequential, if the mapping $x^{T,N} \mapsto f_t^N(x^{T,N})$ is measurable, and $\forall t = 0,1,\ldots,T$,

$$f_t^N(x^{T,N}) = f_t^N(\hat{x}^{T,N}) \in \mathscr{Y}_t^N \ , \ \forall\, x^{T,N} \quad \text{such that}\ x^{t,N} = \hat{x}^{t,N}\ \forall\, t \leq T, \forall T,N \in \mathbb{N} \tag{2.2.29}$$

and the range of each function is at most countable. For a given $t \in \{0,1,\ldots,T\}$, the set of such reproduction coders denoted by $\mathbb{F}_t^N \triangleq \otimes_{n=0}^N F_t^N = \{f_{t,n} \in \mathbb{F}_{t,n} : n = 0,1,\ldots,N\}$, $\mathbb{F}^{T,N} \triangleq \otimes_{t=0}^T F_t^N = \{f_t^N \in \mathbb{F}_t^N : t = 0,1,\ldots,T\}$.

Clearly, a sequential reproduction coder is simply a two parameter generalization of the causal reproduction coder, in which for a fixed time index t, the reproduction symbol $y_t^N \in \mathscr{Y}_t^N$ is allowed to depend on past, present and future spatial symbols $\{X_{t,j} : j = 0,1,\ldots,N\}$. Therefore, the probabilistic equivalent of a sequential reproduction coder is such that it satisfies the following Markov chain.

$$\{X_{t+1,j}, X_{t+2,j}, \ldots, X_{T,j} : j = 0,1,\ldots,N\} \leftrightarrow (X^{t,N}, Y^{t-1,N}) \leftrightarrow Y_t^N \ t \in \{0,1,\ldots,T\} \tag{2.2.30}$$

Christos K. Kourtellaris

Given a source distribution $P_{X^{T,N}}(.)$ and a sequential reproduction coder $\{f_t^N : t=0,1,\ldots,T\}$, the joint distribution $P_{X^{T,N},Y^{T,N}}(.,.)$ is uniquely specified as follows

$$\begin{aligned}
P_{X^{T,N},Y^{T,N}}(dx^{T,N},dy^{T,N}) &= P_{Y^{T,N}|X^{T,N}}(dy^{T,N}|x^{T,N})\otimes P_{X^{T,N}}(dx^{T,N}) \\
&= \otimes_{t=0}^{T} P_{Y_t^N|Y^{t-1,N},X^{T,N}}(dy_t^N|y^{t-1,N}x^{T,N})\otimes P_{X^{T,N}}(dx^{T,N}) \\
&\overset{(a)}{=} \otimes_{t=0}^{T} P_{Y_t^N|Y^{t-1,N},X^{t,N}}(dy_t^N|y^{t-1,N}x^{t,N})\otimes P_{X^{T,N}}(dx^{T,N}) \\
&= \otimes_{t=0}^{T} \delta_{f_t^N(x^{t,N},y^{t-1,N})}(dy_t^N)\otimes P_{X^{T,N}}(dx^{T,N})
\end{aligned}$$

where the equality in (a) is due to the causality constraint of causal reproduction coder which satisfies Markov chain (2.2.30).

The sequential coding scheme consists the following encoding and decoding mappings and reproduction coder.

Definition 2.9. (Zero Delay Sequential Codes)
A sequential $(N,2^{NR_t})$ source code of block length N and normalized rate R_t at time $t\in\{0,1,\ldots,T\}$, consists of the following causal encoding and causal decoding mappings.

Encoding mappings:

$$e_t^N : \mathscr{X}^{t,N} \rightarrow \mathscr{W}_t^N \triangleq \{1,2,..,2^{R_t}\} \text{ and } W_t^N = e_t^N(X^{t,N}),\ \forall\, t\ \in\{0,1,\ldots,T\}$$

Decoding mapping:

$$g_t^N : \mathscr{W}_t^N \times \mathscr{Y}^{t-1,N} \rightarrow \mathscr{Y}_t^N \text{ and } Y_t^N = g_t^N(W_t^N,Y^{t-1,N}),\ \forall\, t\ \in\{0,1,\ldots,T\}$$

Note that the reproduction codes $\{f_t^N = g_t^N\circ e_t^N :\ t=0,1,\ldots,T\}$ are causal.

Sequential codes as introduced by Tatikonda [75] are dealt with mutual information between X^n and Y^n subject to a causal constraint on the reconstruction kernel. However, no closed expression is given for the optimal reproduction distribution $P_{Y_i^N|Y^{i-1,N},X^{i,N}}(dy_i^N|y^{i-1,N},x^{i,N}), i=0,1,\ldots,n$. The coding theorems are derived for two dimensional random precesses $X^{T,N}$, where N denotes the spatial block and T the time block, under the assumption that $P(dX^{T,N}) = \otimes_{n=0}^{N} P(dX_n^T)$, and $\{X_n^T : n=1,\ldots,N\}$ is identically distributed. In sequential codes the time-ordering between the output of the source and the output of the decoder is $X_0^N,Y_0^N,X_1^N,Y_1^N,\ldots$. Next we define the operational definition of a sequential RDF.

Christos K. Kourtellaris

Definition 2.10. (Operational Sequential RDF)
Let $Q^{N,f}_{0,T}(D)$ denote the average distortion or fidelity constraint defined by

$$\begin{aligned} Q^{N,f}_{0,T}(D) &\triangleq \Big\{(f_0^N, f_1^N, \ldots, f_T^N) \in \mathbb{F}^{T,N} : \frac{1}{T+1}\mathbb{E}\Big\{d^{T,N}(X^{T,N}, Y^{T,N}) \leq D\Big\}\Big\}, \\ d^{T,N}(x^{T,N}, y^{T,N}) &\triangleq \sum_{t=0}^{T} \rho_t^N(x^{t,N}, y^{t,N}) \end{aligned} \tag{2.2.31}$$

where $D \geq 0$ and $\mathbb{E}(.)$ denotes expectation with respect to distribution $P_{X^{T,N}}(.)$. Define

$$R^{S,N,O}_{0,T}(D) \triangleq \inf_{(f_0^N, f_1^N, \ldots, f_T^N) \in Q^{N,f}_{0,T}(D)} H(Y_0^N, Y_1^N, \ldots, Y_T^N) \tag{2.2.32}$$

The sequential operational RDF is defined by

$$R^{S,O}_{0,T}(D) \triangleq \lim_{N \longrightarrow \infty} \frac{1}{N+1} R^{S,N,O}_{0,T}(D)$$

provided the limit exists and the infimum in (2.2.32) is finite. If the infimum in (2.2.32) does not exist, then $R^{S,O}_{0,T}(D) = +\infty$.

Remark 2.11. The operational definition of sequential RDF above, is a slight variation of the one given in [75]. Specifically, on [75] there are two formulations. Formulation 1 assumes a pointwise distortion function

$$Q^{N,f,1}_{0,T}(D) \triangleq \Big\{(f_0^N, f_1^N, \ldots, f_T^N) \in \mathbb{F}^{T,N} : \mathbb{E}\{d^N(X_t^N, Y_t^N) \leq D_t\}, t = 0, 1, \ldots, T\Big\}$$

and formulation 2 assumes an average distortion

$$Q^{N,f,2}_{0,T}(D) \triangleq \Big\{(f_0^N, f_1^N, \ldots, f_T^N) \in \mathbb{F}^{T,N} : \frac{1}{T+1}\sum_{t=0}^{T} \mathbb{E}\{\rho_t^N(X_t^N, Y_t^N)\} \leq D\Big\}$$

Next, we describe the test-channel conditional distribution of sequential reproduction coder (i.e. (2.2.30)), and derive a converse coding theorem. By definition of sequential reproduction coder, the test channel is of the form

$$\overrightarrow{P}_{Y^{T,N}|X^{T,N}}(dy^{T,N}|x^{T,N}) \triangleq \otimes_{t=0}^{T} P_{Y_t^N|Y^{t-1,N}, X^{t,N}}(dy_t^N|y^{t-1,N}, x^{t,N})$$

Christos K Kourtellaris

Therefore, given a source distribution defined by $P_{X^{T,N}}$ and the test channel, the joint distribution $P_{X^{T,N},Y^{T,N}}(.,.)$ is defined uniquely by

$$\begin{aligned} P_{X^{T,N},Y^{T,N}}(dx^{T,N},dy^{T,N}) &= \otimes_{t=0}^{T} P_{Y_t^N|Y^{t-1,N},X^{t,N}}(dy_t^N|y^{t-1,N},x^{t,N}) \\ &\quad \otimes P_{X_t^N|X^{t-1,N},Y^{t-1,N}}(dx_t^N|x^{t-1,N},y^{t-1,N}) \\ &= \otimes_{t=0}^{T} P_{Y_t^N|Y^{t-1,N},X^{t,N}}(dy_t^N|y^{t-1,N},x^{t,N}) \otimes P_{X_t^N|X^{t-1,N}}(dx_t^N|x^{t-1,N}) \end{aligned}$$

where the last equality follows because of the definition of causal reproduction coder satisfying (2.2.30). The ensemble of codes should be drawn independently according $P_{Y^{T,N}}(dy^{T,N})$ given by

$$P_{Y^{T,N}}(dy^{T,N}) = \int_{\mathcal{X}^{T,N}} \overrightarrow{P}_{Y^{T,N}|X^{T,N}(dy^{T,N}|x^{T,N})} \otimes P_{X^{T,N}}(dx^{T,N}) \tag{2.2.33}$$

.

Next we define the sequential information RDF as a functional of the source distribution $P_{X^{T,N}}(.)$ and a sequential reproduction conditional distribution

$$Q_{Y^{T,N}|X^{T,N}(dy^{T,N}|x^{T,N})} = \overrightarrow{P}_{Y^{T,N}|X^{T,N}}(dy^{T,N}|X^{T,N})$$

Therefore, the new information measure that should be used is a special case of directed information defined by

$$\begin{aligned} I_{P_X^{T,N}}(X^{T,N} \rightarrow Y^{T,N}) &\triangleq \mathbb{D}(\overrightarrow{P}_{Y^{T,N}|X^{T,N}} \otimes P_{X^{T,N}} || \overrightarrow{P}_{Y^{T,N}} \otimes P_{X^{T,N}}) \\ &\equiv \mathbb{I}_{X^{T,N} \rightarrow Y^{T,N}}(P_{X^{T,N}}, \overrightarrow{P}_{Y^{T,N}|X^{T,N}}) \end{aligned}$$

The functional $\mathbb{I}_{X^{T,N} \rightarrow Y^{T,N}}(.,.)$ indicates the dependence on the distributions $\{P_{X^{T,N}}, \overrightarrow{P}_{Y^{T,N}|X^{T,N}}\}$. This functional is a variant of directed information from $X^{T,N}$ to $Y^{T,N}$ and the sequential information RDF its defined via is infimum over the average distortion constraint.

Definition 2.12. (Sequential Information RDF)
Let $Q_{0,T}^{S,N}(D)$, assuming is non-empty, denote the average distortion or fidelity constraint defined by

$$\begin{aligned} Q_{0,T}^{S,N}(D) \triangleq \{ &\overrightarrow{P}_{Y^{T,N}|X^{T,N}}(.|.) : \frac{1}{T+1} \int_{\mathcal{X}^{T,N},\mathcal{Y}^{T,N}} d_{0,T}^{N}(x^{T,N},y^{T,N}) \\ &\overrightarrow{P}_{Y^{T,N}|X^{T,N}}(dy^{T,N}|x^{T,N}) \otimes P_{X^{T,N}}(dx^{T,N}) \leq D\} \end{aligned}$$

Christos K. Kourtellaris

where $D \geq 0$.
Define

$$R^{S,N}_{0,T}(D) \stackrel{\triangle}{=} \inf_{\overrightarrow{P}_{Y^{T,N}|X^{T,N}} \in Q^{S,N}_{0,T}(D)} \mathbb{I}_{X^{T,N} \rightarrow Y^{T,N}}(P_{X^{T,N}}, \overrightarrow{P}_{Y^{T,N}|X^{T,N}}) \tag{2.2.34}$$

The sequential RDF is defined by

$$R^{S}_{0,T}(D) = \lim_{N \longrightarrow \infty} \frac{1}{N+1} R^{S,N}_{0,T}(D) \tag{2.2.35}$$

provided the limit exists and the infimum in (2.2.34) is finite. If the infimum in (2.2.34) does not exist, then $R^{S}_{0,T}(D) = +\infty$.

Next, we give a converse coding theorem for sequential RDF.

Theorem 2.13. *(Sequential Converse Coding Theorem)*
Let $\{f_{t,n} : t = 0,1,\ldots,T, n = 0,1,\ldots,N\} \in \mathbb{F}^{T,N}$ *be a sequential reproduction coder satisfying the average distortion constraint. Then*

$$R^{S,N,O}_{0,T}(D) \geq R^{S,N}_{0,T}(D)$$

Proof. Consider any sequential reproduction coder. Then

$$\begin{aligned}
\frac{1}{(T+1)(N+1)} \sum_{t=0}^{T} \log\left(e^{R_t}\right) \quad & \geq \frac{1}{(T+1)(N+1)} \sum_{t=0}^{T} H(f^N_t(X^{t,N})) \\
& \geq \frac{1}{(T+1)(N+1)} H(f^N_0(X^{0,N}), f^N_1(X^{1,N}), \ldots, f^N_T(X^{T,N})) \\
& \geq \frac{1}{(T+1)(N+1)} I(X^N_0, X^N_1, \ldots, X^N_T; Y^N_0, Y^N_1, \ldots, Y^N_T) \\
& = \frac{1}{(T+1)(N+1)} I_{P^{T,N}_X}(X^N_0, X^N_1, \ldots, X^N_T; Y^N_0, Y^N_1, \ldots, Y^N_T) \\
& = \frac{1}{(T+1)(N+1)} \mathbb{I}_{X^{T,N} \rightarrow Y^{T,N}}(P_{X^{T,N}}, \overrightarrow{P}_{Y^{T,N}|X^{T,N}})
\end{aligned}$$

Taking the infimum over $\{f_{t,n} : t = 0,1,\ldots,T, n = 0,1,\ldots,N\} \in \mathbb{F}^{T,N}$ which satisfies the average distortion constraint yields

$$\begin{aligned}
\frac{1}{(N+1)(T+1)} R^{S,N,O}_{0,T}(D) \geq \quad & \frac{1}{(N+1)(T+1)} \mathbb{I}_{X^{T,N} \rightarrow Y^{T,N}}(P_{X^{T,N}}, \overrightarrow{P}_{Y^{T,N}|X^{T,N}}), \\
& \forall \overrightarrow{P}_{Y^{T,N}|X^{T,N}} \in Q^{S,N}_{0,T}(D)
\end{aligned}$$

Christos K. Kourtellaris

Further, taking the infimum over $\overrightarrow{P}_{Y^{T,N}|X^{T,N}} \in Q^{S,N}_{0,T}(D)$ the inequality is obtained. □

Next, we introduce specific assumptions which are sufficient to apply the coding theorem derived in [75].

Assumption 2.14.

1. The source is a two dimensional process $\{X_{t,n}:\ t=0,1,\ldots,T,\ n=0,1,\ldots,N\}$ with finite alphabet, $X_{t,n}\in\mathscr{X}_{t,n}\ \forall\ t\ \in\{0,1,\ldots,T\},\ n\ \in\{0,1,\ldots,N\}$, having finite dimensional distributions $P_{X^{T,N}}(dx^{T,N})=\otimes_{n=0}^{N}P_{X_n^T}(dx_n^T)$ and $\{X_n^T:n=0,1,\ldots,N\}$ are identically distributed.
2. The distortion functions $d^{T,N}:\mathscr{X}^{T,N}\times\mathscr{Y}^{T,N}\mapsto[0,\infty)$ is measurable and

$$\begin{aligned}
d^{T,N}(x^{T,N},y^{T,N}) &\triangleq \frac{1}{T+1}\sum_{t=0}^{T}\rho_t^N(x_t^N,y_t^N)\\
&= \frac{1}{(T+1)(N+1)}\sum_{t=0}^{T}\sum_{n=0}^{N}\rho_t(x_{t,n},y_{t,n})\\
&= \frac{1}{N+1}\sum_{n=0}^{N}\rho_{0,T}(x_n^T,y_n^T)
\end{aligned}$$

where,

$$\rho_{0,T}(x_n^T,y_n^T)=\frac{1}{(T+1)}\sum_{t=0}^{T}\rho_t(x_{t,n},y_{t,n})$$

Assumption 2.14.1 states that the random processes $X_n^T\triangleq\{X_{0,n},X_{1,n},\ldots,X_{T,n}\}$ and $X_m^T\triangleq\{X_{0,m},X_{1,m},\ldots,X_{T,m}\}$ are independent $\forall m\neq n,\ m,n\in\{0,1,\ldots,N\}$, and that the processes $\{X_n^T:n=0,1,\ldots,N\}$ are identically distributed. Assumption 2.14.2 states that the distortion function is single letter.

The next theorem establishes that under Assumptions 2.14 it is sufficient to restrict the expression of sequential operational RDF to a single letter with respect to the spatial index.

Theorem 2.15. *Suppose Assumption 2.14.1 holds. Then we have the following*

Christos K. Kourtellaris

1. *The following lower bound holds.*

$$\begin{aligned}\mathbb{I}_{X^{T,N}\rightarrow Y^{T,N}}(P_{X^{T,N}}\overrightarrow{P}_{Y^{T,N}|X^{T,N}}) &= I_{P_{X^{T,N}}}(X^{T,N};Y^{T,N}) \\ &= \sum_{t=0}^{T} I_{P_{X^{T,N}}}(X^{t,N};Y_t^N|Y^{t-1,N}) \qquad (2.2.36)\\ &\geq \sum_{t=0}^{T}\sum_{n=0}^{N} I_{P_{X_n^t}}(X_n^t;Y_{t,n}|Y_n^{t-1}) \\ &= \sum_{n=0}^{N} I_{P_{X_n^T}}(X_n^T;Y_n^T) \qquad (2.2.37)\end{aligned}$$

2. *The lower bound in (2.2.37) holds with equality if and only if the following almost sure condition hold.*

$$P_{Y_t^N|Y^{t-1,N},X_t^N}(dy_t^N|y^{t-1,N}x_t^N) = \otimes_{n=0}^{N} P_{Y_{t,n}|Y_n^{t-1},X_n^t}(dy_{t,n}|y_n^{t-1}x_n^t),\ t=0,1,\ldots,T \quad (2.2.38)$$

Moreover, under the Assumption 2.14.2 the infimum over conditional distributions $\overrightarrow{P}_{Y^{T,N}|X^{T,N}} \in Q_{0,T}^{S,N}$ of $\mathbb{I}_{X^{T,N}\rightarrow Y^{T,N}}(P_{X^{T,N}},\overrightarrow{P}_{Y^{T,N}|X^{T,N}})$ has the property (2.2.38), and

$$R_{0,T}^{S,N}(D) = (N+1)\inf_{\overrightarrow{P}_{Y^T|X^T}\in Q_{0,T}^S(D)} \mathbb{I}_{X^T\rightarrow Y^T}(P_{X^T},\overrightarrow{P}_{Y^T|X^T}) \equiv (N+1)\bar{R}_{0,T}^S(D) \quad (2.2.39)$$

where $\bar{R}_{0,T}^S(D)$ is the single letter with respect to space index $n\in\{0,1,\ldots,N\}$ of the sequential RDF $R_{0,T}^{S,N}(D)$, and $P_{X^T}(.)$ is the single letter source distribution,

$$Q_{0,T}^S(D) = \{\overrightarrow{P}_{Y^T|X^T} : \mathbb{E}\{d_{0,T}(X^T,Y^T)\}\leq D\},\ d_{0,T}(x^T,y^T) \triangleq \frac{1}{T+1}\sum_{t=0}^{T}\rho_t(x_t,y_t)$$

and $\overrightarrow{P}_{Y^T|X^T}(.|x^T)$ is the single letter reproduction channel.

Proof. See [75].

The sequential coding theorem described below is derived in Tatikonda [75].

Theorem 2.16. *(Sequential Coding Theorem)*
Suppose the source alphabet is finite and Assumption 2.14 hold. Then, for any $\varepsilon > 0$ *and*

Christos K. Kourtellaris

finite T, there exists an $N(\varepsilon,T)$ such that for all $N \geq N(\varepsilon,T)$

$$\frac{1}{N+1} R^{S,O}_{0,T} \leq \bar{R}^{S}_{0,T}(D) + \varepsilon$$

Proof. The derivation is based on strong typicality of sequences utilizing the IID assumption of random processes $\{X^T_n : n = 0,1,\ldots,N\}$. □

Finally, note that the sequential RDF is not delayless because it corresponds to block coding. That is, it is causal with respect to the blocks of the data.

2.3 Nonanticipative Information RDF

In this section, we define the nonanticipative information RDF as a functional of a source distribution P_{X^n} and a causal reproduction conditional distribution $\overrightarrow{P}_{Y^n|X^n}(dy^n|x^n)$. This functional is a variant of directed information from the source X^n to the reconstruction Y^n and the nonanticipative information RDF is defined via its infimum over the average distortion constraint.

Definition 2.17. (Nonanticipative Information RDF)
Let $Q^{na}_{0,n}(D)$ (assuming is non-empty) denote the average distortion or fidelity constraint defined by

$$Q^{na}_{0,n}(D) \stackrel{\triangle}{=} \{\overrightarrow{P}_{Y^n|X^n}(.|.) : \frac{1}{n+1}\int_{\mathscr{X}_{0,n},\mathscr{Y}_{0,n}} d_{0,n}(x^n,y^n)\overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) \leq D\} \quad (2.3.40)$$

where $D \geq 0$. Define

$$R^{na}_{0,n}(D) \stackrel{\triangle}{=} \inf_{\overrightarrow{P}_{Y^n|X^n} \in Q^{na}_{0,n}(D)} \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n}) \quad (2.3.41)$$

The nonanticipative RDF is defined by

$$R^{na}(D) = \lim_{n \longrightarrow \infty} \frac{1}{n+1} R^{na}_{0,n}(D) \quad (2.3.42)$$

provided the limit exists and the infimum in (2.3.41) is finite. If the infimum in (2.3.42) does not exist we set $R^{na}(D) = +\infty$.

Christos K. Kourtellaris

An equivalent statement of the Definition 2.17 is that the source is conditional independent on the previous reconstruction symbols given all previous source symbols (see Lemma 2.21). This holds since its reproduction symbol at any time instant is a function of up to current source symbols. The nonanticipative property is necessary for showing operational meaning to the JSCC based on nonanticipative transmission, for delayless communication.

Next, we provide the converse coding theorem for the nonanticipative information RDF.

Theorem 2.18. *(Converse Theorem for Causal Codes)*
The following bounds hold

$$(n+1)r^{c}_{0,n}(D) \geq R^{na}_{0,n}(D) \geq R_{0,n}(D) \tag{2.3.43}$$

where

$$r^{c}_{0,n}(D) \triangleq \inf_{\{f_i(.)\colon\, Y_i=f_i(X^n), f_i(.)\ causal\ i=0,1,\ldots,n\},\ \bar{d}(\mathbf{x},\mathbf{y})\leq D} \frac{1}{n+1}\mathbb{E}\Big\{\ell_n(X^n)\Big\}, \forall\, n\in\mathbb{N}$$

Proof. Consider the joint distribution defined by $P_{X^n}(dx^n)$, and a causal conditioning reproduction distribution $\overrightarrow{P}_{Y^n|X^n}(\cdot|x^n)$. Then, by data processing inequality we have the following bounds.

$$\begin{aligned}\frac{1}{n+1}\mathbb{E}\Big\{\ell_n(X^\infty)\Big\} &\geq \frac{1}{n+1}H(Y^n)\\ &\geq \frac{1}{n+1}\sum_{i=0}^{n}\Big\{H(Y_i|Y^{i-1}) - H(Y_i|Y^{i-1},X^i)\Big\}\\ &\overset{(\alpha)}{=} \frac{1}{n+1}\mathbb{I}_{X^n\rightarrow Y^n}(P_{X^n},\overrightarrow{P}_{Y^n|X^n})\end{aligned} \tag{2.3.44}$$

where (α) follows from the fact that the joint distribution is defined by $P_{X^n}(dx^n)$ and the conditional reproduction distribution $\overrightarrow{P}_{Y^n|X^n}(\cdot|x^n)$. Therefore, given a distortion function $d_{0,n}(x^n,y^n)$ and a distortion level $D\geq 0$, for any finite time $n\in\mathbb{N}$, by using (2.3.44), and taking the infimum over the reproduction codes over randomized reproduction distribution $\overrightarrow{P}_{Y^n|X^n}(\cdot|x^n)\in Q^{na}_{0,n}(D)$ we have the following bounds.

$$(n+1)r^{c}_{0,n}(D) \geq R^{na}_{0,n}(D) \geq R_{0,n}(D),\ \forall n\geq 0 \tag{2.3.45}$$

□

Christos K. Kourtellaris

The bounds in (2.3.45) remain valid if we divide by $\frac{1}{n+1}$ take $\limsup_{n\longrightarrow\infty}$ and then the infimum giving

$$r^c(D) \geq R^{na,+}(D) \triangleq \limsup_{n\longrightarrow\infty} \frac{1}{n+1} R^{na}_{0,n}(D) \geq R^+(D) \triangleq \limsup_{n\longrightarrow\infty} \frac{1}{n+1} R(D) \quad (2.3.46)$$

where

$$r^c(D) \triangleq \limsup_{n\longrightarrow\infty} \frac{1}{n+1} r^c_{0,n}(D) \quad (2.3.47)$$

Therefore, the information nonanticipative RDF, $R^{na}(D)$, is a lower bound on $r^c(D)$, the OPTA by causal codes, and an upper bound to the classical RDF $R(D)$, the OPTA by non-causal codes. These bounds are investigated recently in [22] for quadratic fidelity and stationary Gaussian sources, but due to the complexity of computing $R^{na}(D)$, they introduced additional bounds.

While the expression provided for causal codes [56, 83] is quite attractive, its computation for general sources is very difficult, and no specific examples are computed for sources with memory aside for the case of high resolution [49]. The OPTA by causal codes is bounded below by the expression of nonanticipative RDF rate, which by its turn is bounded below by the expression of classical RDF. The advantage of the proposed nonanticipative RDF, $R^{na}_{0,n}(D)$, is that the optimal reproduction distributions are easily computable.

Remark 2.19. (Letter-by-Letter (LbL) and Coupled Distortion Functions)
Although in our analysis of information nonanticipative RDF we consider an average fidelity set, namely, $Q^{na}_{0,n}(D)$, we can also handle Letter-by-Letter (LbL), and Coupled Letter (CL) Distortion Functions (i.e., the current value of the source depends on the previous and current values of the reproduction) defined by $d_{0,i} : \mathscr{X}_i \times \mathscr{Y}_{0,i} \longmapsto [0,\infty),\ i=0,1,\ldots,n$, with corresponding fidelity set

$$Q^{na,CL}_{0,n}(D_0,\ldots,D_n) \triangleq \Big\{ \overrightarrow{P}_{0,n}(\cdot|x^n) : \mathbb{E}\Big(d_{0,i}(X_i,Y^i)\Big) \leq D_i,\ i=0,1,\ldots,n \Big\} \quad (2.3.48)$$

where $D_i \geq 0,\ i=0,1,\ldots,n$.

Such LbL and CL distortion functions are employed in sequential coding of correlated sources with encoding and/or decoding frame-delays in [50, 80]. However, the average distortion includes thess as special cases.

Christos K. Kourtellaris

2.3.1 Relation to Nonanticipatory ε-Entropy and Message Generation Rates

In this section, we recall Gorbunov-Pinsker's definition of nonanticipatory ε-entropy [34]. Then, we show equivalence of certain statements regarding conditional independence and finally, we show equivalence of the information nonanticipative RDF, information nonanticipative RDF rate, and Gorbunov and Pinsker's definition of nonanticipatory ε-entropy and message generation rates, respectively.

For a given source distribution P_{X^n} and a reproduction $P_{Y^n|X^n} \in Q_{0,n}(D)$, Gorbunov and Pinsker restricted the fidelity set of classical RDF $Q_{0,n}(D)$ to those reproduction distributions which satisfy the following Markov chain (MC).

$$X_{n+1}^{\infty} \leftrightarrow X^n \leftrightarrow Y^n \Longleftrightarrow P_{Y^n|X^\infty}(dy^n|x^\infty) = P_{Y^n|X^n}(dy^n|x^n) - a.a.\ x^\infty \in \mathscr{X}_{0,\infty}, n = 0,1,\ldots \tag{2.3.49}$$

Then, they introduced the nonanticipatory ε-entropy and nonanticipatory message generation rate as follows.

Definition 2.20. (Nonanticipatory ε-entropy and message generation rate)
Consider the fidelity constraint set $Q_{0,n}(D)$. The nonanticipatory ε-entropy is defined by

$$R_{0,n}^{\varepsilon}(D) \stackrel{\triangle}{=} \inf_{\substack{P_{Y^n|X^n} \in Q_{0,n}(D): \\ X_{i+1}^n \leftrightarrow X^i \leftrightarrow Y^i,\ i=0,1,\ldots,n-1}} I(X^n;Y^n) \tag{2.3.50}$$

provided the infimum in (2.3.50) over $Q_{0,n}(D)$ and $X_{i+1}^n \leftrightarrow X^i \leftrightarrow Y^i$, $i = 0,1,\ldots,n-1$, exists; if not, then we set $R_{0,n}^{\varepsilon}(D) = +\infty$. The nonanticipatory message generation rate of the source is defined by

$$R^{\varepsilon}(D) \stackrel{\triangle}{=} \lim_{n \longrightarrow \infty} \frac{1}{n+1} R_{0,n}^{\varepsilon}(D) \tag{2.3.51}$$

provided the limit in the RHS of (2.3.51) exists; if the infimum in (2.3.50) does not exist, we set $R^{\varepsilon}(D) = +\infty$. In addition, we have

$$R^{\varepsilon,+}(D) \stackrel{\triangle}{=} \inf_{\substack{P_{Y^\infty|X^\infty} \in Q_{0,\infty}(D): \\ X_{i+1}^\infty \leftrightarrow X^i \leftrightarrow Y^i,\ i=0,1,\ldots}} \lim_{n \longrightarrow \infty} \frac{1}{n+1} I(X^n;Y^n) \geq R^{\varepsilon}(D) \tag{2.3.52}$$

Christos K. Kourtellaris

The MC constraint (2.3.49) is a probabilistic version of a randomized causal reproduction coder as defined in Definition 2.4. Thus, a source code is called causal if the reproduction code is causal. Since the class of randomized reproduction codes embeds deterministic codes, then probabilistically, a reproduction coder is causal if and only if the following MC holds $X_{i+1}^{\infty} \leftrightarrow X^i \leftrightarrow Y_i, \forall i \in \mathbb{N}$. Therefore, nonanticipatory ε-entropy, $R_{0,n}^{\varepsilon}(D)$, imposes a probabilistic causality constraint on the optimal reproduction distribution.

Gorbunov and Pinsker [36, 60] proceeded further to compute $R^{\varepsilon}(D) \overset{\triangle}{=} \lim_{n\longrightarrow\infty} \frac{1}{n+1} R_{0,n}^{\varepsilon}(D)$, whenever the limit exists, for the class of stationary ergodic scalar Gaussian sources by working on the frequency domain using power spectral densities. Further, in [36, 60] it is also shown that in the limit as $D \rightarrow 0$, the nonanticipatory message generation rate $R^{\varepsilon}(D)$ of stationary Gaussian sources converges to the classical information RDF. Recently in [22], the authors revisited the nonanticipatory ε-entropy for Gaussian sources and a quadratic distortion function to derive several bounds for the OPTA by causal and noncausal codes, using another expression which is an upper bound on $R^{\varepsilon}(D)$.

Now, we are ready to establish the connection between nonanticipatory ε-entropy (2.3.50) (e.g., $R_{0,n}^{\varepsilon}(D)$) and information nonanticipative RDF (e.g., $R_{0,n}^{na}(D)$), and message generation rate of the source (2.3.51) (e.g., $R^{\varepsilon}(D)$) and information nonanticipative RDF rate (e.g., $R^{na}(D)$), which follow directly from the following equivalent statements of MCs.

Lemma 2.21. (*Equivalent Nonanticipative Statements*)

The following statements are equivalent.

MC1: $P_{Y^n|X^n}(dy^n|x^n) = \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i), \ \forall n \in \mathbb{N}.$

MC2: $Y_i \leftrightarrow (X^i, Y^{i-1}) \leftrightarrow (X_{i+1}, X_{i+2}, \ldots, X_n)$ *forms a MC, for each* $i = 0, 1, \ldots, n-1, \ \forall n \in \mathbb{N}$.

MC3: $Y^i \leftrightarrow X^i \leftrightarrow X_{i+1}$ *forms a MC, for each* $i = 0, 1, \ldots, n-1, \ \forall n \in \mathbb{N}$.

MC4: $X_{i+1}^n \leftrightarrow X^i \leftrightarrow Y^i$ *forms a MC, for each* $i = 0, 1, \ldots, n-1, \ \forall n \in \mathbb{N}$.

Proof. See Appendix A.1. □

The fact that **MC1**, **MC2**, **MC3** is obvious. The implication of **MC4** that implies any of **MC1**, **MC2**, **MC3** is also known. What is new in Lemma 2.21 is the equivalence of **MC4** with any of **MC1**, **MC2**, **MC3**. Note that **MC3** of Lemma 2.21 is precisely Granger's definition of temporal causality [71], which is used in econometrics to unravel complex

Christos K. Kourtellaris

relation between macroeconomic variables from a time series observation. It is also applied in bioengineering [47, 71] and more recently in neuroimaging to infer that $\{Y_n : n \in \mathbb{N}\}$ does not cause $\{X_n : n \in \mathbb{N}\}$. Note also that [57] refers to **MC4** as the "weak union" property of conditional independence. For further elaboration on this issue see [62].

In the next theorem, we utilize Lemma 2.21, specifically the fact that **MC4** is equivalent to **MC2** and **MC1**, to show that the extremum of the nonanticipatory ε-entropy (2.3.50), $R^{\varepsilon}_{0,n}(D)$, is equivalent to the extremum of nonanticipative RDF, $R^{na}_{0,n}(D)$.

Theorem 2.22. (*Equivalence of* $R^{na}_{0,n}(D)$ *and* $R^{\varepsilon}_{0,n}(D)$)
Definition 2.17 and Definition 2.20 are equivalent, i.e.,

$$R^{na}_{0,n}(D) = R^{\varepsilon}_{0,n}(D)$$

Proof. By the definition of nonanticipatory ε-entropy $R^{\varepsilon}_{0,n}(D)$, the infimum is taken over the set $Q^{\varepsilon}_{0,n}(D) \stackrel{\triangle}{=} Q_{0,n}(D) \bigcap \{X^n_{i+1} \leftrightarrow X^i \leftrightarrow Y^i,\ i = 0, \ldots, n-1\}$. Using Lemma 2.21 we deduce that the set $Q^{\varepsilon}_{0,n}(D)$ is equivalent to

$$Q_{0,n}(D) \bigcap \{P_{Y^n|X^n} :\ P_{Y^n|X^n}(\cdot|x^n) = \overrightarrow{P}_{Y^n|X^n}(\cdot|x^n)\}$$

Moreover, for any $P_{Y^n|X^n} = \overrightarrow{P}_{Y^n|X^n}$, the mutual information between X^n and Y^n is given by $I(X^n;Y^n) \equiv \mathbb{I}_{P_{X^n}}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n})$.
Since $\{P_{Y_i|Y^{i-1},X^i}(\cdot|y^{n-1},x^n) :\ i = 0, \ldots, n\}$ uniquely defines $\overrightarrow{P}_{Y^n|X^n}(\cdot|x^n)$ and vice-versa, then

$$\begin{aligned} R^{\varepsilon}_{0,n}(D) &= \inf_{\overrightarrow{P}_{Y^n|X^n}(\cdot|x^n): \frac{1}{n+1}\int d_{0,n}(x^n,y^n)\overrightarrow{P}_{Y^n|X^n}(dy^n|x^n)\otimes P_{X^n}(dx^n) \leq D} \mathbb{I}_{X^n \rightarrow Y^n}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n}) \\ &\equiv R^{na}_{0,n}(D) \end{aligned}$$

This completes the derivation. □

Next, we show that Gorbunov and Pinsker's definition although, as stated in [34, I. Introduction], is motivated by real-time applications, such as, "control-related problems," the MC condition $X^n_{i+1} \leftrightarrow X^i \leftrightarrow Y^i,\ i = 0, 1, \ldots, n-1$, imposed in the Definition 2.20 of $R^{\varepsilon}_{0,n}(D)$, rules out any applications to control systems. On the other hand, we show how $R^{na}_{0,n}(D)$ can be generalized to handle control applications.

Christos K. Kourtellaris

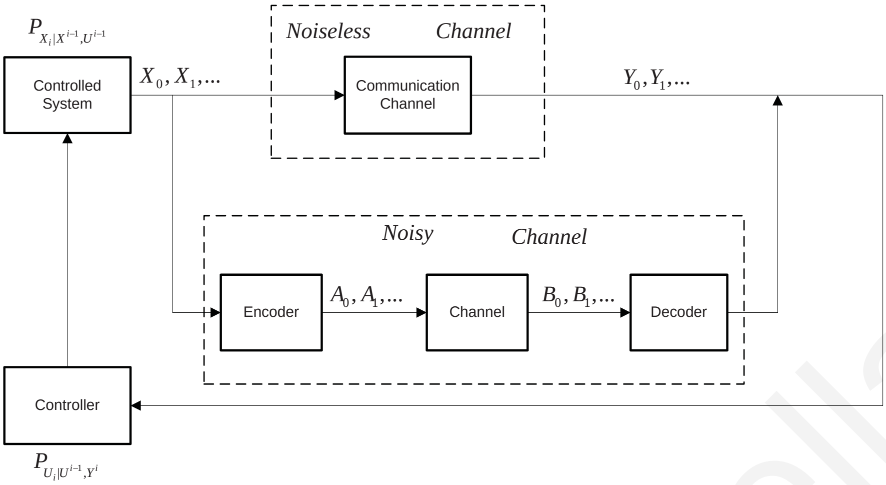

FIGURE 2.3.4: Communication for real-time control processes.

Remark 2.23. Consider the typical block diagram of real-time communication for control over a finite rate channel (which can be noiseless or noisy) illustrated in Figure 2.3.4. Here, $\{X_i: \ i=0,1,\ldots\}$ is the controlled process specified by the conditional distributions $\{P_{X_i|X^{i-1},U^{i-1}}(dx_i|x^{i-1},\upsilon^{i-1}): \ i=0,1,\ldots\}$, $\{U_i: \ i=0,1,\ldots\}$ is the control process specified by the conditional distribution $\{P_{U_i|U^{i-1},Y^i}(d\upsilon_i|\upsilon^{i-1},y^i): \ i=0,1,\ldots\}$, which receives information from a finite rate channel (noisy or noiseless), as shown in Figure 2.3.4.

This is a typical control system subject to rate constraint analysed in [9, 11, 26, 55, 76]. In Figure 2.3.4 the control laws or strategies are randomized (conditional distributions). One may consider regular strategies, i.e., strategies which are measurable functions by letting $\{P_{U_i|U^{i-1},Y^i}(d\upsilon_i|\upsilon^{i-1},y^i): \ i=0,1,\ldots\}$ to be delta measures concentrated at $\{\upsilon_i = \mu_i(\upsilon^{i-1},y^i): \ i=0,1,\ldots\}$. A typical example is the linear controlled system $X_{i+1}=AX_i+BU_i+N_i$, $X_0=x$, where $\{N_i: \ i=0,1,\ldots\}$ is an IID process. In either case, since the control laws or strategies take as inputs previous controls $U^{i-1}=\upsilon^{i-1}$, and past and present reproduction $Y^i=y^i$, if we impose Gorbunov and Pinsker's MC $X_{n+1}^{\infty} \leftrightarrow X^n \leftrightarrow Y^n$, then this MC rules out the dependence of the controlled process distribution $\{P_{X_i|X^{i-1},U^{i-1}}(dx_i|x^{i-1},\upsilon^{i-1}): \ i=0,1,\ldots\}$, on $\{U_i: \ i=0,1,\ldots\}$. On the other hand, given the control strategies, since $P_{X_i|X^{i-1},U^{i-1}} \equiv P_{X_i|X^{i-1},Y^{i-1}}$, then the information nonanticipative RDF is easily extended by

Christos K. Kourtellaris

considering the information measure

$$\mathbb{I}_{X^n\rightarrow Y^n}(P_{X_i|X^{i-1},Y^{i-1}},P_{Y_i|Y^{i-1},X^i}:\ i=0,1,\ldots,n)$$
$$\stackrel{\triangle}{=}\int_{\mathscr{X}_{0,n}\times\mathscr{Y}_{0,n}}\log\Big(\frac{\overleftarrow{P}_{X^n|Y^{n-1}}(dx^n|y^{n-1})\otimes\overrightarrow{P}_{Y^n|X^n}(dy^n|x^n)}{\overleftarrow{P}_{X^n|Y^{n-1}}(dx^n|y^{n-1})\otimes P_{Y^n}(dy^n)}\Big)(\overleftarrow{P}_{X^n|Y^{n-1}}\otimes\overrightarrow{P}_{Y^n|X^n})(dx^n,dy^n) \tag{2.3.53}$$

where $\overleftarrow{P}_{X^n|Y^{n-1}}(dx^n|y^{n-1})\stackrel{\triangle}{=}\otimes_{i=0}^{n}P_{X_i|X^{i-1},Y^{i-1}}(dx_i|x^{i-1},y^{i-1})$, and $P_{Y^n}(\cdot)$ is the marginal of the joint distribution on $\mathscr{Y}_{0,n}$. Clearly, the information measure (2.3.53) is the directed information from X^n to Y^n [53].

For such control system applications, we can define as a measure of performance (distortion function) for control and compression, a generalized distortion function which includes the cost of control and decompression, such as,

$$d_{0,n}(x^n,y^n,u^n)\stackrel{\triangle}{=}\sum_{i=0}^{n}\{||\upsilon_i||^2+||x_i-y_i||^2\} \tag{2.3.54}$$

and then define the corresponding average fidelity set. An interesting problem with practical implication is to minimize over the reproduction distribution and control policies the rate subject to the fidelity defined by (2.3.54), namely,

$$\frac{1}{n+1}\inf_{\mu_i(\upsilon^{i-1},y^i):\ i=0,1,\ldots,n-1}\ \inf_{\overrightarrow{P}_{Y^n|X^n}\in Q_{0,n}^{na}(D)}\mathbb{I}_{X^n\rightarrow Y^n}(\overleftarrow{P}_{X^n|Y^{n-1}},\overrightarrow{P}_{Y^n|X^n})$$

and its limit. This demonstrates our preference in information nonanticipative RDF, $R_{0,n}^{na}(D)$, over $R_{0,n}^{\varepsilon}(D)$.

Clearly, Theorem 2.22 states that information nonanticipative RDF which is a special case of directed information from X^n to Y^n, is equivalent to Gorbunov and Pinsker's nonanticipatory ε-entropy defined via mutual information and Lemma 2.21, **MC4**.

2.3.2 Noisy Coding Theorem and Zero-Delay Codes

The achievability of nonanticipative RDF, can be shown via a noisy coding theorem using delayless codes, by relating the channel capacity to the average distortion obtained from the nonanticipative RDF. Before giving the main results, we state certain data processing inequalities, relating mutual and directed information, which are necessary conditions for

Christos K. Kourtellaris

reliable communications.

Theorem 2.24. *(Data Processing inequalities)*
Consider the basic block diagram of information transmission illustrated in Figure 2.1.1. Then the following hold.

1. *Suppose $X^t \leftrightarrow (A^i, B^{i-1}) \leftrightarrow B_i$ forms a Markov chain for $i = 0, 1, \ldots, n$, $t \leq n$, then*

$$I(X^t; B^n) \leq I(A^n \rightarrow B^n) \tag{2.3.55}$$

2. *Suppose $Y^t \leftrightarrow (X^{i-1}, B^n) \leftrightarrow X_i$ forms a Markov chain for $i = 0, 1, \ldots, t$, $t \leq n$, then*

$$I(X^t; Y^t) \leq I(X^t; B^n) \tag{2.3.56}$$

3. *If the conditions of statements 1), 2) hold, then*

$$I(X^t \rightarrow Y^t) \leq I(X^t; Y^t) \leq I(A^n \rightarrow B^n) \leq I(A^n; B^n), \quad t \leq n \tag{2.3.57}$$

Proof. 1. By the identity of mutual information we have

$$\begin{aligned} I(X^t; B^n) &= H(B^n) - H(B^n|X^t) \\ &= H(B^n) - \sum_{i=0}^{n} H(B_i|X^t, B^{i-1}) \\ &\overset{(\alpha)}{\leq} H(B^n) - \sum_{i=0}^{n} H(B_i|X^t, B^{i-1}, A^i) \\ &\overset{(\beta)}{=} H(B^n) - \sum_{i=0}^{n} H(B_i|B^{i-1}, A^i) \\ &= I(A^n \rightarrow B^n), \ \ \forall t \leq n \end{aligned} \tag{2.3.58}$$

where (α) holds because conditioning does not increase entropy, and (β) follows from the Markov chain.

Christos K. Kourtellaris

2. Similarly,

$$\begin{aligned} I(X^t;Y^t) &= H(X^t)-\sum_{i=0}^{t}H(X_i|X^{i-1},Y^t) \\ &\leq H(X^t)-\sum_{i=0}^{t}H(X_i|X^{i-1},Y^t,B^n) \\ &\overset{(\gamma)}{=} H(X^t)-\sum_{i=0}^{t}H(X_i|X^{i-1},B^n) \\ &= I(X^t;B^n),\ \ \forall t\leq n \end{aligned} \tag{2.3.59}$$

where equality (γ) follows from the Markov chain.

3. The lower bound follows from the fact that $I(X^t;Y^t)=I(X^t\rightarrow Y^t)+I(X^t\leftarrow Y^t)\geq I(X^t\rightarrow Y^t)$, which holds with equality hold if and only $Y^i\leftrightarrow X^i\leftrightarrow X_{i+1}, i=0,1,\ldots,t$ is a Markov chain or equivalently $Y_i\leftrightarrow(X^i,Y^{i-1})\leftrightarrow X_{i+1}^t$ is a Markov chain for $i=0,1,\ldots,t-1$. The upper bound is obtained by (2.3.56) and (2.3.55) □

Therefore, given any communication channel with feedback $\{P_{B_i|B^{i-1},A^i}(db_i|b^{i-1},a^i):\ i=0,1\ldots,n\}$ with a pre-encoder and a post-decoder connected to it, as in Figure 2.3.5, the following theorem is established.

Theorem 2.25. *(Nonanticipative Data Drocessing Inequalities)*
Suppose the following Markov chains hold.

$$\begin{aligned} &X^i\leftrightarrow(A^i,B^{i-1})\leftrightarrow B_i,\ i=0,1,\ldots,n \\ &Y^n\leftrightarrow(X^{i-1},B^i)\leftrightarrow X_i,\ i=0,1,\ldots,n \end{aligned}$$

A necessary condition to achieve end-to-end causal information RDF over the channel $\{P_{B_i|B^{i-1},A^i}(db_i|b^{i-1},a^i):\ i=0,1\ldots,n\}$ *is*

$$R^{na}_{0,n}(D)\leq I(A^n\rightarrow B^n),\qquad \forall\, n\in\mathbb{N} \tag{2.3.60}$$

for all channels $\{P_{B_i|B^{i-1},A^i}(db_i|b^{i-1},a^i)\ :\ i\ =\ 0,1,\ldots,n\}$ *and encoders* $\{P_{A_i|A^{i-1},X^i}(da_i|a^{i-1},x^i):i=0,1,\ldots,n\}$.

Christos K. Kourtellaris

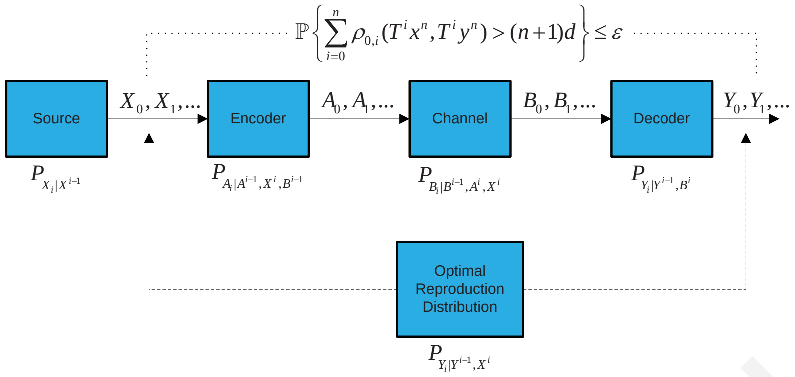
$\mathbb{P}\left\{\sum_{i=0}^{n}\rho_{0,i}(T^ix^n,T^iy^n)>(n+1)d\right\}\leq\varepsilon$
Source
$X_0, X_1, \ldots$
Encoder
$A_0, A_1, \ldots$
Channel
$B_0, B_1, \ldots$
Decoder
$Y_0, Y_1, \ldots$
$P_{X_i|X^{i-1}}$
$P_{A_i|A^{i-1},X^i,B^{i-1}}$
$P_{B_i|B^{i-1},A^i,X^i}$
$P_{Y_i|Y^{i-1},B^i}$
Optimal Reproduction Distribution
$P_{Y_i|Y^{i-1},X^i}$

FIGURE 2.3.5: JSCC based on nonanticipative transmission.

Proof. Similarly to Theorem 2.24, we can show that $I(X^n;Y^n)\leq I(A^n\rightarrow B^n)$. Given any source with distribution $P_{X^n}(x^n)$, a channel $\{P_{B_i|B^{i-1},A^i}(db_i|b^{i-1},a^i):i=0,1,\ldots,n\}$ an encoder $\{P_{A_i|A^{i-1},X^i}(da_i|a^{i-1},x^i):i=0,1,\ldots,n\}$ and a decoder with average distortion $\frac{1}{n+1}\mathbb{E}_{P_{X^n,Y^n}}\{d_{0,n}(X^n,Y^n)\}\leq D$, then by taking the infimum over $P_{Y^n|X^n}$ satisfying the average distortion constraint and the MC: $X_{i+1}^n\leftrightarrow(X^i,Y^{i-1})\leftrightarrow Y_i:i=0,1,\ldots,n-1$, yields (2.3.60). □

We proceed by establishing an operational meaning for the information nonanticipative RDF for sources with memory based on a noisy coding theorem. To this end, we define JSCC with emphasis on nonanticipative coding i.e., the encoder and decoder at each time instant i process samples independently, with memory on past symbols, and without anticipation with respect to symbols occurring at times $j>i$.

We also show that even in the unmatched case, uncoded nonanticipative transmission of sources with memory has an operational meaning, in the sense that the excess distortion probability can be made arbitrarily small, based only on the properties of the information nonanticipative RDF. Figure 2.3.5, describes the block diagram of JSCC using nonanticipative transmission. We assume that the cost of transmitting symbols over the channel is a measurable function

$$c_{0,n}:\mathscr{A}_{0,n}\times\mathscr{B}_{0,n-1}\mapsto[0,\infty),\quad c_{0,n}(a^n,b^{n-1})\triangleq\sum_{i=0}^{n}\gamma(T^ia^n,T^ib^{n-1})\qquad(2.3.61)$$

where T^ib^{n-1} is a measurable function of $\{b_0,b_1,\ldots,b^{i-1}\}$. We use the following definition of a nonanticipative code.

Christos K Kourtellaris

Definition 2.26. (Nonanticipative code)
An (n,d,ε,P) nonanticipative code is a tuple

$$\Big(\mathscr{X}_{0,n},\mathscr{A}_{0,n},\mathscr{B}_{0,n},\mathscr{Y}_{0,n},P_{X^n},\overrightarrow{P}_{A^n|B^{n-1},X^n},\overrightarrow{P}_{B^n|A^n,X^n},\overrightarrow{P}_{Y^n|B^n},d_{0,n},c_{0,n}\Big)$$

where $\{P_{A_i|A^{i-1},B^{i-1},X^i}(\cdot|\cdot,\cdot,\cdot):\forall i\in\mathbb{N}^n\}$, $\{P_{Y_i|Y^{i-1},B^i}\ (\cdot|\cdot,\cdot):\forall i\in\mathbb{N}^n\}$ is the code, $\{P_{B_i|B^{i-1},A^i,X^i}(\cdot|\cdot,\cdot,\cdot):\forall i\in\mathbb{N}^n\}$ is the channel, with excess distortion probability

$$\mathbb{P}\Big\{d_{0,n}(X^n,Y^n)>(n+1)d\Big\}\leq\varepsilon,\quad \varepsilon\in(0,1),\ d\geq 0$$

and transmission cost

$$\frac{1}{n+1}\mathbb{E}\Big\{c_{0,n}(A^n,B^{n-1})\Big\}\leq P,\quad P\geq 0$$

where $\mathbb{P}$ is taken with respect to the joint distribution induced by source-encoder-channel-decoder $P_{X^n,A^n,B^n,Y^n}(dx^n,da^n,db^n,dy^n)$.

An uncoded nonanticipative code, denoted by (n,d,ε), is a subset of an (n,d,ε,P) nonanticipative code in which an encoder and decoder are identity maps, $P_{A_i|A^{i-1},B^{i-1},X^i}\ (da_i|a^{i-1},b^{i-1},x^i)=\delta_{X_i}(da_i)$, $P_{Y_i|Y^{i-1},B^i}(dy_i|y^{i-1},b^i)=\delta_{B_i}(dy_i)$, that is, $A_i=X_i$, $Y_i=B_i$, $i=0,1,\ldots,n$, and the channel $P_{B_i|B^{i-1},A^i}(\cdot|\cdot,\cdot)$ is used without feedback and power constraint.

Next, we define the minimum excess distortion as follows.

Definition 2.27. (Minimum Excess Distortion)
The minimum excess distortion achievable by a nonanticipative code with memory without anticipation (n,d,ε,P) is defined by

$$D^o(n,\varepsilon,P)\overset{\triangle}{=}\inf\Big\{d:\exists(n,d,\varepsilon,P)\ \text{ nonanticipative code}\Big\}\tag{2.3.62}$$

For the uncoded nonanticipative code, (2.3.62) is replaced by

$$\bar{D}^o(n,\varepsilon)\overset{\triangle}{=}\inf\big\{d:\exists(n,d,\varepsilon)\text{ nonanticipative code}\big\}\tag{2.3.63}$$

Christos K. Kourtellaris

Note that in our definition of nonanticipative code (n,d,ε,P) we have assumed indirectly that the finite time information capacity is defined by

$$C_{0,n}(P) \triangleq \sup_{\{P_{A_i|A^{i-1},B^{i-1}}(a_i|a^{i-1},b^{i-1}):\ i=0,1,\ldots,n\}\in\mathscr{P}_{0,n}(P)} \frac{1}{n+1} I(A^n \rightarrow B^n) \tag{2.3.64}$$

where the average power constraint is

$$\mathscr{P}_{0,n}(P) \triangleq \Big\{\{P_{A_i|A^{i-1},B^{i-1}}(a_i|a^{i-1},b^{i-1}):\ i=0,1,\ldots,n\} : \frac{1}{n+1}\mathbb{E}\{c_{0,n}(A^n,B^{n-1})\} \leq P\Big\} \tag{2.3.65}$$

in which $I(A^n \rightarrow B^n)$ is the directed information from A^n to B^n defined by

$$I(A^n \rightarrow B^n) \triangleq \sum_{i=0}^{n} I(A^i;B_i|B^{i-1}) \tag{2.3.66}$$

The information channel capacity is given by

$$C(P) = \lim_{n \longrightarrow \infty} \frac{1}{n+1} C_{0,n}(P) \tag{2.3.67}$$

Thus, we have assumed the supremum (2.3.64) is finite and the limit exists.

Since at a first glance, the probabilistic realization of the optimal nonanticipative reproduction distribution of the information nonanticipative RDF by an encoder-channel-decoder is necessary for probabilistic matching of the source to the channel, we introduce the following definition of realization.

Definition 2.28. (Realization of the nonanticipative RDF)
Given a source $\{P_{X_i|X^{i-1}}\ (dx_i|x^{i-1}) : \forall i \in \mathbb{N}^n\}$, a channel $\{P_{B_i|B^{i-1},A^i,X^i}\ (db_i|b^{i-1},a^i,x^i) : \forall i \in \mathbb{N}^n\}$ is a realization of the optimal reproduction distribution $\{P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) : \forall i \in \mathbb{N}^n\}$, if there exists a pre-channel encoder $\{P_{A_i|A^{i-1},B^{i-1},X^i}\ (da_i|a^{i-1},b^{i-1},x^i) : \forall i \in \mathbb{N}^n\}$ and a post-channel decoder $\{P_{Y_i|Y^{i-1},B^i}\ (dy_i|y^{i-1},b^i) : \forall i \in \mathbb{N}^n\}$ such that

$$\overrightarrow{P}^*_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^{n} P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) = \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) \tag{2.3.68}$$

Christos K. Kourtellaris

where the joint distribution from which (2.3.68) is obtained is precisely

$$P_{X^n,A^n,B^n,Y^n}(dx^n,da^n,db^n,dy^n) = \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},B^i}(dy_i|y^{i-1},b^i) \otimes P_{B_i|B^{i-1},A^i,X^i}(db_i|b^{i-1},a^i,x^i)$$
$$\otimes P_{A_i|A^{i-1},B^{i-1},X^i}(da_i|a^{i-1},b^{i-1},x^i) \otimes P_{X_i|X^{i-1}}(dx_i|x^{i-1})$$

Moreover, we say that $R^{na}(D)$ is realizable if in addition the realization operates with average distortion D and $\lim_{n\longrightarrow\infty}\frac{1}{n+1}\mathbb{I}_{X^n;Y^n}(P_{X^n},\overrightarrow{P}^{*}_{Y^n|X^n}) = R^{na}(D) \overset{\triangle}{=} \lim_{n\longrightarrow\infty}\frac{1}{n+1}R^{na}_{0,n}(D)$.

Using the above definition of realization we now prove achievability of nonanticipative code for sources with memory.

Theorem 2.29. *(Achievability of a nonanticipative Code with Memory Without Anticipation) Suppose the following conditions hold.*

1. *$R^{na}_{0,n}(D)$ has a solution and the optimal reproduction distribution converges to a stationary distribution corresponding to $R^{na}(D)$.*
2. *The encoder and the decoder are unitary maps (no coding), and the channel $P_{B_i|B^{i-1},A^i}$ corresponds to $P_{Y_i|Y^{i-1},X^i}$ (i.e., $A_i = X_i$, $Y_i = B_i$), $i = 0,1,\ldots,n$.*
3. *For a given $D \in [D_{min}, D_{max}]$, $R^{na}(D)$ is finite, and $\lim_{n\rightarrow\infty}\frac{1}{n+1}I(A^n \rightarrow B^n)$ is finite.*

If

$$\mathbb{P}\Big\{\sum_{i=0}^{n}\rho_{0,i}(T^iX^n,T^iY^n) > (n+1)d\Big\} \leq \varepsilon \tag{2.3.69}$$

where $\mathbb{P}$ is taken with respect to $P_{Y^n,X^n}(dy^n,dx^n) = \overrightarrow{P}^{*}_{Y^n|X^n}(dy^n|x^n)\otimes P_{X^n}(dx^n)$, then there exists an uncoded (n,d,ε) nonanticipative code.

Proof. By conditions 1., 2., 3. and the data processing inequality we know that $R^{na}(D) \leq \lim_{n\rightarrow\infty}\frac{1}{n+1}I(A^n \rightarrow B^n) < \infty$. Hence, if (2.3.69) holds, there exists an uncoded (n,d,ε) SbS code. □

Next, we describe several consequences of Theorem 2.29.

Remark 2.30.

1. The method described in Theorem 2.29 is simple; find the optimal reproduction distribution of $R^{na}(D)$, then use this distribution as the channel and ensure that (2.3.69)

Christos K. Kourtellaris

holds, which implies achievability. The only disadvantage is the loss of resources, because in general the channel will have higher capacity than the value of $R^{na}(D)$. Ideally one would like to ensure JSCC so that the channel operates at the supremum of all achievable rates and hence $R^{na}(D)$ is the minimum rate of reproducing source messages at the decoder.

2. In Chapter 4, we will revisit Theorem 2.29 to address the optimal JSCC problem, in which $R^{na}(D) = C(P)$, by designing the encoder, decoder for a specific channel with memory so that matching of the source and the channel is made feasible, often requiring transmission cost constraint imposed on the channel, to reduce the capacity to that of $R^{na}(D)$.

2.4 Optimal Stationary Solution of the Nonaticipative RDF

The goal of this section is to to derive the optimal causal reproduction distribution to characterize the solution of the nonanticipative RDF, for the stationary case.

Given the source $\{P_{X_i|X^{i-1}}(dx_i|x^{i-1}) : i = 0, 1, \ldots, n\}$, and a causal reproduction distribution $\{P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1}, x^i) : i = 0, 1, \ldots, n\}$, the joint measure $P_{X^n,Y^n}(dx^n, dy^n)$ and the marginal measures, $P_{Y^n}(dy^n)$, $P_{X^n}(dx^n)$ are uniquely defined. Hence, the directed information from X^n to Y^n is also defined via

$$\mathbb{I}_{X^n \to Y^n}(X^n \to Y^n) = \sum_{i=0}^{n} \int_{\mathscr{X}_{0,i} \times \mathscr{Y}_{0,i}} \log \Big(\frac{P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1}, x^i)}{P_{Y_i|Y^{i-1}}(dy_i|y^{i-1})} \Big) P_{X^i,Y^i}(dx^i, dy^i) \tag{2.4.70}$$

where

$$P_{X^n,Y^n}(dx^n, dy^n) = \otimes_{i=0}^{n} \Big(P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1}, x^i) \otimes P_{X_i|X^{i-1},Y^{i-1}}(dx_i|x^{i-1}) \Big) \tag{2.4.71}$$

The solution of the nonanticipative RDF can be made precise by first identifying the appropriate spaces on which existence of solution to $R^{na}_{0,n}(D)$ is sought, and equivalence between the constrained and unconstrained problems is shown. This is done in [8] using the weak* convergence topologies, and in [72, 73] using weak topology.

Assumption 2.31. Appropriate conditions are assumed (i.e., [8]) so that an optimal solution exists and the constrained problem $R^{na}_{0,n}(D)$ is equivalent to the unconstrained problem, that

Christos K. Kourtellaris

is,

$$R^{na}_{0,n}(D) = \sup_{s\leq 0} \inf_{\{P_{Y_j|Y^{j-1},X^j}(dy_j|y^{j-1},x^j): j=0,1,\ldots,n\}} \Big\{ \mathbb{I}_{X^n\rightarrow Y^n}(X^n \rightarrow Y^n) \\ -s\Big(\sum_{j=0}^{n} \int_{\mathscr{X}_{0,j}} \int_{\mathscr{Y}_{0,j}} \rho_{0,j}(T^n x^j, T^n y^j) P_{X^j,Y^j}(dx^j,dy^j) - D(n+1)\Big)\Big\} \tag{2.4.72}$$

where $s \leq 0$ is the Lagrangian multiplier and the solution is stationary.

Next, we provide the optimal stationary solution for the nonanticipative RDF, $R^{na}_{0,n}(D)$. An alternative derivation based on Gateaux differential is found in [8, 73], for the case when the source is not affected by past reproduction symbols.

Next, we provide the optimal solution for the nonanticipative RDF.

Theorem 2.32. *Suppose Assumption 2.31 holds.*
The optimal (stationary) reproduction distribution which achieves the infimum, assuming it exists, of the rate distortion function, $R^{na}_{0,n}(D)$, is given by

$$P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) = \frac{e^{s\rho_{0,i}(T^i x^n, T^i y^n)} P^*_{Y_i|Y^{i-1}}(dy_i|y^{i-1})}{\int_{\mathscr{Y}_i} e^{s\rho_{0,i}(T^i x^n, T^i y^n)} P^*_{Y_i|Y^{i-1}}(dy_i|y^{i-1})}, \ i = 0,1,\ldots,n \tag{2.4.73}$$

where $s \leq 0$ and denotes the optimal Lagrange multiplier in (2.4.72), and it is the solution of $s = \frac{d}{dD} R^{na}_{0,n}(D)$.
The information nonanticipative RDF, $R^{na}_{0,n}(D)$, is given by

$$R^{na}_{0,n}(D) = sD(n+1) - \sum_{j=0}^{n} \int_{\mathscr{X}_{0,j}\times\mathscr{Y}_{0,j-1}} \log\Big(\int_{\mathscr{Y}_j} e^{s\rho_{0,j}(T^j x^n, T^j y^n)} P^*_{Y_j|Y^{j-1}}(dy_j|y^{j-1})\Big) \\ P_{X_j|X^{j-1}}(dx_j|x^{j-1}) \otimes P^*_{X^{j-1},Y^{j-1}}(dx^{j-1},dy^{j-1}) \tag{2.4.74}$$

where

$$P^*_{X^{j-1},Y^{j-1}}(dx^{j-1},dy^{j-1}) \triangleq \otimes_{i=0}^{j-1} \Big(P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) \otimes P_{X_i|X^{i-1}}(dx_i|x^{i-1})\Big) \tag{2.4.75}$$

Proof. See [73].

Theorem 2.32 treats the stationary case. The nonstationary case is much more involved and is given in [73].

Christos K. Kourtellaris

Remark 2.33.

1. The optimal stationary reproduction distribution (2.4.73) is causal, hence decoding can be done without waiting to receive the entire sequence x^n before the symbol $y_i, i \leq n$ is reconstructed.

2. From (2.4.73), we deduce that if $\rho_{0,i}(T^i x^n, T^i y^n) = \rho(x_i, y_i)$, then

$$P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1}x^i) = P^*_{Y_i|Y^{i-1},X_i}(dy_i|y^{i-1},x_i) - a.a.(y^{i-1},x^i),\ i = 0,1,\ldots,n$$

   Hence, from (2.4.73) we obtain

$$\begin{aligned} P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) &= P^*_{Y_i|Y^{i-1},X_i}(dy_i|y^{i-1},x_i) \\ &= \frac{e^{s\rho(x_i,y_i)} \int_{\mathscr{X}_i} P^*_{Y_i|Y^{i-1},X_i}(dy_i|y^{i-1},x_i) P_{X_i|Y^{i-1}}(dx_i|y^{i-1})}{\int_{\mathscr{Y}_i} \int_{\mathscr{X}_i} e^{s\rho(x_i,y_i)} P^*_{Y_i|Y^{i-1},X_i}(dy_i|y^{i-1},x_i) P_{X_i|Y^{i-1}}(dx_i|y^{i-1})} \end{aligned} \tag{2.4.76}$$

   where (2.4.76) is obtained by reconditioning. Similarly for other cases.

   However, in general we do not know the length of the sequence $Y^{i-1} \in \mathscr{Y}_{0,i-1}$ on which the optimal reproduction distribution depends on. Properties of the solution are derived for complete separable metric spaces in [73], and they are often very important when solving specific examples like the Gaussian multidimensional process.

3. The optimal reproduction conditional distribution (2.4.73) is implicit because its right hand side term depends on its left side, therefore one has to show existence and possibly uniqueness via fixed point theorems.

2.5 Nonanticipative RDF of a Binary Symmetric Markov Source BSMS(p)

In this section, we compute the optimal reproduction distribution of the information nonanticipative RDF and rate $R^{na}(D)$ for a finite alphabet source with memory, the BSMS(p). The classical RDF for the BSMS(p) is only known for the distortion region $0 \leq D \leq D_c$ [37], while for the remainder of the distortion region only bounds are known [5]. We additionally,

Christos K. Kourtellaris

compare these bounds to the one we proposed based on $R^{na}_{0,n}(D)$ and compute the rate loss of causal codes with respect to noncausal codes, by using the fact that this rate loss is at most $R^{na}(D)-R(D)$ bits/sample, for the region where $R(D)$ is computable. The achievability of the nonanticipative RDF for the BSMS(p) based on nonanticipative transmission is addressed in Section 4.3.4, Chapter 4.

Consider a BSMS(p), with stationary transition probabilities $P_{X_i|X_{i-1}}(x_i=0|x_{i-1}=0)=P_{X_i|X_{i-1}}(x_i=1|x_{i-1}=1)=1-p$ and $P_{X_i|X_{i-1}}(x_i=1|x_{i-1}=0)=P_{X_i|X_{i-1}}(x_i=0|x_{i-1}=1)=p$ and $i\in 0,1,\ldots$. We consider single letter Hamming distortion criterion $\rho(x,y)=0$ if $x=y$ and $\rho(x,y)=1$ if $x\neq y$. The transition probabilities are illustrated via row stochastic matrices.

Theorem 2.34. *The nonanticipative RDF $R^{na}(D)$ for BSMS(p) and single letter Hamming distortion function is given by*

$$R^{na}(D)=\begin{cases} H(m)-H(D) & \text{if } D\leq \frac{1}{2}\\ 0 & \text{otherwise}\end{cases}$$

where $m=1-p-D+2pD$.

Proof. First, we compute the steady state distribution of the source. Since the transition matrix of the BSMS(p) is doubly stochastic with alphabet cardinality 2, the transition probabilities is given by $P_X(0)=P_X(1)=0.5$. The stationary reproduction distribution obtained from Theorem 2.32, is Markov with respect to the source and it is given by

$$\begin{aligned} P^*_{Y_i|Y^{i-1},X^i}(y_i|y^{i-1},x^i) &= P^*_{Y_i|Y^{i-1},X_i}(d_i|y^{i-1},x_i)\\ &= \frac{e^{s\rho(x_i,y_i)}P_{Y_i|Y^{i-1}}(y_i|y^{i-1})}{\sum_{y_i\in\{0,1\}}e^{s\rho(x_i,y_i)}P_{Y_i|Y^{i-1}}(y_i|y^{i-1})},\ i=0,1,\ldots \end{aligned}\tag{2.5.77}$$

For $i=0,1,\ldots$, we calculate $P^*_{Y_i|Y^{i-1}}(dy_i|y^{i-1})$ by reconditioning it on X_i and then substitute it into the RHS of (2.5.77) and solve the systems of equations. Since we do not how much the reproduction distribution depends from the past reproduction symbols Y^{i-1}, we start the iterations from $i=0$. The reconstruction distributions is given by

$$P_{Y_0|X_0}(y_0|x_0)=\frac{e^{sd(x_0,y_0)}P_{Y_0}(y_0)}{\sum_{y_0\in\{0,1\}}e^{sd(x_0,y_0)}P_{Y_0}(y_0)}\tag{2.5.78}$$

Christos K Kourtellaris

where $P_{Y_0}(y_0) = \sum_{X_0} P_{Y_0|X_0}(y_0|x_0)P_{X_0}(x_0)$. Solving the systems of equations yield the following results.

$$P_{Y_0|X_0}(y_0|x_0) = \begin{bmatrix} \frac{1}{1+e^s} & \frac{e^s}{1+e^s} \\ \frac{e^s}{1+e^s} & \frac{1}{1+e^s} \end{bmatrix}$$

while the distribution of the output symbol, at each time instance 0 is IID, and it is given by $P_{Y_0}(y_0) = 0.5, \ y_0 \in \{0,1\}$. Next, we calculate $P_{X_1|Y_0}(x_1|y_0)$, which is necessary in the subsequent iteration, in order to calculate $P_{Y_1|Y_0}(y_1|y_0)$. This is done by reconditioning on X_0.

$$\begin{aligned} P_{X_1|Y_0}(x_1|y_0) &= \sum_{x_0\in\{0,1\}} P_{X_1|Y_0,X_0}(x_1|y_0,x_0)P_{X_0|Y_0}(x_0|y_0) \\ &= \sum_{x_0\in\{0,1\}} P_{X_1|Y_0,X_0}(x_1|y_0,x_0)\frac{P_{Y_0|X_0}(y_0|x_0)P_{X_0}(x_0)}{P_{Y_0}(y_0)} \\ &= \sum_{x_0\in\{0,1\}} P_{X_1|Y_0,X_0}(x_1|y_0,x_0)P_{Y_0|X_0}(y_0|x_0) \\ &\overset{(\alpha)}{=} \sum_{x_0\in\{0,1\}} P_{X_1|X_0}(x_1|x_0)P_{Y_0|X_0}(y_0|x_0) \end{aligned} \tag{2.5.79}$$

where equation (α) holds due to Lemma.2.21.

Then, we proceed to the iteration ($i = 1$), by calculating the conditional probability $P_{Y_1|Y_0}(y_1|y_0) = \sum_{x_0\in\{0,1\}} P_{Y_1|X_0,Y_0}(y_1|x_0,y_0)P_{X_0,Y_0}(x_0|y_0)$, replacing it into

$$P_{Y_1|X_1,Y_0}(y_1|x_1,y_0) = \frac{e^{sd(x_1,y_1)}P_{Y_1|Y_0}(y_1|y_0)}{\sum_{y_1\in\{0,1\}} e^{sd(x_1,y_1)}P_{Y_1|Y_0}(y_1|y_0)}$$

and solve the resulting systems. This procedure yields the following reproduction distribution.

$$P_{Y_1|X_1,Y_0}(y_1|x_1,y_0) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cccc} 0,0 & 0,1 & 1,0 & 1,1 \end{array} \\ \begin{bmatrix} \frac{p}{p+e^s(1-p)} & \frac{1-p}{1-p(1-e^s)} & \frac{e^s p}{1-p(1-e^s)} & \frac{e^s(1-p)}{p+e^s(1-p)} \\ \frac{e^s(1-p)}{p+e^s(1-p)} & \frac{e^s p}{1-p(1-e^s)} & \frac{1-p}{1-p(1-e^s)} & \frac{p}{p+e^s(1-p)} \end{bmatrix} \end{array}$$

Christos K. Kourtellaris

For iteration, $i=2$, we verify that $P_{Y_2|X_2,Y_1,Y_0}=P_{Y_1|X_1,Y_0}$, and that the optimal reproduction distribution is characterized by $P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)=P^*_{Y_i|Y^{i-1},X_i}(dy_i|y^{i-1},x_i)=P^*_{Y_1|X_1,Y_0}(dy_1|x_1,y_0)$, for all $i\geq 2$. The Lagrange multiplier s is found from fidelity constraint as follows.

$$\mathbb{E}\Big\{\rho(x_i,y_i)\Big\}=\frac{e^s}{1+e^s}=D \implies e^s=\frac{D}{1-D},\; D\leq 0.5$$

By substituting the Lagrangian multiplier s, we obtain the expression for the optimal stationary reproduction distribution, given by

$$P^*_{Y_i|X_i,Y_{i-1}}(y_i|x_i,y_{i-1})=\begin{array}{c c} & \begin{array}{cccc} 0,0 & 0,1 & 1,0 & 1,1 \end{array} \\ \begin{array}{c} 0 \\ \\ 1 \end{array} & \left[\begin{array}{cccc} \alpha & \beta & 1-\beta & 1-\alpha \\ & & & \\ 1-\alpha & 1-\beta & \beta & \alpha \end{array}\right] \end{array} \tag{2.5.80}$$

where

$$\alpha=\frac{(1-p)(1-D)}{1-p-D+2pD},\quad \beta=\frac{p(1-D)}{p+D-2pD}$$

It is easy to verify that the distribution of the reproduction sequence is identical to the distribution of the source sequence, and is given by

$$P^*_{Y_i|Y_{i-1}}(y_i|y_{i-1})=\begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} 1-p & p \\ p & 1-p \end{array}\right] \end{array} \tag{2.5.81}$$

while the steady state distribution of the output process is IID, given by

$$P^*_{Y_i}(y_i)=0.5, \quad \forall\, y_i\in\{0,1\} \tag{2.5.82}$$

The distribution of the source symbol given the previous reconstruction symbol is evaluated using (2.5.79), and it is given by

$$P^*_{X_i|Y_{i-1}}(x_i|y_{i-1})=\begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} m & 1-m \\ 1-m & m \end{array}\right] \end{array}, \quad m=1-p-D+2pD \tag{2.5.83}$$

Since the optimal distributions defined by (2.5.80)-(2.5.83) hold for $i=1,2,\ldots,n$, we will evaluate the nonanticipative RDF for $i=1$. This is done by substituting (2.5.80)-(2.5.83) in

Christos K. Kourtellaris

the expression of the nonanticipative RDF, as follows

$$\begin{aligned}
R^{na}(D) &= \lim_{n\longrightarrow\infty} \frac{1}{n+1} I(X^n \rightarrow Y^n) \\
&= \lim_{n\longrightarrow\infty} \left\{ \frac{1}{n+1} I(X^0 \rightarrow Y^0) + \frac{1}{n+1} I(X_1^n \rightarrow Y_1^n) \right\} \\
&= \lim_{n\longrightarrow\infty} \frac{n}{n+1} \sum_{x_1,y_1,y_0} \log \left(\frac{P^*_{Y_1|Y_0,X_1}(y_1|y_0,x_1)}{P^*_{Y_1|Y_0}(y_1|y_0)} \right) P^*_{X_1,Y_1,Y_0}(x_1,y_1,y_0) \\
&= \sum_{x_1,y_1,y_0} \log \left(\frac{P^*_{Y_1|Y_0,X_1}(y_1|y_0,x_1)}{P^*_{Y_1|Y_0}(y_1|y_0)} \right) P^*_{X_1,Y_1,Y_0}(x_1,y_1,y_0) \\
&= \sum_{x_1,y_1,y_0} \log \left(\frac{P^*_{Y_1|Y_0,X_1}(y_1|y_0,x_1)}{P^*_{Y_1|Y_0}(y_1|y_0)} \right) P^*_{Y_1|X_1,Y_0}(y_1|x_1,y_0) P^*_{X_1|Y_0}(x_1|y_0) P^*_{Y_0}(y_0) \\
&= H(p) - mH(\alpha) - (1-m)H(\beta) \\
&= H(m) - H(D) \qquad (2.5.84)
\end{aligned}$$

□

The achievability of the nonanticipative RDF for the BSMS(p), based on the excess distortion probability, is addressed in Section 4.3.4, Chapter 4. There, we show that $R^{na}(D)$ is achievable when the BSMS(p) is transmitted uncoded over a unit memory channel that has the same transition probabilities as the optimal reproduction distribution (2.5.80).

Remark 2.35. The graph of $R^{na}(D)$ is illustrated in Figure 2.5.6. Note that for $p = \frac{1}{2}$, the BSMS(p) reduces to an IID Bernoulli source, $m = 1 - p - D + 2pD = 0.5$, and the nonanticipative RDF is given by $R^{na}(D) = 1 - H(D), D < \frac{1}{2}$, which is equal to the classical RDF of the Bernoulli source, as expected. The methodology to calculate the optimal reproduction distribution and the solution of the nonanticipative RDF, is outlined in Algorithm 1.

Christos K. Kourtellaris

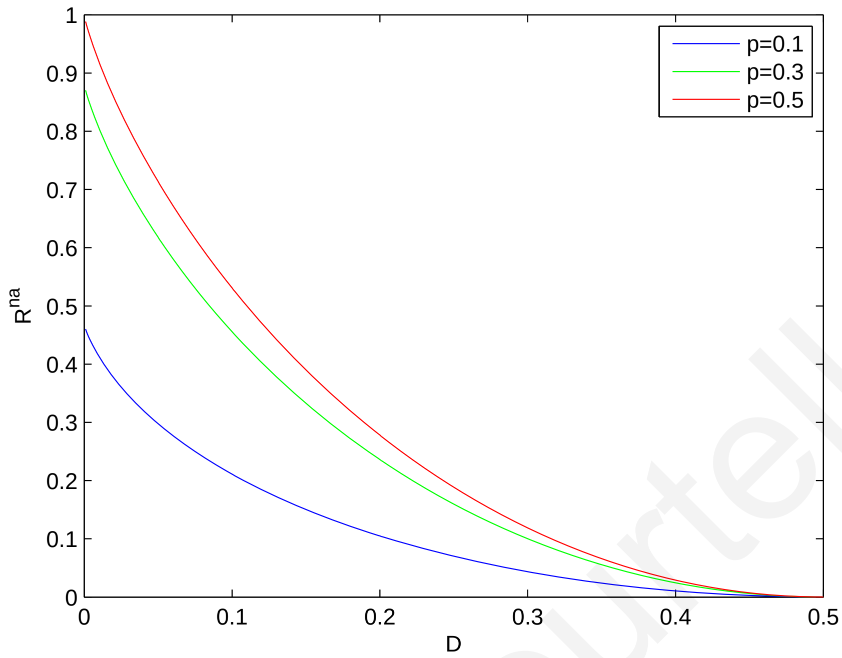

FIGURE 2.5.6: $R^{na}(D)$ for different values of parameter p.

Algorithm 1: Calculation of the Nonanticipative RDF for a Markov source.

```
Data: $d_{0,n}(x^n,y^n)$, $P_{X^n}(x^n)$ : steady state distribution, end← 1, i← 0
while end=1 do
    Apply $P_{X_i|Y^{i-1}}$ in $P_{Y_i|Y^{i-1}}$;
    Replace $P_{Y_i|Y^{i-1}}$ in $P_{Y_i|X^i,Y^{i-1}}$;
    Solve system equations in (2.5.77) and calculate $P_{Y_i|X^i,Y^{i-1}}$;
    Calculate $P_{X_{i+1}|Y^i}$;
    if $P_{Y_i|X^i,Y^{i-1}} = P_{Y_{i-1}|X^{i-1},Y^{i-2}}$ then
        end←0;
    else
        i←i+1;
    end
end
Calculate the Lagrangian $s$ from $\mathbb{E}[\frac{1}{n+1}d_{0,n}(x^n,y^n)] = D$;
Replace on $P_{Y_i|X^i,Y^{i-1}}$, $P_{X_i|Y^{i-1}}$, $P_{Y_i|Y^{i-1}}$;
Result: Apply the distributions on $R^{na}(D)$;
```

Christos K. Kourtellaris

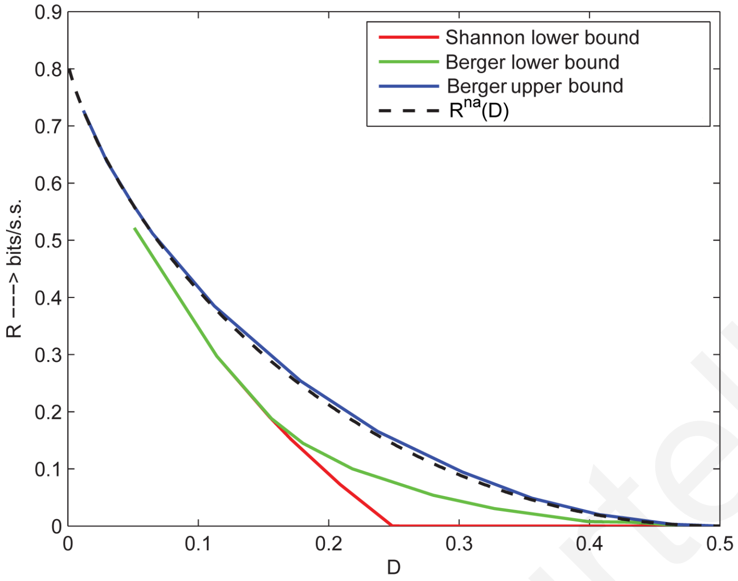

FIGURE 2.5.7: $R(D)$ for BSMS(p) for $0 \leq D \leq D_c$ and Bounds for $p = 0.25$.

In the next section we apply the result of the nonanticipative RDF to provide an upper bound for the classical RDF, and to compute the OPTA by causal codes with respect to that of noncausal codes.

2.5.1 Evaluation of Bounds

The classical RDF for the BSMS(p) is only known for the distortion region $0 \leq D \leq D_c$ [37], and is given by

$$R(D) = H(p) - H(D) \text{ if } D \leq D_c = \frac{1}{2}\Big(1 - \sqrt{1 - (\frac{p}{q})^2}\Big), \; p \leq 0.5 \qquad (2.5.85)$$

For the remainder of the distortion region $D > D_c$ only bounds are known [5]. It is also shown in [37] that (2.5.85) provides a lower bound for the classical RDF for $D > D_c$[3]. Our expression of the nonanticipative RDF provides an upper bound on the classical RDF for all possible values of D, $0 \leq D \leq 0.5$. Next, we compare the upper and lower bounds, derived by Berger in [5], which hold for $0 \leq D \leq \frac{1}{2}$ and we show that the upper bound in [5] is not as tight as the one obtained via $R^{na}(D)$. For the BSMS(p), we also compute the Rate Loss

[3] In Chapter 5 we will derive Gray's lower bound [37] by using the nonanticipative RDF with feedforward information.

Christos K. Kourtellaris

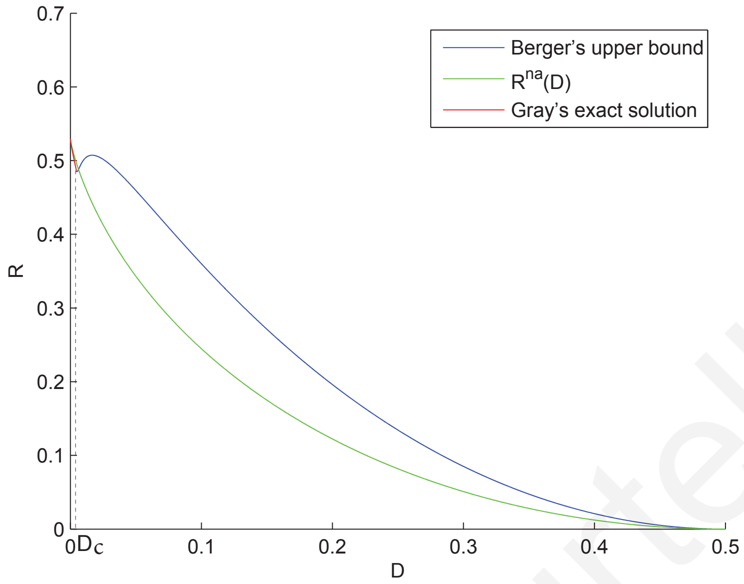

FIGURE 2.5.8: Comparison of the functional behaviour of $R^{na}(D)$ and $R(D)$ for BSMS(p) with $p = 0.12$.

(RL) of causal codes with respect to noncausal codes, by using the fact that this RL is at most $R^{na}(D) - R(D), \forall D \le D_c$ bits/sample.

Figure 2.5.7 shows the graph of $R(D)$ for $0 \le D \le D_c$, Berger's lower and upper bound [5], Shannon's lower bound and the upper bound based on $R^{na}(D)$. We observe that for $p = 0.25$, the upper bound based on $R^{na}(D)$ does slightly better than Berger's upper bound. Moreover, since $R^{na}(D)$ is nonincreasing and convex as a function of D, and nonincreasing for all values of $p \in [0, 0.5]$ (these are easily shown), then the upper bound based on $R^{na}(D)$ is convex, when compared to Berger's upper bound which is not necessarily convex and nonincreasing. This observation is illustrated in Figure 2.5.8. Finally, we use the bound $R(D) \le R^{na}(D) \le r^{c,+}(D)$ to deduce that the (RL) of causal codes for the BSMS(p) cannot exceed

$$RL = R^{na}(D) - R(D) \le \begin{cases} H(m) - H(p) & \text{if } 0 \le D \le p \\ H(m) - H(D) & \text{if } D_c < D \le 0.5 \end{cases}$$

This bound on the rate loss is illustrated in Figure 2.5.9 which demonstrates the fluctuation of the RL for $p \in [0, 0.5]$. It is interesting to see that the maximum value of the RL is 0.2144 and corresponds to $(p = 0.1012, D = 0.1012)$. This bound is exact for $D \le D_c \le p$. For

Christos K. Kourtellaris

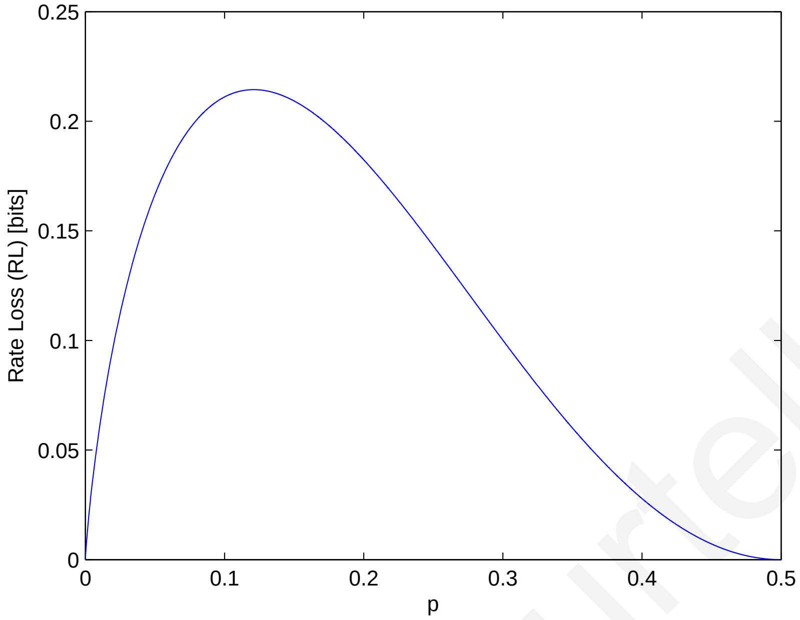

FIGURE 2.5.9: Comparison of the Rate Loss (RL) for $p \in [0, 0.5]$.

high resolution ($D \rightarrow 0$), the classical rate distortion function and the nonanticipative rate distortion function are equivalent and equal to $H(p)$.

2.6 Multidimensional Partially Observed Gaussian Process

Here, we consider a multidimensional partially observed Gaussian-Markov process and we compute the closed form expression of the information nonanticipative RDF, $R^{na}(D)$.
Consider the following discrete-time multidimensional partially observed linear Gauss-Markov system described by

$$\begin{cases} Z_{t+1} = AZ_t + BW_t, \ Z_0 = z, \ t \in \mathbb{N} \\ X_t = CZ_t + NV_t, \ t \in \mathbb{N} \end{cases} \tag{2.6.86}$$

where $Z_t \in \mathbb{R}^m$ is the state (unobserved) process and $X_t \in \mathbb{R}^p$ is the information source, obtained from noisy measurements of CZ_t. The model in (2.6.86) is often encountered in applications where the process $\{Z_t : t \in \mathbb{N}\}$ is not directly observed; instead, what is directly observed is the process $\{X_t : t \in \mathbb{N}\}$ which is a noisy version of it. This is a realistic model for any sensor which collects information for the underlying process CZ_t, since the sensor is a measurement device often subject to additive Gaussian noise. Hence, in this application

Christos K. Kourtellaris

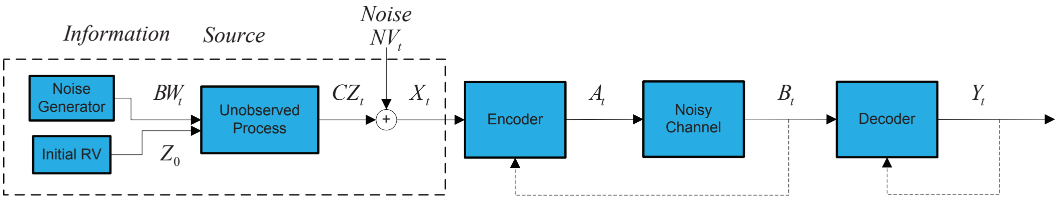

FIGURE 2.6.10: Communication System.

the objective is to compress the sensor data, which is the only observable information. Next, we introduce certain assumptions which are sufficient for existence of the limit, $R^{na}(D) \triangleq \lim_{n \longrightarrow \infty} \frac{1}{n+1} R^{na}_{0,n}(D)$.

(G1) (C,A) is detectable and $(A, \sqrt{BB^{tr}})$ is stabilizable, $(N \neq 0)$ [7];

(G2) The state and observation noise $\{(W_t,V_t) : t \in \mathbb{N}^n\}$ are Gaussian IID vectors $W_t \in \mathbb{R}^k$, $V_t \in \mathbb{R}^d$, mutually independent with parameters $N(0, I_{k\times k})$ and $N(0, I_{d\times d})$, independent of the Gaussian RV Z_0, with parameters $N(\bar{z}_0, \bar{\Sigma}_0)$.

(G3) The distortion function is single letter defined by $d_{0,n}(x^n, y^n) \triangleq \sum_{t=0}^{n} ||x_t - y_t||^2_{\mathbb{R}^p}$.

For the fully observed scalar case corresponding to $X_t = Z_t \in \mathbb{R}$, the reconstruction of $\{Z_t : \ t \in \mathbb{N}^n\}$ and its realization over a scalar additive white Gaussian noise (AWGN) channel is discussed in [76], while the partially observed (2.6.86) for the scalar case $X_t \in \mathbb{R}$, is discussed in [9] via indirect methods. However, as pointed out in [22], the computation of the nonanticipative RDF for the vector Gaussian process is unsolved. Here, we show that the conjecture stated in [12] is indeed true. To this end, we provide a closed form expression to the nonanticipative RDF for the vector Gaussian process.

According to Theorem 2.32, the optimal stationary reproduction distribution is given by

$$\begin{aligned} P^*_{Y_t|Y^{t-1},X^t}(dy_t|y^{t-1},x^t) &= \frac{e^{s||y_t - x_t||^2_{\mathbb{R}^p}} P^*_{Y_t|Y^{t-1}}(dy_t|y^{t-1})}{\int_{\mathcal{Y}_t} e^{s||y_t - x_t||^2_{\mathbb{R}^p}} P^*_{Y_t|Y^{t-1}}(dy_t|y^{t-1})}, \ s \leq 0 \\ &\equiv P^*_{Y_t|Y^{t-1},X^t}(dy_t|y^{t-1},x_t) - a.a. \ (y^{t-1}, x_t). \end{aligned} \tag{2.6.87}$$

Hence, from (2.6.87), it follows that the optimal reproduction is Markov with respect to the process $\{X_t : \ t \in \mathbb{N}\}$. Moreover, since the exponential term $||y_t - x_t||^2_{\mathbb{R}^p}$ in the RHS of (2.6.87) is quadratic in (x_t, y_t), and $\{Z_t : \ i \in \mathbb{N}\}$ is Gaussian then $\{(Z_t, X_t) : \ t \in \mathbb{N}\}$ are jointly Gaussian, and it follows that a Gaussian distribution $P_{Y_t|Y^{t-1},X_t}(\cdot|y^{t-1}, x_t)$ (for a fixed realization of (y^{t-1}, x_t)), and Gaussian distribution $P_{Y_t|Y^{t-1}}(\cdot|y^{t-1})$ can match the left and right side of (2.6.87). Therefore, at any time $t \in \mathbb{N}$, the output Y_t of the optimal

Christos K. Kourtellaris

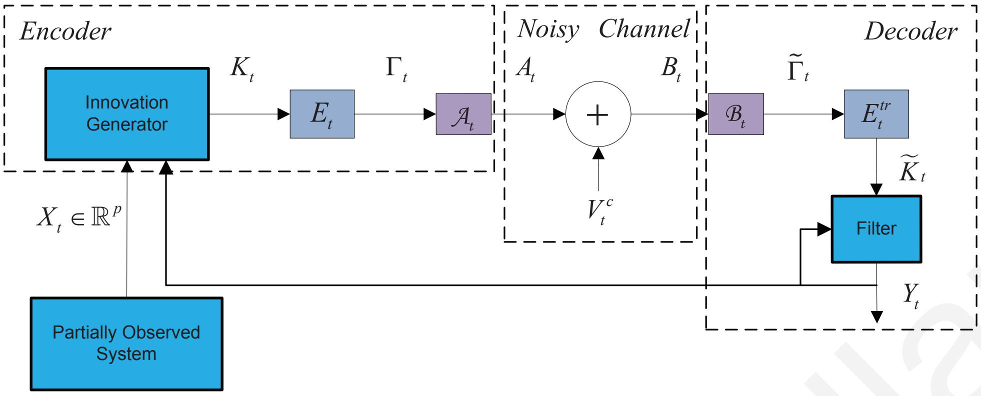

FIGURE 2.6.11: Realization of the optimal stationary reproduction distribution.

reconstruction channel depends on X_t and the previous outputs Y^{t-1}, and its conditional distribution is Gaussian. Hence, the channel connecting $\{X_t : t \in \mathbb{N}\}$ to $\{Y_t : t \in \mathbb{N}\}$ has the general form

$$Y_t = \bar{A}X_t + \bar{B}Y^{t-1} + V_t^c,\ t \in \mathbb{N} \tag{2.6.88}$$

where $\bar{A} \in \mathbb{R}^{p\times p}$, $\bar{B} \in \mathbb{R}^{p\times tp}$, and $\{V_t^c :\ t \in \mathbb{N}\}$ is an independent sequence of Gaussian vectors $N(0;Q)$.

Introduce the error estimate $\{K_t :\ t \in \mathbb{N}\}$, and its covariance $\{\Lambda_t :\ t \in \mathbb{N}\}$, defined by

$$K_t \triangleq X_t - \widehat{X}_{t|t-1},\ \widehat{X}_{t|t-1} \triangleq \mathbb{E}\Big\{X_t|\sigma\{Y^{t-1}\}\Big\},\ \Lambda_t \triangleq \mathbb{E}\{K_tK_t^{tr}\},\ t \in \mathbb{N} \tag{2.6.89}$$

where $\sigma\{Y^{t-1}\}$ is the σ-algebra generated by the sequence $\{Y^{t-1}\}$. The covariance is diagonalized by introducing a unitary transformation $\{E_t : t \in \mathbb{N}\}$ such that

$$E_t\Lambda_tE_t^{tr} = diag\{\lambda_{t,1},\ldots\lambda_{t,p}\},\ \Gamma_t \triangleq E_tK_t,\ t \in \mathbb{N}. \tag{2.6.90}$$

Note that although $\{\Gamma_t :\ t \in \mathbb{N}\}$ has independent Gaussian components, each component is correlated. Analogously, introduce to the process $\{\tilde{K}_t :\ t \in \mathbb{N}\}$ defined by

$$\tilde{K}_t \triangleq Y_t - \widehat{X}_{t|t-1},\ \tilde{\Gamma}_t = E_t\tilde{K}_t,\ t \in \mathbb{N}. \tag{2.6.91}$$

We shall compute the information nonanticipative RDF by considering the realization shown in Fig. 2.6.11, where $\{V_c^t :\ t = 0,1,\ldots\}$ is Gaussian $N(0;Q)$, and $\{\mathscr{A}_t, \mathscr{B}_t :\ t = 0,1,\ldots\}$ are to be determined. Note that the square error fidelity criterion $d_{0,n}(\cdot,\cdot)$ is not affected by the preprocessing and post processing of $\{(X_t,Y_t) :\ t \in \mathbb{N}\}$, since $d_{0,n}(X^n,Y^n) = d_{0,n}(K^n,\tilde{K}^n) =$

Christos K. Kourtellaris

$\sum_{t=0}^{n}||\tilde{K}_t - K_t||_{\mathbb{R}^p}^2 = \sum_{t=0}^{n}||\tilde{\Gamma}_t - \Gamma_t||_{\mathbb{R}^p}^2 = d_{0,n}(\Gamma^n, \tilde{\Gamma}^n)$. Using basic properties of conditional entropy, if necessary, we can show the following expressions are equivalent.

$$\begin{aligned} R^{na}(D) &= \lim_{n \longrightarrow \infty} R_{0,n}^{na,K^n,\tilde{K}^n}(D) \stackrel{\triangle}{=} \lim_{n \longrightarrow \infty} \inf_{\overrightarrow{P}_{\tilde{K}^n|K^n}:\, \mathbb{E}\left\{d_{0,n}(K^n,\tilde{K}^n) \leq D\right\}} \frac{1}{n+1}\sum_{t=0}^{n} I(K_t;\tilde{K}_t|\tilde{K}^{t-1}) \\ &= \lim_{n \longrightarrow \infty} R_{0,n}^{na,\Gamma^n,\tilde{\Gamma}^n}(D) \stackrel{\triangle}{=} \lim_{n \longrightarrow \infty} \inf_{\overrightarrow{P}_{\tilde{\Gamma}^n|\Gamma^n}:\, \mathbb{E}\left\{d_{0,n}(\Gamma^n,\tilde{\Gamma}^n) \leq D\right\}} \frac{1}{n+1}\sum_{t=0}^{n} I(\Gamma_t;\tilde{\Gamma}_t|\tilde{\Gamma}^{t-1}). \end{aligned} \tag{2.6.92}$$

Next, we state the main results.

Theorem 2.36. *($R^{na}(D)$ of multidimensional partially observed Gaussian source) Under Assumptions (**G1**)-(**G3**), the information nonanticipative RDF rate for the multidimensional partially observed Gaussian source (2.6.86) is given by*

$$R^{na}(D) = \frac{1}{2}\sum_{i=1}^{p} \log\left(\frac{\lambda_{\infty,i}}{\delta_{\infty,i}}\right) \tag{2.6.93}$$

where $diag\{\lambda_{\infty,1},\ldots,\lambda_{\infty,p}\} = \lim_{t \longrightarrow \infty} E_t \Lambda_t E_t^{tr} = E_\infty \Lambda_\infty E_\infty^{tr}$,

$$\begin{aligned} \Lambda_\infty &= \lim_{t \longrightarrow \infty} \mathbb{E}\Big\{\Big(C\big(Z_t - \mathbb{E}\{Z_t|\sigma\{Y^{t-1}\}\}\big) + NV_t\Big)\Big(C\big(Z_t - \mathbb{E}\{Z_t|\sigma\{Y^{t-1}\}\}\big) + NV_t\Big)^{tr}\Big\} \\ &= C \lim_{t \longrightarrow \infty} \Sigma_t C^{tr} + NN^{tr} = C\Sigma_\infty C^{tr} + NN^{tr} \end{aligned} \tag{2.6.94}$$

$$\delta_{\infty,i} \stackrel{\triangle}{=} \begin{cases} \xi_\infty & \text{if} \quad \xi_\infty \leq \lambda_{\infty,i} \\ \lambda_{\infty,i} & \text{if} \quad \xi_\infty > \lambda_{\infty,i} \end{cases}, \ i = 2,\ldots,p \tag{2.6.95}$$

and ξ_∞ is chosen such that $\sum_{i=1}^{p}\delta_{\infty,i} = D$. Moreover, Σ_∞ is the steady state covariance of the error $Z_t - \mathbb{E}\{Z_t|Y^{t-1}\} \sim N(0,\Sigma_\infty)$, $\widehat{Z}_{t|t-1} \stackrel{\triangle}{=} \mathbb{E}\{Z_t|Y^{t-1}\}$, of the Kalman filter given by

$$\begin{aligned} \widehat{Z}_{t+1|t} &= A\widehat{Z}_{t|t-1} \\ &\quad + A\Sigma_\infty(E_\infty^{tr}H_\infty E_\infty C)^{tr}M_\infty^{-1}\big(Y_t - C\widehat{Z}_{t|t-1}\big),\ \hat{Z}_0 = \mathbb{E}\{Z_0|Y^{-1}\}, Z_0 - \hat{Z}_0 \sim N(0,\Sigma_\infty) \end{aligned} \tag{2.6.96}$$

$$\Sigma_\infty = A\Sigma_\infty A^{tr} - A\Sigma_\infty(E_\infty^{tr}H_\infty E_\infty C)^{tr}M_\infty^{-1}(E_\infty^{tr}H_\infty E_\infty C)\Sigma_\infty A^{tr} + BB_\infty^{tr} \tag{2.6.97}$$

$$M_\infty = E_\infty^{tr}H_\infty E_\infty C\Sigma_\infty(E_\infty^{tr}H_\infty E_\infty C)^{tr} + E_\infty^{tr}H_\infty E_\infty NN^{tr}(E_\infty^{tr}H_\infty E_\infty)^{tr} + E_\infty^{tr}\mathscr{B}_\infty Q\mathscr{B}_\infty^{tr}E_\infty \tag{2.6.98}$$

Christos K. Kourtellaris

and

$$H_\infty = \lim_{t\longrightarrow\infty} H_t,\ H_t \stackrel{\triangle}{=} diag\{\eta_{t,1},\ldots,\eta_{t,p}\},\ \eta_{t,i} = 1-\frac{\delta_{t,i}}{\lambda_{t,i}},\ i=1,\ldots,p,\ t\in\mathbb{N} \tag{2.6.99}$$

$$\mathscr{B}_\infty = \lim_{t\longrightarrow\infty} \mathscr{B}_t = \sqrt{H_\infty\Delta_\infty Q^{-1}},\ \mathscr{B}_t \triangleq \sqrt{H_t\Delta_t Q^{-1}},\ t\in\mathbb{N} \tag{2.6.100}$$

$$\Delta_\infty = \lim_{t\longrightarrow\infty} \Delta_t,\ \Delta_t = diag\{\delta_{t,1},\ldots,\delta_{t,p}\}, t\in\mathbb{N}. \tag{2.6.101}$$

Proof. See [73]. □

In the next remark, we confirm that Theorem 2.36 gives, as a special case, the value of $R^{na}(D)$ for scalar Gaussian stationary source found in [22, Theorem 3].

Remark 2.37. Consider the special case of first-order (scalar) Gaussian-Markov source [22, Theorem 3]

$$X_{t+1} = \alpha X_t + \sigma_w W_t,\ W_t \sim N(0,1).$$

This corresponds to the dynamical system (2.6.86) with $m=p=1$, $C=1$, $N=0$, $A=\alpha$, $B=\sigma_w$, i.e., $\sigma_w W_t \sim N(0,\sigma_w^2)$, hence $X_t = Z_t$. Clearly, $\Lambda_\infty = \Sigma_\infty$, $\Delta_\infty = D$, where $H_\infty = 1-\frac{D}{\Sigma_\infty}$ and $E_\infty = 1$.

Using (2.6.98), we have

$$M_\infty = \Sigma_\infty H_\infty^2 + H_\infty D = H_\infty\big(\Sigma_\infty H_\infty + D\big) = H_\infty\big(\Sigma_\infty(1-\frac{D}{\Sigma_\infty}) + D\big) = \Sigma_\infty H_\infty. \tag{2.6.102}$$

Also, by (2.6.97), we get

$$\begin{aligned}\Sigma_\infty &= \alpha^2\Sigma_\infty - \alpha^2\Sigma_\infty^2 H_\infty^2 M^{-1} + \sigma_w^2 \stackrel{(a)}{=} \alpha^2\Sigma_\infty - \alpha^2\Sigma_\infty^2 H_\infty^2 H_\infty^{-1}\Sigma_\infty^{-1} + \sigma_w^2\\ &= \alpha^2\Sigma_\infty - \alpha^2\Sigma_\infty H_\infty + \sigma_w^2 = \alpha^2\Sigma_\infty - \alpha^2\Sigma_\infty\big(1-\frac{D}{\Sigma_\infty}\big) + \sigma_w^2 = \alpha^2 D + \sigma_w^2\end{aligned} \tag{2.6.103}$$

where (a) follows from (2.6.102). Finally, by substituting (2.6.103) in the expression of the nonanticipative RDF (2.6.93) we obtain

$$R^{na}(D) = \frac{1}{2}\log\frac{|\Lambda_\infty|}{|\Delta_\infty|} = \frac{1}{2}\log\frac{\Sigma_\infty}{D} = \frac{1}{2}\log\Big(\frac{\alpha^2 D+\sigma_w^2}{D}\Big) = \frac{1}{2}\log\Big(\alpha^2 + \frac{\sigma_w^2}{D}\Big). \tag{2.6.104}$$

which is the expression derived in [22, Theorem 3]. Hence, Theorem 2.36 generalizes previous work to multidimensional (vector) Gaussian-Markov stationary process.

Christos K. Kourtellaris

In the following lemma, we show that $\{\tilde{K}_t :\ t \in \mathbb{N}\}$ is the innovation process of $\{Y_t :\ t \in \mathbb{N}\}$, and hence the two processes generate the same σ-algebras (they contain the same information).

Lemma 2.38. *(Equivalence of σ-algebras)*
The following hold.

$$\mathscr{F}^{Y}_{0,t} \stackrel{\triangle}{=} \sigma\{Y_s :\ s = 0,1,\ldots,t\} = \mathscr{F}^{\tilde{K}}_{0,t} \stackrel{\triangle}{=} \sigma\{\tilde{K}_s :\ s = 0,1,\ldots,t\},\ \forall t \in \mathbb{N}.$$

that is, $\mathscr{F}^{Y}_{0,t} \subseteq \mathscr{F}^{\tilde{K}}_{0,t}$ *and* $\mathscr{F}^{\tilde{K}}_{0,t} \subseteq \mathscr{F}^{Y}_{0,t}$, $\forall t \in \mathbb{N}$.

Proof. Since $\tilde{K}_s = Y_s - \mathbb{E}\big\{X_s|Y^{s-1}\big\}, 0 \leq s \leq t$, then $\mathscr{F}^{\tilde{K}}_{0,t} \subseteq \mathscr{F}^{Y}_{0,t}$, $\forall t \in \mathbb{N}$. Hence, we need to show that $\mathscr{F}^{Y}_{0,t} \subseteq \mathscr{F}^{\tilde{K}}_{0,t}$, $\forall t \in \mathbb{N}$. The innovation process of $\{Y_t :\ t \in \mathbb{N}\}$ is by definition (see Fig. 2.6.11, (2.6.91)).

$$\begin{aligned}
I_t &= Y_t - \mathbb{E}\big\{Y_t|Y^{t-1}\big\} \\
&= E^{tr}_{\infty} H_{\infty} E_{\infty}\Big(X_t - \mathbb{E}\big\{X_t|Y^{t-1}\big\}\Big) + E^{tr}_{\infty}\mathscr{B}_{\infty}V^{c}_{t} + \mathbb{E}\big\{X_t|Y^{t-1}\big\} - \mathbb{E}\big\{X_t|Y^{t-1}\big\} \\
&= E^{tr}_{\infty} H_{\infty} E_{\infty}\Big(X_t - \mathbb{E}\big\{X_t|Y^{t-1}\big\}\Big) + E^{tr}_{\infty}\mathscr{B}_{\infty}V^{c}_{t} = \tilde{K}_t. \qquad (2.6.105)
\end{aligned}$$

Since the innovation process $\{I_s :\ 0 \leq s \leq t\}$ and the optimal reproduction process $\{Y_s :\ 0 \leq s \leq t\}$ generates the same $\sigma-$algebras, then $\mathscr{F}^{I}_{0,t} \subseteq \mathscr{F}^{Y}_{0,t}$, $\mathscr{F}^{Y}_{0,t} \subseteq \mathscr{F}^{I}_{0,t}$, i.e., $\mathscr{F}^{Y}_{0,t} = \mathscr{F}^{I}_{0,t}$, and hence, by (2.6.105) we also obtain $\mathscr{F}^{Y}_{0,t} \subseteq \mathscr{F}^{\tilde{K}}_{0,t}$. This completes the proof. □

We now observe the following consequence of Lemma 2.38.

Remark 2.39. By Lemma 2.38, all conditional expectations with respect to the process $\{Y_t : t = 0,1,\ldots\}$ can be replaced by conditional expectations with respect to the independent process $\{\tilde{K}_t : t = 0,1,\ldots\}$. Hence, the process $\{K_t : t = 0,1,\ldots\}$ can be written as $K_t = X_t - \mathbb{E}\big\{X_t|\sigma\{Y^{t-1}\}\big\} = X_t - \mathbb{E}\big\{X_t|\sigma\{\tilde{K}^{t-1}\}\big\}$, and its reconstruction is given by

$$\tilde{K}_t = E^{tr}_{\infty} H_{\infty} E_{\infty}\Big(X_t - \mathbb{E}\big\{X_t|\tilde{K}^{t-1}\big\}\Big) + E^{tr}_{\infty}\mathscr{B}_{\infty}V^{c}_{t} = E^{tr}_{\infty} H_{\infty} E_{\infty} K_t + E^{tr}_{\infty}\mathscr{B}_{\infty}V^{c}_{t},\ t = 0,1,\ldots.$$

Moreover, by Lemma 2.38, K_t and $\tilde{K}_t$ are independent of $Y_0,\ldots,Y_{t-1}$, and $\tilde{K}_0,\ldots,\tilde{K}_{t-1}$, $t = 0,1,\ldots$. This property is analogous to the JSCC of a scalar RV over a scalar additive Gaussian noise channel with feedback [41, Theorem 5.6.1].

Christos K. Kourtellaris

2.7 Conclusions

In this chapter we provide a framework based on nonanticipative RDF for general sources with memory, which is suitable for nonanticipative transmission. We describe the connection of this new information measure with Gorbunov and Pinsker nonanticipatory $\varepsilon-$entropy and with the classical RDF. A noisy coding theorem is derived and the optimal reproduction distribution as well as the solution of the nonanticipative RDF are calculated.

We then apply these theoretical results, to calculate the nonanticipative RDF and the optimal reproduction distribution of the binary symmetric Markov source. We compare our results with other existing bounds on the classical RDF (i.e., OPTA by noncausal codes), and we presented the rate loss of causal codes with respect to noncausal codes.

We kept the mathematical sophistication simple, by concentrating on the importance of this nonanticipative RDF and its application meanings. In [73] all mathematical issues are addressed for general abstract spaces, complete separable metric spaces, while the multidimensional Gaussian example is also derived.

Christos K. Kourtellaris

Chapter 3

Structural Properties of Extremum Problems of Capacity

3.1 Introduction

Channel capacity coding theorems are often classified into Discrete Memoryless Channels (DMC), and channels with memory, with or without feedback. For channels with memory and feedback the information measure often employed is the mutual information from the source output to the channel output [41]. It is related to the so-called directed information, which accounts for causality and direction of information flow, introduced by Masssey [52] and subsequently applied by Kramer [46]; this directional measure of information is attributed to Marko [51].

Historically, Shannon [66] and Dobrushin [23] derived formulas for capacity of DMC and established coding theorems, while Ebert [24] and Cover and Pombra [19] characterized the capacity of Gaussian channels with memory and feedback, showing that memory can increase capacity. Chen and Berger [18] analysed limited memory channels with feedback, when the channel output and the channel input-output pair are first-order Markov models, presented a formulae for channel capacity in terms of directed information, derived sufficient conditions under which coding theorems hold, and applied dynamic programming to analyze the capacity achieving distribution of the unit memory channel. Tatikonda [75] applied information spectrum methods to derive coding theorems for general finite alphabet channels with memory and feedback. Moreover, in [85] dynamic programming is used to describe the capacity of certain types of channels. However, aside from certain memoryless and Gaussian

Christos K. Kourtellaris

channels, computing the capacity achieving input distribution and the capacity of channels with memory, with and without feedback, are open questions.

Recently, Shayevitz and Feder [68–70] introduced the so-called Posterior Matching Scheme (PMS), a recursive encoding scheme that achieves the capacity of DMC with feedback, provided one knows the capacity achieving input distribution. This scheme goes back to the idea put forward by Horstein [40], who designed encoders that achieve the capacity of discrete memoryless symmetric channels with feedback. The PMS is further investigated by Gorantla and Coleman [33] for DMC with feedback.

In this chapter we consider general channels with memory and feedback and general sources with memory and feedback, and we derive results along the following directions.

1. Structural encoder properties which maximize the directed information from the source to the channel output, and tight bounds on the converse to the coding theorem.
2. Structural properties of capacity achieving distribution.
3. Dynamic programming recursions to compute the encoder and the achieving distribution in 1. and 2., respectively.
4. Generalize PMS for designing encoders, which achieve the information capacity for channels with memory and feedback.
5. Capacity and optimal input distribution of the Binary State Symmetric Channel (BSSC) with or without feedback, with and without transmission cost constraints.

The material on structural encoder properties and capacity achieving distribution, generalizes current and past research in the area of capacity of channels with memory and feedback. Specifically, the material on the structural encoder properties of the capacity achieving distribution, and the tight bounds on the converse to the coding theorem, state that for general sources and channels, maximizing directed information over all encoder strategies which are non-Markov with respect to the source is equivalent to maximizing it over Markov encoding strategies. These results generalize the previous work found in [6, 18], for memoryless channels.

The material on PMS describes coding schemes which achieve the capacity of channels with memory and feedback. These results generalize previous work found in [33, 68–70] from memoryless channels to channels with memory and feedback.

Christos K. Kourtellaris

The material in applying dynamic programming to extremum problems of capacity is motivated from optimal stochastic control theory with partial information, in which separated strategies are employed [13, 14]. Here, an information state is identified which carries all the information available in any channel output sequence. The material on maximizing directed information from the channel input to its output over all channel input distributions with specific structural properties, via dynamic programming, simplifies the previous dynamic programming described in [85].

Throughout this chapter we do not present the direct channel coding theorem, because such theorems are derived in [18, 43, 59, 77], for finite alphabet channels, and they can easily extended to abstract alphabets. In addition, we do not address existence of solutions to extremum problems because such results follow directly form [10].

3.2 Problem Formulation

In this section we introduce the various blocks of the communication system of Figure 3.2.1. We assume all processes (introduced below) are defined on a complete probability space $(\Omega, \mathbb{F}, \mathbb{P})$ with filtration $\{\mathbb{F}_t : t \in \mathbb{N}\}$. The alphabets of the source output, channel input, channel output and decoder output are assumed to be sequences of Polish spaces (complete separable metric spaces) $\{\mathscr{X}_t : t = 0, 1, \ldots, n\}$, $\{\mathscr{A}_t : t = 0, 1, \ldots, n\}$, $\{\mathscr{B}_t : t = 0, 1, \ldots, n\}$, and $\{\mathscr{Y}_t : t = 0, 1, \ldots, n\}$, respectively. Moreover, we associate these alphabets with their corresponding measurable spaces $(\mathscr{X}_t, \mathbb{B}(\mathscr{X}_t))$, $(\mathscr{A}_t, \mathbb{B}(\mathscr{A}_t))$, $(\mathscr{B}_t, \mathbb{B}(\mathscr{B}_t))$ and $(\mathscr{Y}_t, \mathbb{B}(\mathscr{Y}_t))$ (e.g., $\mathbb{B}(\mathscr{A}_t)$ is a Borel $\sigma-$algebra of subsets of the set $\mathscr{A}_t$ generated by closed sets). Thus, we identify sequences with the product measurable spaces as follows.

$$
\begin{aligned}
(\mathscr{X}_{0,n}, \mathbb{B}(\mathscr{X}_{0,n})) &\stackrel{\triangle}{=} \times_{k=0}^{n} (\mathscr{X}_k, \mathbb{B}(\mathscr{X}_k)) \\
(\mathscr{A}_{0,n}, \mathbb{B}(\mathscr{A}_{0,n})) &\stackrel{\triangle}{=} \times_{k=0}^{n} (\mathscr{A}_k, \mathbb{B}(\mathscr{A}_k)) \\
(\mathscr{B}_{0,n}, \mathbb{B}(\mathscr{B}_{0,n})) &\stackrel{\triangle}{=} \times_{k=0}^{n} (\mathscr{B}_k, \mathbb{B}(\mathscr{B}_k)) \\
(\mathscr{Y}_{0,n}, \mathbb{B}(\mathscr{Y}_{0,n})) &\stackrel{\triangle}{=} \times_{k=0}^{n} (\mathscr{Y}_k, \mathbb{B}(\mathscr{Y}_k))
\end{aligned}
$$

Christos K. Kourtellaris

We denote the source output, channel input, channel output, and decoder output by the following processes.

$$\begin{aligned} X^n &\triangleq \{X_t : t = 0,1,\ldots,n\}, && X : \{t\} \times \Omega \mapsto \mathscr{X}_t \\ A^n &\triangleq \{A_t : t = 0,1,\ldots,n\}, && A : \{t\} \times \Omega \mapsto \mathscr{A}_t \\ B^n &\triangleq \{B_t : t = 0,1,\ldots,n\}, && B : \{t\} \times \Omega \mapsto \mathscr{B}_t \\ Y^n &\triangleq \{Y_t : t = 0,1,\ldots,n\}, && Y : \{t\} \times \Omega \mapsto \mathscr{Y}_t \end{aligned}$$

We denote the set of probability measures on any measurable space $(\mathscr{Z}, \mathbb{B}(\mathscr{Z}))$ by $\mathscr{M}_1(\mathscr{Z})$.

Often, we describe conditional distributions by stochastic Kernels and Markov chains by conditional independence, defined below.

Definition 3.1. Consider the measurable spaces $(\mathscr{A}, \mathbb{B}(\mathscr{A}))$, $(\mathscr{B}, \mathbb{B}(\mathscr{B}))$. A stochastic Kernel is a mapping $q : \mathbb{B}(\mathscr{B}) \times \mathscr{A} \to [0,1]$ satisfying the following two properties:

1) For every $a \in \mathscr{A}$, the set function $q(\cdot; a)$ is a probability measure (possibly finitely additive) on $\mathbb{B}(\mathscr{B})$.

2) for every $F \in \mathbb{B}(\mathscr{B})$, the function $q(F; \cdot)$ is $\mathbb{B}(\mathscr{A})$-measurable.

The set of all such stochastic Kernels is denoted by $\mathscr{K}(\mathscr{B}; \mathscr{A})$.

Next we introduce the definition of conditionally independence.

Definition 3.2. Consider a probability space $(\Omega, \mathbb{F}, \mathbb{P})$ and the measurable spaces $(\mathscr{A}, \mathbb{B}(\mathscr{A}))$, $(\mathscr{B}, \mathbb{B}(\mathscr{B})), (\mathscr{Z}, \mathbb{B}(\mathscr{Z}))$ on it. The σ-algebra $\mathbb{B}(\mathscr{Z})$ is called conditionally independent of $\mathbb{B}(\mathscr{A})$ given $\mathbb{B}(\mathscr{B})$ if and only if

$$\mathbb{P}(Z|A,B) = \mathbb{P}(Z|B), \forall\, Z \in \mathbb{B}(\mathscr{Z}), \text{ for almost all} A \in \mathscr{A}, B \in \mathscr{B}$$

If $(\mathscr{A}, \mathbb{B}(\mathscr{A})), (\mathscr{B}, \mathbb{B}(\mathscr{B})), (\mathscr{Z}, \mathbb{B}(\mathscr{Z}))$ are associated with R.V.'s $A : (\Omega, \mathbb{F}) \mapsto (\mathscr{A}, \mathbb{B}(\mathscr{A}))$, $B : (\Omega, \mathbb{F}) \mapsto (\mathscr{B}, \mathbb{B}(\mathscr{B}))$, $Z : (\Omega, \mathbb{F}) \mapsto (\mathscr{Z}, \mathbb{B}(\mathscr{Z}))$, for $q(.;.,.) \in \mathscr{K}(\mathscr{Z}; \mathscr{A} \times \mathscr{B})$ then the above definition is equivalent to $q(dz; a, b) = q(dz; b)$, for almost all $a \in \mathscr{A}, b \in \mathscr{B}$. Such conditional independence is denoted by $A \leftrightarrow B \leftrightarrow Z$ forms a Markov chain in both directions. Note that, $A \leftrightarrow B \leftrightarrow Z$ if and only if $q(da, dz; b) = q(da; b) \otimes q(dz; b)$, for almost all $b \in \mathscr{B}$.

Christos K. Kourtellaris

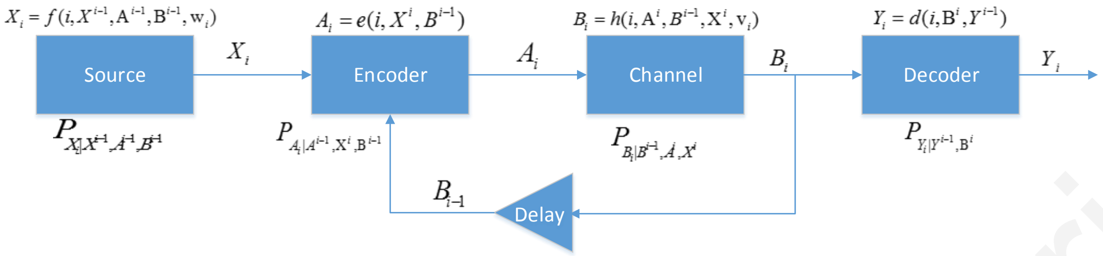
$X_i = f(i, X^{i-1}, A^{i-1}, B^{i-1}, w_i)$
$A_i = e(i, X^i, B^{i-1})$
$B_i = h(i, A^i, B^{i-1}, X^i, v_i)$
$Y_i = d(i, B^i, Y^{i-1})$
Source
X_i
Encoder
A_i
Channel
B_i
Decoder
Y_i
$P_{X_i|X^{i-1}, A^{i-1}, B^{i-1}}$
$P_{A_i|A^{i-1}, X^i, B^{i-1}}$
$P_{B_i|B^{i-1}, A^i, X^i}$
$P_{Y_i|Y^{i-1}, B^i}$
B_{i-1}
Delay

FIGURE 3.2.1: General Communication System with Feedback

The general communication system is described by two random processes. The information source process $\{X_t : t \in \mathbb{N}\}$ and the output of the channel $\{B_t : t \in \mathbb{N}\}$. The rest of the system components are the encoder generating the encoder process $\{A_t : t \in \mathbb{N}\}$, and the decoder generating the decoder process, $\{Y_t : t \in \mathbb{N}\}$, by manipulating the functions of the encoder and decoder, which are functions of the processes $\{(X_t, B_t) : t \in \mathbb{N}\}$. Specifically, the communication system design consists of selecting the encoder and decoder variables based on the available information at each time instant in such a way to achieve a desired performance for the overall communication system.

Suppose the information process satisfies the recursive dynamics[1]

$$X_t = f(t, X^{t-1}, B^{t-1}, w_t), \quad w_0 = \omega_0,\ t \in \mathbb{N} \tag{3.2.1}$$

and the channel process satisfies the recursive dynamics.

$$B_t = h(t, B^{t-1}, A^t, v_t) \quad t \in \mathbb{N} \tag{3.2.2}$$

where $f(.,.,.)$, $h(.,.,.,.)$ are measurable functions. The primitive RV's (source of randomness) are

$$\omega = \{w_0, v_0, \ldots, w_n, v_n\} \quad \forall\, n \geq 0 \tag{3.2.3}$$

which may or may be not be a vector of independent RV's. The spaces of $\{w_t : t = 0, 1, \ldots\}$ and $\{v_t : t = 0, 1, \ldots\}$ are $\{\mathscr{W}_t : t = 0, 1, \ldots\}$ and $\{\mathscr{V}_t : t = 0, 1, \ldots\}$. The distribution of the primitive RV's is defined by their joint probability on a probability space $(\Omega, \mathbb{F}, \mathbb{P})$, where $\omega \in \Omega$, with ω given by (3.2.3).

[1] We allow dependence on previous channel outputs, B^{t-1}, to avoid excluding control sources.

Christos K. Kourtellaris

Consider a feedback encoder, with information available at the encoder at time t given by

$$X^t \stackrel{\triangle}{=} (X_0, X_1, \ldots, X_t), \;\; B^{t-1} \stackrel{\triangle}{=} (B_0, B_1, \ldots, B_{t-1}), \quad t = 0, 1, \ldots \tag{3.2.4}$$

Then, the encoder process is described by

$$A_t = e(t, X^t, B^{t-1}, A^{t-1}) \equiv A_t(X^t, B^{t-1}, A^{t-1}), \quad t = 0, 1, \ldots \tag{3.2.5}$$

where $e(t, ., .)$ is a measurable function called the deterministic encoder function at time t. If the encoder does not have feedback then the encoder process is described by

$$A_t = e(t, X^t) \equiv A_t(X^t), \quad t = 0, 1, \ldots \tag{3.2.6}$$

Consider a decoder with information available at the decoder at time t given by B^t. Then the decoder process is described by

$$Y_t = d(t, Y^{t-1}, B^t) \equiv Y_t(Y^{t-1}, B^t), \quad t = 0, 1, \ldots \tag{3.2.7}$$

where $d(t, .)$ is a measurable function called the deterministic decoder function at time t.

At time $t = 0$, we may consider two scenarios for the information available to the encoder. The first scenario assumes no available feedback information so that A_0 is a deterministic function of X_0, i.e. $A_0 = A_0(X_0)$, and hence $\mathbb{B}(B^{-1}) = \{\emptyset, \Omega\}$ is the trivial σ-field. The second scenario assumes that A_0 is a deterministic function of X_0 and B^{-1} is stationary process having an invariant measure. One may also consider a third scenario, in which $A_0 = A_0(x_0, y_0)$, where y_0 is a fixed realization of Y_0.

Therefore, the measurable function $f(t, ., .)$ and $h(t, ., .)$ together with the alphabet spaces $\mathscr{X}_t, \mathscr{A}_t, \mathscr{B}_t, \mathscr{Y}_t$ describe the communication system of Figure 3.2.1. Often it is desirable to describe the encoder function by

$$\begin{aligned} e^t(.) &\stackrel{\triangle}{=} \Big\{e(0, ., .), e(1, ., .), \ldots, e(t, ., .)\Big\} \\ &\equiv \Big\{A_0(.), A_1(.), \ldots, A_t(.)\Big\} \equiv A^t(.), \quad t = 0, 1, \ldots \end{aligned} \tag{3.2.8}$$

Christos K. Kourtellaris

called the encoder laws or strategies, and the decoder function by

$$\begin{aligned} d^t(.) &\triangleq \Big\{d(0,.,.), d(1,.,.), \ldots, d(t,.,.)\Big\} \\ &\equiv \Big\{Y_0(.), Y_1(.), \ldots, Y_t(.)\Big\} \equiv Y^t(.),\ t = 0,1,\ldots \end{aligned} \tag{3.2.9}$$

based on the available information entering as inputs to the strategies. The available information to the encoder and the decoder is often called “information structure”.

For a given encoder strategy

$$A^t(.) \triangleq \{A_t : t = 0,1,\ldots\}$$

and decoder strategy

$$Y^t(.) \triangleq \{Y_t : t = 0,1,\ldots\}$$

the processes

$$X_t, A_t, B_t, Y_t : t \in \mathbb{N}$$

can be expressed by successive substitution of equations (3.2.1), (3.2.2), as follows.

$$\begin{aligned} X_0 &= w_0 \\ A_0 &= e(0, w_0, B^{-1}) \equiv E_0(w_0, B^{-1}) \\ B_0 &= h(0, B^{-1}, A_0, v_0) \equiv H_0(w_0, B^{-1}, v_0) \\ Y_0 &= d(0, Y^{-1}, B_0, B^{-1}) \equiv D_0(Y^{-1}, H_0(w_0, B^{-1}, v_0), B^{-1}) \\ X_1 &= f(1, w_0, B^{-1}, H_0(w_0, B^{-1}, v_0), w_1) \equiv F^{A_0}(w_0, B^{-1}, v_0, w_1) \\ A_1 &= e(1, w_0, f(1, w_0, B^{-1}, H_0(w_0, B^{-1}, v_0), w_1), B^{-1}, B_0, A_0) \equiv E_1^{A_0}(w^1, v_0, B^{-1}) \\ &\ \vdots \end{aligned} \tag{3.2.10}$$

The probability measure induced on the space $(\mathcal{X}_{0,t} \times \mathcal{A}_{0,t} \times \mathcal{B}_{0,t} \times \mathcal{Y}_{0,t},\ \mathbb{B}(\mathcal{X}_{0,t}) \times \mathbb{B}(\mathcal{A}_{0,t}) \times \mathbb{B}(\mathcal{B}_{0,t}) \times \mathbb{B}(\mathcal{Y}_{0,t}))$ by the functions of recursions (3.2.10) for $i = 0,1,\ldots,t$ is indicated by $\mathbb{P}^{A^t,Y^t}$ or $\mathbb{P}^{e^t,d^t}$. Expectation with respect to the measure $\mathbb{P}^{A^t,Y^t}$ will be indicated by $\mathbb{E}^{A^t,Y^t}[.]$ or $\mathbb{E}^{e^t,d^t}[.]$. We shall often omit the index ”t” and write $\mathbb{E}^{A,Y}[.]$, and when clear from the context omit the dependence of the measures on the strategies. Specifically, for a measure function $\ell : \mathcal{X}_t \times \mathcal{B}_t \mapsto \mathbb{R}$ we write

$$\mathbb{E}^{A^t(.),Y^t(.)}[\ell(X_t, B_t)] = \int_{\mathcal{X}_t \times \mathcal{B}_t} \ell(x_t, b_t) P^{A,Y}_{X_t,B_t}(dx_t, db_t) = \int_{\Omega} \ell(F^A_t(\omega), H^A_t(\omega)) \mathbb{P}(\omega)$$

Christos K. Kourtellaris

Clearly under the strategy A^t, the source and channel model are used to define the conditional probability

$$\begin{aligned}&\mathbb{P}\{\omega : X_t(\omega) \in J, B_t(\omega) \in K | X^{t-1}, B^{t-1}\}\\ &\qquad \equiv P^A_{X_t,B_t|X^{t-1},B^{t-1}}\Big(X_t \in J, B_t \in K | X^{t-1}, B^{t-1}\Big),\ \ \forall\, J \in \mathbb{B}(\mathscr{X}_t),\ K \in \mathbb{B}(\mathscr{B}_t)\end{aligned}$$

The following illustrates how stochastic Kernels will be used in connection to conditional distribution.

$$\begin{aligned}P^A_{X_{t+1},B_{t+1}|X^t,B^t}(X_{t+1} \in J, B_{t+1} \in K | X^t, B^t) &= P_{t+1}(J \times K; X^t, B^t, A^t(X^t, B^{t-1})),\ \forall\, J \in \mathbb{B}(\mathscr{X}_{t+1}),\\ &\qquad K \in \mathbb{B}(\mathscr{B}_{t+1})\\ P^A_{X_{t+1}|X^t,B^t}(X_{t+1} \in J | X^t, B^t) &= P_{t+1}(J; X^t, B^t), \qquad \forall J \in \mathbb{B}(\mathscr{X}_{t+1})\\ P^A_{B_{t+1}|X^{t+1},B^t}(B_{t+1} \in K | X^{t+1}, B^t) &= P_{t+1}(K; X^{t+1}, B^t, A^{t+1}(X^{t+1}, B^t)), \forall K \in \mathbb{B}(\mathscr{B}_{t+1})\end{aligned}$$

We often describe the source, encoder, channel, decoder by stochastic Kernels.

3.2.1 Definition of Subsystems

Given the communication block diagram of Figure 3.2.1, we define its different blocks below, by emphasizing on the information structures of each processing block.

Generalized Information Source:
The generalized information source is a sequence of stochastic Kernels

$$\Big\{P_j(dx_j; x^{j-1}, b^{j-1}, a^{j-1}) \in \mathscr{K}(\mathscr{X}_j; \mathscr{X}_{0,j-1} \times \mathscr{B}_{0,j-1} \times \mathscr{A}_{0,j-1}) :\ \ j \in \mathbb{N}^n\Big\} \quad (3.2.11)$$

In most communication applications the following Markov chain holds.

$$(B^{j-1}, A^{j-1}) \leftrightarrow X^{j-1} \leftrightarrow X_j,\ \ \forall j \in \mathbb{N}^n$$

The reason we consider the general definition of (3.2.11) is to include controlled sources, in which the control process is applied using feedback, either from channel output or decoder output.

Christos K. Kourtellaris

Channel Encoder:

The encoder is a sequence of stochastic Kernels

$$\Big\{P_j(da_j;a^{j-1},x^j,b^{j-1})\ \in \mathscr{K}(\mathscr{A}_j;\mathscr{A}_{0,j-1}\times\mathscr{X}_{0,j}\times\mathscr{B}_{0,j-1}):\quad j\in\mathbb{N}^n\Big\} \tag{3.2.12}$$

Based on the information structure available at the encoder, the encoder strategies are classified as follows.

Definition 3.3. (Encoder Strategies)

1. *Randomized Feedback.*
   The set of randomized feedback encoders is denoted by $\mathscr{E}^{RF}[0,n]\subseteq\{\mathscr{K}(\mathscr{A}_j;\mathscr{A}_{0,j-1}\times\mathscr{X}_{0,j}\times\mathscr{B}_{0,j-1}):j=0,1,\ldots,n\}$. For each time $j\in\mathbb{N}^n$, the randomized encoder depends on the entire history of the source symbol $X^j=x^j$, in addition to $B^{j-1}=b^{j-1},A^{j-1}=a^{j-1}$.

2. *Randomized Markov.*
   The set of randomized Markov encoder strategies is denoted by $\mathscr{E}^{RM}[0,n]\subseteq\{\mathscr{K}(\mathscr{A}_j;\mathscr{X}_j\times\mathscr{B}_{0,j-1}):j=0,1,\ldots,n\}$. For each time $j\in\mathbb{N}^n$, the randomized encoder depends on the symbol $X_j=x_j$ in addition to $B^{j-1}=b^{j-1}$. Thus, such encoders satisfy

$$P_j(da_j;a^{j-1},x^j,b^{j-1})=P_j(da_j;x_j,b^{j-1})-a.a\ (a^{j-1},x^j,b^{j-1}),\ \ \forall j\in\mathbb{N}^n$$

3. *Randomized Open loop.*
   The set of randomized open loop encoder strategies is denoted by $\mathscr{E}^{ROL}[0,n]\subseteq\{\mathscr{K}(\mathscr{A}_j;\mathscr{A}^{j-1},\mathscr{X}^j):j=0,1,\ldots,n\}$. For each time $j\in\mathbb{N}^n$, the randomized encoder depends on the symbols $X^j=x^j$, $A^{j-1}=a^{j-1}$ and not on $B^{j-1}=b^{j-1}$. Note that the Randomized Open Loop Markov strategies with respect to the source is a subclass of $\mathscr{E}^{ROL}[0,n]$.

   Deterministic encoders are sequences of delta measures and hence they are identified by sequences of measurable functions

$$\Big\{e_j:\mathscr{A}_{0,j-1}\times\mathscr{X}_{0,j}\times\mathscr{B}_{0,j-1}\mapsto\mathscr{A}_j:a_j=e_j(a^{j-1},x^j,b^{j-1}),\ \ j\in\mathbb{N}^n\Big\}$$

Christos K. Kourtellaris

4. *Deterministic Feedback.*

   The set of deterministic feedback encoder strategies is denoted by $\mathscr{E}^{DF}[0,n] \subseteq \mathscr{K}^{RF}$ $[0,n]$. For each time $j \in \mathbb{N}^n$, $A_j(.)$ is $\mathbb{B}(X^j) \times \mathbb{B}(A^{j-1}) \times \mathbb{B}(B^{j-1})$ measurable. Thus, for each realization $B^{j-1} = b^{j-1}, A^{j-1} = a^{j-1}$ the encoder strategy $e_j(\cdot,\cdot,\cdot)$ is a measurable function of the past realizations $X^j = x^j$. Thus, such an encoder is of the form

   $$\Big\{e_j : \mathscr{X}_{0,j} \times \mathscr{A}_{0,j-1} \times \mathscr{B}_{0,j-1} \mapsto \mathscr{A}_j : a_j = e_j(x^j, a^{j-1}, b^{j-1}), \quad j \in \mathbb{N}^n\Big\}$$

5. *Deterministic Markov.*

   The set of deterministic Markov encoder strategies is denoted by $\mathscr{E}^{DM}[0,n] \subseteq \mathscr{K}^{RM}$ $[0,n]$. For each time $j \in \mathbb{N}^n$, $A_j(.)$ is $\mathbb{B}(X_j) \times \mathbb{B}(B^{j-1})$ measurable. Thus, such an encoder is of the form

   $$\{e_j : \mathscr{X}_j \times \mathscr{B}_{0,j-1} \mapsto \mathscr{A}_j : a_j = e_j(x_j, b^{j-1}), \quad j \in \mathbb{N}^n\}$$

6. *Deterministic Open loop.*

   The set of deterministic open loop encoder strategies is denoted by $\mathscr{E}^{DOL}[0,n] \subseteq \mathscr{K}^{ROL}[0,n]$. For each time $j \in \mathbb{N}^n$, A_j is $\mathbb{B}(X_j)$ measurable. Thus, such an encoder is a sequence of measurable functions of the form

   $$\Big\{e_j : \mathscr{X}_{0,j} \times \mathscr{A}_{0,j-1} \mapsto \mathscr{A}_j : a_j = e_j(x^j, a^{j-1}), \quad j \in \mathbb{N}^n\Big\}$$

   Note that, deterministic Open Loop Markov with respect to the source is a subset of $\mathscr{E}^{DOL}[0,n]$, of the form

   $$\Big\{e_j : \mathscr{X}_j \mapsto \mathscr{A}_j : a_j = e_j(x_j), \quad j \in \mathbb{N}^n\Big\}$$

Feedback strategies, randomized or deterministic, can be used when the channel allows feedback between its output and its input, while open loop strategies, randomized or deterministic are used when no channel feedback is allowed.

Communication Channel with Memory:

A communication channel is a sequence of stochastic Kernels

$$\Big\{P_j(db_j; b^{j-1}, a^j, x^j) \in \mathscr{Q}(\mathscr{B}_j; \mathscr{B}_{0,j-1} \times \mathscr{A}_{0,j} \times \mathscr{X}_{0,j}) : \quad j \in \mathbb{N}^n\Big\} \tag{3.2.13}$$

Christos K. Kourtellaris

Note that, often the channel takes the simplified form

$$P_j(db_j;b^{j-1},a^j,x^j)=P_j(db_j;b^{j-1},a^j)-a.a.\ \ (b^{j-1},a^j,x^j),\quad j\in\mathbb{N}^n$$

especially for channels in which the information capacity is defined between the input and the output of the channel.

A channel, with or without feedback, is called memoryless channel if and only if the following Markov chain holds.

$$(B^{j-1},X^j,A^{j-1})\leftrightarrow A_j\leftrightarrow B_j,\quad \forall j\in\mathbb{N}^n \tag{3.2.14}$$

Any channel with finite input and output alphabets satisfying (3.2.14) is called Discrete Memoryless Channel (DMC).

Channel Decoder:
The decoder is a sequence of stochastic Kernels

$$\Big\{P_j(dy_j;y^{j-1},b^j)\ \in\mathscr{K}(\mathscr{Y}_j:\mathscr{Y}_{0,j-1}\times\mathscr{B}_{0,j}):\quad j\in\mathbb{N}^n\Big\} \tag{3.2.15}$$

Deterministic decoders are sequences of delta measures identified by sequences of measurable functions

$$\Big\{d_j:\mathscr{Y}_{0,j-1}\times\mathscr{B}_{0,j}\mapsto\mathscr{Y}_j:y_j=d_j(y^{j-1},b^j),\quad j\in\mathbb{N}^n\Big\}$$

Next, we give the definition of a channel code.

Definition 3.4. An $\{(n,M_n,\varepsilon_n):n=0,1,\ldots\}$ code sequence for the channel with feedback consists of the following.

1. A set of messages $\mathscr{M}_n\overset{\triangle}{=}\Big\{1,2,\ldots,M_n\Big\}$ and a class of encoders (deterministic or random) measurable mappings $\Big\{\varphi_i:\mathscr{M}_n\times\mathscr{B}^{i-1}\mapsto A_i:i=0,1,\ldots,n-1\Big\}$ that transforms each message $X\in\mathscr{M}_n$ into a channel input $A^{n-1}\in\mathscr{A}_{0,n-1}$ of length n. For example, $\varphi\in\mathscr{E}^{DM}[0,n-1]$ is the set of encoding strategies $\{\varphi_i:i=0,1,\ldots,n-1\}$ such that $\{A_i=\varphi_i(X,B^{i-1}):i=0,1,\ldots,n-1\}$. Note that the more general strategies satisfy $\{\bar{\varphi}_i(X,A^{i-1},B^{i-1}):i=0,1,\ldots,n-1\}=\{\varphi_i(X,B^{i-1}):i=0,1,\ldots,n-1\}$. For $x\in\mathscr{M}_n$ we call $u_x\in\mathscr{A}_{0,n}$, $u_x=(\varphi_0(x,b^{-1}),\varphi_1(x,b^0),\varphi_2(x,b^1),\ldots,\varphi_n(x,b^{n-1}))$ the codeword for message $x\in\mathscr{M}_n$ and code $\mathscr{C}_n=(u_1,u_2,\ldots,u_{M_n})$ the code. Thus, when the transmitter wishes to send the message $x\in\mathscr{M}_n$, it transmits the codeword u_x of the current message x.

Christos K. Kourtellaris

2. A class of decoder measurable mappings $d^n : \mathscr{B}_{0,n-1} \to \mathscr{M}_n$, $Y = d^n(B^{n-1})$, such that the average probability of decoding error satisfies

$$P_e^n \triangleq \frac{1}{M_n} \sum_{x \in \mathscr{M}_n} Prob(Y \neq x | X = x) = \varepsilon_n$$

Thus, the receiver which has access to the realization $b^{n-1} \in \mathscr{B}_{0,n-1}$ can partition $\mathscr{B}_{0,n-1}$ into $\mathscr{M}_n$ disjoint subsets, $\mathscr{B}_{0,n} = \mathscr{D}_1 \cup \mathscr{D}_2 \cup \ldots \mathscr{D}_{M_n}$, $\mathscr{D}_i \cap \mathscr{D}_j = \emptyset, \forall i \neq j$, before the start of the transmission operation, and then decide that message $x \in \mathscr{M}_n$ is transmitted if $\mathscr{B}^{n-1} \in \mathscr{D}_x$. Hence, $\mathscr{D}_x$, $x \in \mathscr{M}_n$ is the decoding region of message $x \in \mathscr{M}_n$, which may be specified via the typical set decoding, maximum-likelihood set decoding, e.t.c.. With respect to this decoder, the average probability of error is also expressed as

$$P_e^n = \frac{1}{M_n} \sum_{x \in \mathscr{M}_n} \text{Prob}(b^{n-1} \in \mathscr{D}_x^c | u_x) \tag{3.2.16}$$

Next, we give the definition of achievable rate.

Definition 3.5 (Operational Capacity)**.**

(a) R is an achievable rate if there exists an $\{(n, M_n, \varepsilon_n) : n = 0, 1, \ldots\}$ code sequence satisfying $\lim_{n\to\infty} \varepsilon_n = 0$ and $\liminf_{n\to\infty} \frac{1}{n} \log M_n \geq R$. The supremum of all achievable rates R is defined as the capacity.

(b) R is an ε-achievable rate if there exists an $\{(n, M_n, \varepsilon_n) : n = 0, 1, \ldots\}$ code sequence satisfying $\limsup_{n\to\infty} \varepsilon_n \leq \varepsilon$ and $\liminf_{n\to\infty} \frac{1}{n} \log M_n \geq R$. The supremum of all ε achievable rates R for all $0 \leq \varepsilon < 1$ is defined as the ε-channel capacity.

Direct and converse coding theorems that link the operational Definition 3.5 to its informational definition are derived in [39] for channels without feedback using mutual information. For channels with feedback, that link the operational Definition 3.5 to its informational definition are derived in [41] and [19] using mutual information between X^n and B^n; when $X^n \leftrightarrow (A^n, B^{n-1}) \leftrightarrow B_n$, $n = 0, 1, \ldots$ holds, coding theorems are derived using directed information from A^n to B^n in [18, 75].

Over the years Walrand-Varaiya [81] and Teneketzis [78] treated the problem of optimizing a given pay-off, for various classes of sources and channels, over encoder-decoder strategies. However the considered pay-offs are not related to any of the information theoretic measures, while they often assume DMC.

Christos K. Kourtellaris

3.2.2 Directed Information

Given a source, a channel, and encoder and decoder strategies, we define the joint probability measure on $\mathscr{X}_{0,n} \times \mathscr{A}_{0,n} \times \mathscr{B}_{0,n} \times \mathscr{Y}_{0,n}$ using stochastic Kernels as follows.

$$\begin{aligned}P_{0,n}(dx^n, da^n, db^n, dy^n) = \otimes_{i=0}^n \Big(& P_i(dy_i; y^{i-1}, b^i) \otimes P_i(db_i; b^{i-1}, a^i, x^i) \\ & \otimes P_i(da_i; b^{i-1}, a^{i-1}, x^i) \otimes P_i(dx_i; x^{i-1}, b^{i-1}, a^{i-1})\Big)\end{aligned} \tag{3.2.17}$$

To obtain (3.2.17), we have assumed the following Markov chains hold.

$$Y^{i-1} \leftrightarrow (X^{i-1}, B^{i-1}, A^{i-1}) \leftrightarrow X_i, \quad \forall\, i \in \mathbb{N}^n \tag{3.2.18}$$

$$Y^{i-1} \leftrightarrow (X^{i}, B^{i-1}, A^{i-1}) \leftrightarrow A_i, \quad \forall\, i \in \mathbb{N}^n \tag{3.2.19}$$

$$Y^{i-1} \leftrightarrow (X^{i}, B^{i-1}, A^{i}) \leftrightarrow B_i, \quad \forall\, i \in \mathbb{N}^n \tag{3.2.20}$$

$$(X^{i}, A^{i}) \leftrightarrow (Y^{i-1}, B^{i}) \leftrightarrow Y_i, \quad \forall\, i \in \mathbb{N}^n \tag{3.2.21}$$

The right hand side of (3.2.17) is further simplified by considering specific channels and sources, and specific information structures for the encoder and the decoder. For example, it is often the case that the channel satisfies

$$P_i(db_i; b^{i-1}, a^i, x^i) = P_i(db_i; b^{i-1}, a^i) - a.a.\ \ (b^{i-1}, a^i, x^i), \ \ \forall\, i \in \mathbb{N}^n$$

However, there are examples in which capacity cannot be defined from the channel input to the channel output [19].

Feedback channels, and in general network information theory utilizes information theoretic measures which are directional. Here, we provide an elaborate discussion on directional information starting with definition introduced by Marko [51], and subsequently developed by Massey [52].

Suppose we are given the two distributions $P_{X^n}(dx^n)$ and $P_{B^n|X^n}(db^n|dx^n)$, which uniquely define $P_{X^n,B^n}(dx^n, db^n)$ and $P_{B^n}(db^n)$. The definition of Shannon's self-mutual information $i(X^n; B^n)$ between two sequences X^n and B^n, is defined via the information density (logarithm of a Radon-Nykodym derivative) by

$$i(x^n; b^n) \triangleq \log \frac{P_{X^n,B^n}(dx^n, db^n)}{P_{X^n}(dx^n) \otimes P_{B^n}(db^n)} \tag{3.2.22}$$

Christos K. Kourtellaris

Note that $P_{B^n|X^n} << P_{B^n} \times P_{X^n}$ if and only if $P_{B^n|X^n}(.|x^n) << P_{B^n}(.)$-a.s., thus

$$\frac{P_{X^n,B^n}(dx^n,db^n)}{P_{X^n}(dx^n)\otimes P_{B^n}(db^n)} = \frac{P_{B^n|X^n}(db^n|x^n)}{P_{B^n}(db^n)} \text{ a.s.} \tag{3.2.23}$$

By taking the average of $i(x^n;b^n)$ over all realizations with respect to the joint distribution $P_{X^n,B^n}(dx^n,db^n)$, we obtain the expression of mutual information between X^n and B^n, as follows:

$$\begin{aligned} I(X^n;B^n) &= \mathbb{D}(P_{X^n,B^n}||P_{X^n}\times P_{B^n}) \\ &= \int_{\mathscr{X}_{0,n}\times\mathscr{B}_{0,n}} \log\Big(\frac{P_{B^n|X^n}(db^n|x^n)}{P_{B^n}(db^n)}\Big) P_{B^n|X^n}(db^n|x^n)\otimes P_{X^n}(dX^n) \\ &\equiv \mathbb{I}_{X^n;B^n}(P_{X^n},P_{B^n|X^n}) \end{aligned} \tag{3.2.24}$$

Hence, mutual information is a functional of two distributions $\{P_{B^n|X^n}, P_{X^n}\}$, and thus the adopted notation $\mathbb{I}_{X^n;B^n}(P_{X^n},P_{B^n|X^n})$. Since $i(X^n = x^n, B^n = b^n)$ is interpreted as the information provided about $X^n = x^n$ by observing $B^n = b^n$, then $I(X^n;B^n)$ is the average information that B^n provides about X^n with respect to being independent process. In view of the symmetry $I(X^n;B^n) = I(B^n;X^n)$, mutual information is also the average information B^n provides about X^n over the channel $X^n \Rightarrow P_{B^n|X^n} \Rightarrow B^n$, or X^n provides about B^n over the channel $B^n \Rightarrow P_{X^n|B^n} \Rightarrow X^n$.

Suppose we are given the families of conditional distributions $\{P_{B_j|B^{j-1},X^j}(db_j|b^{j-1},x^j) : j = 0,1,\ldots\}$ and $\{P_{X_j|X^{j-1},B^{j-1}}(dx_j|x^{j-1},b^{j-1}) : j = 0,1,\ldots\}$, which uniquely define the joint and marginal distributions $P_{X^n,B^n}(dx^n,db^n)$, $P_{X^n}(dx^n)$ and $P_{B^n}(db^n)$. Clearly, the self-mutual information admits the decomposition

$$\begin{aligned} i(X^n;B^n) &= \log\Big(\otimes_{j=0}^{n} \frac{P_{B_j|B^{j-1},X^j}(db_j|b^{j-1},x^j)\otimes P_{X_j|X^{j-1},B^{j-1}}(dx_j|x^{j-1},b^{j-1})}{P_{B_j|B^{j-1}}(db_j|b^{j-1})\otimes P_{X_j|X^{j-1}}(dx_j|x^{j-1})}\Big) \\ &= \log\Big(\otimes_{j=0}^{n} \frac{P_{B_j|B^{j-1},X^j}(db_j|b^{j-1},x^j)}{P_{B_j|B^{j-1}}(db_j|b^{j-1})}\Big) \\ &\quad +\log\Big(\otimes_{j=0}^{n} \frac{P_{X_j|X^{j-1},B^{j-1}}(dx_j|x^{j-1},b^{j-1})}{P_{X_j|X^{j-1}}(dx_j|x^{j-1})}\Big) \\ &= i(X^n \rightarrow B^n) + i(X^n \leftarrow B^n) \end{aligned}$$

Christos K. Kourtellaris

where

$$
\begin{aligned}
i(X^n \rightarrow B^n) &\triangleq \log\Big(\otimes_{j=0}^{n} \frac{P_{B_j|B^{j-1},X^j}(db_j|b^{j-1},x^j)}{P_{B_j|B^{j-1}}(db_j|b^{j-1})}\Big) && (3.2.25)\\
i(X^n \leftarrow B^n) &\triangleq \log\Big(\otimes_{j=0}^{n} \frac{P_{X_j|X^{j-1},B^{j-1}}(dx_j|x^{j-1},b^{j-1})}{P_{X_j|X^{j-1}}(dx_j|x^{j-1})}\Big) && (3.2.26)
\end{aligned}
$$

Define

$$
\begin{aligned}
I(X^n \rightarrow B^n) &\triangleq \sum_{i=0}^{n} I(X^i;B_i|B^{i-1}) && (3.2.27)\\
I(X^n \leftarrow B^n) &\triangleq \sum_{i=0}^{n} I(B^{i-1};X_i|X^{i-1}) && (3.2.28)
\end{aligned}
$$

Taking the expectation with respect to the joint distribution yields

$$
\begin{aligned}
I(X^n;B^n) &= \mathbb{E}\Big\{i(X^n \rightarrow B^n)\Big\} + \mathbb{E}\Big\{i(X^n \leftarrow B^n)\Big\} && (3.2.29)\\
&= I(X^n \rightarrow B^n) + I(X^n \leftarrow B^n) && (3.2.30)
\end{aligned}
$$

Note that $I(X^n \rightarrow B^n)$ is the directed information in the direction $X^n \rightarrow B^n$ over a sequence of causal channels $(X^i,B^{i-1}) \Rightarrow P_{B_i|B^{i-1},X^i} \Rightarrow B_i,\ i=0,1,\ldots,n$, the feedforward information. On the other hand, $I(X^n \leftarrow B^n)$ is the directed information in the direction $X^n \leftarrow B^n$ over a sequence of causal channels $(X^{i-1},B^{i-1}) \Rightarrow P_{X_i|X^{i-1},B^{i-1}} \Rightarrow X_i,\ i=0,1,\ldots,n$, the called feedback information.

Next, we give three interpretations of (3.2.27) and (3.2.28) which are important in our subsequent analysis.

• *Representation 1*
The next representation is given in [51, 52] .

$$
\begin{aligned}
I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} \int \log \frac{P_{B_i|B^{i-1},X^i}(db_i;b^{i-1},x^i)}{P_{B_i|B^{i-1}}(db_i;b^{i-1})} P_{B^i,X^i}(db^i,dx^i)\\
&= \sum_{i=0}^{n} \int \mathbb{D}(P_{B_i|B^{i-1},X^i}(.|b^{i-1},x^i)||P_{B_i|B^{i-1}}(.|b^{i-1}))P_{X_i|X^{i-1},B^{i-1}}(dx_i|x^{i-1},b^{i-1})\\
&\quad \otimes_{j=0}^{i-1}\Big(P_{B_j|B^{j-1},X^j}(b_j|b^{j-1},x^j) \otimes P_{X_j|X^{j-1},B^{j-1}}(dx_j|x^{j-1},b^{j-1})\Big)\\
&\equiv \mathbb{I}_{X^n \rightarrow B^n}(P_{X_i|X^{i-1},B^{i-1}}, P_{B_i|B^{i-1},X^i} : i \in \mathbb{N}^n) && (3.2.31)
\end{aligned}
$$

Christos K. Kourtellaris

This representation shows that $I(X^n \rightarrow B^n)$ is a function of two causal conditional distributions, $\{P_{X_i|X^{i-1},B^{i-1}}, P_{B_i|B^{i-1},X^i} : i \in \mathbb{N}^n\}$, and it is consistent with the interpretation given above. Note, that unlike mutual information, $\mathbb{I}_{X^n;Y^n}(P_{X^n}, P_{Y^n|X^n})$, which is a functional of two distributions, $\{P_{X^n}, P_{Y^n|X^n}\}$, and inherits several of its properties from the properties of relative entropy, such as lower semicontinuity with respect to P_{X^n} for fixed $P_{Y^n|X^n}$ and vice versa, convexity with respect to $P_{Y^n|X^n}$ for fixed P_{X^n}, and concavity with respect to P_{X^n} for fixed $P_{Y^n|X^n}$, these properties are not easily extended to the directed information functional $\mathbb{I}_{X^n \rightarrow B^n}(P_{X_i|X^{i-1},B^{i-1}}, P_{B_i|B^{i-1},X^i} : i \in \mathbb{N}^n)$. However, in [10, 73, 74] all these properties are extended to the directional information using the following alternative definition.

Define the $(n+1)$-fold convolution measures by

$$\overrightarrow{P}_{B^n|X^n}(db^n|x^n) \overset{\triangle}{=} \otimes_{i=0}^{n} P_{B_i|B^{i-1},X^i}(db_i|b^{i-1},x^i) \tag{3.2.32}$$

$$\overleftarrow{P}_{X^n|B^{n-1}}(dx^n|b^{n-1}) \overset{\triangle}{=} \otimes_{i=0}^{n} P_{X_i|X^{i-1},B^{i-1}}(dx_i|x^{i-1},b^{i-1}) \tag{3.2.33}$$

It is known [10, 73, 74] that $\overrightarrow{P}_{B^n|X^n}(db^n|x^n)$ uniquely defines $\{P_{B_i|B^{i-1},X^i}(db_i|b^{i-1},x^i) : i \in \mathbb{N}^n\}$ and vice-versa, and similarly for $\overleftarrow{P}_{X^n|B^{n-1}}(dx^n|b^{n-1})$. Then, an equivalent expression for $I(X^n \rightarrow B^n)$ is the following.

$$\begin{aligned} I(X^n \rightarrow B^n) &= \int \log\left(\frac{\overrightarrow{P}_{B^n|X^n}(db^n|x^n)}{P_{B^n}(db^n)}\right) \overrightarrow{P}_{B^n|X^n}(db^n|x^n) \otimes \overleftarrow{P}_{X^n|B^{n-1}}(dx^n|b^{n-1}) \\ &\equiv \mathbb{I}_{X^n \rightarrow B^n}(\overleftarrow{P}_{X^n|B^{n-1}}, \overrightarrow{P}_{B^n|X^n}) \end{aligned}$$

The functional $\mathbb{I}_{X^n \rightarrow B^n}(\overleftarrow{P}_{X^n|B^{n-1}}, \overrightarrow{P}_{B^n|X^n})$ inherits all properties of mutual information, because the subset of conditional distributions on $(\mathscr{B}_{0,n}, \mathbb{B}(\mathscr{B}_{0,n}))$ and $(\mathscr{X}_{0,n}, \mathbb{B}(\mathscr{X}_{0,n}))$ defined by (3.2.32) and (3.2.33), respectively, are convex sets. These results are found in [10, 73, 74].

The representation (3.2.31) has a very interesting interpretation in terms of a controlled conditional distribution as follows. Suppose the channel depends on the input via the most recent symbol, that is

$$P_i(db_i; b^{i-1}, x^i) = P_i(db_i; b^{i-1}, x_i) - a.a. \ \ (b^{i-1}, x^i), \ \ \forall i \in \mathbb{N}^n \tag{3.2.34}$$

Christos K. Kourtellaris

Then, $I(X^n \rightarrow B^n)$ reduces to

$$\begin{aligned} I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} I(X_i;B_i|B^{i-1}) \\ &= \sum_{i=0}^{n} \int \log\left(\frac{P_{B_i|B^{i-1},X_i}(db_i|b^{i-1},x_i)}{P_{B_i|B^{i-1}}(db_i|b^{i-1})}\right) P_{B_i|B^{i-1},X_i}(db_i|b^{i-1},x_i) \\ &\quad \otimes P_{X_i|B^{i-1}}(dx_i|b^{i-1}) \otimes P_{B^{i-1}}(db^{i-1}) \end{aligned} \tag{3.2.35}$$

Consider the case when the sequence of channels $\{P_{B_i|B^{i-1},X_i}(db_i|b^{i-1},x_i) : i \in \mathbb{N}^n\}$ is fixed, and $\{P_{X_i|B^{i-1}}(dx_i|b^{i-1}) : i \in \mathbb{N}^n\}$ is the variable designed to maximize (3.2.35). Then, by Bayes rule

$$\begin{aligned} P_{B_i|B^{i-1}}(db_i|b^{i-1}) &= \int_{\mathscr{X}_i} P_{B_i|B^{i-1},X_i}(db_i|b^{i-1},x_i) \otimes P_{X_i|B^{i-1}}(dx_i|b^{i-1}) \\ &= \int_{\mathscr{X}_i} P_{B_i|B^{i-1},X_i}(db_i|b^{i-1},x_i) \otimes \pi_i(dx_i|b^{i-1}) \\ &\equiv P^{\pi_i}_{B_i|B^{i-1}}(db_i|b^{i-1}) \end{aligned} \tag{3.2.36}$$

Clearly, (3.2.36) demonstrates that $\{P^{\pi_i}_{B_i|B^{i-1}}(db_i|b^{i-1}) : \forall i \in \mathbb{N}^n\}$ is the controlled process controlled by the conditional distribution $\{\pi_i(dx_i|b^{i-1}) : \forall i \in \mathbb{N}^n\}$.

The interpretation is that, in the calculation of channel capacity via maximization of directed information, the probability distribution of B_i given past channel outputs B^{i-1}, namely $P_{B_i|B^{i-1}}(db_i|b^{i-1})$ is the controlled process, and the probability distribution of X_i given the past channel outputs B^{i-1} namely $P_{X_i|B^{i-1}}(dx_i|b^{i-1}) = \pi_i(dx_i|b^{i-1})$ is the control process $\forall i \in \mathbb{N}^n$. The process $\{\pi_i(dx_i;b^{i-1}) : i = 0,\ldots,n\}$ is induced by the channel input distribution in a specific way, depending on whether the channel is used with or without feedback. Therefore, by the additivity property of the pay-off (3.2.35), we can derive a dynamic programming equation to determine the sequence of $\{\pi_i(dx_i|b^{i-1}) : \forall i \in \mathbb{N}^n\}$ which maximizes directed information. We shall revisit this observation in subsequent sections.

- *Representation 2*

$$\begin{aligned} I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} \int_{\mathscr{B}_{0,i}} \mathbb{D}(P_{X^i|B^i}(.|b^i)||P_{X^i|B^{i-1}}(.|b^{i-1}))P_{B^i}(db^i) \\ &= \sum_{i=0}^{n} \mathbb{E}\left\{\log \frac{P_{X^i|B^i}(dx^i|b^i)}{P_{X^i|B^{i-1}}(dx^i|b^{i-1})}\right\} \end{aligned} \tag{3.2.37}$$

Christos K. Kourtellaris

This representation shows that each term in (3.2.37) can be expressed as a relative entropy distance between the á posteriori distributions $\{P_{X^i|B^i}(dx^i|b^i), P_{X^i|B^{i-1}}(dx^i|b^{i-1})\}$, averaged over the distribution $P_{B^i}(db^i)$, $\forall\ i \in \mathbb{N}^n$. At each instant of time, $\log P_{X^i|B^i}(dx^i|b^i)$ $-\log P_{X_i|B^{i-1}}(dx^i|b^{i-1})$ may be viewed as the new information gained about the random variable X^i, by receiving an additional observation $B_i = b_i$, given all previous observations $B^{i-1} = b^{i-1}$. Moreover by writing equation (3.2.37) as

$$I(X^n \rightarrow B^n) = \sum_{i=0}^{n} \int_{\mathscr{B}_{0,i}} \mathbb{D}(P_{X^i|B^i}(.|b^i)||P_{X^i|B^{i-1}}(.|b^i))P_{B^i}(db^i) \tag{3.2.38}$$

then each term in the right hand side of equation (3.2.38) can be viewed as the averaged new information measured in a sequence of relative entropies between $P_{X^i|B^i}(.|b^i)$ and $P_{X^i|B^{i-1}}(.|b^{i-1})$ about X^i, by receiving an additional observation $B_i = b_i$ given all passed observations $B^{i-1} = b^{i-1}$, for all $i \in \mathbb{N}^n$.

Note that by assuming a channel of the form (3.2.34), then we obtain the simplified form

$$\begin{aligned} I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} I(X_i; B_i|B^{i-1}) \\ &= \sum_{i=0}^{n} \int_{\mathscr{B}_{0,i}} \mathbb{D}(P_{X_i|B^i}(.|b^i)||P_{X_i|B^{i-1}}(.|b^{i-1}))P_{B^i}(db^i) \end{aligned} \tag{3.2.39}$$

where

$$\begin{aligned} P_{B^i}(db^i) &= \otimes_{j=0}^{i} P_{B_j|B^{j-1}}(db_j|b^{j-1}) = \otimes_{j=0}^{i} \int_{\mathscr{X}_j} P_{B_j|B^{j-1},X_j}(db_j|b^{j-1}, x_j) \otimes \pi_j(dx_j|b^{j-1}) \\ &\equiv \otimes_{j=0}^{i} P^{\pi_j}_{B_j|B^{j-1}}(db_j|b^{j-1}) \end{aligned} \tag{3.2.40}$$

Thus, when the channels $\{P_{B_j|B^{j-1},X_j}(db_j \mid b^{j-1}, x_j) : i = 0, \ldots, n\}$ are fixed and $\{P_{X_j|B^{j-1}}(dx_j| b^{j-1}) : i = 0, \ldots, n\}$ are variable over which (3.2.39) is maximized, then the former is the controlled process and latter is the control process.

Christos K. Kourtellaris

• *Representation 3*

$$\begin{align} I(X^n \to B^n) &= \sum_{i=0}^{n} \Big\{ H(B_i|B^{i-1}) - H(B_i|B^{i-1},X^i) \Big\} \nonumber \\ &\leq \sum_{i=0}^{n} H(B_i|B^{i-1}) - \sum_{i=0}^{n} H(B_i|B^{i-1},X^n) \tag{3.2.41} \\ &= I(X^n;B^n) \tag{3.2.42} \end{align}$$

The interpretation of this representation is similar to the one above, since it denotes the reduction of uncertainty at each step given the message until up to that step. Note that inequality in (3.2.41) holds with equality if and only if the following Markov chain holds.

$$(X_{i+1},\ldots,X_n) \leftrightarrow (X^i,B^{i-1}) \leftrightarrow B_i, \qquad i = 0,1,\ldots \tag{3.2.43}$$

Since (3.2.43) is often valid (unless the source is affected by the past channel outputs, (such as control applications), its is clear that the information measure can be either $I(X^n;B^n)$ or $I(X^n \to B^n)$, and in this case

$$I(X^n;B^n) = I(X^n \to B^n) = \mathbb{I}_{X^n \to B^n}(P_{X^n}, \overrightarrow{P}_{B^n|X^n}) \tag{3.2.44}$$

Clearly, (3.2.44) is the information measure utilized to define $R^{na}_{0,n}(D)$ in Chapter 2. If this channel is used without feedback then (3.2.43) holds and (3.2.44) is valid.

3.3 Structural Properties of Encoders

In this section, we address the following issue.

- Identify general structural properties of encoders, for a given class of sources and channels with memory and feedback, which maximize the directed information from the source to the channel output.

Hence, this problem addresses the design of encoders and their properties, when the information capacity has an operational meaning.
The problem is stated below.

Problem 3.6. (*Maximizing Directed Information*)

Christos K. Kourtellaris

(a) Randomized Encoders
Given an admissible randomized class of encoders $\mathscr{E}^{RF}[0,n]$, find the structural properties of $\{P_j^*(da_j;a^{j-1},x^j,b^{j-1}): \ j \in \mathbb{N}^n\} \in \mathscr{E}^{RF}[0,n]$ which maximizes directed information

$$J_{0,n}^R(P_j^*: j=0,1,\ldots,n) \triangleq \sup_{\{P_j: j=0,1,\ldots,n\}\in\mathscr{E}^{RF}[0,n]} I(X^n \rightarrow B^n)$$

(b) Deterministic Encoders
Given an admissible deterministic class of encoders $\mathscr{E}^{DF}[0,n]$, find the structural properties of $\{e_j^*(a^{j-1},x^j,b^{j-1}): \forall j \ \in \mathbb{N}^n\} \in \mathscr{E}^{DF}[0,n]$ which maximizes directed information

$$J_{0,n}^D(e_j^*: j=0,1,\ldots,n) \triangleq \sup_{\{e_j: j=0,1,\ldots,n\}\in\mathscr{E}^{DF}[0,n]} I(X^n \rightarrow B^n)$$

We make the following comments regarding Problem 3.6.

1. Problem 3.6.(a) with $J_{0,\infty}^R \triangleq \liminf_{n\rightarrow\infty} \frac{1}{n+1} J_{0,n}^R$ is an infinite horizon encoder design, with respect to randomized strategies, under the assumption that the corresponding channel has operational meaning
2. Problem 3.6.(b) with $J_{0,\infty}^D \triangleq \liminf_{n\rightarrow\infty} \frac{1}{n+1} J_{0,n}^D$ is an infinite horizon encoder design with respect to deterministic strategies, under the assumption that the corresponding channel has operational meaning.

The reason we introduce randomized strategies is due to the fact that often existence of deterministic strategies is difficult to ensure [1]. Therefore, our aim is to understand the structural properties of the encoder, such as, symbol by symbol transmission is optimal, which implies that nothing can be gained by designing encoder which operates on block of source symbols at each transmission.

Consider Problem 3.6.(a) of maximizing directed information over the class of randomized strategies $\{P_{A_j;A^{j-1},X^j,B^{j-1}}(da_j;a^{j-1},x^j,b^{j-1}): j=0,1,\ldots,n\} \in \mathscr{E}^{RF}[0,n]$. An interesting question is to determine for a given channel and source, the information structure of the encoder over which $J_{0,n}^R$ should be optimized. To address such questions consider the pay-off expressed utilizing Representation 2 (3.2.37) of directed information as follows.

$$J_{0,n}^R(P_j: j=0,1,\ldots,n) = \sum_{i=0}^{n} \mathbb{E}\Big\{ \log \frac{P_{X^i|B^i}(dx^i|b^i)}{P_{X^i|B^{i-1}}(dx^i|b^{i-1})} \Big\}$$

Christos K. Kourtellaris

Since the pay-off is additive, to apply dynamic programming the first question is whether

$$\begin{aligned}&\mathbb{E}\Big\{\log\frac{P_{X^i|B^i}(dx^i|b^i)}{P_{X^i|B^{i-1}}(dx^i|b^{i-1})}\Big|X^i,A^i,B^{i-1}\Big\}\\ \overset{?}{=}\,&\mathbb{E}\Big\{\log\frac{P_{X^i|B^i}(dx^i|b^i)}{P_{X^i|B^{i-1}}(dx^i|b^{i-1})}\Big|X_i,P_{X^{i-1}|B^{i-1}}(dx^{i-1}|b^{i-1}),P_{X^{i-1}|B^{i-2}}(dx^{i-1}|b^{i-2})\Big\}\end{aligned} \tag{3.3.45}$$

If (3.3.45) holds, and the joint process $\{X_i,P_{X^i|B^i}(dx^i|b^i),P_{X^i|B^{i-1}}(dx^i|b^{i-1}):i=0,1,\ldots,n\}$ is jointly Markov process, controlled by $\{A_i:i=0,1,\ldots,n\}$, then we can proceed by deriving a dynamic programming equation. In view of this, it is natural to derive the recursive equations for $\{P_{X^i|B^i}(dx^i|b^i),P_{X^i|B^{i-1}}\ (dx^i|b^{i-1}):i=0,1,\ldots,n\}$ to determine whether they are jointly Markov process, and hence (3.3.45) is valid. Moreover, if this is the case then the optimal encoder has the property that for each time i, it is a functional of

$$\{X_i,P_{X^i|B^i}(.|b^i),P_{X^i|B^{i-1}}(.|b^{i-1})\},\ \forall\, i\in\mathbb{N}^n \tag{3.3.46}$$

The above discussion also applies to *Representation 1*.

Consider Problem 3.6.(b). Then an interesting question is under what conditions on the channel and source distributions, the maximizing encoders are Markov with respect to the source, i.e., $\{a_j=e_j(x_j,y^{j-1}):j=0,1,\ldots,n\}$. If this is the case there is no additional gain using encoders that depend on the whole past of the source, such as $\{a_j=e_j(a^{j-1},x^j,y^{i-1}):$ $j=0,1,\ldots,n$. Suppose that for a given channel the information definition of capacity is also operational. Then, we know that there exists an encoder-decoder which achieves capacity, that is, a direct channel coding theorem is shown, and the problem of finding the encoder strategy, is equivalently formulated as an optimization problem, of maximizing the information rate from the source to the channel output, by choosing the optimal encoder strategy, among all permissible encoders.

Problem 3.6.(b)

We start by addressing Problem 3.6.(b), of maximizing the directed information over the class of deterministic encoder strategies $\{e_j(a^{j-1},x^j,b^{j-1}):\ j=0,1,...n\}\in\mathscr{E}^{DF}[0,n]$. The information structure of the encoder at any time j is $\{a^{j-1},x^j,b^{j-1}\}$ and a specific strategy

Christos K. Kourtellaris

$\{e_0,....,e_n\} \in \mathscr{E}^{DF}[0,n]$ is given by

$$\begin{aligned} a_j = e_j(a^{j-1},x^j,b^{j-1}) = e_j\Big(e_0(a^{-1},x_0,b^{-1}),\\ e_1(a^0,x^1,b^0),...,e_{j-1}(a^{j-2},x^{j-1},b^{j-2}),e_j(a^{j-1},x^j,b^{j-1})\Big), \ \forall \ j \in \mathbb{N}^n \end{aligned}$$

This is the most general class of deterministic encoder strategies $\mathscr{E}^{DF}[0,n]$, since no assumptions are imposed either on the source or the channel.

Our goal is to identify general conditions on the source and channel distributions so that maximizing $I(X^n \rightarrow B^n)$ over the class of encoders $\mathscr{E}^{DF}[0,n]$ with information structure $\{(a^{j-1},x^j,b^{j-1}) : \ j \in \mathbb{N}^n\}$ is equivalent to maximizing $I(X^n \rightarrow B^n)$ over an encoder that belongs to $\mathscr{E}^{DM}[0,n]$, i.e., having an information structure $\{(x_j,b^{j-1}) : \ j \in \mathbb{N}^n\}$, and hence the encoder strategies are of the form $\{e_j(x_j,b^{j-1}) : \ j \in \mathbb{N}^n\} \in \mathscr{E}^{DF}[0,n]$.

Encoder structures of the form $\mathscr{E}^{DM}[0,n]$ are of particular interest in real-time communication applications, because it implies Symbol-by-Symbol (SbS) transmission is optimal (i.e., only one symbol is encoded at each transmission instant) and hence, no additional gain can be obtained by block coding of source sequences, in terms of performance, thus reducing the complexity of block coding.

The following conditions are important to prove that encoder structures $\mathscr{E}^{DM}[0,n]$ are indeed optimal.

Assumption 3.7. The information source is restricted to a sequence of stochastic Kernels

$$P_j(dx_j;x^{j-1},b^{j-1},a^{j-1}) = P_j(dx_j;x_{j-1},b^{j-1},a_{j-1}) - \text{a.a } (x^{j-1},a^{j-1},b^{j-1}), \ \forall \ j \in \mathbb{N}^n \tag{3.3.47}$$

Assumption 3.8. The communication channel is restricted to a sequence of stochastic Kernels

$$P_j(db_j;b^{j-1},a^j,x^j) = P_j(db_j;b^{j-1},a_j,x_j), -\text{a.a } (x^j,a^j,b^{j-1}), \ \forall \ j \in \mathbb{N}^n \tag{3.3.48}$$

We make the following observations regarding Assumptions 3.7, 3.8.

1. The condition on the source described by equation (3.3.47) allows feedback dependence on the channel output history and Markovian dependence on the previous source

Christos K. Kourtellaris

and encoder outputs. The reason we treat this general source is motivated by control applications, in which the source is a controlled process, controlled by a control process which uses information from the channel or decoder outputs. For communication application, (3.3.47) should be replaced by

$$P_j(dx_j;x^{j-1},b^{j-1},a^{j-1}) = P_j(dx_j;x_{j-1}) - \text{a.a } (x^{j-1},a^{j-1},b^{j-1}),\ \forall\ j \in \mathbb{N}^n \tag{3.3.49}$$

2. The condition on the channel described by (3.3.48) allows dependence on the channel output history and current source and encoder outputs. Such channels are generalized versions of those often used when feedback increases capacity as in [19]. For channels in which the information capacity is defined between the input and the output of the channel, then (3.3.48) is replaced by

   $$P_j(db_j;b^{j-1},a^j,x^j) = P_j(db_j;b^{j-1},a_j) - \text{a.a } (x^{j-1},a^{j-1},b^{j-1}),\ \forall\ j \in \mathbb{N}^n \tag{3.3.50}$$

   Most communication channels analysed in literature assume the form (3.3.50).

3. Assumptions 3.7, 3.8, can be further generalized to sources and channels which depend on previous source and encoders symbols having limited memory, by introducing additional variables into the formulation.

The first main result, which appeared in [17] on structural properties of the encoder is given in the next theorem.

Theorem 3.9. *Consider Problem.3.6 under Assumptions 3.7, 3.8.*
Then we have the following.
(a) Randomized Encoders ($\mathscr{E}^{RF}[0,n]$)
1. For a given $\{P_j(da_j;a^{j-1},x^j,b^{j-1}) : j = 0,1,\dots,n\} \in \mathscr{E}^{RF}[0,n]$ the directed information $I(X^n \rightarrow B^n)$ is given by

$$I(X^n \rightarrow B^n) = \int \log\Big(\frac{\overrightarrow{P}_{0,n}(db^n;x^n)}{P_{0,n}(db^n)}\Big)(\overrightarrow{P}_{0,n} \otimes \overleftarrow{P}_{0,n})(dx^n,db^n) \tag{3.3.51}$$

Christos K. Kourtellaris

where

$$P_{0,n}(dx^n,db^n)=\int_{\mathscr{A}_{0,n}}\otimes_{j=0}^{n}\Big(P_j(db_j;b^{j-1},a_j,x_j)\otimes P_j(da_j;a^{j-1},x^j,b^{j-1})$$
$$\otimes P_j(dx_j;x_{j-1},b^{j-1},a_j)\Big) \tag{3.3.52}$$
$$\overleftarrow{P}_{0,n}(dx^n;b^{n-1})=\otimes_{i=0}^{n}P_i(dx_i;x^{i-1},b^{i-1}) \tag{3.3.53}$$

and the marginals are constructed from the joint distribution.

2. The sequence of optimal encoder strategies maximizing $I(X^n\rightarrow B^n)$ over $\mathscr{E}^{RF}[0,n]$ has the form

$$P_j^*(da_j;x^j,b^{j-1},a^{j-1})=P_j^*(da_j;x_j,b^{j-1})-a.a\ (x^j,a^{j-1},b^{j-1}),\ \forall\ j\in\mathbb{N}^n \tag{3.3.54}$$

and

$$J_{0,n}^{R}(P_j^*:j=0,1,\ldots,n)\overset{\triangle}{=}\sup_{\{P_j(da_j;a^{j-1},x^j,b^{j-1}):j=0,1,\ldots,n\}\in\mathscr{E}^{RF}[0,n]}I(X^n\rightarrow B^n) \tag{3.3.55}$$
$$=\sup_{\{P_j(da_j;x_j,b^{j-1}),j=0,1,\ldots,n\}\in\mathscr{E}^{RM}[0,n]}I(X^n\rightarrow B^n) \tag{3.3.56}$$

where

$$I(X^n\rightarrow B^n)=\sum_{i=0}^{n}\int\log\Big(\frac{P_i(db_i;b^{i-1},x_i)}{P_i(db_i;b^{i-1})}\Big)P_{0,i}(dx^i,db^i) \tag{3.3.57}$$
$$P_i(db_i;b^{i-1},x_i)=\int_{\mathscr{A}_i}P_i(db_i;b^{i-1},a_i,x_i)\otimes P_i(da_i;x_i,b^{i-1}),\quad i=0,1,\ldots,n \tag{3.3.58}$$
$$P_i(db_i,b^{i-1})=\int_{\mathscr{A}_i\times\mathscr{X}_i}P_i(db_i;b^{i-1},x_i,a_i)\Big(P_i(da_i;x_i,b^{i-1})\otimes P_i(dx_i;b^{i-1})\Big),\quad i=0,1,\ldots,n \tag{3.3.59}$$

Christos K. Kourtellaris

(b) Deterministic Encoders ($\mathscr{E}^{DF}[0,n]$)

1. For a given $e \in \mathscr{E}^{DF}[0,n]$ *the directed information* $I(X^n \rightarrow B^n)$ *is given by*

$$I(X^n \rightarrow B^n) \overset{\triangle}{=} \sum_{i=0}^{n} \mathbb{E}^e \Big\{ \log \Big(\frac{P_i(dB_i;B^{i-1},X^i)}{P(dB_i;B^{i-1})} \Big) \Big\} \tag{3.3.60}$$

$$= \sum_{i=0}^{n} \mathbb{E}^e \Big\{ \log \Big(\frac{P_i(dB_i;B^{i-1},X_i,e_i(X^i,B^{i-1}))}{P(dB_i;B^{i-1})} \Big) \Big\} \tag{3.3.61}$$

$$= \sum_{i=0}^{n} I(X_i,A_i;B_i|B^{i-1}) \Big|_{A_i=e_i(X^i,B^{i-1})} \tag{3.3.62}$$

where

$$P_i(db_i,b^{i-1}) = \int_{\mathscr{X}_{0,i}} P_i(db_i;b^{i-1},x_i,e_i(x^i,b^{i-1})) \otimes P_i(dx^i;b^{i-1}), \quad i=0,1,\ldots,n \tag{3.3.63}$$

2. The sequence of optimal encoder strategies maximizing $I(X^n \rightarrow B^n)$ *over* $\mathscr{E}^{DF}[0,n]$ *has the form*

$$e_j^*(a^{j-1},x^j,b^{j-1}) = g_j^*(x_j,b^{j-1}), \ \forall \ j \in \mathbb{N}^n \tag{3.3.64}$$

and

$$J_{0,n}^D(e_j^* : j=0,1,\ldots,n) \overset{\triangle}{=} \sup_{\{e_j(x^j,a^{j-1},b^{j-1}):j=0,1,\ldots,n\} \in \mathscr{E}^{DF}[0,n]} I(X^n \rightarrow B^n) \tag{3.3.65}$$

$$= \sup_{\{g_j(x_j,b^{j-1}):j=0,1,\ldots,n\} \in \mathscr{E}^{DM}[0,n]} I(X^n \rightarrow B^n) \tag{3.3.66}$$

where for $g \in \mathscr{E}^{DM}[0,n]$

$$I(X^n \rightarrow B^n) = \sum_{i=0}^{n} \mathbb{E}^g \Big\{ \log \Big(\frac{P_i(dB_i;B^{i-1},X_i,g_i(X_i,B^{i-1}))}{P(dB_i;B^{i-1})} \Big) \Big\} \tag{3.3.67}$$

$$= \sum_{i=0}^{n} I(X_i,A_i;B_i|B^{i-1}) \Big|_{A_i=g_i(X_i,B^{i-1})} \tag{3.3.68}$$

$$= \sum_{i=0}^{n} I(X_i;B_i|B^{i-1}) \tag{3.3.69}$$

$$= \sum_{i=0}^{n} \int \log \Big(\frac{P_i(db_i;b^{i-1},x_i,g_i(x_i,b^{i-1}))}{P(db_i;b^{i-1})} \Big) P_i(db_i,db^{i-1},dx_i) \tag{3.3.70}$$

$$P_i(db_i;b^{i-1}) = \int_{\mathscr{X}_i} P_i(db_i;b^{i-1},x_i,g_i(x_i,b^{i-1})) \otimes P_i(dx_i;b^{i-1}), \quad i=0,1,\ldots,n \tag{3.3.71}$$

Christos K. Kourtellaris

Proof. see Appendix B.1. □

Remark 3.10. The following observations are consequences of the previous theorem.

1. Theorem 3.9 states that under Assumptions 3.7, 3.8, maximizing directed information over non-Markov strategies (with respect to the source) is equivalent to maximizing it over Markov (with respect to the source) strategies, for both deterministic and randomized strategies. This is a surprising result because it implies that no source block coding can give better performance.

2. The optimal encoder $g \in \mathscr{E}^{DM}[0,n]$ has the property that

$$\begin{aligned} I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} I(X_i, A_i; B_i | B^{i-1}) \Big|_{A_i = g_i(X_i, B^{i-1})} \\ &= \sum_{i=0}^{n} \Big(I(X_i; B_i | B^{i-1}) + I(A_i; B_i | B^{i-1}, X_i) \Big) \Big|_{A_i = g_i(X_i, B^{i-1})} \\ &= \sum_{i=0}^{n} I(X_i; B_i | B^{i-1}) \end{aligned} \tag{3.3.72}$$

   and

$$\begin{aligned} I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} I(X_i, A_i; B_i | B^{i-1}) \Big|_{A_i = g_i(X_i, B^{i-1})} \\ &= \sum_{i=0}^{n} \Big(I(A_i; B_i | B^{i-1}) + I(X_i; B_i | B^{i-1}, A_i) \Big) \Big|_{A_i = g_i(X_i, B^{i-1})} \end{aligned} \tag{3.3.73}$$

3. According to (3.3.71) the control process $\{P_i(db_i; b^{i-1}) : i = 0, 1, \ldots, n\}$ is a linear functional of the a posteriori distribution $\{P_i(dx_i; b^{i-1}) : i = 0, 1, \ldots, n\}$ and these are controlled by the control process $\{A_i : i = 0, 1, \ldots, n\}$ via the policies $\{g_i(x_i; b^{i-1}) : i = 0, 1, \ldots, n\}$.

4. The encoder structural properties will be used together with structural properties of the capacity achieving channel input distribution to identify structural properties when designing systems based on SbS (JSCC) transmission.

Next, we derive the consequences of Theorem 3.9, for the generalized unit memory channel, which is not independent of the source.

Christos K. Kourtellaris

Corollary 3.11. *Suppose Assumptions 3.7, 3.8 are replaced by*

$$P_j(dx_j; x^{j-1}, b^{j-1}, a^{j-1}) = P_j(dx_j; x_{j-1}) - a.a.\ (x^{j-1}, b^{j-1}, a^{j-1}),\ \forall\ j \in \mathbb{N}^n \tag{3.3.74}$$

$$P_j(db_j; b^{j-1}, x^j, a^j) = P_j(db_j; b_{j-1}, x_j, a_j) - a.a.\ (x^j, b^{j-1}, a^j),\ \forall\ j \in \mathbb{N}^n \tag{3.3.75}$$

Define the restricted policies $\mathscr{E}^{DMM}[0,n] \subseteq \mathscr{E}^{DM}[0,n]$ *by*

$$\mathscr{E}^{DMM}[0,n] \stackrel{\triangle}{=} \Big\{ g \in \mathscr{E}^{DM}[0,n] : g_i(x_i, b^{i-1}) = g_i^M(x_i, b_{i-1}), i = 0, 1, \ldots, n \Big\} \tag{3.3.76}$$

Then, the optimal deterministic encoder maximizing $I(X^n \rightarrow B^n)$ *over* $\mathscr{E}^{DF}[0,n]$ *has the property*

$$e_j^*(a^{j-1}, x^j, b^{j-1}) = g_j^M(x_j, b_{j-1}), \forall\ \ j \in \mathbb{N}^n \tag{3.3.77}$$

the process $\{B_i : i = 0, 1, \ldots\}$ *is a first-order Markov, that is,*

$$P_i(db_i; b^{i-1}) = P_i(db_i; b_{i-1}) - a.a.\ \ b^{i-1},\ \ i = 0, 1, \ldots, n \tag{3.3.78}$$

and

$$\sup_{\{e_j(x^j, a^{j-1}, b^{j-1}) : j = 0,1,\ldots,n\} \in \mathscr{E}^{DF}[0,n]} I(X^n \rightarrow B^n)$$

$$= \sup_{\{g_j(x_j, b^{j-1}) : j = 0,1,\ldots,n\} \in \mathscr{E}^{DM}[0,n]} \sum_{i=0}^n \mathbb{E}^g \Big\{ \log \Big(\frac{P_i(dB_i; B_{i-1}, X_i, g_i(X_i, B^{i-1}))}{P(dB_i; B^{i-1})} \Big) \Big\} \tag{3.3.79}$$

$$= \sup_{\{g_j(x_j, b^{j-1}) : j = 0,1,\ldots,n\} \in \mathscr{E}^{DM}[0,n]} \sum_{i=0}^n I(X_i, A_i; B^{i-1}) \Big|_{A_i = g_i(X_i, B^{i-1})} \tag{3.3.80}$$

$$= \sup_{\{g_j^M(x_j, b_{j-1}) : j = 0,1,\ldots,n\} \in \mathscr{E}^{DMM}[0,n]} \sum_{i=0}^n \mathbb{E}^{g^M} \Big\{ \log \Big(\frac{P_i(dB_i; B_{i-1}, X_i, g_i^M(X_i, B_{i-1}))}{P(dB_i; B^{i-1})} \Big) \Big\} \tag{3.3.81}$$

$$= \sup_{\{g_j^M(x_j, b_{j-1}) : j = 0,1,\ldots,n\} \in \mathscr{E}^{DMM}[0,n]} \sum_{i=0}^n I(X_i, A_i; B_i | B_{i-1}) \Big|_{A_i = g^M(X_i, B_{i-1})} \tag{3.3.82}$$

$$\equiv \sup_{\{g_j^M(x_j, b_{j-1}) : j = 0,1,\ldots,n\} \in \mathscr{E}^{DMM}[0,n]} \sum_{i=0}^n I(X_i; B_i | B_{i-1}) \tag{3.3.83}$$

$$\tag{3.3.84}$$

Christos K. Kourtellaris

where

$$P_i(db_i;b_{i-1}) = \int_{\mathscr{X}_i} P_i(db_i;b_{i-1},x_i,g_i^M(x_i,b_{i-1})) \otimes P_i(dx_i;b_{i-1}), \quad i=0,1,\ldots,n \tag{3.3.85}$$

$$P_i(db_i,db_{i-1},dx_i) = P_i(db_i;b_{i-1},x_i,g_i^M(x_i,b_{i-1})) \otimes P_i(dx_i;b_{i-1}) \otimes P_i(db_{i-1}), i=0,1,\ldots,n \tag{3.3.86}$$

$$P_i(dx_i;b_{i-1}) = \int_{\mathscr{X}_{i-1}} P_i(dx_i;x_{i-1}) \otimes P_{i-1}(dx_{i-1};b_{i-1}), \quad i=0,1,\ldots,n \tag{3.3.87}$$

Proof. By Theorem 3.9, the maximization occurs over the set $\mathscr{E}^{DM}[0,n]$. For a $g \in \mathscr{E}^{DM}[0,n]$, $\{a_i = g_i(x_i,b^{i-1}) : i=0,\ldots,n\}$ then

$$\begin{aligned} I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} I(X^i;B_i|B^{i-1}) \\ &= \sum_{i=0}^{n} \mathbb{E}^g \left\{ \log \frac{P_i(dB_i;B^{i-1},X^i)}{P_i(dB_i;B^{i-1})} \right\} \\ &= \sum_{i=0}^{n} \mathbb{E}^g \left\{ \log \frac{P_i(dB_i;B^{i-1},X^i,A^i)}{P_i(dB_i;B^{i-1})} \right\} \\ &= \sum_{i=0}^{n} \mathbb{E}^g \left\{ \log \frac{P_i(dB_i;B_{i-1},X_i,A_i)}{P_i(dB_i;B^{i-1})} \right\} \\ &= \sum_{i=0}^{n} I(X_i,A_i;B_i|B^{i-1}) \Big|_{A_i=g_i(X_i,B^{i-1})} \end{aligned} \tag{3.3.88}$$

Hence,

$$\begin{aligned} \sup_{g \in \mathscr{E}^{DF}[0,n]} I(X^n \rightarrow B^n) &= \sup_{g \in \mathscr{E}^{DF}[0,n]} \sum_{i=0}^{n} I(X_i,A_i;B_i|B^{i-1}) \Big|_{A_i=g_i(X^i,B^{i-1})} \\ &= \sup_{g \in \mathscr{E}^{DM}[0,n]} \sum_{i=0}^{n} I(X_i,A_i;B_i|B^{i-1}) \Big|_{A_i=g_i(X_i,B^{i-1})} && (3.3.89) \\ &\geq \sum_{i=0}^{n} I(X_i,A_i;B_i|B^{i-1}) \Big|_{A_i=g_i(X_i,B^{i-1})}, \forall\, g \in \mathscr{E}^{DM}[0,n] && (3.3.90) \end{aligned}$$

Take a $g^M \in \mathscr{E}^{DMM}[0,n]$ such that $\{B_i : i=0,\ldots,n\}$ is a first order Markov process, *i.e.*, $P(db_i;b^{i-1}) = P(db_i;b_{i-1}) - a.a.\ b^{i-1}, i=0,\ldots,n$, defined by

$$\mathscr{E}^{DMM,*}[0,n] \triangleq \{g \in \mathscr{E}^{DMM}[0,n] : P_i(dx_i;b^{i-1}) = P_i(dx_i;b_{i-1}) - a.a.\ b^{i-1}, i=0,\ldots,n\}$$

Christos K. Kourtellaris

Then for $g^M \in \mathscr{E}^{DMM,*}[0,n]$ we have

$$\begin{aligned}
P(db_i;b^{i-1}) &= \int_{\mathscr{X}_i} P(db_i;b^{i-1},x_i,g_i^M(x_i,b_{i-1})) \otimes P_i(dx_i;b^{i-1}) \\
&= \int_{\mathscr{X}_i} P(db_i;b_{i-1},x_i,g_i^M(x_i,b_{i-1})) \otimes P_i(dx_i;b^{i-1}) \\
&= P_i(db_i;b_{i-1}) \\
&= \int_{\mathscr{X}_i} P(db_i;b_{i-1},x_i,g_i^M(x_i,b_{i-1})) \otimes P_i(dx_i;b_{i-1}) \qquad (3.3.91)
\end{aligned}$$

For such a $g^M \in \mathscr{E}^{DMM,*}[0,n] \subseteq \mathscr{E}^{DM}[0,n]$, by (3.3.89), (3.3.90) and (3.3.91) we have

$$\begin{aligned}
\sup_{g\in\mathscr{E}^{DF}[0,n]} I(X^n \rightarrow B^n) &= \sup_{g\in\mathscr{E}^{DM}[0,n]} \sum_{i=0}^{n} I(X_i,A_i;B_i|B^{i-1}) \\
&\geq \sum_{i=0}^{n} \mathbb{E}^{g^M}\left\{\log \frac{P_i(dB_i;B_{i-1},X_i,g_i^M(X_i,B_{i-1}))}{P_i(dB_i;B^{i-1})}\right\} \\
&= \sum_{i=0}^{n} \mathbb{E}^{g^M}\left\{\log \frac{P_i(dB_i;B_{i-1},X_i,g_i^M(X_i,B_{i-1}))}{P_i(dB_i;B_{i-1})}\right\} \\
&= \sum_{i=0}^{n} I(X_i,A_i;B_i|B_{i-1})\Big|_{A_i=g_i^M(X_i,B_{i-1})}, \quad \forall g^M \in \mathscr{E}^{DMM,*}[0,n] \qquad (3.3.92)
\end{aligned}$$

On the other hand for a given $g \in \mathscr{E}^{DM}[0,n]$, by (3.3.88), we have

$$\begin{aligned}
I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} I(X_i,A_i;B_i|B^{i-1})\Big|_{A_i=g_i(X_i,B^{i-1})} \\
&= \sum_{i=0}^{n} H(B_i|B^{i-1}) - \sum_{i=0}^{n} H(B_i|B_{i-1},X_i,g_i(X_i,B^{i-1})) \\
&\leq \sum_{i=0}^{n} H(B_i|B_{i-1}) - \sum_{i=0}^{n} H(B_i|B_{i-1},X_i,g_i(X_i,B^{i-1})) \qquad (3.3.93) \\
&= \sum_{i=0}^{n} I(X_i,g_i(X_i,B^{i-1});B_i|B_{i-1})
\end{aligned}$$

where the inequality follows from the fact that conditioning does not increase entropy. The inequality (3.3.93) holds with equality if $g_i \in \mathscr{E}^{DMM,*}[0,n]$, that is, (3.3.91) holds. By (3.3.93) taking the supremum of the left hand side over $\mathscr{E}^{DM}[0,n]$, we have

$$\begin{aligned}
\sup_{g\in\mathscr{E}^{DM}[0,n]} I(X^n \rightarrow B^n) &= \sup_{g\in\mathscr{E}^{DM}[0,n]} \sum_{i=0}^{n} I(X_i,A_i;B_i|B^{i-1})\Big|_{A_i=g_i(X_i,B_{i-1})} \\
&\leq \sum_{i=0}^{n} I(X_i,A_i;B_i|B_{i-1})\Big|_{A_i=g_i(X_i,B^{i-1})}, \ \forall g \in \mathscr{E}^{DM}[0,n] \ (3.3.94)
\end{aligned}$$

Christos K. Kourtellaris

Since (3.3.94) holds $\forall g \in \mathscr{E}^{DM}[0,n]$, evaluating at $g_i \in \mathscr{E}^{DMM,*}[0,n]$, we have

$$\sup_{g\in\mathscr{E}^{DM}[0,n]} \sum_{i=0}^{n} I(X_i;B_i|B^{i-1}) \leq \sum_{i=0}^{n} I(X_i,A_i;B_i|B_{i-1})\Big|_{A_i=g_i(X_i,B_{i-1})}, \ \forall g \in \mathscr{E}^{DMM,*}[0,n] \tag{3.3.95}$$

Combining (3.3.92) and (3.3.95), we deduce that the supremum, if it exists it is achieved in $g^M \in \mathscr{E}^{DMM,*}[0,n]$. This completes the derivation. □

Remark 3.12. The previous corollary, includes as a special case, the channel with feedback satisfying

$$P_j(db_j;b^{j-1},a^j,x^j) = P_j(db_j;b_{j-1},a_j) - a.a.(b^{j-1},a^j), i=0,1,\ldots,n \tag{3.3.96}$$

which does not have more information capacity. The capacity achieving distribution of this simplified channel with unit memory, is discussed in the Shannon Lecture by Berger in [6], where it is conjectured that the capacity achieving distribution is of the form[2]

$$P_j(da_j;a^{j-1},b^{j-1}) = P_j(da_j;b_{j-1}) - a.a.(a^{i-1},b^{j-1}), i=0,1,\ldots,n \tag{3.3.97}$$

and that the finite-time capacity is given by

$$\sup_{P_i(da_i;b_{i-1}):i=0,1,\ldots,n} \sum_{i=0}^{n} I(A_i;B_i|B_{i-1}) \tag{3.3.98}$$

Unfortunately, we did not managed to find any derivation of property (3.3.97) for the unit memory channel (3.3.96). Nevertheless, the previous general Corollary can be used to derive property (3.3.97) as follows.

By Corollary 3.11 and assuming (3.3.96) we have the following identities.

[2] The derivation in [6] is insufficient because the author utilized mutual information instead of the directed information as the information measure.

Christos K. Kourtellaris

$$\sup_{e\in\mathscr{E}^{DF}[0,n]} I(X^n\rightarrow B^n) = \sup_{g\in\mathscr{E}^{DM}[0,n]}\sum_{i=0}^{n} I(X_i,A_i;B_i|B^{i-1})\Big|_{A_i=g_i(X_i,B^{i-1})} \tag{3.3.99}$$

$$= \sup_{g\in\mathscr{E}^{DM}[0,n]}\sum_{i=0}^{n}\mathbb{E}^{g}\left\{\log\frac{P_i(dB_i;B_{i-1},X_i,g_i(X_i,B^{i-1}))}{P_i(dB_i;B^{i-1})}\right\} \tag{3.3.100}$$

$$= \sup_{g^M\in\mathscr{E}^{DMM}[0,n]}\sum_{i=0}^{n}\mathbb{E}^{g^M}\left\{\log\frac{P_i(dB_i;B_{i-1},X_i,g_i^M(X_i,B_{i-1}))}{P_i(dB_i;B_{i-1})}\right\} \tag{3.3.101}$$

$$= \sup_{g^M\in\mathscr{E}^{DMM}[0,n]}\sum_{i=0}^{n} I(X_i,A_i;B_i|B_{i-1})\Big|_{A_i=g_i^M(X_i,B_{i-1})} \tag{3.3.102}$$

$$= \sup_{g^M\in\mathscr{E}^{DMM}[0,n]}\sum_{i=0}^{n}\mathbb{E}^{g^M}\left\{\log\frac{P_i(dB_i;B_{i-1},X_i,g_i^M(X_i,B_{i-1}))}{P_i(dB_i;B_{i-1})}\right\} \tag{3.3.103}$$

$$= \sup_{g^M\in\mathscr{E}^{DMM}[0,n]}\sum_{i=0}^{n} I(A_i;B_i|B^{i-1})\Big|_{A_i=g_i^M(X_i,B_{i-1})} \tag{3.3.104}$$

$$= \sup_{P_i(da_i;b_{i-1}):i=0,1,\ldots,n}\sum_{i=0}^{n} I(A_i;B_i|B_{i-1}) \tag{3.3.105}$$

where

(3.3.99) follows from Theorem 3.9;

(3.3.100) follows by definition;

(3.3.101) follows from Corollary 3.11;

(3.3.102) follows by definition;

(3.3.103) follows (3.3.96) (the MC);

(3.3.104) follows by definition;

(3.3.105) follows from the fact that maximizing over randomized strategies does not increase the pay-off (since in our case the information structure is classical).

The previous structural properties of encoders imply that the maximization can be done using stochastic control and dynamic programming techniques, of partially observed systems.

In stochastic optimal control of partially observed systems, the control process is a functional of the observations. Often, such problems are converted to equivalent stochastic control problems, in which the controlled process is the á posteriori distribution, and the control process is replaced by a function of this distribution, called "separated controls". Next, we describe an analogous procedure, by first introducing the definition of separated encoder

Christos K. Kourtellaris

strategies.

Definition 3.13. (Separated Encoder Strategies)
Consider a source and a channel, specified by Assumptions 3.7, 3.8. Given a set of encoder strategies ($\mathscr{E}^{RM}[0,n]$), define the á posteriori conditional distribution $\Pi_j^a(dx_j;b^{j-1}) \triangleq Prob(X_j \in dx_j | B^{j-1} = b^{j-1}), \forall\ j \in \mathbb{N}^n$.

(a) Randomized Encoders:

A randomized encoder $\{P_j : j \in \mathbb{N}^n\} \in \mathscr{E}^{RM}[0,n]$ is called separated if $P_j(da_j;x_j,b^{j-1})$ depends on $B^{j-1} = b^{j-1}$ only through the conditional distribution $\Pi_j^a(dx_j;b^{j-1})$, $\forall\ j \in \mathbb{N}^n$. The set of separated randomized encoder strategies is denoted by $\mathscr{E}^{sep,RM}[0,n]$.

(b) Deterministic Encoders:

A deterministic encoder $\{g_j : j \in \mathbb{N}^n\} \in \mathscr{E}^{DF}[0,n]$ is called separated if $a_j = g_j(x_j,b^{j-1})$ depends on $B^{j-1} = b^{j-1}$ only through the conditional distribution $\Pi^a(dx_j;b^{j-1})$, $\forall\ j \in \mathbb{N}^n$. The set of separated deterministic encoder strategies is denoted by $\mathscr{E}^{sep,DM}[0,n]$.

Thus, for any $\{g_j : j \in \mathbb{N}^n\} \in \mathscr{E}^{sep,DM}[0,n]$ then the encoder strategy at time j is of the form $a_j = g_j(x_j,b^{j-1}) = g_j(x_j,\Pi^a(dx_j;b^{j-1}))$. Such separated encoder strategies are well analyzed in stochastic control problems with partial information [13, 14]. The connection to stochastic control is established as follows.

Although, one starts with a partially observable stochastic control problem, by identifying some information state (a quantity that carries the same information as the observations), in this case, conditional distribution, then the partially observable problem is converted into a fully observable problem with pay-off expressed as a functional of the information state. The resulting equivalent optimization problem is to control the information state, via separated control strategies in order to incur the best possible performance. The important assumption to utilize separate strategies is that the information state, for example, the á posteriori distribution, is a Markov process. Mathematically this is equivalent to the following. For any bounded continuous test function $\Phi : \mathscr{X}_j \mapsto \mathbb{R}$, the following should hold.

$$\begin{aligned} &\mathbb{E}\Big\{\int_{\mathscr{X}_j} \Phi(x)\Pi_j^a(dx|B^{j-1})\Big|B^{j-2}\Big\} \\ &= \mathbb{E}\Big\{\int_{\mathscr{X}_j} \Phi(x)\Pi_j^a(dx|B^{j-1})\Big|\Pi_{j-1}^a(dx|B^{j-2}) = \pi_{j-1}(B^{j-2})\Big\} \end{aligned} \tag{3.3.106}$$

Christos K. Kourtellaris

If (3.3.106) is satisfied, then B^{j-2} carries the same information as $\Pi^a_{j-1}(dx|b^{j-2})$. Consider for example the case where the conditional distribution satisfies the following recursion.

$$\Pi^a_j(dx|b^{j-1}) = T_j(b_{j-1}, \Pi^a_{j-1}(.|b^{j-2}))$$

where $T_j(.,.)$ is a mapping form (b_{j-1}, Π^a_{j-1}) to $\Pi^a_j, j = 0,1,\ldots,n$. Then (3.3.106) holds. However, if the mapping is replaced by $T_j(b_{j-1}, b_{j-2}, \Pi^a_{j-1}(.|b^{j-2}))$, then (3.3.106) holds provided the conditioning includes the additional information B_{j-2}.

By analogy with stochastic control problems, one can express the directed information $I(X^n \rightarrow B^n)$ in terms of the information state, $\{\Pi^a_j(dx_j|b^{j-1}) : j = 0,1,\ldots,n\}$ and then employ separated encoder strategies to maximize it, subject to a dynamic recursion satisfied by the information state. In principle, and under certain assumptions, this methodology will lead to a principle of optimality and an associated dynamic programming satisfied by the optimal cost-to-go.

3.3.1 Encoder Design via Dynamic Programming

The objective in this section is to derive recursive equations for the information available to the encoder, and then introduce dynamic programming recursions to characterize the optimal encoders which maximize directed information.

As before, all processes are initially defined on a complete probability space $(\Omega, \mathbb{F}(\Omega), \mathbb{P}^a)$. Define the following complete σ algebras (completion is with respect to the null sets of measure zero of $(\Omega, \mathbb{F}(\Omega), \mathbb{P}^a)$):

$$\begin{aligned}
\mathscr{G}_{0,n} &\triangleq \sigma\{X_0, X_1, \ldots, X_n, B_0, B_1, \ldots, B_n\} \\
\mathscr{F}_{0,n} &\triangleq \sigma\{X_0, X_1, \ldots, X_n, B_0, B_1, \ldots, B_{n-1}\} \\
\mathscr{J}_{0,n} &\triangleq \sigma\{B_0, B_1, \ldots, B_n\}
\end{aligned} \tag{3.3.107}$$

where $\sigma\{X\}$ denotes the σ-algebra generated by R.V. X. Clearly, a deterministic encoder $e_j \in \mathscr{E}^{DM}[0,n]$ for each $j \in \mathbb{N}^n$, is an $\mathscr{F}_{0,j}-$ measurable function.

Christos K. Kourtellaris

Christos K. Kourtellaris

Consider Problem 3.6.(b) of maximizing the directed information over encoder class $\mathscr{E}^{DF}[0,n]$. Define the conditional pay-off on the interval $[k,n]$ by

$$\bar{J}^{D}_{k,n}(\{e_j : j=k,\ldots,n\},\mathscr{F}_{0,k}) \triangleq \mathbb{E}^e\Big\{\sum_{i=k}^{n}\log\Big(\frac{P_i(dB_i;B^{i-1},e^i(X^i,B^{i-1}),X^i)}{P_i^e(dB_i;B^{i-1})}\Big)\Big|\mathscr{F}_{0,k}\Big\}$$

By the smoothing property of conditional expectation we have

$$\begin{aligned} J^{D}_{k,n}(\{e_j : j=k,\ldots,n\}) &\triangleq \mathbb{E}^e\Big\{\mathbb{E}^e\Big[\sum_{i=k}^{n}\log\Big(\frac{P_i(dB_i;B^{i-1},e^i(X^i,B^{i-1}),X^i)}{P_i^e(dB_i;B^{i-1})}\Big)\Big|\mathscr{F}_{0,k}\Big]\Big\} \\ &= \mathbb{E}^e\Big\{\bar{J}^{D}_{k,n}(\{e_j : j=k,\ldots,n\},\mathscr{F}_{0,k})\Big\} \end{aligned}$$

Further,

$$\begin{aligned} &\max_{\{e_j:j=k,\ldots,n\}\in\mathscr{E}^{DF}[k,n]} J^{D}_{k,n}(\{e_j : j=k,\ldots,n\}) \\ &= \mathbb{E}^e\Big\{\max_{\{e_j:j=k,\ldots,n\}\in\mathscr{E}^{DF}[k,n]} \bar{J}_{k,n}(\{e_j : j=k,\ldots,n\},\mathscr{F}_{0,k})\Big\} \end{aligned} \tag{3.3.108}$$

Define the value function by

$$V_k(\mathscr{F}_{0,k}) = \max_{\{e_j:j=k,\ldots,n\}\in\mathscr{E}^{DF}[k,n]} \bar{J}^{D}_{k,n}(\{e_j : j=k,\ldots,n\},\mathscr{F}_{0,k})\Big\} \tag{3.3.109}$$

Next, we present the dynamic programming recursion satisfied by (3.3.109).

Theorem 3.14. *Suppose there exists encoder strategies* $\{e_j^* : j=0,1,\ldots,n\}\in\mathscr{E}^{DF}[0,n]$ *and a function* $V_k(\mathscr{F}_{0,k})$ *which satisfies the dynamic programming recursion:*

$$\begin{aligned} V_k(\mathscr{F}_{0,k}) &= \max_{e_k\in\mathscr{E}^{DF}[k,k]} \mathbb{E}^e\Big\{\log\Big(\frac{P_k(dB_k;B^{k-1},e^k(X^k,B^{k-1}),X^k)}{P_k^e(dB_k;B^{k-1})}\Big) \\ &\quad +V_{k+1}(\mathscr{F}_{0,k+1})|\mathscr{F}_{0,k}\Big\},\ k=0,1,\ldots,n-1 \end{aligned} \tag{3.3.110}$$

$$V_n(\mathscr{F}_{0,n}) = \int_{\mathscr{B}_n}\log\Big(\frac{P_n(dB_n;B^{n-1},e^n(X^n,B^{n-1}),X^n)}{P_n^e(dB_n;B^{n-1})}\Big)P_n(dB_n;B^{n-1},e^n(X^n,B^{n-1}),X_n) \tag{3.3.111}$$

Then $\{e^*_j : j = 0,1,\ldots,n\} \in \mathscr{E}^{DF}[0,n]$ *obtained from the solution of (3.3.110), (3.3.111) is an optimal encoder strategy and*

$$J^D_{0,n}(\{e_j\}^n_{j=0}) = E\Big\{V_0(\mathscr{F}_{0,0})\Big\} \tag{3.3.112}$$

Proof. Follows from dynamic programming (see Peter Caines book [7]). □

Remark 3.15. The dynamic programming recursion given in Theorem 3.14 is quite general; no assumptions are introduced on the channel or the source. Theorem 3.14 simplifies considerably if we assume Assumptions 3.7, 3.8, and then use Theorem 3.9(b), i.e., $e_j(x^i, b^{j-1}) = g_j(x_i, b^{j-1}),\ j = 0,1,\ldots,n.$ For the unit memory channel of Corollary 3.11, (3.3.110) and (3.3.111) become

$$\begin{aligned} V_k(x_k, b_{k-1}) = &\max_{g^M_k \in \mathscr{E}^{DMM}[k,k]} \mathbb{E}^{g^M}\Big\{\log\Big(\frac{P_k(dB_k; B_{k-1}, g^M_k(X_k, B_{k-1}), X_k)}{P^{g^M}_k(dB_k; B_{k-1})}\Big) \\ &+ V_{k+1}(x_{k+1}, b_k) | X_k = x_k, B_{k-1} = b_{k-1}\Big\} \end{aligned} \tag{3.3.113}$$

$$V_n(x_n, b_{n-1}) \;=\; \int_{\mathscr{B}_n} \log\Big(\frac{P_n(dB_n; B_{n-1}, g^M_n(X_n, B_{n-1}), X_n)}{P^{g^M}_n(dB_n; B_{n-1})}\Big) P_n(dB_n; B_{n-1}, g^M_n(X_n, B_{n-1}), X_n) \tag{3.3.114}$$

Clearly, (3.3.113) and (3.3.114) are easy to compute. Similar recursions also hold for $\mathscr{E}^{DM}[0,n]$.

3.4 Structural Properties of Capacity Achieving Distribution

In this section, we address the following issue.

- Identify structural properties of capacity achieving input distribution, for channels with memory and feedback.

Thus, this problem addresses the calculation of the capacity achieving distribution, when the information capacity has an operational meaning.

Christos K. Kourtellaris

The problem is stated below.

Problem 3.16. (*Achieving Information Capacity*)

(a) Given an admissible set of source and channel input distributions defined by

$$\overrightarrow{P}_{0,n}(dx^n,da^n;b^{n-1}) \triangleq \otimes_{i=0}^{n}\big(P_i(da_i;a^{i-1},x^i,b^{i-1})\otimes P_i(dx_i;x^{i-1},b^{i-1},a^{i-1})\big) \tag{3.4.115}$$

and the definition of information capacity over a finite horizon defined by

$$C_{0,n}(P) \triangleq \sup_{\overrightarrow{P}_{0,n}(dx^n,da^n;b^{n-1})\in\mathscr{P}_{0,n}(P)} I(X^n\rightarrow B^n) \tag{3.4.116}$$

where, for a given power $P>0$, the cost constraint is defined by

$$\mathscr{P}_{0,n}(P) \triangleq \Big\{\overrightarrow{P}_{0,n}(dx^n,da^n;b^{n-1}) : \frac{1}{n+1}\mathbb{E}\Big\{\sum_{j=0}^{n} e_j(X^j,A^{j-1},B^{j-1})\Big\}\leq P\Big\} \tag{3.4.117}$$

find the structural properties of $\overrightarrow{P}_{0,n}(dx^n,da^n;b^{n-1})\in\mathscr{P}_{0,n}(P)$ which achieves the information capacity $C_{0,n}(P)$.

(b) Given an admissible set of channel input distributions $\{P_i(da_i;a^{i-1},b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{A}_{0,i-1}\times\mathscr{B}_{0,i-1}) : i\in\mathbb{N}^n\}$, and the definition of information capacity over a finite horizon defined by

$$C_{0,n}(P) \triangleq \sup_{\{P_i(da_i;a^{i-1},b^{i-1}):i=0,1,\ldots,n\}\in\mathscr{P}_{0,n}(P)} I(A^n\rightarrow B^n) \tag{3.4.118}$$

where, for a given $P>0$ the cost constraint is defined by

$$\mathscr{P}_{0,n}(P) \triangleq \Big\{\{P_i(da_i;a^{i-1},b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{A}_{0,i-1}\times\mathscr{B}_{0,i-1}) : i=0,1,\ldots,n\} : \\ \frac{1}{n+1}\mathbb{E}\Big\{\sum_{j=0}^{n} e_j(A^j,B^{j-1})\Big\}\leq P\Big\} \tag{3.4.119}$$

find the structural properties of input distribution which achieves the information capacity. Similarly, the unconstraint information capacity is defined over the distributions $\overrightarrow{P}_{0,n}(dx^n, da^n;b^{n-1})$ and $\{P_i(da_i;a^{i-1},b^{i-1}) : i=0,1,\ldots,n\}$, respectively, and it is denoted by $C_{0,n}$.

Remark 3.17. Note that (3.4.115) is only required in control applications since the source $\{X_i=0,1,\ldots,n\}$ is affected by the channel outputs (or decoder), and encoder law. For

Christos K. Kourtellaris

communication applications it is assumed that (3.4.115) is replaced by

$$\overrightarrow{P}_{0,n}(da^n, db^n; b^{n-1}) \triangleq \otimes_{i=0}^{n} \big(P_i(da_i; a^{i-1}, x^i, b^{i-1}) \otimes P_i(dx_i; x^{i-1})\big) \tag{3.4.120}$$

If in addition, $P_i(da_i; a^{i-1}, x^i, b^{i-1}) = P_i(da_i; a^{i-1}, b^{i-1})$-a.a. $(a^{i-1}, x^i, b^{i-1}), i = 0, 1, \ldots, n$, then (3.4.115) is replaced by (3.4.118).

We make the following comments regarding Problem 3.16. Problem 3.16.(a) with $C_{0,\infty}(P) \triangleq \liminf_{n \rightarrow \infty} \frac{1}{n+1} C_{0,n}(P)$, together with coding theorems, implies that this quantity has an operational meaning, hence $C_{0,\infty}(P)$ is the supremum of all achievable rates. Therefore, our interest is to identify the structural properties of the capacity achieving distribution. Similarly, Problem 3.16.(b) identifies structural properties of the achieving input distribution, for the case when the following MC holds.

$$X^i \leftrightarrow (A^i, B^{i-1}) \leftrightarrow B_i, \ i = 0, 1, \ldots, n$$

First we clarify certain issues concerning the proper information measures for addressing capacity of channels with memory and feedback, which also apply to networks. Recall that the mutual information between two random sequences X^n and B^n, denoted by $I(X^n; B^n)$ is a measure of average information the sequence B^n conveys to the sequence X^n. Since it is symmetric then it is also a measure of average information the sequence X^n conveys to the sequence B^n.

$$\begin{aligned} I(X^n; B^n) &\triangleq \mathbb{E}\Big\{ \log \frac{P_{B^n|X^n}(db^n|x^n)}{P_{X^n}(dx^n)} \Big\} \equiv \mathbb{I}_{X^n;B^n}(P_{X^n}, P_{B^n|X^n}) \\ &= I(X^n \rightarrow B^n) + I(X^n \leftarrow B^n) \\ &= \mathbb{I}_{X^n;B^n}\big(P_{X_i|X^{i-1},B^{i-1}}, P_{B_i|B^{i-1},X^i} : i = 0, 1, \ldots, n\big) \end{aligned} \tag{3.4.121}$$

where the terms in (3.4.121) are given by (3.2.25), (3.2.26).

Christos K. Kourtellaris

The notation $\mathbb{I}_{X^n;B^n}(P_{X_i|X^{i-1},B^{i-1}}, P_{B_i|B^{i-1},X^i} : i = 0,1,\ldots,n)$ indicates the functional dependence of the mutual information on the two sequences of stochastic Kernels $\{P_{X_i|X^{i-1},B^{i-1}}, P_{B_i|B^{i-1},X^i} : i = 0,1,\ldots,n\}$, and the notation $\mathbb{I}_{X^n;B^n}(P_{X^n},\ P_{B^n|X^n})$ its dependence on $\{(P_{X^n},\ P_{B^n|X^n}\}$. The quantity $I(X^n \rightarrow B^n)$ represents the average information in the direction $X^n \rightarrow B^n$ (feedforward) via the sequence of channels $\{P_{B_i|B^{i-1},X^i}(db_i|b^{i-1},x^i) : i = 0,1,\ldots,n\}$ while $I(X^n \leftarrow B^n)$ represents the average information in the direction $X^n \leftarrow B^n$ (feedback) via the sequence of channels $\{P_{X_i|X^{i-1},B^{i-1}}(dx_i|x^{i-1},b^{i-1}) : i = 0,1,\ldots,n\}$.

Therefore, it will be a mistake to use $I(X^n;B^n)$ to define capacity of the channel connecting X^n to B^n, unless some assumptions are introduced. We will explain this issue shortly.

Define the $(n+1)$ convolution measures

$$\overrightarrow{P}_{0,n}(db^n;x^n) \stackrel{\triangle}{=} \otimes_{i=0}^n P_i(db_i;b^{i-1},x^i) \equiv \otimes_{i=0}^n P_{B_i|B^{i-1},X^i}(db_i|b^{i-1},x^i)$$

$$\overleftarrow{P}_{0,n}(dx^n;b^{n-1}) \stackrel{\triangle}{=} \otimes_{i=0}^n P_i(dx_i;x^{i-1},b^{i-1}) \equiv \otimes_{i=0}^n P_{X_i|X^{i-1},B^{i-1}}(dx_i|x^{i-1},b^{i-1})$$

Then,

$$\begin{aligned} I(X^n \rightarrow B^n) &= \int \Big(\log \frac{\overrightarrow{P}_{0,n}(dy^n;x^n)}{P_{0,n}(dy^n)}\Big) \overrightarrow{P}_{0,n}(db^n;x^n) \otimes \overleftarrow{P}_{0,n}(dx^n;b^{n-1}) \\ &\equiv I_{X^n \rightarrow B^n}(\overrightarrow{P}_{0,n}, \overleftarrow{P}_{0,n}) \end{aligned} \tag{3.4.122}$$

Hence, if the source $\{X_i : i = 0,1,\ldots\}$ is affected by $\{Y_i : i = 0,1,\ldots\}$, only the directed information expression (3.4.122) should be used to define capacity of the channel.

The next theorem helps clarify the implications of the Markov chain $B^{i-1} \leftrightarrow X^{i-1} \leftrightarrow X_i$, $i = 0,1,\ldots,n$, on various notions of information capacity for which operational meanings can be sought, and often derived in the literature.

Theorem 3.18. *The following statements are equivalent.*

1. $P_{0,n}(db^n;x^n) = \overrightarrow{P}_{0,n}(db^n;x^n)$, $\forall\, n \in \mathbb{N}$.

2. $B_j \leftrightarrow (X^j, B^{j-1}) \leftrightarrow X_{j+1}^n$, $\forall\, j \in \mathbb{N}^{n-1}$, $\forall n \in \mathbb{N}$.

3. $I(X^n;B^n) = \mathbb{I}_{X^n;B^n}\big(P_{0,n}(dx^n), \overrightarrow{P}_{0,n}(db^n;x^n)\big)$.

4. $I(X^n \leftarrow B^n) = 0$, $\forall\, n\ \in \mathbb{N}$.

Christos K. Kourtellaris

5. $B^j \leftrightarrow X^j \leftrightarrow X_{j+1}$, $\forall\, j \in \mathbb{N}^{n-1}$.

6. $X^n_{j+1} \leftrightarrow X^j \leftrightarrow B^j$, $\forall\, j \in \mathbb{N}^{n-1}$, $\forall\, n \ \in \ \mathbb{N}$

Proof. The proof is similar to the proof of Lemma 2.21 (Appendix A.1). □

Remark 3.19. Note that for any source-channels then $B^{i-1} \leftrightarrow X^{i-1} \leftrightarrow X_i$, $i = 0,1,\dots,n$, forms a Markov chain if and only if $I(X^n \leftarrow B^n) = 0$. Thus, under the Markov chain $B^{i-1} \rightarrow X^{i-1} \rightarrow X_i$, $i = 0,1,\dots,n$, the information capacity of channels with memory and feedback can be defined via mutual information or directed information because of the following identity.

$$\begin{aligned} I(X^n;B^n) &= I_{X^n;B^n}\big(P_{0,n}(dx^n), \overrightarrow{P}_{0,n}(db^n;x^n)\big) && (3.4.123)\\ &= \mathbb{I}_{X^n\rightarrow B^n}(P_i(dx_i;x^{i-1}), P_i(db_i;b^{i-1},x^i) : \ i = 0,1,\dots,n) && (3.4.124)\end{aligned}$$

Actually (3.4.123), (3.4.124) are special cases of mutual and directed information, because the joint distribution is obtained via the source $P(dx^n)$ and the channel $\overrightarrow{P}_{0,n}(db^n;x^n)$.

We state the next assumption because it is often hidden on the definitions of capacity of channels with memory and feedback.

Assumption 3.20. $(B^{i-1},A^{i-1}) \leftrightarrow X^{i-1} \leftrightarrow X_i, i = 0,1,\dots,n$, equivalently, $P_i(dx_i;x^{i-1},b^{i-1},a^{i-1}) = P_i(dx_i;x^{i-1}) - a.a.(x^{i-1},b^{i-1},a^{i-1})$, $i = 0,1,\dots,n$, $\forall\, n \in \mathbb{N}$.

Clearly, for an encoder $A_i = e_i(X^i,B^{i-1})$, $i = 0,1,\dots,n$, any channel of the form $B_i = g_i(X^i,B^{i-1},A^i) + f_i(X^{i-1},B^{i-1},A^i)V_i$, $i \in \mathbb{N}^n$, where $\{V_i : i \in \mathbb{N}^n\}$ is any noise such that $P_i(dx_i;x^{i-1},\ v^{i-1}) = P_i(dx_i;x^{i-1})$, $i = 0,1,\dots,n$, satisfies Assumption 3.20. For example, a random process $\{X_i : \ i \in \mathbb{N}^n\}$ defined via $X_{i+1} = f(X^i,B^i,W_i)$, $i \in \mathbb{N}^n$, X_0 a R.V., in which $\{\{V_i : i \in \mathbb{N}^n\}, \{W_i : i \in \mathbb{N}^n\}, X_0\}$ are mutually independent satisfies Assumption 3.20.

Therefore, given a general channel with memory and feedback, which does not satisfy Assumption 3.20, the quantity which should be used to give operational meaning of capacity is the information capacity $I(X^n \rightarrow B^n)$.

Definition 3.21. The finite time information capacity is defined by

$$C^1_{0,n} \stackrel{\triangle}{=} \sup_{\overrightarrow{P}_{0,n}(dx^n,da^n;b^{n-1})} I(X^n \rightarrow B^n) \qquad (3.4.125)$$

Christos K. Kourtellaris

The information capacity is defined by

$$C^1_\infty \overset{\triangle}{=} \liminf_{n\to\infty} \sup_{\overrightarrow{P}_{0,n}(dx^n,da^n;b^{n-1})} \frac{1}{n+1} I(X^n \to B^n) \tag{3.4.126}$$

Similarly for $C^1_{0,n}(P)$, $C^1_\infty(P)$.
If Assumption.3.20 holds, then in (3.4.125), (3.4.126) the directed information is equal to the mutual information, i.e., $I(X^n \to B^n) = I(X^n;B^n) \equiv \mathbb{I}_{X^n\to B^n}(P(dx_i;x^{i-1}),\, P_i(db_i;\, b^{i-1},x^i) : i = 0,1,\ldots,n)$.

The classical paper by Pombra and Cover [19], utilizes a special form of (3.4.125) and (3.4.126), with directed information replaced by mutual information, to find the capacity of Gaussian channels with memory and feedback, which satisfy Assumption 3.20, because $I(X^n;B^n) = I(X^n \to B^n) \overset{\triangle}{=} \sum_{i=0}^n I(X^i;B_i|B^{i-1})$ $= \mathbb{I}_{X^n\to B^n}(P_i(dx_i;x^{i-1}), P_i(db_i;b^{i-1},x^i) : i = 0,1,\ldots,n)$.

Next, we recall a fundamental inequality of mutual information, which is often used together with Fano's inequality [28] to derive upper bounds on the information capacity using the converse coding theorem. Let X,Y,Z be real-valued RV's and Z a measurable function of Y, defined by $Z = f(Y)$. Then

$$I(X;Y) \geq I(X;Z), \;\; Z = f(Y) \tag{3.4.127}$$

and if $f(.)$ is a bijection and the inverse $f^{-1}(.)$ is also measurable, then the equality holds in (3.4.127). Using (3.4.127) for an encoder structure $A_i = \bar{e}_i(A^{i-1},X^i,B^{i-1}) = e_i(X^i,B^{i-1}),\ i =$

Christos K. Kourtellaris

$0,1,\ldots,n$, then

$$\begin{aligned}
I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} I(X^i;B_i|B^{i-1}) && (3.4.128)\\
&= \sum_{i=0}^{n} \Big\{H(B_i|B^{i-1}) - H(B_i|B^{i-1},X^i)\Big\} && (3.4.129)\\
&= \sum_{i=0}^{n} \Big\{H(B_i|B^{i-1}) - H(B_i|B^{i-1},X^i,A^i)\Big\} && (3.4.130)\\
&= \sum_{i=0}^{n} I(X^i,A^i;B_i|B^{i-1}) && (3.4.131)\\
&= \sum_{i=0}^{n} I(X^i;B_i|B^{i-1},A^i) + \sum_{i=0}^{n} I(A^i;B_i|B^{i-1}) && (3.4.132)\\
&\overset{(\alpha)}{=} \sum_{i=0}^{n} I(A^i;B_i|B^{i-1}) && (3.4.133)\\
&= I(A^n \rightarrow B^n) && (3.4.134)
\end{aligned}$$

and (α) holds if $e_i(.,B^{i-1})$ is a bijection and the inverse $e_i^{-1}(.,B^{i-1})$ is measurable $\forall i \in \mathbb{N}^n$. Note that if the encoder is constructed based on a measurable function of the error, i.e., $A_i = e_i(X^i - \sigma(B^{i-1}))$ and $e_i(.,\sigma(B^{i-1}))$ is a bijection and its inverse measurable, then the equality holds.

Remark 3.22. For a source $P(dx_i;x^{i-1},a^i,b^{i-1})$, $i=0,1,\ldots,n$, a channel $P(db_i;b^{i-1},a^i,x^i)$, $i=0,1,\ldots,n$, by Theorem 2.24, if the MC $X^i \leftrightarrow (A^i,B^{i-1}) \leftrightarrow B_i$, $i=0,1,\ldots,n$, holds, then

$$I(X^n \rightarrow B^n) \leq I(A^n \rightarrow B^n) \qquad (3.4.135)$$

Moreover, if $A_i = e_i(X^i,B^{i-1})$, $e_i(.)$ is measurable, and $e_i(.,B^{i-1})$ is a bijection, and $e_i^{-1}(.,B^{i-1})$ is measurable, for $i=0,1,\ldots,n$, then

$$I(X^n \rightarrow B^n) = I(A^n \rightarrow B^n) \qquad (3.4.136)$$

Remark 3.23. Under Assumption 3.20, in the case of no feedback, $A_i = g_i(X^i)$ in Definition 3.21 the supremum is replaced by $P_{0,n}(dx^n,da^n)$. Since by Theorem 2.24, under the MC $X^i \leftrightarrow (A^i,B^{i-1}) \leftrightarrow B_i$, $I(X^n;B^n) = I(X^n \rightarrow B^n) \leq I(A^n \rightarrow B^n)$, a possible choice of achieving equality in $I(X^n;B^n) = I(A^n \rightarrow B^n)$ is $X_i = A_i, i=0,1,\ldots,n$, and consequently the supremum over $P_{0,n}(dx^n,da^n) \in \mathscr{M}_1(\mathscr{X}_{0,n} \times \mathscr{A}_{0,n})$ reduces to the supremum over $A^n \in \mathscr{M}_1(\mathscr{A}_{0,n})$ of $I(A^n;B^n) = I(A^n \rightarrow B^n) = I(X^n \rightarrow B^n) = I(X^n;B^n)$.

Christos K. Kourtellaris

Next, we describe the case when the operational capacity is defined via $I(A^n \rightarrow B^n)$ (i.e., from the channel input to the channel output).

Assumption 3.24. $X^i \leftrightarrow (B^{i-1}, A^i) \leftrightarrow B_i$ is a MC, equivalently $P_i(db_i; b^{i-1}, a^i, x^i) = P_i(db_i; b^{i-1}, a^i)$, a.a. x^i, a^i, b^{i-1} for $i = 0, 1, \ldots, n$, $\forall n \in \mathbb{N}$.

Note, that any channel of the form $B_i = f_i(B^{i-1}, A^i, V_i)$, $i = 0, 1, \ldots$ and encoder of the form $\{A_i = e_i(A^{i-1}, X^i, B^{i-1}) : i = 0, 1, \ldots\} \in \mathscr{E}^{DF}[0,n]$ for which $\{V_i : i = 0, 1, \ldots\}$ is independent of $\{X^i : i = 0, 1, \ldots\}$, satisfies Assumption 3.24. However, if $\{V_i : i = 0, 1, \ldots\}$ is correlated with $\{X^i : i = 0, 1, \ldots\}$ then assumption 3.24 might fail, and capacity cannot be defined using $I(A^n \rightarrow B^n)$.

Next, we relate directed information $I(X^n \rightarrow B^n)$ to $I(A^n \rightarrow B^n)$. Under Assumption 3.24, given an encoder strategy $\{A_i = e_i(X^i, A^{i-1}, B^{i-1}) : i = 0, 1, \ldots, n\} \in \mathscr{E}^{DF}[0,n]$, then

$$\begin{aligned} I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} I(X^i; B_i | B^{i-1}) \\ &= \sum_{i=0}^{n} \mathbb{E}\left\{ \log\left(\frac{P_i(dB_i; B^{i-1}, X^i)}{P_i(dB_i; B^{i-1})} \right) \right\} \\ &= \sum_{i=0}^{n} \mathbb{E}\left\{ \log\left(\frac{P_i(dB_i; B^{i-1}, X^i, A^i)}{P_i(dB_i; B^{i-1})} \right) \right\} \\ &= \sum_{i=0}^{n} I(X^i, A^i; B_i | B^{i-1}) \Big|_{A_i = e_i(A^{i-1}, X^i, B^{i-1})} \\ &\overset{(a)}{=} \sum_{i=0}^{n} \mathbb{E}\left\{ \log\left(\frac{P_i(dB_i; B^{i-1}, A^i)}{P_i(dB_i; B^{i-1})} \right) \right\} \\ &\overset{(b)}{=} I(A^n \rightarrow B^n) \end{aligned} \tag{3.4.137}$$

where (a) holds due to Assumption 3.24 and (b) follows by definition. The sequence of equations leading to equation (3.4.137) demonstrates the assumptions required so that information capacity can be defined using the channel input and the channel Kernel, independently of the source output.

Definition 3.25. Suppose Assumption 3.24 hold. The finite time information capacity is defined by

$$C^2_{0,n} = \sup_{\{P_i(da_i; a^{i-1}, b^{i-1}) \in \mathscr{K}(\mathscr{A}_i; \mathscr{A}_{0,i-1} \times \mathscr{B}_{0,i-1}) : i = 0, 1, \ldots, n\}} I(A^n \rightarrow B^n)$$

Christos K. Kourtellaris

Moreover, the information capacity is defined by

$$C_\infty^2 = \liminf_{n\to\infty} \sup_{\{P_i(da_i;a^{i-1},b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{A}_{0,i-1}\times\mathscr{B}_{0,i-1}):i=0,1,\ldots,n\}} \frac{1}{n+1} I(A^n \to B^n)$$

Similarly for $C_{0,n}^2(P)$, $C_\infty^2(P)$.

A coding theorem for C_∞^2 of Definition 3.25 is derived in [75]. The source coding theorem states that if the source entropy per unit time is less than the channel capacity per unit time, then the error probability can be arbitrary small using encoding and decoding. The converse states that if the source entropy is greater than the capacity, arbitrary small probability of error cannot be achieved. Next we prove the converse coding theorem, which is analogous to those found in literature, without feedback.

Consider a source $\{X_i : i \in \mathbb{N}^k\}$ of length $k+1$ connected to a decoder over a sequence of $n+1$ channel outputs $\{B_i : i \in \mathbb{N}^n\}$, $n \geq k$. The next theorem relates the $k+1$ source outputs and $n+1$ channel outputs to the time interval between each source letter, τ_s and the time interval between each channel letter τ_c.

Theorem 3.26. *Consider a source $\{X_i : i \in \mathbb{N}^k\}$ and its reproduction $\{Y_i : i \in \mathbb{N}^k\}$ having finite alphabet of cardinality M. Define the error at the jth transmission $j \in \mathbb{N}^k$ by*

$$P_{e,j}^{(k)} \triangleq Prob(Y_j \neq X_j)$$

and the average probability of error

$$P_e^{(k)} \triangleq \frac{1}{k+1}\sum_{j=0}^{k} P_{e,j}^{(k)} \tag{3.4.138}$$

Assume

1. *For $k \leq n$, $X_i \leftrightarrow (X^{i-1}, B^n) \leftrightarrow Y^k$, $i = 0,1,\ldots,k$.*
2. *For $k \leq n$, $X^k \leftrightarrow (A^i, B^{i-1}) \leftrightarrow B_i$, $i = 0,1,\ldots,n$.*
3. *The source produces letter at a rate of one letter each τ_s seconds, and the channel has capacity C_∞^2 and it is used at a rate of one letter each τ_c seconds.*

Christos K. Kourtellaris

If the source sequence is connected to the decoder through $(n+1)$ *channel uses, where*

$$n+1=\left\lfloor\frac{(k+1)\tau_s}{\tau_c}\right\rfloor$$

then for any length $(k+1)$*, the average error probability* $P_e^{(k)}$ *satisfies*

$$P_e^{(k)}\log(M-1)+H(P_e^k)\geq\limsup_{k\to\infty}\frac{1}{k+1}H(X^k)-\frac{\tau_s}{\tau_c}C_\infty^2 \tag{3.4.139}$$

Proof. The following inequalities hold

$$\begin{aligned}P_e^{(k)}\log(M-1)+H(P_e^{(k)}) &\overset{(a)}{\geq} \frac{1}{k+1}H(X^k|Y^k)\\ &\overset{(b)}{=} \frac{1}{k+1}H(X^k)-\frac{1}{k+1}I(X^k;Y^k)\\ &\overset{(c)}{\geq} \frac{1}{k+1}H(X^k)-\frac{1}{k+1}I(X^k;B^n)\\ &\overset{(d)}{\geq} \frac{1}{k+1}H(X^k)-\frac{\tau_c}{\tau_s(k+1)}\frac{\tau_s}{\tau_c}I(A^n\rightarrow B^n)\end{aligned}$$

where (a) follows from Theorem 4.3.2, of Gallager [28], (b) is an identity, (c) follows from 1. and in view of Theorem 2.24.2, finally (d) follows from 3., in view of Theorem 2.24.2. Taking the lim inf and lim sup retains the above identity. □

The above inequality shows that no matter what coding is done, the average probability of error per source symbol must satisfy (3.4.139). This average probability of error is bounded away from zero if

$$\limsup_{k\to\infty}\frac{1}{k+1}H(X^k)>\frac{\tau_s}{\tau_c}C_\infty^2$$

Remark 3.27. Note that SbS transmission Theorem 3.26 corresponds to $n=k$ and the MC1 is replaced by $Y^{i-1}\leftrightarrow(B^{i-1},X^{i-1})\leftrightarrow X_i$, $i=0,1,\ldots,n$, (i.e. such as in controlled sources), so converse to coding theorem holds. However, one should show achievability of C_∞^1, C_∞^2 as well without imposing this MC. This will be done using SbS transmission, based on the duality to Theorem 2.29.

A simplified information definition is obtained by invoking the following assumption.

Christos K. Kourtellaris

Assumption 3.28. $A^{i-1} \leftrightarrow (A_i, B^{i-1}) \leftrightarrow B_i$, $i = 0, 1, \ldots, n$, equivalently $P_i(db_i; b^{i-1}, a^i) = P_i(db_i; b^{i-1}, a_i)$, $i = 0, 1, \ldots, n$, $\forall n \in \mathbb{N}$.

Assumption 3.28 is the analogue of the one used in Theorem 3.9, to derive structural encoder properties.

Lemma 3.29. *Suppose Assumption 3.28 holds.*
Then

$$I(A^i; B_i | B^{i-1}) = I(A_i; B_i | B^{i-1}) = \mathbb{E}\left\{ \log \frac{P_i(dB_i; B^{i-1}, A_i)}{P_i(dB_i; B^{i-1})} \right\}, \qquad \forall\, i \in \mathbb{N} \quad (3.4.140)$$

and

$$\sum_{i=0}^{n} I(A^i; B_i | B^{i-1}) = \sum_{i=0}^{n} I(A_i; B_i | B^{i-1}), \qquad \forall\, n \in \mathbb{N} \qquad (3.4.141)$$

Proof. This is a direct consequence of the Assumption 3.28. □

Definition 3.30. Suppose Assumptions 3.24 and 3.28 hold. The finite time information capacity is defined by

$$C^3_{0,n} = \sup_{\{P_i(da_i; a^{i-1}, b^{i-1}) \in \mathscr{K}(\mathscr{A}_i; \mathscr{A}_{0,i-1} \times \mathscr{B}_{0,i-1}) : i=0,1,\ldots,n\}} \sum_{i=0}^{n} I(A_i; B_i | B^{i-1}) \qquad (3.4.142)$$

The information capacity is defined by

$$C^3_{\infty} = \liminf_{n \to \infty} \sup_{\{P_i(da_i; a^{i-1}, b^{i-1}) \in \mathscr{K}(\mathscr{A}_i; \mathscr{A}_{0,i-1} \times \mathscr{B}_{0,i-1}) : i=0,1,\ldots,n\}} \frac{1}{n+1} \sum_{i=0}^{n} I(A_i; B_i | B^{i-1}) (3.4.143)$$

Similarly for $C^3_{0,n}(P)$, $C^3_{\infty}(P)$.

Next, we prove the structural form of the capacity achieving distribution for $C^3_{0,n}$.

Theorem 3.31. *Suppose Assumptions 3.24 and 3.28 hold.*
Then

$$I(A^n \to B^n) = \sum_{i=0}^{n} I(A_i; B_i | B^{i-1}) \equiv \sum_{i=0}^{n} \mathbb{E}\left\{ \log \left(\frac{P_i(dB_i; A_i, B^{i-1})}{P_i(dB_i; B^{i-1})} \right) \right\}, \qquad \forall\, n \in \mathbb{N}^n \qquad (3.4.144)$$

Christos K. Kourtellaris

The sequence of optimal conditional distributions $\{P_i(da_i;a^{i-1},b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{A}_{0,i-1}\times\mathscr{B}_{0,i-1}):$ $i=0,1,\ldots,n\}$ *that maximize (3.4.144) (assuming they exist) have the form*

$$P_i^*(da_i;a^{i-1},b^{i-1})=\pi_i^*(da_i;b^{i-1})-\ a.a.\quad (a^{i-1},b^{i-1}),\quad i=0,1,\ldots,n \tag{3.4.145}$$

and

$$\begin{aligned} C_{0,n}^3 &= \max_{\{P_i(da_i;a^{i-1},b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{A}_{0,i-1}\times\mathscr{B}_{0,i-1}):i=0,1,\ldots,n\}} \sum_{i=0}^n \mathbb{E}\Big\{\log\Big(\frac{P_i(dB_i;B^{i-1},A_i)}{P_i(dB_i;B^{i-1})}\Big)\Big\} \\ &= \max_{\{P_i(da_i;b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}):i=0,1,\ldots,n\}} \sum_{i=0}^n \mathbb{E}\Big\{\log\Big(\frac{P_i(dB_i;B^{i-1},A_i)}{P_i(dB_i;B^{i-1})}\Big)\Big\} \end{aligned} \tag{3.4.146}$$

Proof. The equality in (3.4.144) follows from Assumptions 3.24 and 3.28. Note that the right side of (3.4.144) depends on the channel, and the conditional distribution $\{P_i(db_i;b^{i-1}):$ $i=0,1,\ldots,n\}$. Define $\pi_i(da_i;b^{i-1})\stackrel{\triangle}{=}P_i(da_i;b^{i-1})$, $i=0,1,\ldots,n$. Then,

$$\begin{aligned} P_i(db_i;b^{i-1}) &= \int_{\mathscr{A}_i} P_i(db_i;a_i,b^{i-1})P_i(da_i;b^{i-1}) \\ &\equiv \int_{\mathscr{A}_i} P_i(db_i;a_i,b^{i-1})\pi_i(da_i;b^{i-1}), \\ &\equiv P_i^{\pi_i}(db_i;b^{i-1})\quad i=0,1,\ldots,n \end{aligned} \tag{3.4.147}$$

where the superscript in (3.4.147) inficates the dependence of $P_i(db_i;b^{i-1})$ on $\pi_i(da_i;b^{i-1})$, $i=0,1,\ldots,n$. For a fixed $B^{i-1}=b^{i-1}$, $\{P_i(db_i;b^{i-1})\equiv P_i^{\pi_i}(db_i;b^{i-1}):i=0,1,\ldots,n\}$ is the controlled process controlled by the input distribution $\{P_i(da_i;b^{i-1})\equiv\pi_i(da_i;b^{i-1}):i=0,1,\ldots,n\}$, the control process. Thus, $\{P_i^{\pi_i}(db_i;b^{i-1}):i=0,1,\ldots,n\}$ is the state of the system controlled by $\{\pi_i(da_i;b^{i-1}):i=0,1,\ldots,n\}$. By (3.4.147), we have

$$\begin{aligned} &\mathbb{E}\Big\{\sum_{i=0}^n\log\Big(\frac{P_i(dB_i;B^{i-1},A_i)}{P_i(dB_i;B^{i-1})}\Big)\Big\} \\ &= \mathbb{E}\Big\{\sum_{i=0}^n\mathbb{E}\Big(\log\frac{P_i(dB_i;B^{i-1},A_i)}{P_i(dB_i;B^{i-1})}\Big|B^{i-1},A_i,\pi_i(dA_i;B^{i-1}),\{P_j(dA_j;A^{j-1},B^{j-1}):j=0,\ldots,i\}\Big)\Big\} \\ &\overset{(\alpha)}{=} \mathbb{E}\Big\{\sum_{i=0}^n\mathbb{E}\Big(\log\frac{P_i(dB_i;B^{i-1},A_i)}{P_i(dB_i;B^{i-1})}\Big|B^{i-1},A_i,\pi_i(dA_i;B^{i-1})\Big)\Big\} \\ &= \mathbb{E}\Big\{\sum_{i=0}^n\ell\Big(B^{i-1},A_i,P_i^{\pi}(dA_i;B^{i-1}),\pi_i(dA_i;B^{i-1})\Big)\Big\} \end{aligned} \tag{3.4.148}$$

where the equality (α) follows from Assumption 3.28. Since $P_i(db_i;b^{i-1})$ depends on b^{i-1} and the control process $\pi_i(da_i;b^{i-1})$. Then, the maximization of (3.4.148) over $\{P_i(da_i;a^{i-1},$ $b^{i-1}):i=0,1,\ldots,n\}$ is equivalent to that over the smaller set $\{\pi_i(da_i;b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}):$

Christos K. Kourtellaris

$i=0,1,\ldots,n\}$. This maximization is done by choosing $\{\pi_i(da_i;b^{i-1}) : i=0,1,\ldots,n\}$ to control $\{P_i^{\pi}(db_i;b^{i-1}) : i=0,1,\ldots,n\}$, which depend on b^{i-1} and the control distribution $\pi_i(da_i;b^{i-1})$, hence the maximizing distribution has the form $\{P_i^*(da_i;a^{i-1},b^{i-1}) = \pi_i^*(da_i;b^{i-1}) : i=0,1,\ldots,n\}$. This completes the derivation □

Next we make some observations concerning the statements of Theorem 3.31.

Remark 3.32. The following statements are consequences of Theorem thccach and Theorem 3.31.

1. By Theorem 3.31 (under Assumptions 3.24, 3.28) we deduce that the class of capacity achieving distributions has the following conditional independence property.

$$P_i(da_i,a^{i-1},b^{i-1}) = \pi_i(da_i;b^{i-1}) - a.a.(a^{i-1},b^{i-1}), i=0,\ldots,n$$

2. By Theorem 3.9 (under Assumptions 3.7, 3.8), we have the following identities.

$$\sup_{e\in\mathscr{E}^{DF}[0,n]} I(X^n\rightarrow B^n) \overset{(\alpha)}{=} \sup_{g\in\mathscr{E}^{DM}[0,n]} \sum_{i=0}^{n} \mathbb{E}^g\left\{\log\frac{P_i(dB_i;B^{i-1},X_i,g_i(X_i,B^{i-1}))}{P_i(dB_i;B^{i-1})}\right\} \tag{3.4.149}$$

$$\overset{(\beta)}{=} \sup_{g\in\mathscr{E}^{DM}[0,n]} \sum_{i=0}^{n} I(X_i,A_i;B_i|B^{i-1})\Big|_{A_i=g_i(X_i,B^{i-1})} \tag{3.4.150}$$

$$\overset{(\gamma)}{=} \sup_{g\in\mathscr{E}^{DM}[0,n]} \sum_{i=0}^{n} I(A_i;B_i|B^{i-1})\Big|_{A_i=g_i(X_i,B^{i-1})} \tag{3.4.151}$$

$$\overset{(\delta)}{=} \sup_{\{\pi_i(da_i;b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}):i=0,\ldots,n\}} \sum_{i=0}^{n} I(A_i;B_i|B^{i-1}) \tag{3.4.152}$$

$$\overset{(\varepsilon)}{=} \sup_{\{\pi_i(da_i;b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}):i=0,\ldots,n\}} \sum_{i=0}^{n} \mathbb{E}^{\pi}\left\{\log\frac{P_i(dB_i;B^{i-1},A_i)}{P_i^{\pi}(dB_i;B^{i-1})}\right\} \tag{3.4.153}$$

$$\overset{(\zeta)}{=} \sup_{\{\pi_i(da_i;a^{i-1},b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{A}_{0,i-1}\times\mathscr{B}_{0,i-1}):i=0,\ldots,n\}} \sum_{i=0}^{n} I(A_i;B_i|B^{i-1}) \tag{3.4.154}$$

where

(α) follows from Theorem 3.9.

(β) follows from Theorem 3.9.

(γ) holds if the MC, $X^i \leftrightarrow (B^{i-1},A^i) \leftrightarrow B_i$ (Assumption 3.24), is satisfied, or $g_i(.,b^{i-1})$ is one-to-one and onto measurable, and its inverse also measurable, $i=0,1,\ldots,n$,

Christos K. Kourtellaris

because by the chain rule

$$I(X_i,A_i;B_i|B^{i-1})\Big|_{A_i=g_i(X_i,B^{i-1})}=I(X_i;B_i|B^{i-1},A_i)\Big|_{A_i=g_i(X_i,B^{i-1})}+I(A_i;B_i|B^{i-1}),\\ i=0,1,\ldots,n$$

the first right hand side term is zero if $A_i=g_i(X_i,B^{i-1})$ has the stated property.
(δ) follows from the fact that randomized strategies do not incur a higher pay-off (the information structures are classical).
(ε) and (ζ) are the statements of Theorem 3.31.

3. By Corollary 3.11 and statement 2 above, for the unit memory channel defined by

$$P_i(db_i;b^{i-1},x^i,a^i)=P_i(db_i;b_{i-1},a_i)-a.a.(b^{i-1},x^i,a^i),\ i=0,\ldots,n \tag{3.4.155}$$

we deduce the following equalities.

$$\begin{aligned}\sup_{g\in\mathscr{E}^{DM}[0,n]} I(X^n\rightarrow B^n) &\overset{(\alpha)}{=} \sup_{g^M\in\mathscr{E}^{DMM}[0,n]}\sum_{i=0}^{n} I(X_i,A_i;B_i|B_{i-1})\\ &\overset{(\beta)}{=} \sup_{g^M\in\mathscr{E}^{DMM}[0,n]}\sum_{i=0}^{n} I(A_i;B_i|B_{i-1})\Big|_{A_i=g_i^M(X_i,B_{i-1})}\\ &\overset{(\gamma)}{=} \sup_{\{\pi^M(da_i;b_{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{B}_{i-1}):i=0,\ldots,n\}}\sum_{i=0}^{n} I(A_i;B_i|B_{i-1})\\ &\overset{(\delta)}{=} \sup_{\{\pi(da_i;b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}):i=0,\ldots,n\}}\sum_{i=0}^{n} I(A_i;B_i|B^{i-1})\end{aligned}$$

where $(\alpha),(\beta)$ follows from Theorem 3.9, (γ) follows from the fact that maximizing over randomized strategies yields the same pay-off, and (δ) is the equivalent of (α).

This property of capacity achieving distribution, that is,

$$P_i(db_i;b^{i-1})=P_i(db_i;b_{i-1})-a.a.\ b^{i-1},i=0,1,\ldots,n \tag{3.4.156}$$

$$P_i(da_i;a^{i-1},b^{i-1})=\pi_i^M(da_i;b_{i-1})-a.a.\ (a^{i-1},b^{i-1}),i=0,1,\ldots,n \tag{3.4.157}$$

is first addressed in Shannon Lecture [6] using mutual information as the information measure. In our opinion this is the first complete derivation of properties (3.4.156) and (3.4.157) using directed information. Recently the authors in [58] consider a special case of the unit memory channel, with binary input and output alphabets, which breaks into the S-channel

Christos K. Kourtellaris

Christos K. Kourtellaris

and the Z-channel and derived (3.4.156)-(3.4.157) using existing capacity results of the S-channel and the Z-channel.

Next we provide an independent derivation of the structural properties of the unit memory channel, which is the analogue of Corollary 3.11. We shall make use of the variational equality of directed information derived in [10].

Given the general causal conditioned distribution $\overrightarrow{P}_{0,n}(db^n|a^n) = \otimes_{i=0}^{n} P_i(db_i; b^{i-1}, a^i)$ and $\overleftarrow{P}_{0,n}(da^n|b^{n-1}) = \otimes_{i=0}^{n} P_i(da_i; a^{i-1}, b^{i-1})$, the joint measure $P(da^n, db^n) = (\overrightarrow{P}_{0,n} \otimes \overleftarrow{P}_{0,n})(da^n, db^n)$ and the marginal measure $P_{0,n}(db^n) = \int_{\mathscr{A}_{0,n}} (\overrightarrow{P}_{0,n} \otimes \overleftarrow{P}_{0,n})(da^n, db^n)$, then the following variational equality holds

$$
\begin{aligned}
I(A^n \rightarrow B^n) &\triangleq \int_{\mathscr{A}_{0,n} \times \mathscr{B}_{0,n}} \log\left(\frac{\overrightarrow{P}_{0,n}(db^n|a^n)}{P_{0,n}(db^n)}\right)(\overrightarrow{P}_{0,n} \otimes \overleftarrow{P}_{0,n})(da^n, db^n) \\
&= \inf_{\nu_{0,n}(db^n) \in \mathscr{M}_1(\mathscr{B}_{0,n})} \int_{\mathscr{A}_{0,n} \times \mathscr{B}_{0,n}} \log\left(\frac{\overrightarrow{P}_{0,n}(db^n|a^n)}{\nu_{0,n}(db^n)}\right)(\overrightarrow{P}_{0,n} \otimes \overleftarrow{P}_{0,n})(da^n, db^n)
\end{aligned}
$$

where the infimum is achieved at

$$
\nu_{0,n}(db^n) = \int_{\mathscr{A}_{0,n}} (\overrightarrow{P}_{0,n} \otimes \overleftarrow{P}_{0,n})(da^n, db^n) \equiv P_{0,n}(db^n)
$$

Corollary 3.33. *Suppose the following holds*

$$
P_i(db_i; b^{i-1}, a^i, x^i) = P_i(db_i; b_{i-1}, a_i) - a.a.\ (b^{i-1}, a^i, x^i), i = 0, 1, \ldots, n \quad (3.4.158)
$$

Define the policies $\mathscr{E}^{RMM}[0,n]$ *by*

$$
\begin{aligned}
\mathscr{E}^{RMM}[0,n] \triangleq \Big\{ &\pi(da_i; b^{i-1}) \in \mathscr{K}(\mathscr{A}_i; \mathscr{B}_{0,i-1}) : \pi(da_i; b^{i-1}) = \pi^M(da_i; b_{i-1}) - a.a. \\
&b^{i-1}, i = 0, 1, \ldots, n \Big\} \subseteq \mathscr{E}^{RM}[0,n] \triangleq \Big\{ P_i(da_i; b^{i-1}) \in \mathscr{K}(\mathscr{A}_i; \mathscr{B}_{0,i-1}) : \\
&i = 0, 1, \ldots, n \Big\}
\end{aligned}
$$

Then the optimal capacity achieving channel input distribution $\{P_i(da_i; a^{i-1}, b^{i-1}) \in \mathscr{K}(\mathscr{A}_i; \mathscr{A}_{0,i-1} \times \mathscr{B}_{0,i-1}) : i = 0, 1, \ldots, n\}$ *maximizing* $I(A^n \rightarrow B^n)$ *has the following form*

$$
P_i(da_i; a^{i-1}, b^{i-1}) = \pi_i(da_i; b^{i-1}) = \pi_i^M(da_i; b_{i-1}) - a.a.\ (a^{i-1}, b^{i-1}), i = 0, 1, \ldots, n \quad (3.4.159)
$$

and

$$\begin{aligned}
&\sup_{P_i(da_i;a^{i-1},b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{A}_{0,i-1}\times\mathscr{B}_{0,i-1})} I(A^n\rightarrow B^n)\\
&=\sup_{\pi\in\mathscr{E}^{RM}[0,n]}\sum_{i=0}^{n}\mathbb{E}\left\{\log\left(\frac{P_i(dB_i;B_{i-1},A_i)}{P_i^{\pi}(dB_i;B^{i-1})}\right)\right\} && (3.4.160)\\
&=\sup_{\pi^M\in\mathscr{E}^{RMM}[0,n]}\sum_{i=0}^{n}\mathbb{E}\left\{\log\left(\frac{P_i(dB_i;B_{i-1},A_i)}{P_i^{\pi^M}(dB_i;B_{i-1})}\right)\right\} && (3.4.161)\\
&=\sup_{\pi^M\in\mathscr{E}^{RMM}[0,n]}\sum_{i=0}^{n}I(A_i;B_i|B_{i-1}) && (3.4.162)
\end{aligned}$$

where

$$\begin{aligned}
P_i^{\pi}(db_i;b^{i-1})&=\int_{\mathscr{A}_i}P_i(db_i;b_{i-1},a_i)\pi_i(da_i;b^{i-1}),\ i=0,1,\ldots,n && (3.4.163)\\
P_i^{\pi^M}(db_i;b_{i-1})&=\int_{\mathscr{A}_i}P_i(db_i;b_{i-1},a_i)\pi_i^M(da_i;b_{i-1}),\ i=0,1,\ldots,n && (3.4.164)
\end{aligned}$$

and the resulting process $\{B_i : i=0,1,\ldots,n\}$ *which achieves the supremum, is a first order Markov process.*

Proof. By Theorem 3.31, since $\mathscr{E}^{RMM}[0,n]\subseteq\mathscr{E}^{RM}$, then

$$\begin{aligned}
\sup_{\pi^M\in\mathscr{E}^{RMM}[0,n]}\sum_{i=0}^{n}I(A_i;B_i|B_{i-1}) &\equiv \sup_{\pi^M\in\mathscr{E}^{RMM}[0,n]}\sum_{i=0}^{n}\int\log\left(\frac{P_i(db_i;b_{i-1},a_i)}{P_i^{\pi^M}(db_i;b_{i-1})}\right)P_i^{\pi^M}(db_i,db_{i-1},da_i)\\
&\leq \sup_{\pi\in\mathscr{E}^{RM}[0,n]}\sum_{i=0}^{n}\int\log\left(\frac{P_i(db_i;b_{i-1},a_i)}{P_i^{\pi}(db_i;b^{i-1})}\right)P_i^{\pi}(db_i,db^{i-1},da_i)\\
&\equiv \sup_{\pi\in\mathscr{E}^{RM}[0,n]}\sum_{i=0}^{n}I(A_i;B_i|B^{i-1}) && (3.4.165)
\end{aligned}$$

where $\{(P_i^{\pi}(db_i;b^{i-1}),P_i^{\pi^M}(db_i;b_{i-1})) : i=0,\ldots,n\}$; they are induced by the channel and $\{\pi_i(da_i;b^{i-1}),\pi_i^M(db_i;b_{i-1}) : i=0,\ldots,n\}$. Next, we show that the right hand side of (3.4.165) is bounded above as follows.

$$\sup_{\pi\in\mathscr{E}^{RM}[0,n]}\sum_{i=0}^{n}I(A_i;B_i|B^{i-1})\leq\sup_{\pi^M\in\mathscr{E}^{RMM}[0,n]}\sum_{i=0}^{n}I(A_i;B_i|B_{i-1}) \qquad (3.4.166)$$

Christos K. Kourtellaris

Using the variational equality, for policies $\pi \in \mathscr{E}^{RM}[0,n]$, we have

$$\begin{aligned}
&\sup_{\pi \in \mathscr{E}^{RM}[0,n]} \sum_{i=0}^{n} I(A_i;B_i|B^{i-1}) \\
&= \sup_{\pi \in \mathscr{E}^{RM}[0,n]} \inf_{\nu_{0,n}(db^n) \in \mathscr{M}_1(\mathscr{B}_{0,n})} \int_{\mathscr{A}_{0,n} \times \mathscr{B}_{0,n}} \log\left(\frac{\overrightarrow{P}_{0,n}(db^n|a^n)}{\nu_{0,n}(db^n)}\right) P^{\pi}(da^n, db^n)
\end{aligned}$$

where $P^{\pi}(da^n, db^n)$ is the one defined by the channel satisfying (3.4.158), $\pi \in \mathscr{E}^{RM}[0,n]$, and $\nu_{0,n}(db^n) \in \mathscr{M}_1(\mathscr{B}_{0,n})$ is any arbitrary distribution. Consequently, since $\nu_{0,n}(db^n)$ is arbitrary, we take the one induced by

$$\nu_{0,n}(db^n) = \otimes_{i=0}^{n} \nu_i^{\pi^M}(db_i; b_{i-1}) \equiv \nu_{0,n}^{\pi^M}(db^n) \tag{3.4.167}$$

$$\nu_i^{\pi^M}(db_i; b_{i-1}) = \int_{\mathscr{A}_i} P_i(db_i; b_{i-1}, a_i) \pi_i^M(da_i; b_{i-1}), \;\; i = 0, 1, \ldots, n \tag{3.4.168}$$

to obtain the upper bound

$$\begin{aligned}
\sup_{\pi \in \mathscr{E}^{RM}[0,n]} \sum_{i=0}^{n} I(A_i;B_i|B^{i-1}) &\overset{(\alpha)}{\leq} \sup_{\pi \in \mathscr{E}^{RM}[0,n]} \int \log\left(\frac{\overrightarrow{P}_{0,n}(db^n|a^n)}{\nu_i^{\pi^M}(db_i; b_{i-1})}\right) P^{\pi}(da^n, db^n) \\
&\overset{(\beta)}{=} \sup_{\pi^M \in \mathscr{E}^{RMM}[0,n]} \int \log\left(\frac{\overrightarrow{P}_{0,n}(db^n|a^n)}{P_i^{\pi^M}(db_i; b_{i-1})}\right) P_i^{\pi^M}(db_i, db_{i-1}, da_i)
\end{aligned} \tag{3.4.169}$$

where the equality in (β) follows since $\nu_i^{\pi^M}(db_i; b_{i-1})$ is a Markov process conditioned on $\{b_{i-1}, \pi_i^M(da_i; b_{i-1})\}$. Combining (3.4.165) and (3.4.169) we deduce that the $\sup_{\pi \in \mathscr{E}^{RM}} \sum_{i=0}^{n} I(A_i;B_i|B^{i-1})$ is achieved in $\pi^M \in \mathscr{E}^{RMM}$, which implies (3.4.159) and the identities (3.4.160)-(3.4.162). This completes the proof.

□

Remark 3.34. We make the following observation regarding Theorem 3.31.

1. Suppose Assumption 3.28 is replaced by the following MC.

$$A^{i-2} \leftrightarrow (A_{i-1}, A_i, B^{i-1}) \leftrightarrow B_i, \; i = 0, 1, \ldots, n, \; n \in \mathbb{N} \tag{3.4.170}$$

Christos K. Kourtellaris

Then under the Assumptions 3.24, and validity of (3.4.170) we have

$$\begin{aligned} I(A^n \rightarrow B^n) &= \sum_{i=0}^{n} I(A_i, A_{i-1}; B_i | B^{i-1}), \ \forall n \in \mathbb{N} \\ &= \mathbb{E}\left\{ \log \frac{P_i(dB_i; B^{i-1}, A_i, A_{i-1})}{P_i(dB_i; B^{i-1})} \right\} \qquad (3.4.171) \end{aligned}$$

Moreover,

$$\begin{aligned} P_i(db_i; b^{i-1}) &= \int_{\mathscr{A}_i \times \mathscr{A}_{i-1}} P_i(db_i; a_i, a_{i-1}, b^{i-1}) P_i(da_i; a_{i-1}, b^{i-1}) P_{i-1}(da_{i-1}; b^{i-1}) \\ &\equiv \int_{\mathscr{A}_i \times \mathscr{A}_{i-1}} P_i(db_i; a_i, a_{i-1}, b^{i-1}) \pi_i(da_i; a_{i-1}, b^{i-1}) \pi_{i-1}(da_{i-1}; b^{i-1}) \\ &\equiv P_i^{\pi_i, \pi_{i-1}}(db_i; b^{i-1}), i = 0, 1, \ldots, n \qquad (3.4.172) \end{aligned}$$

Then, the derivation of Theorem 3.31 with (3.4.140) replaced by (3.4.171) is repeated, and we have the following generalization

$$\begin{aligned} &\sup_{\{P_i(da_i; a^{i-1}, b^{i-1}) \in \mathscr{K}(\mathscr{A}_i; \mathscr{A}_{0,i-1} \times \mathscr{B}_{0,i-1}) : i=0,1,\ldots,n\}} \sum_{i=0}^{n} \mathbb{E}\left\{ \log \left(\frac{P_i(dB_i; B^{i-1}, A_i, A_{i-1})}{P_i(dB_i; B^{i-1})} \right) \right\} \\ = &\sup_{\{P_i(da_i; a_{i-1}, b^{i-1}) \in \mathscr{K}(\mathscr{A}_i; \mathscr{A}_{i-1} \times \mathscr{B}_{0,i-1}) : i=0,1,\ldots,n\}} \\ &\left\{ \sum_{i=0}^{n} \mathbb{E}\left\{ \log \left(\frac{P_i(dB_i; B^{i-1}, A_i, A_{i-1})}{P_i(dB_i; B^{i-1})} \right) \right\} \right\} \end{aligned}$$

Moreover one can also define $\bar{a} = (a_i, a_{i-1})$ so that the optimization is over $\{P_i(d\bar{a}_i; b^{i-1}) : i = 0, 1, \ldots, n\}$.

2. The above conclusion also holds for channels satisfying the MC

$$A^{i-m-1} \leftrightarrow (A_{i-m}, \ldots, A_{i-1}, A_i, B^{i-1}) \leftrightarrow B_i, \ i = 0, 1, \ldots, n, \ n \in \mathbb{N}$$

and optimizing $I(A^n \rightarrow B^n) = \sum_{i=0}^{n} I(A_{i-m}^{i}; B_i | B^{i-1})$ over $P_i(da_i, a^{i-1}; b^{i-1}) : i = 0, 1, \ldots, n$ is the same as that over $\{P_i(da_i; a_{i-m}^{i-1}, b^{i-1}) : i = 0, 1, \ldots, n\}$. Indeed, several conclusions hold if the channel has limited memory of the form

$$P_i(db_i; b_{i-1}, b_{i-2}, \ldots, b_{i-k}, a_i, a_{i-1}, \ldots, a_{i-m}), \ k \in \{1, 2, \ldots, K\}, \ m \in \{0, 1, \ldots, M\}$$

3. The main point to be made here is that the $\{P_i(db_i; b^{i-1}) : i = 0, 1, \ldots, n\}$ is the controlled process and $\{P_i(da_i; a^{i-1}, b^{i-1}) : i = 0, 1, \ldots, n\}$ is the control process.

Christos K Kourtellaris

4. The structural properties of channel input distribution which maximize information capacity also hold for information capacity with transmission cost constraints defined by

$$\mathscr{P}_{0,n}(P) \triangleq \Big\{P_i(da_i;a^{i-1},b^{i-1}) \in \mathscr{K}(\mathscr{A}_i;\mathscr{A}_{0,i-1}\times\mathscr{B}_{0,i-1}) : i=0,1,\ldots,n : \\ \frac{1}{n+1}\mathbb{E}\Big\{\sum_{j=0}^{n} e_i(a^j,b^{j-1})\Big\} \leq P\Big\}$$

provided the following condition holds.

- Condition 1.
  If for every $F\in\mathbb{B}(\mathscr{B}_i)$, the channel $P_i(F;.,.)$ is $\mathbb{B}(I_i)$-measurable then $e_i(.,.)$ is also $\mathbb{B}(I_i)$-measurable for $i=0,1,\ldots,n$.

For example Condition 1 holds if the channel satisfies

$$P_i(db_i;b^{i-1},a^i,x^i) = P_i(db_i;b_{i-1},a_i,a_{i-1}) - a.a.\ (b^{i-1},a^i,x^i),\ i=0,1,\ldots,n$$

and the cost function is

$$e_i(a^i,b^{i-1}) = \bar{e}(a_i,a_{i-1},b_{i-1}),\ i=0,1,\ldots,n$$

Remark 3.35. Theorem 3.31 can be generalized to channels of the form

$$P_j(db_j;b^{j-1},x^j,a^j) = P_j(db_j;b^{j-1},a_{j-k},a_{j-k+1},\ldots,a_j) -\ a.a.\ (x^j,b^{j-1},a^j), \quad j=0,1,\ldots,n$$

and even to channels which also depend on $\{x_i : i=0,\ldots,n\}$.

Theorem 3.31 gives as a special case the results stated in [6, 18] for the unit memory channel defined by

$$P_j(db_j;b^{j-1},x^j,a^j) = P_j(db_j;b_{j-1},a_j) -\ a.a.\ (x^j,b^{j-1},a^j), \quad j=0,1,\ldots,n$$

Christos K Kourtellaris

3.5 Equivalence of Encoder Design and Capacity Achieving Distributions

In this section, we put together the structural properties of extremum problems of designing encoders and finding the capacity achieving distribution to show that these optimization problems are equivalent. Therefore, one does not need to treat these problems separately, but rather one can indeed synthesize JSCC and source-channel matching based on nonanticipative transmission.

Here, we consider the special case when the information capacity is defined via $C^3_{0,n}$ and we show how to design encoders so that directed information including the encoder but not the decoder, is precisely equal to $C^3_{0,n}$, and it is achievable via SbS code.

The results of this section are valid under the following assumption.

Assumption 3.36. Assumption 3.7 holds and assumption 3.8 is replaced by

$$P_j(db_j;b^{j-1},a^j,x^j) = P_j(db_j;b^{j-1},a_j) - a.a.\ (b^{j-1},a^j,x^j) : j = 0,1,\dots,n \tag{3.5.173}$$

Remark 3.37. The reason we consider the source given by Assumption 3.7 instead of

$$P_j(dx_j;x^{j-1},a^{j-1},b^{j-1}) = P_j(dx_j;x_{j-1}) - a.a.(x^{j-1},a^{j-1},b^{j-1}) : j = 0,1,\dots,n$$

is to allow controlled sources as well, hence we do not impose Assumption 3.20. Condition (3.5.173) implies that the information capacity is that of Definition 3.30.

We state the following corollary of Theorem 3.9.

Theorem 3.38. *Under Assumption 3.36 for any encoder from the class $\mathscr{E}^{DF}[0,n]$ we have*

$$\sup_{e\in\mathscr{E}^{DF}[0,n]} I(X^n \to B^n) \stackrel{\triangle}{=} \sup_{e\in\mathscr{E}^{DF}[0,n]} \sum_{i=0}^{n} \mathbb{E}^e \left\{ \log\left(\frac{P^e_{B_i|B^{i-1},X^i}(dB_i|B^{i-1},X^i)}{P^e_{B_i|B^{i-1}}(dB_i|B^{i-1})} \right) \right\} \tag{3.5.174}$$

$$= \sup_{g\in\mathscr{E}^{DM}[0,n]} \sum_{i=0}^{n} I(A_i;B_i|B^{i-1})|_{A_i=g_i(X_i,B^{i-1})} \tag{3.5.175}$$

Christos K. Kourtellaris

Proof. The following identities hold.

$$
\begin{aligned}
I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} \mathbb{E}^e \left\{ \log \left(\frac{P_i(dB_i|B^{i-1}, X^i)}{P_i(dB_i|B^{i-1})} \right) \right\} \\
&= \sum_{i=0}^{n} \mathbb{E}^e \left\{ \log \left(\frac{P_i(dB_i|B^{i-1}, X^i, A^i)}{P_i(dB_i|B^{i-1})} \right) \Big|_{A_j = e_j(X^j, B^{j-1}) : j=0,1,\ldots,i} \right\} \\
&= \sum_{i=0}^{n} \mathbb{E}^g \left\{ \log \left(\frac{P_i(dB_i;B^{i-1}, A_i)}{P_i(dB_i|B^{i-1})} \right) \Big|_{A_i = g_i(X_i, B^{i-1})} \right\} && (3.5.176) \\
&= \sum_{i=0}^{n} I(A_i;B_i|B^{i-1}) && (3.5.177)
\end{aligned}
$$

where (3.5.176) follows from the MC of Assumption 3.36. Therefore, by Theorem 3.9 maximizing $I(X^n \rightarrow B^n)$ over $\mathscr{E}^{DF}[0,n]$ is equivalent to maximizing $\sum_{i=0}^{n} I(g_i(X_i, B^{i-1}); B_i|B^{i-1})$ over $\mathscr{E}^{DM}[0,n]$, and (3.5.175) is obtained .

□

By the MC of Assumption 3.36, and the structural properties of the channel input distributions $\{P_i(da_i; a^{i-1}, b^{i-1}) \in \mathscr{K}(\mathscr{A}_i; \mathscr{A}_{0,i-1} \times \mathscr{B}_{0,i-1}) : i = 0, 1, \ldots, n\}$ which maximize directed information $I(A^n \rightarrow B^n)$, we derive a converse coding theorem to identify a tight upper bound for which an operational meaning will be saught, which is compatible to the expression (3.5.175).

Then, we show equivalence between the problem of computing capacity and the problem of designing the capacity achieving encoder.

Theorem 3.39. *Suppose Assumptions 3.36 hold.*

1. Any achievable rate R satisfies

$$
\begin{aligned}
R &\leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log M_n \\
&\leq \liminf_{n \rightarrow \infty} \sup_{\{P_i(da_i; a^{i-1}, b^{i-1}) \in \mathscr{K}(\mathscr{A}_i; \mathscr{A}_{0,i-1} \times \mathscr{B}_{0,i-1}) : i=0,1,\ldots,n-1\}} \frac{1}{n} C^3_{0,n-1} && (3.5.178) \\
&= \liminf_{n \rightarrow \infty} \sup_{\{P_i(da_i; b^{i-1}) \in \mathscr{K}(\mathscr{A}_i; \mathscr{B}_{0,i-1}) : i=0,1,\ldots,n-1\}} \frac{1}{n} \sum_{i=0}^{n-1} I(A_i; B_i|B^{i-1}) \equiv C^3_{\infty} && (3.5.179)
\end{aligned}
$$

Christos K. Kourtellaris

2. The encoder design and the calculation of capacity are related by

$$\sup_{g\in\mathscr{E}^{DM}[0,n]}\frac{1}{n}\sum_{i=0}^{n-1}I(A_i;B_i|B^{i-1})|_{A_i=g_i(X_i,B^{i-1})}$$
$$=\sup_{\{P_i(da_i;b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}):i=0,1,\ldots,n-1\}}\frac{1}{n}\sum_{i=0}^{n-1}I(A_i;B_i|B^{i-1}) \tag{3.5.180}$$

where $\{X_i : i=0,1,\ldots,n-1\}$ *is a jointly uniform random process in* $[0,1]^n$.

Proof. 1. This part is given in Appendix B.2. 2. Since for general complete separable metric spaces any randomized strategy from the class $\{P_i(da_i;a^{i-1},b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{A}_{0,i-1}\times\mathscr{B}_{0,i-1}) : i=0,1,\ldots,n-1\}$ can be realized by deterministic strategies from the class $g\in\mathscr{E}^{DM}[0,n]$, then (3.5.180) holds. □

The point to be made regarding Theorem 3.39 is that by (3.5.174), the upper bound (3.5.179) is tight, since under very general conditions randomized strategies can be realized by deterministic strategies, via the solution of the maximizing encoder problem (3.5.175). Moreover, this upper bound could not be obtained without knowing the structural properties of encoders maximizing directed information from the source to the channel output, and the structural properties of the information capacity achieving distributions of $C^3_{0,n}$.
Also, for any rate R violating the bound (3.5.179), the probability of decoding error can be arbitrarily near to 1.
Therefore, the information measure for which an achivable SbS code should be saught is the one defined below.

Definition 3.40. Suppose Assumption 3.36 hold.
The finite time information capacity is defined by

$$C^3_{0,n}=\sup_{\{P_i(da_i;b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}):i=0,1,\ldots,n\}}\sum_{i=0}^{n}I(A_i;B_i|B^{i-1}) \tag{3.5.181}$$

The information capacity is

$$C^3_{\infty}=\liminf_{n\to\infty}\sup_{\{P_i(da_i;b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}):i=0,1,\ldots,n\}}\frac{1}{n+1}\sum_{i=0}^{n}I(A_i;B_i|B^{i-1}) \tag{3.5.182}$$

Similarly for $C^3_{0,n}(P)$, $C^3_{\infty}(P)$.

Christos K. Kourtellaris

Therefore, to need to determine the value of the optimization problem $C^3_{0,n}$ and C^3_∞ and the achieving distribution, $\{P^*_i(da;b^{i-1}) : i = 0,1,\ldots,n\}$, which is also equivalent via (3.5.180) to the optimization problem over deterministic encoder policies.

Using Theorem 3.39 we can purse two methods for designing encoder so that the directed information including the encoder, $I(X^n \rightarrow B^n)$ is precisely equal to the supremum in C^3_∞. One method is via dynamic programing to find the optimal encoder, and then the operational capacity C^3_∞ via (3.5.180). The other method is to solve (3.5.181) to find the capacity achieving distribution $\{P^*_i(da_i;b^{i-1}) \in \mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}) : i = 0,1,\ldots,n\}$. We describe the later.

Let $\{P^*_i(da_i;b^{i-1}) \in \mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}) : i = 0,1,\ldots,n\}$ be the sequence of stochastic Kernels which achieves the supremum of $C^3_{0,n}$, and let $F^*_{A_i|B^{i-1}}(a_i)$ be its corresponding conditional distribution.

Consider a separated encoder of the form

$$\Big\{a^*_i = g^*_i(x_i,b^{i-1}) = g^{*,sep}_i(x_i,P_i(dx_i;b^{i-1})) : i = 0,1,\ldots,n\Big\} \in \mathscr{E}^{sep,DM}[0,n]$$

where $P_i(dx_i;b^{i-1}) \in \mathscr{K}(\mathscr{X}_i;\mathscr{B}_{0,i-1})$ is a stochastic Kernel, and denote by $F_{X_i|B^{i-1}}(x_i), i = 1,\ldots,n$ its corresponding conditional distribution function. Define the *á* posterior matching scheme by

$$\Big\{A^*_i = g^{*,sep}_i(X_i,F_{X_i|B^{i-1}}(X_i)) = F^{*,-1}_{A_i|B^{i-1}} \circ F_{X_i|B^{i-1}}(X_i) : i = 0,1,\ldots,n\Big\} \tag{3.5.183}$$

This scheme corresponds to an encoder transmitting at each $i \in \mathbb{N}^n$ the symbol A^*_i via the mapping $g^{*,sep}_i(\cdot,B^{i-1})$. The following hold at each $i \in \mathbb{N}^n$.

1. For a fixed $B^{i-1} = b^{i-1}$, $F_{X_i|B^{i-1}}(x_i)$ is a random variable uniformly distributed on the interval $[0,1)$. Hence, it is independent of b^{i-1}.
2. For a fixed $B^{i-1} = b^{i-1}$, $F^{*,-1}_{A_i|B^{i-1}}(\cdot)$ is the inverse of a distribution function, applied to a uniformly distributed random variable. Hence, it transforms the uniform random variable $U_i = F_{X_i|B^{i-1}}(x_i)$ into a random variable A^*_i having the finite capacity achieving distribution $F^*_{A_i|B^{i-1}}(a_i)$. That is, $F^{*,-1}_{A_i|B^{i-1}} \circ F_{X_i|B^{i-1}}(x_i)$ for a fixed $B^{i-1} = b^{i-1}$ transforms A^*_i into a RV distributed according to $F^*_{A_i|B^{i-1}}$.

Christos K. Kourtellaris

Substituting the above PMS into $I(X^n \to B^n) = \sum_{i=0}^n I(X^i;B_i|B^{i-1})$, by (3.5.175) we have

$$\begin{aligned} I(X^n \to B^n) &= \sum_{i=0}^{n} I(A_i^*;B_i|B^{i-1})|_{A_i^*=g_i^{*,sep}(X_i,B^{i-1})} \\ &= \sup_{\{P_i(da_i;b^{i-1})\in\mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}):i=0,1,\ldots,n-1\}} \sum_{i=0}^{n-1} I(A_i;B_i|B^{i-1}) \end{aligned} \tag{3.5.184}$$

Remark 3.41. The left hand side of (3.5.180) when solved, determines $g \in \mathscr{E}^{DM}[0,n]$ which transforms via the uniform random process $\{X_i : i = 0,1,\ldots,n-1\}$, $\{A_i = g_i(B^{i-1},X_i) : i = 0,1,\ldots,n-1\}$ into the capacity achieving distribution.

Remark 3.42. Although, we have shown, PMS holds under the general Assumption 3.36, we need to ensure this scheme has an operational meaning, in terms of decoding error probability.

This can be done via the Dual of the SbS code achievability of Theorem 2.29 as follows.

1. In Definition 2.26 of SbS code, (n,d,ε,P), the channel $P_i(db_i;b^{i-1},a^i,x^i), i = 0,\ldots,n$ is fixed and the capacity achieving PMS encoder is found and fixed (and Assumption 3.36 hold).

2. In Definition 2.27 of minimum excess distortion specify the decoding error function, for example, precisely as the excess distortion probability.

3. In Definition 4.9 of realization of $R^{na}(D)$, fix the channel and the capacity achieving PMS encoder, and find the decoder which realizes the optimal reproduction distribution of the nonanticipative RDF, $R^{na}(D)$, and realizes $R^{na}(D)$.

4. For fixed source-encoder-channel-decoder, and for a given P find a $D \in [D_{min}, D_{max}]$, so that $R^{na}(D)$ is finite, and the excess decoding error probability is satisfied.

Then the SbS code is achievable (JSCC) with respect to decoding error probability and $\lim_{n\to\infty} \frac{1}{n+1} C_{0,n}(P) \geq \lim_{n\to\infty} \frac{1}{n+1} R_{0,n}^{na}(D), \forall n \in \mathbb{N}$.

Since in general $C_{0,n}(P) \geq R_{0,n}^{na}(D), \forall n \in \mathbb{N}$ this does not correspond to the minimum rate of reproducing the source at the decoder, that is, it does not correspond to a source-channel matching code.

The achievable SbS code corresponds to the source-channel matching for a given P, if there exists a $D \in [D_{min}, D_{max}]$, so that $R^{na}(D)$ is finite and $\lim_{n\to\infty} \frac{1}{n+1} C_{0,n}(P) = \lim_{n\to\infty} \frac{1}{n+1} R_{0,n}^{na}(D)$.

Christos K. Kourtellaris

Remark 3.43.

1. The previous PMS can be generalized to information capacity formulae without imposing Assumptions 3.36, for certain general channels of the form

$$P_j(db_j;b^{j-1},a^j_{j-K},x^j_{j-m}), \qquad j=0,1,\ldots,n, \;\; K,M finite \tag{3.5.185}$$

For DMCs with feedback the capacity achieving channel input distribution satisfies $P_i^*(da_i;b^{i-1}) = P^*(da) \equiv F^*_{A_i|B^{i-1}}(a_i) = F^*_A(a), \forall i \in \mathbb{N}$, the joint process $\{(A_i,B_i) : i \in \mathbb{N}^n\}$ is independent and in the limit ergodic, that is, $F^*_A(.)$ is the distribution which achieves the supremum of the single letter expression $I(A;B)$, and the PMS is

$$\Big\{A_i^* = g_i^{*,sep}(X_i, F_{X_i|B^{i-1}}(X_i)) = F_A^{*,-1} \circ F_{X_i|B^{i-1}}(X_i) : i=0,1,\ldots,n\Big\} \tag{3.5.186}$$

2. The achievability for the PMS will be revisited Chapter 4 to address source-channel maching so that $\lim_{n\to\infty}\frac{1}{n+1}C_{0,n}(P) = \lim_{n\to\infty}\frac{1}{n+1}R^{na}_{0,n}(D)$.

3.6 The Binary State Symmetric Channel

In this section we apply the previous results regarding the structural properties of encoders to a specific channel with memory, with or without feedback. Chen and Berger [18] derived coding theorems for channels with unit memory and feedback, in which the channel output $\{B_i : i=0,1,\ldots\}$ and the channel input-output pair $\{(A_i,B_i) : i=0,1,\ldots\}$ is assumed to be first order Markov process, while Berger in his Shannon lecture [6], conjectured the form of the capacity achieving input distribution. The Binary State Symmetric channel $BSSC(\alpha_1,\beta_1)$, or POST channel [3], is a special case of the limited memory channel [6], defined by the transition probabilities given below.

$$P_{B_i|A_i,B_{i-1}}(b_i|a_i,b_{i-1}) = \begin{array}{ccccc} & 0,0 & 0,1 & 1,0 & 1,1 \\ 0 & \alpha_1 & \beta_1 & 1-\beta_1 & 1-\alpha_1 \\ 1 & 1-\alpha_1 & 1-\beta_1 & \beta_1 & \alpha_1 \end{array}$$

Recently Asnani, Permuter and Weissman [3, 58], derived a closed form expression for the unconstraint feedback capacity of the Previous Output State (POST) channel, while they

Christos K. Kourtellaris

proved the surprising result that feedback does not increase the capacity of this channel. In their approach they consider the previous channel output as the state of the channel.

Our work which is motivated by SbS joint source channel coding of a binary source with memory via the $BSSC(\alpha_1, \beta_1)$ (Chapter 4), compliments the recent work of Asnani, Permuter and Weissman [3, 58]. In our analysis, we define the state of the channel as the modulo2 addition of the current channel input and the previous channel output, which breaks the channel into two binary symmetric channels. We additionally impose a "natural" cost constraint on the channel which is necessary for JSCC elaborated in Chapter 4. In general the cost function is an important element to achieve optimality of source-channel communication system [6, 29].

Our main contributions are the analytical expressions for the constraint capacity with feedback, and the optimal input and output distributions. Moreover, we provide the optimal input distribution for the no feedback capacity, constraint or unconstraint, where we show that a first order Markovian input distribution induces the optimal output distribution of the feedback case, and we give an analytical expression for the no feedback capacity. Finally, from the constraint capacity with feedback, we obtain as a special case the unconstraint feedback capacity formulae, derived in [3, 58]. However, our capacity formulae highlights the optimal time sharing among the two binary symmetric channels (states of the general unit memory channel), and provide the capacity achieving channel input distribution.

The unit memory channel is defined by

$$\overrightarrow{P}_{B^n|A^n}(db^n|a^n) \stackrel{\triangle}{=} \otimes_{i=1}^{n} P_{B_i|A_i,B_{i-1}}(db_i|a_i,b_{i-1}) \tag{3.6.187}$$

Definition 3.44. (Cost constraint for the unit memory channel)
The cost of transmitting a specific symbol over the unit memory channel, defined by a measurable function $\gamma_n : \mathscr{A}^n \times \mathscr{B}^{n-1} \mapsto [0,\infty)$,

$$\gamma_n(a^n, b^{n-1}) \stackrel{\triangle}{=} \sum_{i=1}^{n} c_i(a_i, b_{i-1})$$

Definition 3.45. (Capacity achieving distribution with feedback)
The form of the capacity achieving distribution for the unit memory channel with feedback is given by (follows from Theorem 3.31)

$$\overleftarrow{P}_{A^n|B^{n-1}}(da^n|b^{n-1}) \stackrel{\triangle}{=} \otimes_{i=1}^{n} P_{A_i|B_{i-1}}(da_i|b_{i-1}) \tag{3.6.188}$$

Christos K. Kourtellaris

Definition 3.46. (Capacity achieving distribution without feedback)
The form of the capacity achieving distribution for the unit memory channel without feedback is given by $\{P_{A_i|A^{i-1}}(da_i|a^{i-1}) : i \in \mathbb{N}^n\}$

$$P_{A^n}(A^n) = \otimes_{i=1}^n P_{A_i|A^{i-1}}(da_i|a^{i-1})$$

The Binary State Symmetric Channel (BSSC) is a special case of the unit memory channel, and is defined via the transition probabilities

$$P_{B_i|A_i,B_{i-1}}(b_i|a_i,b_{i-1}) = \begin{array}{c|cccc} & 0,0 & 0,1 & 1,0 & 1,1 \\ \hline 0 & \alpha_1 & \beta_1 & 1-\beta_1 & 1-\alpha_1 \\ 1 & 1-\alpha_1 & 1-\beta_1 & \beta_1 & \alpha_1 \end{array} \tag{3.6.189}$$

We define the state of the channel, at any time instant $i \in \mathbb{N}^n$, as the modulo2 addition of the input symbol a_i and the previous output symbol b_{i-1}, $s_i = a_i \oplus b_{i-1}$. Due to the invertability of the state, given the channel state and any of the channel input or previous output symbol, the remaining symbol is uniquely defined. Thus, we may transform the channel $P_{B_i|A_i,B_{i-1}}$ to its equivalent form $P_{B_i|A_i,S_i}$. The channel then breaks down into two symmetric binary channels with crossover probabilities $(1-\alpha_1)$ and $(1-\beta_1)$, given by

$$P_{B_i|A_i,S_i}(b_i|a_i,s_i=0) = \begin{bmatrix} \alpha_1 & 1-\alpha_1 \\ 1-\alpha_1 & \alpha_1 \end{bmatrix} = P_{B_i|S_i,B_{i-1}}(b_i|s_i=0,b_{i-1})$$

$$P_{B_i|A_i,S_i}(b_i|a_i,s_i=1) = \begin{bmatrix} \beta_1 & 1-\beta_1 \\ 1-\beta_1 & \beta_1 \end{bmatrix} = P_{B_i|S_i,B_{i-1}}(b_i|s_i=1,b_{i-1})$$

We define the Binary Symmetric Channel with crossover probability $(1-\alpha_1)$, $BSC(1-\alpha_1)$, as the “state zero” channel, and the Binary Symmetric Channel with crossover probability $(1-\beta_1)$, $BSC(1-\beta_1)$, as the “state one” channel. Next, we define a cost constraint on the channel that has the following physical interpretation. Assume $\alpha_1 > \beta_1 \geq 0.5$. Then the capacity of the state zero channel $(1-H(\alpha_1))$, is greater than the capacity of the state one channel $(1-H(\beta_1))$. With “abuse” of terminology, we interpret the $(BSC(1-\alpha_1))$ as the “good channel” and the $(BSC(1-\beta_1))$ as the bad channel. It is further reasonable to assume

Christos K. Kourtellaris

that we pay a larger fee to use the “good channel” and a smaller fee to use the “bad channel”. We quantify this policy by assigning a binary pay-off to each of the channels. Hence, we assign a cost equal to 1 for the good channel, and a cost equal to 0 for the bad channel, $\forall i \in \mathbb{N}$, given by

$$c_i(a_i,b_{i-1}) = \begin{cases} 1 & \text{if } a_i = b_{i-1}, \text{ or } s_i = 0 \\ 0 & \text{if } a_i \neq b_{i-1}, \text{ or } s_i = 1 \end{cases}$$

The letter-by-letter average cost constraint is given by

$$\mathbb{E}\{c(A_i,B_{i-1})\} = P_{A_i,B_{i-1}}(0,0) + P_{A_i,B_{i-1}}(1,1) = P_{S_i}(0)$$

The binary form of the constraint does not downgrade the problem, since it can be easily upgraded to more complex forms, without affecting the proposed methodology (i.e. $(1-\delta)$, δ, where $\delta = \mathit{constant}$). Additionally, if $\beta_1 > \alpha_1 \geq 0.5$ we reverse the cost, while if α_1 and/or β_1 are less than 0.5 we flip the respective channel input.

3.6.1 Capacity of the BSSC with Feedback

In this section we provide closed form expressions for the feedback capacity and the optimal input distributions, both for the constrained case and the unconstrained case. The feedback capacity of the $BSSC(\alpha_1,\beta_1)$ without cost constraint is given by,

$$\begin{aligned} C^{fb} &\triangleq \lim_{n\to\infty} \max_{\overleftarrow{P}_{A^n|B^{n-1}}} \frac{1}{n} I(A^n \to B^n) \\ &= \lim_{n\to\infty} \max_{\overleftarrow{P}_{A^n|B^{n-1}}} \frac{1}{n} \sum_{i=1}^{n} \{H(B_i|B^{i-1}) - H(B_i|Bi-1, A^i)\} \\ &\overset{(\alpha_1)}{=} \lim_{n\to\infty} \max_{\{P(A_i|B_{i-1})\}_{i=1}^{n}} \frac{1}{n} \sum_{i=1}^{n} \{H(B_i|B_{i-1}) - H(B_i|A_i,B_{i-1})\} \end{aligned} \tag{3.6.190}$$

where (α_1) holds from the property of the input distributions that generates a Markov process at the output (Theorem. 3.31). By the form of the input distribution we have

1. $\{B_i : i = 1,2,\ldots,n\}$ is a first order Markov chain.
2. $\{(A_i,B_i) : i = 1,2,\ldots,n\}$ is a first order Markov chain.

Christos K. Kourtellaris

If we further assume that $\{P_{A_i|B_{i-1}} : i = 1,2,\ldots\}$ is stationary, then the output process $\{B_i : i = 1,2,\ldots\}$ is a stationary Markov chain (by reconditioning). Moreover, it is shown in [18] the input distribution convergence to a stationary distribution as $n \longrightarrow \infty$. Using [18], then

$$C^{fb} = \max_{\{P(A_i|B_{i-1})\}} \{H(B_i|B_{i-1}) - H(B_i|A_i,B_{i-1})\} \tag{3.6.191}$$

Expression (3.6.191) is also shown in [58].

3.6.1.1 Constrained Capacity with Feedback

The constraint capacity of the $BSSC(\alpha_1,\beta_1)$ with feedback, is defined by (3.6.190), where the input distribution additionally satisfies the average cost constraint, as follows

$$\begin{aligned} C^{fb} &\stackrel{\triangle}{=} \lim_{n\to\infty} \max_{\overleftarrow{P}_{A^n|B^{n-1}}:\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{c(A_i,B_{i-1})\}=\kappa} \frac{1}{n} I(A^n \rightarrow B^n) \\ &= \lim_{n\to\infty} \max_{\overleftarrow{P}_{A^n|B^{n-1}}:\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{c(A_i,B_{i-1})\}=\kappa} \frac{1}{n}\sum_{i=1}^{n} \{H(B_i|B_{i-1}) - H(B_i|A_i,B_{i-1})\} \end{aligned} \tag{3.6.192}$$

Due to the form of the cost constraint, it can be shown that the derivation in [18] remains valid, and and that the limiting channel input distribution is stationary, i.e., $\{P_{A_i|B_{i-1}} : i = 1,2,\ldots,n\}$ convergence to a stationary distribution as $n \longrightarrow \infty$, as in the unconstraint case.

While for the unconstraint case the input distribution $P_{A_i|B_{i-1}}(a_i|b_{i-1}) \in [0,1], \forall(a_i,b_{i-1}) \in \{0,1\}$, the imposed average cost constrain $\mathbb{E}\{c(a_i,b_{i-1})\} = \kappa$, $\kappa \in [0,1]$, restricts the set of input distributions $P_{A_i|B_{i-1}}(a_i|b_{i-1})$ to those that satisfy

$$P_{A_i|B_{i-1}}(0|0)P_{B_{i-1}}(0) + P_{A_i|B_{i-1}}(1|1)P_{B_{i-1}}(1) = \kappa \tag{3.6.193}$$

The constraint rate of the BSSC with feedback is illustrated in Figure 3.6.2. The projection on the distribution plane, denoted by the black dotted line, shows all possible pairs of input distributions that satisfy the cost constraint defined in (3.6.193).

Next, we state the main theorem regarding the constraint capacity of the $BSSC(\alpha_1,\beta_1)$ with feedback.

Christos K. Kourtellais

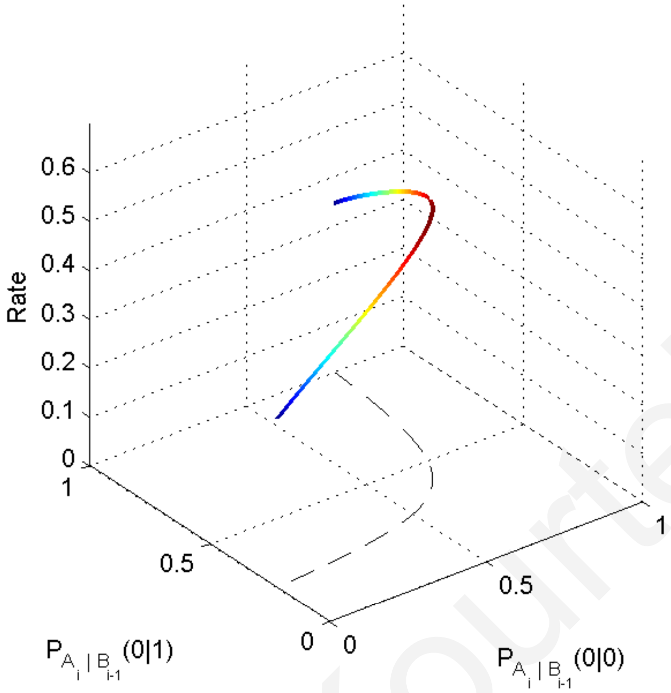

FIGURE 3.6.2: Rate of the BSSC with feedback subject to the cost constraint for $\alpha_1 = 0.92$, $\beta_1 = 0.79$ and $\kappa = 0.71$.

Christos K. Kourtellaris

Theorem 3.47. *(constraint capacity of the BSSC(α_1,β_1) with feedback)*
The feedback capacity of the BSSC(α_1,β_1) subject to the average cost constraint $\mathbb{E}\{c(A_i,B_{i-1})\}$ $= k$ is given by

$$C^{fb}(\kappa) = H(\lambda) - \kappa H(\alpha_1) - (1-\kappa)H(\beta_1) \tag{3.6.194}$$

where $\lambda = \alpha_1\kappa + (1-\kappa)(1-\beta_1)$. The optimal input and output distributions are given by

$$P^*_{A_i|B_{i-1}}(a_i|b_{i-1}) = \begin{array}{c} 0 \\ 1 \end{array}\begin{bmatrix} \kappa & 1-\kappa \\ 1-\kappa & \kappa \end{bmatrix} \tag{3.6.195}$$

$$P^*_{B_i|B_{i-1}}(b_i|b_{i-1}) = \begin{array}{c} 0 \\ 1 \end{array}\begin{bmatrix} \lambda & 1-\lambda \\ 1-\lambda & \lambda \end{bmatrix} \tag{3.6.196}$$

Proof. The second term of the RHS of (3.6.192) is fixed by the cost constraint, and it is given by

$$H(B_i|B_{i-1},A_i) = \kappa H(\alpha_1) + (1-\kappa)H(\beta_1) \tag{3.6.197}$$

We proceed by calculating $H(B_i|B_{i-1})$, with respect to the input distributions that satisfy the cost constraint. The conditional distribution of the output is given by

$$P_{B_i|B_{i-1}} = \sum_{A_i} P_{B_i|A_i,B_{i-1}} P_{A_i|B_{i-1}} \tag{3.6.198}$$

For computational reasons it is preferable to find the maxima with respect to the distribution of the output process that satisfy the average distortion constrain. By combining (3.6.193) and (3.6.198) we rewrite the cost constraint as a function of $P_{B_i|B_{i-1}}$, as shown below,

$$P_{B_i|B_{i-1}}(0|0)P_{B_i}(0) + P_{B_i|B_{i-1}}(1|1)(1-P_{B_i}(0)) = \lambda, \tag{3.6.199}$$

$$P_{B_i}(0)\text{=}P_{B_i|B_{i-1}}(0|0)P_{B_i}(0)\text{+}(1\text{-}P_{B_i|B_{i-1}}(1|1))(1\text{-}P_{B_i}(0)) \tag{3.6.200}$$

where $\lambda = \alpha_1\kappa + (1-\beta_1)(1-\kappa)$. Manipulating (3.6.199) and (3.6.200), we obtain the following expressions for $P_{B_i|B_{i-1}}(0|0)$ and $P_{B_i}(0)$, as functions of $P_{B_i|B_{i-1}}(1|1)$ and constants $\alpha_1, \beta_1, \kappa$.

$$P_{B_i}(0) = \frac{1+\lambda-2P_{B_i|B_{i-1}}(1|1)}{2(1-P_{B_i|B_{i-1}}(1|1))}, \tag{3.6.201}$$

$$P_{B_i|B_{i-1}}(0|0) = \frac{2\lambda-(1+\lambda)P_{B_i|B_{i-1}}(1|1)}{1+\lambda-2P_{B_i|B_{i-1}}(1|1)} \tag{3.6.202}$$

To simplify the notation, we set $q_b \stackrel{\triangle}{=} P_{B_i|B_{i-1}}(1|1)$. The conditional entropy $H(B_i|B_{i-1})$ is then equal to

$$\begin{aligned} H(B_i|B_{i-1}) &= -\sum_{B_i,B_{i-1}} \left(\log(P_{B_i|B_{i-1}})\right) P_{B_i|B_{i-1}} P_{B_{i-1}} \\ &= -\frac{2\lambda-(1+\lambda)q_b}{2(1-q_b)} \log\left(\frac{2\lambda-(1+\lambda)q_b}{1+\lambda-2q_b}\right) \\ &\quad -\frac{1-m}{2}\left(\log(1-q_b)+\frac{q_b}{1-q_b}\log q_b\right) \\ &\quad -\frac{1-m}{2}\log\left(\frac{(1-m)(1-q_b)}{1+m-2q_b}\right) \end{aligned} \tag{3.6.203}$$

Maximizing (3.6.203) with resect to q_b, yields

Christos K. Kourtellaris

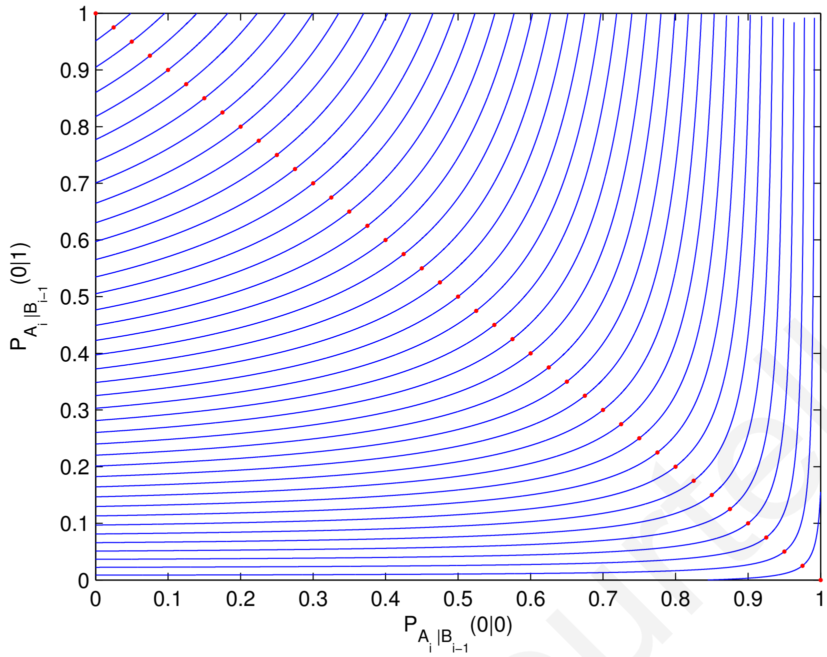

FIGURE 3.6.3: Input distributions and the respective capacity for $\kappa = 0, 0.025, 0.05 \ldots, 1$

$$\begin{aligned}
\frac{dH(B_i|B_{i-1})}{dq_b} &= \frac{1-\lambda}{2(qb-1)^2}\log\left(\frac{2\lambda-(1+\lambda)q_b}{1+\lambda-2q_b}\right) - \frac{(\lambda-1)^2}{2(q_b-1)(1+\lambda-2q_b)} - \frac{1-\lambda}{2(q_b-1)} \\
&\quad - \frac{1-\lambda}{2(q_b-1)^2}\log q_b + \frac{(1-\lambda)q_b}{2(q_b-1)q_b} + \frac{(\lambda-1)^2}{2(q_b-1)(1+\lambda-2q_b)} = 0 \\
&\Rightarrow \frac{1-\lambda}{2(qb-1)^2}\left(\log\left(\frac{2\lambda-(1+\lambda)q_b}{1+\lambda-2q_b}\right) - \log q_b\right) = 0 \\
&\Rightarrow \log\left(\frac{2\lambda-(1+\lambda)q_b}{(1+\lambda-2q_b)q_b}\right) = 0 \\
&\Rightarrow q_b = \lambda \ \text{ and } \ q_b = 1 \ \text{(trivial solution)}
\end{aligned} \tag{3.6.204}$$

By substituting the non-trivial solution of (3.6.204) into the single letter expression of the constraint capacity we obtain (3.6.207), (3.6.208). □

The capacity of the BSSC for various values of κ is illustrated in Figure 3.6.3. Each curve illustrates the input distributions that satisfy the average constraint for a fixed κ, while the red mark illustrates the pair of optimal input distribution that achieve the capacity. These conditional distributions are projected on the line $P_{A_i|B_{i-1}}(0|0) + P_{A_i|B_{i-1}}(0|1) = 1$. which in context with (3.6.197) yield the already proven distribution of the output symbol.

Christos K. Kourtellaris

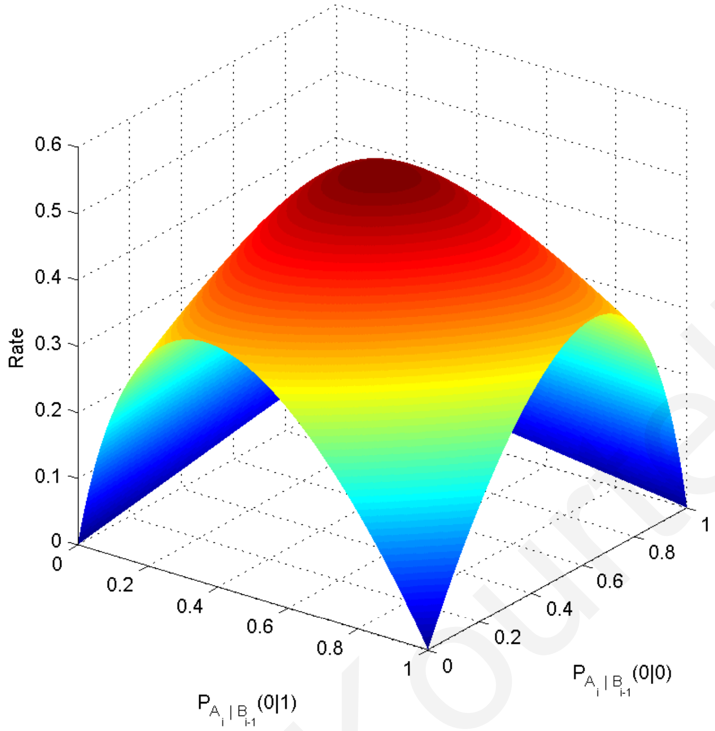

FIGURE 3.6.4: Rate of the BSSC with feedback, for $\alpha_1 = 0.92$ and $\beta_1 = 0.79$.

3.6.1.2 Unconstraint Capacity with Feedback

While the constraint capacity restricts the set of input distributions $P_{A_i|B_{i-1}}(0|0), P_{A_i|B_{i-1}}(0|1)$ to those that satisfy the average cost constraint, the unconstraint capacity allows any values of the input distributions on the set $[0,1] \times [0,1]$. The rate for the unconstraint BSSC with feedback, for fixed α_1, β_1 and κ, is illustrated in Figure 3.6.4.

Next, we show how to obtain the unconstraint capacity with feedback, and thus verify existing results in [58].

Theorem 3.48. *(Unconstraint capacity of the BSSC(α_1, β_1) with feedback)*
The unconstraint feedback capacity of the BSSC(α_1, β_1) is given by

$$C^{fb} = H(\lambda^*) - \kappa^* H(\alpha_1) - (1-\kappa^*) H(\beta_1) \tag{3.6.205}$$

Christos K. Kourtellaris

where

$$\lambda^* = \alpha_1 \kappa^* + (1-\kappa^*)(1-\beta_1), \quad \kappa^* = \frac{\beta_1(1+2^{\frac{H(\beta_1)-H(\alpha_1)}{\alpha_1+\beta_1-1}})-1}{(\alpha_1+\beta_1-1)(1+2^{\frac{H(\beta_1)-H(\alpha_1)}{\alpha_1+\beta_1-1}})} \tag{3.6.206}$$

The optimal input and output distributions are given by

$$P^*_{A_i|B_{i-1}}(a_i|b_{i-1}) = \begin{bmatrix} \kappa^* & 1-\kappa^* \\ 1-\kappa^* & \kappa^* \end{bmatrix} \tag{3.6.207}$$

$$P^*_{B_i|B_{i-1}}(b_i|b_{i-1}) = \begin{bmatrix} \lambda^* & 1-\lambda^* \\ 1-\lambda^* & \lambda^* \end{bmatrix} \tag{3.6.208}$$

Proof. The unconstraint capacity is defined as the maximization of directed information over all possible channel input distributions. Alternatively, it may be defined via the double maximization of the input distributions that satisfy the average cost constraint maximized over all possible values of the constraint, via

$$C = \max_{\kappa} C^{fb}(\kappa) \tag{3.6.209}$$

The second approach can be easily computed by maximizing the already known expression of the constraint capacity, with respect to k, as follows.

$$\begin{aligned} \frac{dC(\kappa)}{d\kappa} &= (\alpha_1+\beta_1-1)\log(b-\kappa^*(\alpha_1+\beta_1-1)) \\ &\quad -(\alpha_1+\beta_1-1)\log(1-b+\kappa^*(\alpha_1+\beta_1-1)) + H(\beta_1) - H(\alpha_1) \\ &= 0 \\ &\Rightarrow \kappa^*(\alpha_1+\beta_1-1)(1+2^{\frac{H(\beta_1)-H(\alpha_1)}{\alpha_1+\beta_1-1}}) = \beta_1(1+2^{\frac{H(\beta_1)-H(\alpha_1)}{\alpha_1+\beta_1-1}})-1 \end{aligned}$$

Therefore the optimal average cost constraint the maximizes the capacity is given by

$$\kappa^* = \frac{\beta_1(1+2^{\frac{H(\beta_1)-H(\alpha_1)}{\alpha_1+\beta_1-1}})-1}{(\alpha_1+\beta_1-1)(1+2^{\frac{H(\beta_1)-H(\alpha_1)}{\alpha_1+\beta_1-1}})}$$

□

Christos K. Kourtellaris

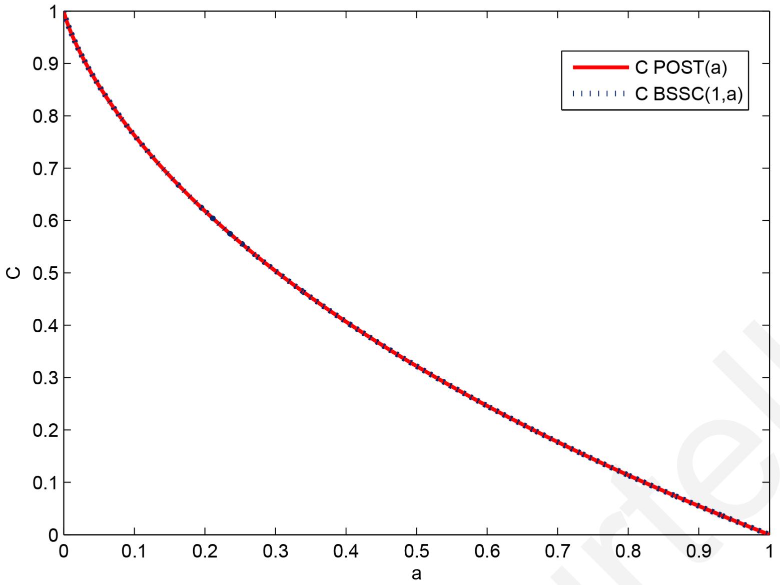

FIGURE 3.6.5: Comparison of the results for the POST(a) channel and the BSCC(1,a).

The result of the unconstraint feedback capacity may be physically interpreted as the optimal time sharing (κ^*) among the two symmetric states of the channel.

In Figure 3.6.5 we show the graphs of the unconstraint capacity of the BSSC$(1,\alpha)$ with feedback, and that derived in [58], to illustrate that they are identical as expected. For all values chosen the graph is identical.

3.6.2 Capacity of the BSSC without Feedback

In general, feedback increases the capacity of a channel with memory. This statement does no hold for the BSSC [3, 58]. Feedback and no feedback capacity are the same if there exists an input distribution P_{A^n} which induces the optimal output distribution, $P^*_{B^n}$, and joint distribution $P^*_{A^n,B^n}$ of the channel with feedback.

Next we introduce an intermediate step to additionally guarantee that the average cost constraint is satisfied. This approach applies to the cost constraint case as well.

Lemma 3.49. *Assume there exists an input distribution of the form* $P^*_{A^n} = \otimes_{i=0}^n P_{A_i|A^{i-1}}$ *that induces the optimal input distribution* $\overleftarrow{P}^*_{A^n|B^n} = \otimes_{i=0}^n P^*_{A_i|A^{i-1},B^{i-1}}$ *and optimal output distribution* $P^*_{B^n}$ *of the feedback case.*

Christos K. Kourtellaris

Then, the feedback capacity is achieved via a no feedback input distribution and the average constraint is satisfied.

Proof. The feedback capacity is a functional of $\{\overleftarrow{P}_{A^n|B^{n-1}}, \overrightarrow{P}_{B^n|A^n}\}$, which define uniquely the joint distribution $\{P_{A^n,B^n}\}$ and the marginal distribution P_{B^n}. In the no feedback case, the capacity is a functional of $\{P_{A^n}, \overrightarrow{P}_{B^n|A^n}\}$ and the maximization is over P_{A^n}. However, if there exists an input distribution for the no feedback case, $P^*_{A^n}$, which induces the optimal input and marginal distribution of the feedback case, $\overleftarrow{P}^*_{A^n|B^n}, P^*_{B^n}$, then the expression of the no feedback capacity is the same as the feedback capacity. □

Theorem 3.50. *For the BSSC(α_1, β_1), the first-order Markovian input distribution $P^*_{A^n} = \otimes_{i=0}^n P^*_{A_i|A_{i-1}}$, which induces the optimal input and output distribution of the feedback case given by $P^*_{A_i|B_{i-1}}$ and $P^*_{B^n}$ respectively, is given by*

$$P^*_{A_i|A_{i-1}}(a_i|a_{i-1}) = \begin{bmatrix} \frac{1-\kappa-\gamma}{1-2\gamma} & \frac{\kappa-\gamma}{1-2\gamma} \\ \frac{\kappa-\gamma}{1-2\gamma} & \frac{1-\kappa-\gamma}{1-2\gamma} \end{bmatrix} \tag{3.6.210}$$

where $\gamma = \alpha_1\kappa + \beta_1(1-\kappa)$.
For the unconstraint case $\kappa = \kappa^*$ and $\gamma = \gamma^*$.

Proof. To prove the claims, we need to show that a Markovian input distribution achieves the capacity achieving channel input distribution with feedback. Consider the following identities.

$$\begin{aligned} P^*_{A_i|B_{i-1}} &= \sum_{A_{i-1}} P_{A_i|A_{i-1},B_{i-1}} P_{A_{i-1}|B_{i-1}} \\ &= \sum_{A_{i-1}} P_{A_i|A_{i-1},B_{i-1}} \frac{P_{B_{i-1}|A_{i-1}} P_{A_{i-1}}}{P_{B_{i-1}}} \\ &= \sum_{A_{i-1}} \frac{P_{A_i|A_{i-1}} P_{A_{i-1}}}{P_{B_{i-1}}} \sum_{B_{i-2}} P_{B_{i-1}|A_{i-1},B_{i-2}} P_{B_{i-2}|A_{i-1}} \\ &= \sum_{A_{i-1}} \frac{P_{A_i|A_{i-1}}}{P_{B_{i-1}}} \sum_{B_{i-2}} P_{B_{i-1}|A_{i-1},B_{i-2}} P_{A_{i-1}|B_{i-2}} P_{B_{i-2}} \end{aligned} \tag{3.6.211}$$

Thus, we search for an input distribution without feedback $P_{A_i|A_{i-1},B_{i-1}} = P_{A_i|A_{i-1}}$ that satisfies (3.6.211). Solving iteratively this system of equations yields the values of the optimal input

Christos K. Kourtellaris

distribution without feedback given by (3.6.210). Since $P^*_{B_i|B_{i-1}} = \sum_{A_i} P^*_{B_i|B_{i-1},A_i} P^*_{A_i|B_{i-1}}$ and $P^*_{A_i|A_{i-1}}$ given by (3.6.210) induce $P^*_{A_i|B_{i-1}}$, then it also induce $P^*_{A_i|B_{i-1}}$. □

If we assume that we begin from the steady state distribution of the output, $P_{B_0}(0) = 0.5$, then (3.6.210) holds for $i \geq 1$. Moreover, if we assume that we begin from an arbitrary initial state, then the optimal input distributions at any time instant is process which converges to a stationary symmetric Markov form given by (3.6.210), in finite number of steps. In Figure 3.6.6, we illustrate that the input distribution converges for the worst case scenario, in terms of convergence, where $\alpha_1 = 1$ and $\beta_1 = 0.5$. The terms $P_0(.|.)$ and $P_1(.|.)$ indicate the conditional input distribution without feedback given $B_0 = 0$ and $B_0 = 1$, respectively, while the terms $P_0(.)$ and $P_1(.)$ indicate the distribution of the output symbol given $B_0 = 0$ and $B_0 = 1$, respectively.

If any $P_{A_i|A_{i-1}}(a_i|a_{i-1})$ induces $P^*_{A_i|B_{i-1}}(a_i|b_{i-1})$, then it also induces the optimal output process $P^*_{B_i|B_{i-1}}(b_i|b_{i-1})$ and $P^*_{A_i|A_{i-1}}(a_i|a_{i-1}) = P_{A_i|A_{i-1}}(a_i|a_{i-1})$. The capacity for the unconstraint case is then equal to

$$\begin{aligned} C &= \lim_{n\to\infty} \max_{P_{A^n}} \frac{1}{n} I(A^n \to B^n) \\ &= \max_{P_{A_i|A_{i-1}}} I(A_i;B_i|B_{i-1}) = C^{fb} \end{aligned} \tag{3.6.212}$$

and similarly for the constraint $C(\kappa) = C^{fb}(\kappa)$.

3.6.3 Special Cases of the BSSC with and without Feedback

Next, we provide two special cases of the BSSC where we apply our results and evaluate the unconstraint capacity, the optimal input distributions, with or without feedback, and the optimal output distributions.

I. Memoryless BBSC ($\alpha_1 = \beta_1 = 1 - \varepsilon$):

Consider the case where $\alpha_1 = \beta_1 = 1 - \varepsilon$. This reduces the BSSC to the memoryless Binary Symmetric Channel (BSC) since both of the channels are binary symmetric channels with cross over probability ε. Then, $\kappa^* = 0.5$ and $\lambda^* = 0.5$, thus the optimal input and output

Christos K. Kourtellaris

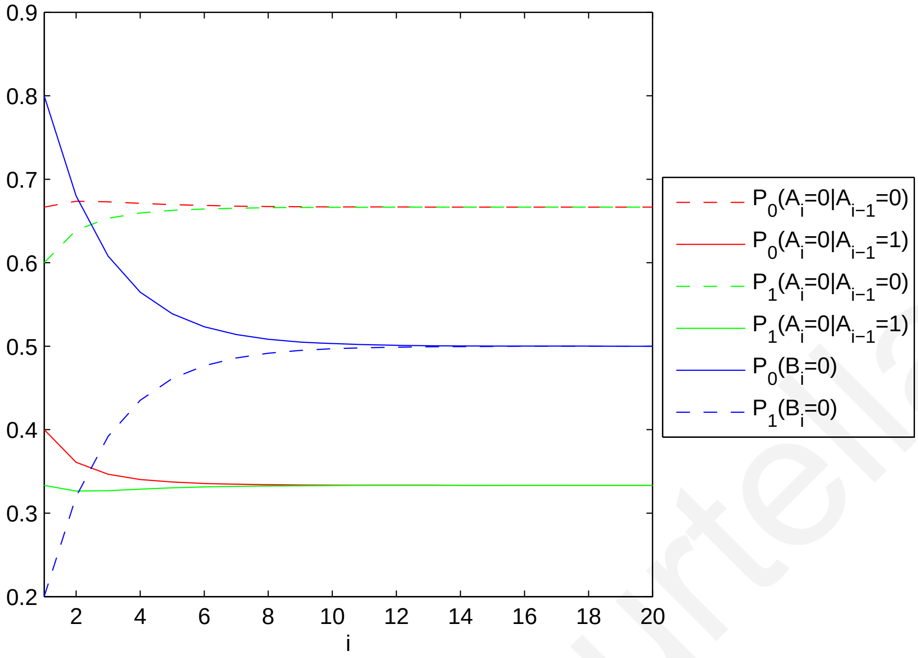

FIGURE 3.6.6: Convergence of the input distribution without feedback for the worst case scenario $(\alpha_1 = 1, \beta_1 = 0.5)$. The index i denotes the time index.

distributions are IID processes. The capacity is then equal to

$$\begin{aligned} C &= H((1-\varepsilon)(1-0.5)+\varepsilon 0.5) - 0.5H(\varepsilon) - 0.5H(\varepsilon) \\ &= 1 - H(\varepsilon) \end{aligned}$$

These results are consistent to the known results of the memoryless BSC.

II. Best and Worst BBSC $(\alpha_1 = 1, \beta_1 = 0.5)$*:*

Consider the case where $\alpha_1 = 1$ and $\beta_1 = 0.5$. In this case the "state zero" channel behaves as a perfect channel, where the output equal to the input, while the "state one" channel behaves as a bad channel where the output is equal to the input with probability 0.5. By applying equations (3.6.206)-(3.6.208) and (3.6.210), we obtain $\kappa^* = 0.6$, $\lambda^* = 0.8$. Therefore the capacity is equal to

$$C = H(0.2) - 0.6H(1) - 0.4H(0.5) = 0.3219$$

The optimal input distributions, with or without feedback, are given by

Christos K. Kourtellaris

$$P^*_{A_i|B_{i-1}}(a_i|b_{i-1}) = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$$

and

$$P^*_{A_i|A_{i-1}}(a_i|a_{i-1}) = \begin{bmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix}$$

while the optimal output distribution for both is given by

$$P^*_{B_i|B_{i-1}}(b_i|b_{i-1}) = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}$$

3.7 Conclusions

In this chapter we studied two fundamental optimization problems for general channels with memory and feedback.

The first problem investigates the structural properties of encoders, for a given source and channel which maximize the directed information from the source to the channel output. For this problem we focus on the design of encoders, and their structural properties, when the information capacity has an operational meaning. We gave the first complete derivation of the capacity achieving distribution for the unit memory channel.

The second problem investigates the structural properties of capacity achieving distribution. For this problem we focus on the calculation of the capacity achieving distribution, and its structural properties, when the information capacity has an operational meaning.

In addition we presented dynamic programming which can be used to determine the encoders and the capacity achieving distribution. We have used the encoder structural properties and capacity achieving distribution to show how to design encoders which achieve the capacity.

We also generalize the Posterior Matching of designing encoders and decoders which achieve capacity for general channels with memory and feedback.

Christos K. Kourtellaris

Finally, we applied the theoretical framework to the unit memory channel, and we showed that the optimal encoder with feedback depends only on the channel output and not the whole sequence. Subsequently, we calculated closed form expressions for the capacities and the optimal input distributions, in the presence or absence of feedback and cost constraint. Our methodology highlights the symmetric form of the channel, and is interpreted as the optimal time sharing among the two states.

Christos K. Kourtellaris

Chapter 4

Nonanticipative Joint Source Channel Coding for Real-Time Transmission

Christos K. Kourtellaris

Shannon in his seminal paper [65] showed that source coding and channel coding may be treated separately without affecting the optimality of the overall design. Over the years, reliable communication analysis is divided into two parts; optimal source compression and optimal channel coding. Although this separation procedure introduces significant advantages with respect to the theoretical analysis and practical design of the codes, it applies mostly to point to point ergodic communications systems, and it does not apply to multiuser channels or non-ergodic systems. Even for the case of ergodic point-to-point communication link, the drawback of this approach is that it introduces delays by assuming arbitrarily large codeword lengths. Thus, it is not suitable for delay sensitive communication schemes. Additionally, it increases the complexity due to the complex structure of encoders and decoders, and leads to excess demands of resources (memory, computational power, power consumption).

Similarly, classical Joint Source Channel Coding (JSCC) [67], although capable of solving more complex communication problems, such as, sources and channels with memory, or even network communication problems, fails to deal with delays, since its performance is also evaluated in the limit of large blocklengths. JSCC over finite blocklengths, on the other hand, may reduce delays while its error exponent outperforms that of separate source channel coding [21]. Additionally, achievable bounds exists for certain kind of sources with fidelity constraints and channels with cost constraint pairs.

Coding over infinitely large blocklengths, although optimal under certain conditions, it is not the only optimal choice [29, 31]. Two well known source-channel pairs, the IID Bernoulli source with single letter Hamming distortion criterion transmitted via a binary symmetric channel, and the Gaussian source with mean square error distortion transmitted via a Gaussian channel, reinforce the belief that the encoder and the decoder can be jointly designed optimally, processing symbols in real time. This optimal transmission scheme is very simple compared to the complexity of separated source and channel coding. For the IID Bernoulli source, this is achieved by uncoded transmission (the encoder and the decoder are identity maps) over a binary symmetric channel, and this design eliminates the delay and the complexity of the overall scheme. For the IID Gaussian source, it is achieved via semi-coded transmission, to meet the power constraint, over an additive Gaussian noisy channel, with or without feedback. The overall transmission scheme is delayless, while the complexity reduces to minimum, due to the simple form of the encoder-decoder. Therefore, nonanticipative JSCC, uncoded or semi-coded, is an optimal coding approach for these two examples [29, 31].

The objective of this chapter is to put forward a framework for optimal performance and reliable communication based on nonanticipative JSCC for sources and channels with memory, with or without feedback. The necessary theoretical framework builds on the material of the previous two chapters; the nonanticipative rate distortion for sources with memory, and the capacity of channels subject to cost constraint with memory, with and without feedback, which are elaborated extensively in Chapter 2 and Chapter 3. After we introduce the mathematical framework for nonanticipative transmission, we apply it to the Binary Symmetric Markov source with crossover probability p, $(BSMS(p))$ transmitted over the Binary State Symmetric Channel $(BSSC(\alpha_1,\beta_1))$, and we show optimality for the overall design. To evaluate the performance of the nonanticipative transmission of the overall system, we apply the average distortion which evaluates the performance in the limit, and the minimum excess distortion which evaluates the performance for finite number of transmissions.

This chapter consists of the following parts.

- Definitions of nonanticipative and Symbol-by-Symbol (SbS) code [1], and definition of the minimum excess distortion.
- Realization of the optimal non-anticipative reproduction distribution and achievability of the nonanticipative code via a noisy channel.

[1] Recall that SbS code encodes causally the current source symbol.

Christos K. Kourtellaris

- Nonanticipative and SbS JSCC scheme for the Binary Symmetric Markov Source ($BSMS(p)$), via a Binary State Symmetric Channel, ($BSSC(\alpha_1,\beta_1)$). We discuss both feedback and no feedback realizations as well as the unmatched case, where the capacity of the channel is greater that the nonanticipative RDF of the source. The performance is evaluated by the excess distortion probability.

4.1 Problem Formulation

In this section we define the elements of a nonanticipative and SbS code in an abstract setting. Let $\mathbb{N} \stackrel{\triangle}{=} \{0,1,\ldots\}$, $\mathbb{N}^n \stackrel{\triangle}{=} \{0,1,\ldots,n\}$. Let $\mathscr{X},\mathscr{A},\mathscr{B},\mathscr{Y}$ denote the source output, channel input, channel output, and decoder output alphabets, respectively, which are assumed to be complete separable metric spaces (Polish spaces) to avoid excluding continuous alphabets. We define their product spaces by $\mathscr{X}_{0,n} \stackrel{\triangle}{=} \times_{i=0}^n \mathscr{X}$, $\mathscr{A}_{0,n} \stackrel{\triangle}{=} \times_{i=0}^n \mathscr{A}$, $\mathscr{B}_{0,n} \stackrel{\triangle}{=} \times_{i=0}^n \mathscr{B}$, $\mathscr{Y}_{0,n} \stackrel{\triangle}{=} \times_{i=0}^n \mathscr{Y}$. Let $x^n \stackrel{\triangle}{=} \{x_0,x_1,\ldots,x_n\} \in \mathscr{X}_{0,n}$ denote the source sequence of length n, and similarly for channel input, channel output, decoder (reproduction) output sequences, $a^n \in \mathscr{A}_{0,n}$, $b^n \in \mathscr{B}_{0,n}$, $y^n \in \mathscr{Y}_{0,n}$, respectively. We associate the above product spaces by their measurable spaces $(\mathscr{X}_{0,n},\mathbb{B}(\mathscr{X}_{0,n}))$, $(\mathscr{A}_{0,n},\mathbb{B}(\mathscr{A}_{0,n}))$, $(\mathscr{B}_{0,n},\mathbb{B}(\mathscr{B}_{0,n}))$, $(\mathscr{Y}_{0,n},\mathbb{B}(\mathscr{Y}_{0,n}))$. Next, we introduce the various distributions of the blocks appearing in Figure 4.1.1.

Definition 4.1. (Source) The source is a sequence of conditional distributions defined by

$$P_{X^n}(dx^n) \stackrel{\triangle}{=} \otimes_{i=0}^n P_{X_i|X^{i-1}}(dx_i|x^{i-1}), \qquad n \in \mathbb{N}$$

- The source is called Markov if

$$P_{X_i|X^{i-1}}(dx_i|x^{i-1}) = P_{X_i|X_{i-1}}(dx_i|x_{i-1}) - a.a\, x^{i-1}, \quad \forall\, i \in \mathbb{N}^n$$

Definition 4.2. (Encoder) The encoder is a sequence of conditional distributions defined by

$$\overrightarrow{P}_{A^n|B^{n-1},X^n}(da^n|b^{n-1},x^n) \stackrel{\triangle}{=} \otimes_{i=0}^n P_{A_i|A^{i-1},B^{i-1},X^i}(da_i|a^{i-1},b^{i-1},x^i), \qquad n \in \mathbb{N}$$

Thus, the encoder assumes feedback from the output of the channel, and it is nonanticipative in the sense that at each time $i \in \mathbb{N}^n$, $P_{A_i|A^{i-1},B^{i-1},X^i}(da_i|a^{i-1},b^{i-1},x^i)$ is a measurable function

Christos K. Kourtellaris

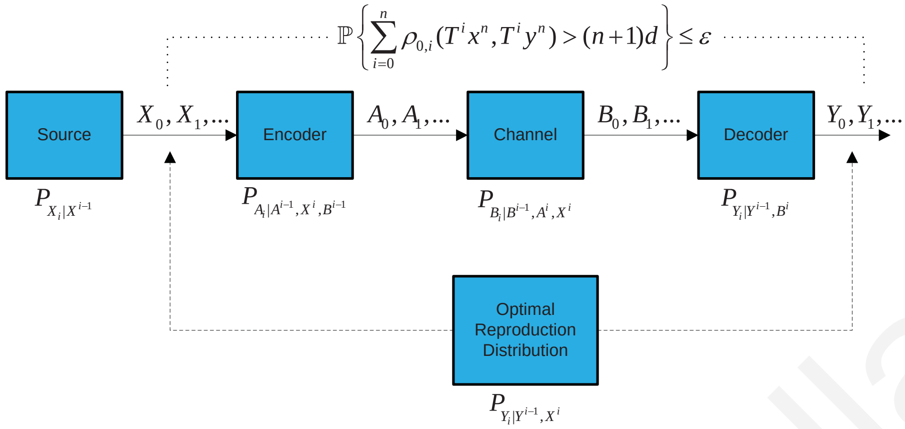
$\mathbb{P}\left\{\sum_{i=0}^{n}\rho_{0,i}(T^i x^n, T^i y^n) > (n+1)d\right\} \leq \varepsilon$
Source
$X_0, X_1, \ldots$
Encoder
$A_0, A_1, \ldots$
Channel
$B_0, B_1, \ldots$
Decoder
$Y_0, Y_1, \ldots$
$P_{X_i|X^{i-1}}$
$P_{A_i|A^{i-1}, X^i, B^{i-1}}$
$P_{B_i|B^{i-1}, A^i, X^i}$
$P_{Y_i|Y^{i-1}, B^i}$
Optimal Reproduction Distribution
$P_{Y_i|Y^{i-1}, X^i}$

FIGURE 4.1.1: Real-Time Communication scheme with feedback.

on past and previous source symbols $x^i \in \mathscr{X}_{0,i}$, past channel input symbols $a^i \in \mathscr{A}_{0,i}$, and past channel output symbols $y^i \in \mathscr{Y}_{0,i}$.

- The encoder is called Markov with respect to the source if

$$P_{A_i|A^{i-1},B^{i-1},X^i}(da_i|a^{i-1},b^{i-1},x^i) = P_{A_i|A^{i-1},B^{i-1},X_i}(da_i|a^{i-1},b^{i-1},x_i) - a.a.$$
$$(a^{i-1},b^{i-1},x^i),\ \forall\, i \in \mathbb{N}^n$$

Markov encoders are SbS encoders because at each time the encoded symbol depends on the last source symbol and not the whole past of the source.

Definition 4.3. (Channel) The channel is a sequence of conditional distributions defined by

$$\overrightarrow{P}_{B^n|A^n,X^n}(db^n|a^n,x^n) \triangleq \otimes_{i=0}^{n} P_{B_i|B^{i-1},A^i,X^i}(db_i|b^{i-1},a^i,x^i), \qquad n \in \mathbb{N}^n$$

Thus the channel has memory, feedback and it also depends nonanticipatively on the source sequence.

- The channel is called Markov with respect to the source if

$$P_{B_i|B^{i-1},A^i,X^i}(db_i|b^{i-1},a^i,x^i) = P_{B_i|B^{i-1},A_i,X_i}(db_i|b^{i-1},a_i,x_i) - a.a.\ (a^i,b^{i-1},x^i),\ \forall\, i \in \mathbb{N}^n$$

- The channel is defined from the input to the channel, A^n, to the output of the channel, B^n, if

$$P_{B_i|B^{i-1},A^i,X^i}(db_i|b^{i-1},a^i,x^i) = P_{B_i|B^{i-1},A^i}(db_i|b^{i-1},a^i) - a.a.\ (a^i,b^{i-1},x^i),\ \forall\, i \in \mathbb{N}^n$$

Christos K. Kourtellaris

- The channel is called first order Markov if

$$P_{B_i|B^{i-1},A^i,X^i}(db_i|b^{i-1},a^i,x^i) = P_{B_i|B_{i-1},A_i,X_i}(db_i|b_{i-1},a_i,x_i) - a.a.\ (a^i,b^{i-1},x^i),\ \forall\, i \in \mathbb{N}^n$$

- The channel is called memoryless if

$$P_{B_i|B^{i-1},A^i,X^i}(db_i|b^{i-1},a^i,x^i) = P_{B_i|A_i}(db_i|a_i) - a.a.\ (a^i,b^{i-1},x^i),\ \forall\, i \in \mathbb{N}^n$$

Definition 4.4. (Decoder) The decoder is a sequence of conditional distributions defined by

$$\overrightarrow{P}_{Y^n|B^n}(dy^n|b^n) \stackrel{\triangle}{=} \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},B^i}(dy_i|y^{i-1},b^i), \quad n \in \mathbb{N}$$

- The decoder is called Markov with respect to the channel output if

$$P_{Y_i|Y^{i-1},B^i}(dy_i|y^{i-1},b^i) = P_{Y_i|Y^{i-1},B_i}(dy_i|y^{i-1},b_i) - a.a.\ (y^{i-1},b^i),\ \forall\, i \in \mathbb{N}^n$$

The above Definitions 4.1-4.4, of source-encoder-channel-decoder are general, they have memory and feedback without anticipation, hence we call the encoder-decoder code, a nonanticipative code.

Given a source, an encoder, a channel, and a decoder, one can define uniquely the joint measure on $(\mathscr{X}_{0,n} \times \mathscr{A}_{0,n} \times \mathscr{B}_{0,n} \times \mathscr{Y}_{0,n}, \mathbb{B}(\mathscr{X}_{0,n}) \times \mathbb{B}(\mathscr{A}_{0,n}) \times \mathbb{B}(\mathscr{B}_{0,n}) \times (\mathbb{B}(\mathscr{Y}_{0,n}))$ by

$$\begin{aligned}
P_{X^n,A^n,B^n,Y^n}(dx^n,da^n,db^n,dy^n) &= \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},B^i,A^i,X^i}(dy_i|y^{i-1},b^i,a^i,x^i) \\
&\quad \otimes P_{B_i|B^{i-1},Y^{i-1},A^i,X^i}(db_i|b^{i-1},y^{i-1},a^i,x^i) \\
&\quad \otimes P_{A_i|A^{i-1},B^{i-1},Y^{i-1},X^i}(da_i|a^{i-1},b^{i-1},y^{i-1},x^i) \\
&\quad \otimes P_{X_i|X^{i-1},A^{i-1},B^{i-1},Y^{i-1}}(dx_i|x^{i-1},a^{i-1},b^{i-1},y^{i-1}) \\
&= \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},B^i}(dy_i|y^{i-1},b^i) \\
&\quad \otimes P_{B_i|B^{i-1},A^i,X^i}(db_i|b^{i-1},a^i,x^i) \\
&\quad \otimes P_{A_i|A^{i-1},B^{i-1},X^i}(da_i|a^{i-1},b^{i-1},x^i) \\
&\quad \otimes P_{X_i|X^{i-1}}(dx_i|x^{i-1})
\end{aligned}$$

Christos K. Kourtellaris

Christos K. Kourtellaris

The previous equality holds if and only if the following Markov chains (MCs) hold.

$$(A^{i-1},B^{i-1},Y^{i-1}) \leftrightarrow X^{i-1} \leftrightarrow X_i, \quad \forall i \in \mathbb{N}^n \tag{4.1.1}$$

$$Y^{i-1} \leftrightarrow (A^{i-1},B^{i-1},X^i) \leftrightarrow A_i, \quad \forall i \in \mathbb{N}^n \tag{4.1.2}$$

$$Y^{i-1} \leftrightarrow (A^{i},B^{i-1},X^i) \leftrightarrow B_i, \quad \forall i \in \mathbb{N}^n \tag{4.1.3}$$

$$(A^{i},X^i) \leftrightarrow (B^{i},Y^{i-1}) \leftrightarrow Y_i, \quad \forall i \in \mathbb{N}^n. \tag{4.1.4}$$

Next, we introduce the distortion function between the source and its reproduction, and the cost function of the channel. The quality of reproducing at each time instant $i \in \mathbb{N}^n$, of x_i by y_i is evaluated by the measurable distortion function

$$d_{0,n} : \mathscr{X}_{0,n} \times \mathscr{Y}_{0,n} \mapsto [0,\infty),\ d_{0,n}(x^n,y^n) \triangleq \sum_{i=0}^{n} \rho(T^i x^n, T^i y^n)$$

where $(T^i x^n, T^i y^n)$ are causal mapping (i.e., for each $i \in \mathbb{N}^n$, $T^i x^n$ and $T^i y^n$ depend on their past and their current symbols respectively). For a single letter distortion function we take $\rho(T^i x^n, T^i y^n) = \rho(x_i, y_i)$. The cost of transmitting a specific symbol over the channel is a measurable function

$$c_{0,n} : \mathscr{A}_{0,n} \times \mathscr{B}_{0,n-1} \mapsto [0,\infty),\ c_{0,n}(a^n,b^{n-1}) \triangleq \sum_{i=0}^{n} \gamma(T^i a^n, T^i b^{n-1})$$

where at each time instant $t \in \mathbb{N}^n$, $T^i b^{n-1}$ depends on $b_0,\ldots,b_{i-1}$. Next, we state the definition of a nonanticipative code with respect to the excess distortion probability.

Definition 4.5. (Nonanticipative code) Let $d \geq 0$, $\varepsilon \in (0,1)$ and $P \geq 0$. An (n,d,ε,P) nonanticipative code for $(\mathscr{X}_{0,n}, \mathscr{A}_{0,n}, \mathscr{B}_{0,n}, \mathscr{Y}_{0,n}\ , P_{X^n}, \overrightarrow{P}_{B^n|A^n,X^n}, d_{0,n}, c_{0,n})$ is a source-channel code $\{P_{A_i|A^{i-1},B^{i-1},X^i}(\cdot|\cdot) : i \in \mathbb{N}^n\}$, $\{P_{Y_i|Y^{i-1},B^i}(\cdot|\cdot) : i \in \mathbb{N}^n\}$ with excess distortion probability

$$\mathbb{P}\Big\{d_{0,n}(x^n,y^n) > (n+1)d\Big\} \leq \varepsilon \tag{4.1.5}$$

and average transmission cost

$$\frac{1}{n+1}\mathbb{E}\Big\{c_{0,n}(a^n,b^{n-1})\Big\} \leq P \tag{4.1.6}$$

Such a code is by definition nonanticipative. A SbS code is a nonanticipative code which satisfies $P_{A_i|B^{i-1},X^i} = P_{A_i|B^{i-1},X_i}$. Moreover, such a code is called SbS if the encoder is Markov with respect to the source if, that is,

$$P_{A_i|B^{i-1},X^i}(a_i|b^{i-1},x^i) = P_{A_i|B^{i-1},X_i}(a_i|b^{i-1},x_i) - a.a.\ (b^{i-1},x^i), \quad \forall i \in \mathbb{N}^n$$

and similarly for the decoder.

The objective of the chapter is to design nonanticipative and SbS codes and show achievability. Next, we define the minimum achievable excess distortion.

Definition 4.6. (Minimum Excess Distortion) The minimum excess distortion achievable by a nonanticipative code (n,d,ε,P) is defined by

$$D^o(n,\varepsilon,P) \overset{\triangle}{=} \inf\Big\{d : \exists(n,d,\varepsilon,P)\ \text{ Nonanticipative code}\Big\}$$

This performance measure is suitable for nonanticipative and SbS transmission of finite length, since it is able to bound the probability of error for a fixed n. Another performance measure which is suitable for nonanticipative transmission in the limit as $n \rightarrow \infty$, is the average distortion.

Clearly, our definition of nonanticipative code is randomized, hence it embeds deterministic codes as a special case. Note that in the absence of a cost constraint on the channel, a nonanticipative code is denoted by (n,d,ε), and we set $D^o(n,\varepsilon,P) = D^o_1(n,\varepsilon)$.

Next, we give an alternative definition of achievability by defining a nonanticipative code via the outage rate probability and the outage capacity.

Definition 4.7. (Nonanticipative code via outage probability) An alternative definition of achievability is obtained by considering an (n,R,ε,D) nonanticipative code with outage rate probability

$$\mathbb{P}\left\{(A^n,B^n) : \frac{1}{n+1}\log\frac{P_{B^n|A^n}(b^n|a^n)}{P_{B^n}(b^n)} < R\right\} \leq \varepsilon \tag{4.1.7}$$

and average fidelity constraint

$$\frac{1}{n+1}\mathbb{E}\Big\{d_{0,n}(X^n,Y^n)\Big\} \leq D \tag{4.1.8}$$

Christos K. Kourtellaris

Definition 4.8. (Outage capacity) The outage capacity achievable by a nonanticipative code (n,R,ε,P) is defined by

$$C^o(n,\varepsilon,P) \stackrel{\triangle}{=} \sup\Big\{R : \exists(n,d,\varepsilon,P) \ \text{Nonanticipative code}\Big\}$$

4.2 Coding Theorems

In this section we show achievability of nonanticipative code.

The realization of the optimal reproduction distribution by an encoder-channel-decoder such that the reproduction of the sequence X^n by the Y^n matches the nonanticipative minimizing reproduction distribution, is necessary for probabilistic matching of the source and the channel. Moreover, if the realization satisfies the fidelity constraint and $\lim_{n\rightarrow\infty}\frac{1}{n+1}I_{P_{X^n}}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n}) = R^{na}(D)$, then $R^{na}(D)$ has an operational meaning. Next, we give the precise definition of the realization. Such a realization is not optimal because the channel in general has higher capacity.

Definition 4.9. (Realization) Given a source $\{P_{X_i|X^{i-1}}\ (dx_i|x^{i-1}) : i = 0,1,\ldots,n\}$, a general channel $\{P_{B_i|B^{i-1},A^i,X^i}\ (db_i|b^{i-1},a^i,x^i) : i = 0,1,\ldots,n\}$ is a realization of the optimal reproduction distribution $\{P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) : i = 0,1,\ldots,n\}$ obtained from the solution of nonanticipative RDF, if there exists a pre-channel encoder $\{P_{A_i|A^{i-1},B^{i-1},X^i}\ (da_i|a^{i-1},b^{i-1},x^i) : i = 0,1,\ldots,n\}$ and a post-channel decoder $\{P_{Y_i|Y^{i-1},B^i}\ (dy_i|y^{i-1},b^i) : i = 0,1,\ldots,n\}$ such that

$$\overrightarrow{P}^*_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^n P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) \tag{4.2.9}$$

where the right hand side of (4.2.9) is obtained from the joint distribution of the source, encoder, channel, decoder, given by

$$\begin{aligned} P_{X^n,A^n,B^n,Y^n}(dx^n,da^n,db^n,dy^n) \quad = \quad & \otimes_{i=0}^n P_{Y_i|Y^{i-1},B^i}(dy_i|y^{i-1},b^i) \\ & \otimes P_{B_i|B^{i-1},A^i,X^i}(db_i|b^{i-1},a^i,x^i) \\ & \otimes P_{A_i|A^{i-1},B^{i-1},X^i}(da_i|a^{i-1},b^{i-1},x^i) \\ & \otimes P_{X_i|X^{i-1}}(dx_i|x^{i-1}) \end{aligned} \tag{4.2.10}$$

Christos K. Kourtellaris

Moreover we say that $R^{na}(D)$ is realizable if in addition the realization operates with average distortion D and $\lim_{n\rightarrow\infty}\frac{1}{n+1}I_{P_{X^n}}(P_{X^n},\overrightarrow{P}_{Y^n|X^n}) = R^{na}(D) \leq \infty$.

Clearly, (4.2.10) holds if and only if the MCs (4.1.1)-(4.1.4) hold. Moreover, if the optimal reproduction distribution is realizable according to Definition 4.9, then the following data processing inequality holds (see Chapter 2, Theorem 2.24).

$$I_{X^n\rightarrow Y^n}(P_{X^n},\overrightarrow{P}_{Y^n|X^n}) \leq I(X^n \rightarrow B^n), \quad \forall n \in \mathbb{N} \tag{4.2.11}$$

Note that (4.2.9) and (4.2.10) are very general and include as a special case the joint distribution

$$\begin{aligned} P_{X^n,A^n,B^n,Y^n}(dx^n,da^n,db^n,dy^n) \quad = \quad & \otimes_{i=0}^{n} P_{Y_i|Y^{i-1},B^i}(dy_i|y^{i-1},b^i) \\ & \otimes P_{B_i|B^{i-1},A^i}(db_i|b^{i-1},a^i) \\ & \otimes P_{A_i|A^{i-1},B^{i-1},X^i}(da_i|a^{i-1},b^{i-1},x^i) \\ & \otimes P_{X_i|X^{i-1}}(dx_i|x^{i-1}) \end{aligned} \tag{4.2.12}$$

Equation (4.2.12) holds if and only if (4.1.3) is replaced by

$$(Y^{i-1},X^i) \leftrightarrow (A^i,B^{i-1}) \leftrightarrow B_i, \;\; \forall i \in \mathbb{N}^n \tag{4.2.13}$$

Moreover if (4.2.13) holds, then we have the data processing inequality (see Chapter 2, Theorem 2.24)

$$I_{X^n\rightarrow Y^n}(P_{X^n},\overrightarrow{P}_{Y^n|X^n}) \leq I(A^n \rightarrow B^n), \quad \forall n \in \mathbb{N} \tag{4.2.14}$$

If $R^{na}(D)$ is realizable according to Definition 4.9, then the source is not necessarily matched to the channel, but $R^{na}(D)$ can be given an operational meaning, based on Definition 4.5.

Now, we are ready to prove achievability of a nonanticipative code. To this end, we need to introduce the operational definition of channel capacity.
Consider the average cost constraint defined by

$$\mathscr{P}_{0,n}(P) \triangleq \Big\{\overrightarrow{P}_{X^n,A^n|B^{n-1}} : \frac{1}{n+1}\mathbb{E}\{c_{0,n}(A^n,B^{n-1})\} \leq P\Big\}$$

Since we consider the general scenario that (4.1.1)-(4.1.4) hold (e.g., MC (4.2.13) is not assumed), we define the finite-time information channel capacity from the source to the

Christos K. Kourtellaris

channel output [19], by

$$C_{0,n}(P) \overset{\triangle}{=} \sup_{\overrightarrow{P}_{X^n,A^n|B^{n-1}} \in \mathscr{P}_{0,n}(P)} I(X^n \rightarrow B^n)$$

The information channel capacity (provided sup is finite and the limit exists) is defined by

$$C(P) = \lim_{n \rightarrow \infty} \frac{1}{n+1} C_{0,n}(P)$$

We have the following achievability theorem.

Theorem 4.10. *(Achievability of nonanticipative code).*
A. Instantaneous. Suppose the following conditions hold for any finite n.

1. *$R^{na}_{0,n}(D)$ has a solution;*
2. *$C_{0,n}(P)$ has a solution;*
3. *The optimal reproduction distribution $\overrightarrow{P}^*_{Y^n|X^n}(dy^n|x^n)$ is realizable, and (R_n, D_n) is realizable;*
4. *For a given $D_n \in [D_{min}, D_{max}]$ there exists a P such that the realization gives $R^{na}_{0,n}(D_n) = C_{0,n}(P) = I(X^n \rightarrow B^n)$.*

B. Limiting. Suppose the following conditions hold.

1. *$R^{na}(D)$ has a solution;*
2. *$C(P)$ has a solution;*
3. *The optimal reproduction distribution $\overrightarrow{P}^*_{Y^\infty|X^\infty}(dy^\infty|x^\infty)$ for $R^{na}(D)$ is stationary and realizable, and (R, D) is realizable;*
4. *For a given $D \in [D_{min}, D_{max}]$ there exists a P such that the realization gives $R^{na}(D) = C(P) = \lim_{\longrightarrow\infty} \frac{1}{n+1} I(X^n \rightarrow B^n)$.*

If $\mathbb{P}^*_{X^n,Y^n}\Big\{\sum_{i=0}^n \rho(T^iX^n, T^iY^n) > (n+1)d\Big\} \leq \varepsilon, d > D$, where $\mathbb{P}^*$ is taken with respect to $P^*_{Y^n,X^n}(dy^n, dx^n) = \otimes_{i=0}^n \Big(P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1}, x^i) \otimes P_{X_i|X^{i-1},Y^{i-1}}(dx_i|x^{i-1}, y^{i-1})\Big)$ then there exists an (n, d, ε, P) nonanticipative code.

Proof. Part B: If conditions $B.1. - B.3.$ hold then the optimal stationary reproduction distribution is realizable, and this realization achieves $R^{na}(D)$, and $C(P)$. By 4. the source is

Christos K. Kourtellaris

matched to the channel so that the excess distortion probability of a nonanticipative code with memory without anticipation satisfies the excess distortion. □

Note that, if we replace the excess distortion by average distortion, an example for a Gaussian RV transmitted over a memoryless additive Gaussian channel is given in [5].

Remark 4.11. An equivalent definition achievability for a nonanticipative code can be shown by applying the concepts of outage rate probability and outage capacity, as giben in Definition 4.7 and Definition 4.8.

Next, recall that when the source is Markov, and the channel is Markov with respect to the source, nothing can be gained by considering encoders which at each time instant i depend on the entire past block of the source symbols X^i. This will imply that the optimal code is not only nonanticipative but it is also a SbS code, Markov with respect to the source.

We introduce the following assumption, to simplify the search for source-channel matching, in terms of the encoder, optimal reproduction distribution and decoder.

Assumption 4.12. (Markov Source and Channel Markov w.r.t. the Source)
The distortion, transmission cost, and source and channel conditional distributions satisfy the following conditions

1. $\rho(T^ix^n,T^iy^n)=\overline{\rho}(x_i,T^iy^n), \qquad \gamma(T^ia^n,T^ib^{n-1})=\overline{\gamma}(a_i,b^{i-1}) \qquad \forall i\in\mathbb{N}^n$
2. $P_{X_i|X^{i-1}}(dx_i|x^{i-1})=P_{X_i|X_{i-1}}(dx_i|x_{i-1})-a.a.(x^{i-1}), \qquad \forall i\in\mathbb{N}^n$
3. $P_{B_i|B^{i-1},A^i,X^i}(db_i|b^{i-1},a^i,x^i)=P_{B_i|B^{i-1},A_i,X_i}(db_i|b^{i-1},a_i,x_i)-a.a.(b^{i-1},a^i,x^i), \qquad \forall i\in\mathbb{N}^n.$

By Assumption 4.12 and Theorem 4.10, the optimal reproduction distribution is of the form

$$P_{Y_i|X^i,Y^{i-1}}(dy_i|x^i,y^{i-1})=P_{Y_i|X_i,Y^{i-1}}(dy_i|x_i,y^{i-1})-a.a.\ (X^i,Y^{i-1}), \qquad \forall i\in\mathbb{N}^n$$

In view of Assumption 4.12 we have

$$I(X^n\rightarrow B^n)\triangleq\sum_{i=0}^{n}I(X^i;B_i|B^{i-1})=\sum_{i=0}^{n}\mathbb{E}\left\{\log\frac{P_{B_i|B^{i-1},X_i}(db_i|b^{i-1},x_i)}{P_{B_i|B^{i-1}}(db_i|b^{i-1})}\right\}$$

Theorem 4.13. *Under Assumption 4.12, maximizing directed information over non-Markovian encoders with respect to the source is equivalent to maximizing it over Markovian encoders*

Christos K. Kourtellaris

with respect to the source, that is,

$$\sup_{\substack{P_{A_i|A^{i-1},B^{i-1},X^i}:\ i=0,1,\ldots,n\\ \frac{1}{n+1}\mathbb{E}\{c_{0,n}(A^n,B^{n-1})\}\leq P}} I(X^n\rightarrow B^n)=\sup_{\substack{P_{A_i|B^{i-1},X_i}:\ i=0,1,\ldots,n\\ \frac{1}{n+1}\mathbb{E}\{c_{0,n}(A^n,B^{n-1})\}\leq P}} I(X^n\rightarrow B^n) \tag{4.2.15}$$

Moreover, the maximization in (4.2.15) with respect to deterministic encoders $\{a_i=e_i(x^i,a^{i-1},y^{i-1}):i=1,\ldots,n\}$ *is equivalent to the maximization with respect to encoders* $\{a_i=e_i(x_i,y^{i-1}):i=1,\ldots,n\}$.

Next we introduce the following Assumptions, which allow us to define the information capacity between the input of the channel A^n, and the output of the channel B^n.

Assumption 4.14. The following MC holds

$$X^i\leftrightarrow(A^i,B^{i-1})\leftrightarrow B_i,\ \forall i\in\mathbb{N}^n$$

Under Assumptions 4.12 and 4.14, the following identity holds

$$I(X^n\rightarrow B^n)=I(A^n\rightarrow B^n)=\sum_{i=0}^{n}I(A_i;B_i|B^{i-1}) \tag{4.2.16}$$

Therefore, we have the following variation of Theorem 4.13 (see Theorem 3.31, Chapter 3).

Theorem 4.15. *Suppose assumptions 4.12, 4.14 hold.*
The finite-time information capacity satisfies

$$C_{0,n}(P)\stackrel{\triangle}{=}\sup_{\substack{P_{A_i|A^{i-1},B^{i-1}}:\ i=0,1,\ldots,n\\ \frac{1}{n+1}\mathbb{E}\{c_{0,n}(A^n,B^{n-1})\}\leq P}} I(A^n\rightarrow B^n)=\sup_{\substack{P_{A_i|B^{i-1}}:\ i=0,1,\ldots,n\\ \frac{1}{n+1}\mathbb{E}\{c_{0,n}(A^n,B^{n-1})\}\leq P}} I(A^n\rightarrow B^n)$$

4.3 Joint Source Channel Matching of a BSMS(p) via a Unit Memory Channel

In this section, we apply Theorem 4.10 to show that SbS joint source channel coding for the Binary Symmetric Markov Source ($BSMS(p)$) transmitted over the Binary State Symmetric

Christos K. Kourtellaris

Channel ($BSSC(\alpha_1,\beta_1)$) with cost constraint, is indeed feasible. For the sake of completeness, we begin by recalling the results of the nonanticipative RDF of a BSMS(p) and the capacity of the state symmetric channel with feedback, elaborated explicitly in Chapters 2 (Section 2.5) and Chapter 3 (Section 3.6), respectively. We provide the necessary conditions for JSCC coding schemes, with and without feedback, which operate optimally.

Moreover, we show the surprising result that even in the unmatched case of $C(P) \geq R^{na}(D)$ reliable communication is still feasible, with respect to the excess distortion probability. Finally, we prove that finite length SbS transmission (transmission of a finite number of symbols) is possible by providing bounds for the excess distortion probability for this schemes.

4.3.1 Nonanticipative RDF of Binary Symmetric Markov Source and Capacity of the Binary State Symmetric Channel

We begin by stating the main results of the nonanticipative RDF of the binary symmetric Markov source (Section 2.5, Chapter 2), for further use. The nonanticipative RDF of the BSMS(p) and single letter Hamming distortion criterion is given by

$$R^{na}(D) = \begin{cases} H(m) - H(D) = H(p) - mH(\alpha) - (1-m)H(\beta) & \text{if } D \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

where

$$m = 1 - p - D + 2pD$$

The optimal reproduction distribution depends only on the current source symbol and the previous reproduction symbol, the conditional distribution of the source symbol given all previous reproduction symbols depends only on the last reproduction symbol, while the distribution of the reproduction symbols also depends only on the previous reproduction symbol. These distributions are given below.

$$P^*_{Y_i|X_i,Y_{i-1}}(y_i|x_i,y_{i-1}) = \begin{array}{c} \\ 0 \\ 1 \end{array} \!\!\begin{array}{c} \begin{array}{cccc} 0,0 & 0,1 & 1,0 & 1,1 \end{array} \\ \left[\begin{array}{cccc} \alpha & \beta & 1-\beta & 1-\alpha \\ 1-\alpha & 1-\beta & \beta & \alpha \end{array}\right] \end{array} \tag{4.3.17}$$

Christos K. Kourtellaris

$$P^*_{X_i|Y_{i-1}}(x_i|y_{i-1}) = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} m & 1-m \\ 1-m & m \end{array}\right] \end{array}, \quad P^*_{Y_i|Y_{i-1}}(y_i|y_{i-1}) = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} p & 1-p \\ 1-p & p \end{array}\right] \end{array}$$

where

$$\alpha = \frac{(1-p)(1-D)}{m}, \quad \beta = \frac{p(1-D)}{1-m}$$

Next, we define the capacity of the BSSC, subject to a cost constraint. The channel over which source symbols are transmitted is chosen to have the form of the optimal reproduction distribution, given by

$$P_{B_i|A_i,B_{i-1}}(b_i|a_i,b_{i-1}) = \begin{array}{c c} & \begin{array}{cccc} 0,0 & 0,1 & 1,0 & 1,1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cccc} \alpha_1 & \beta_1 & 1-\beta_1 & 1-\alpha_1 \\ 1-\alpha_1 & 1-\beta_1 & \beta_1 & \alpha_1 \end{array}\right] \end{array} \tag{4.3.18}$$

By applying the one to one and onto, hence invertible, transformation $s_i = a_i \oplus b_{i-1}$, we rewrite the transition probability matrix as follows.

$$P_{B_i|A_i,S_i}(b_i|a_i,0) = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} \alpha_1 & 1-\alpha_1 \\ 1-\alpha_1 & \alpha_1 \end{array}\right] \end{array}, \; P_{B_i|A_i,S_i}(b_i|a_i,1) = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} \beta_1 & 1-\beta_1 \\ 1-\beta_1 & \beta_1 \end{array}\right] \end{array} \tag{4.3.19}$$

This transformation highlights the symmetric form of channel defined by the transition probability matrix (4.3.18), and provides an interpretation for the final expression of the capacity. We call this channel "state symmetric channel", since subject to the transformation s_i, it reduces to two symmetric channels with crossover probabilities $(1-\alpha_1)$ and $(1-\beta_1)$.

In Chapter 3, we also introduced an average cost constraint for the BSSC of the following form.

$$\begin{aligned} \mathbb{E}\{c(A_i,B_{i-1})\} &= 1.\{P_{A_i,B_{i-1}}(0,0)+P_{A_i,B_{i-1}}(1,1)\}+0.\{P_{A_i,B_{i-1}}(0,1)+P_{A_i,B_{i-1}}(1,0)\} \\ &= 1.P_{S_i}(0)+0.P_{S_i}(1)=\mathbb{E}\{\overline{S_i})\}=P_{S_i}(0)\equiv\kappa \end{aligned} \tag{4.3.20}$$

where $\kappa \in [0,1]$ is a given constant.

Christos K. Kourtellaris

Feedback does not increase the capacity of this channel. The capacity subject to the predefined binary cost constraint of the state symmetric channel is given by

$$C(S) = H(\beta_1(1-\kappa)+(1-\alpha_1)\kappa) - \kappa H(\alpha_1) - (1-\kappa)H(\beta_1) \tag{4.3.21}$$

The optimal input distributions, with and without feedback, are given by

$$P^*_{A_i|B_{i-1}}(a_i|b_{i-1}) = \begin{array}{c} 0 \\ 1 \end{array}\begin{bmatrix} \kappa & 1-\kappa \\ 1-\kappa & \kappa \end{bmatrix}, \quad P^*_{A_i|A_{i-1}}(a_i|a_{i-1}) = \begin{array}{c|cc} & 0 & 1 \\ \hline 0 & \frac{1-\kappa-\gamma}{1-2\gamma} & \frac{\kappa-\gamma}{1-2\gamma} \\ 1 & \frac{\kappa-\gamma}{1-2\gamma} & \frac{1-\kappa-\gamma}{1-2\gamma} \end{array}$$

where $\gamma = \alpha_1\kappa + \beta_1(1-\kappa)$.

Recall also that the unconstraint capacity is given by

$$C = H(\beta_1(1-\kappa^*)+(1-\alpha_1)\kappa^*) - \kappa^* H(\alpha_1) - (1-\kappa^*)H(\beta_1) \tag{4.3.22}$$

where the value of κ that maximizes the capacity is defined by κ^* and is equal to

$$\kappa^* = \frac{\beta_1(1+2^{\frac{H(\beta_1)-H(\alpha_1)}{\alpha_1+\beta_1-1}})-1}{(\alpha_1+\beta_1-1)(1+2^{\frac{H(\beta_1)-H(\alpha_1)}{\alpha_1+\beta_1-1}})}$$

Of course, this will result in a value which is at least greater than the capacity of the constraint case. The unconstraint capacity will be applied to the unmatched realization, while its exact calculation will be used to evaluate the rate loss, $(C - R^{na}(D))$, of uncoded transmission over a channel with capacity larger than the nonanticipative RDF.

Another equivalent representation of the state symmetric channel is also possible by conditioning over the state and the previous channel output, since $P_{B_i|A_i,B_{i-1}}(b_i|a_i,b_{i-1})$ uniquely defines $P_{B_i|S_i,B_{i-1}}(b_i|s_i,b_{i-1})$, given by

$$P_{B_i|S_i,B_{i-1}}(b_i|s_i,b_{i-1}) = \begin{array}{c|cccc} & 0,0 & 0,1 & 1,0 & 1,1 \\ \hline 0 & \alpha_1 & 1-\alpha_1 & 1-\beta_1 & \beta_1 \\ 1 & 1-\alpha_1 & \alpha_1 & \beta_1 & 1-\beta_1 \end{array} \tag{4.3.23}$$

and vice versa. This alternative equivalent representation of the channel defined by (4.3.17), is given by (4.3.18). We will apply this equivalent channel to construct the encoder-decoder

Christos K. Kourtellaris

scheme for the state symmetric channel in the presence of noiseless feedback.

Next, we prove that the capacity of the channel defined by (4.3.18), subject to average cost constraint defined by (4.3.20), is equal to the capacity of the channel defined by (4.3.23) over the same cost constraint.

Lemma 4.16. *The capacity of the state symmetric channel with feedback subject to a cost constraint* $\mathbb{E}\{c(A_i,B_{i-1})\} = \mathbb{E}\{\overline{S_i}\} = \kappa$*, where* $s_i = a_i \oplus b_{i-1}$*, is expressed by the following equivalent representations.*

$$C(S) = \max_{P_{A_i|B_{i-1}}:\mathbb{E}[A_i\oplus \bar{B}_{i-1}]=\kappa} I(A_i;B_i|B_{i-1}) = \max_{P_{S_i|B_{i-1}}:\mathbb{E}[\bar{S}_i]=\kappa} I(S_i;B_i|B_{i-1}), \quad \forall\, i \in \mathbb{N}^n$$

Proof. The conditional distribution of the channel defined by (4.3.18) uniquely defines (4.3.23) and vice-versa, while the conditional distribution $P_{S_i|B_{i-1}}$ is uniquely defined by $P_{A_i|B_{i-1}}$ and vice-versa. Additionally, the average cost constraint defined by $\{P_{A_i|B_{i-1}} : \mathbb{E}[A_i \oplus \bar{B}_{i-1}] = \kappa\}$ uniquely defines $\{P_{S_i|B_{i-1}} : \mathbb{E}[\bar{S}_i] = \kappa\}$. Thus,

$$\max_{P_{A_i|B_{i-1}}:\mathbb{E}[A_i\oplus \bar{B}_{i-1}]=\kappa} (A_i;B_i|B_{i-1}) = \max_{P_{S_i|B_{i-1}}:\mathbb{E}[\bar{S}_i]=\kappa} I(S_i;B_i|B_{i-1}), \quad \forall\, i \in \mathbb{N}^n$$

□

The optimal input distribution $P^*_{S_i|B_{i-1}}$, is given by

$$P^*_{S_i|B_{i-1}}(s_i|b_{i-1}) = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} \kappa & \kappa \\ 1-\kappa & 1-\kappa \end{array}\right] \end{array} \tag{4.3.24}$$

4.3.2 Joint Source Channel Matching: Information Matching and Realizations

The joint source channel matching itself consists of two parts; the information matching of the RDF and the capacity, and the realization scheme. The information matching is achieved if the nonanticipative RDF of the source (based on the fidelity constraint) is equal to the capacity of the channel subject to a cost constraint (i.e., Theorem 4.10 holds).

Christos K. Kourtellaris

The second part, and in most cases the most challenging one, is to construct an actual encoder-decoder scheme that fulfils the following conditions:

- Achieves the information matching.
- Satisfies the average distortion of the source.
- Satisfies the average cost constraint of the channel.

In this section, we show how to achieve information matching between the nonanticipative RDF of a $BSMS(p)$ with single letter Hamming distortion measure, and the capacity of the state symmetric channel with cross over probability of each state $(1-\alpha)$ and $(1-\beta)$, and with average cost constraint that satisfies $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\{c(A_i,B_{i-1})\}=\kappa$. Subsequently, we present the realization schemes, with or without feedback, that achieve the information matching and satisfy both the average distortion and the average cost constraint.

4.3.2.1 Information Matching

Comparing the nonanticipative RDF of the BSMS(p) with single letter Hamming distortion measure, and the capacity of the state symmetric channel with a binary cost constraint, we observe that these two information measures become equal by setting the cost constraint $\kappa=m$, and the channel parameters $\alpha_1=\alpha$ and $\beta_1=\beta$. Therefore,

$$\begin{aligned}C(S) &= H(\beta_1(1-\kappa)+(1-\alpha_1)\kappa)-\kappa H(\alpha_1)-(1-\kappa)H(\beta_1)\\ &= H(\beta(1-m)+(1-\alpha)m)-mH(\alpha)-(1-m)H(\beta)\\ &= H(p)-mH(\alpha)-(1-m)H(\beta)=R^{na}(D)\end{aligned}$$

Remark 4.17. For $p=0.5$, the source reduces to an IID Bernoulli source and the nonanticipative RDF is equal $1-H(D)$ (RDF of the IID Bernoulli source with single letter distortion criterion). Moreover, consider the case where the channel parameters (α_1,β_1) are both equal to $1-D$, and that there is no cost constraint on the channel. Then, the channel distribution is the same for the two possible states $s=0$ and $s=1$, and it reduces to a memoryless binary symmetric channel with crossover probability D ($BSC(D)$). The unconstraint capacity is achieved at $\kappa=0.5$, and it is equal to $1-H(D)$, which is equal to the capacity of the memoryless $BSC(D)$. Thus, our general nonanticipative RDF of the Markov source and a channel capacity with memory reduces to the well known case of the joint source channel

Christos K. Kourtellaris

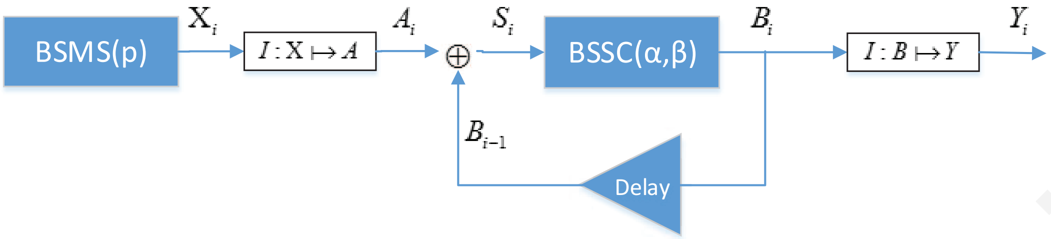

FIGURE 4.3.2: Feedback communication scheme for JSCC

.

matching of the IID Bernoulli source with single letter distortion criterion over the binary symmetric channel, which is achieved via uncoded transmission [29].

4.3.2.2 Feedback Realization

The communication scheme is illustrated in Figure 4.3.2. The source and the channel are fixed while the encoder and the decoder must be designed in such a manner that JSCM is achieved. The encoder consists of two blocks, the identity transformation $I : X \mapsto A$ (or the pre-encoder) and the S-Encoder (or th modulo2 encoder). The proposed encoder-decoder scheme is the following:

- *Pre-encoder:* This pre-encoder performs a unitary transformation on the source data, $(a_i = x_i)$. While practically it might be omitted, it is useful since it generates the variable a_i that defines the cost constraint of the channel. Its utilization keeps the notation clean since it generates a cost constraint for the channel that does not depend on the source.
- *S-Encoder:* Generates the input of the channel by performing a modulo2 addition between the output of the pre-encoder and the previous channel output $(s_i = a_i \oplus b_{i-1})$.
- *Decoder:* This decoder performs an identity transformation on the channel output data, thus $(y_i = b_i)$.

Summarizing, the encoder-decoder scheme, the only active block is the S-Encoder, while the two other blocks (pre-encoder and decoder) perform identity transformation on their inputs. Thus, practically the transmission of the data is characterized as semi-uncoded. To verify its optimality in terms of channel capacity, we first need to show that the channel input distribution is equal to the source output distribution, and that the average cost constraint is satisfied. We begin our analysis from the average distortion.

Christos K. Kourtellaris

Suppose that the symmetric binary Markov source is transmitted via the proposed semi-uncoded transmission scheme. The average distortion, Δ, is computed by evaluating the following expression.

$$\begin{aligned}
\Delta &= \mathbb{E}[\rho(X_i,Y_i)] \\
&= \mathbb{E}[\rho(A_i,B_i)] \\
&= \mathbb{E}[\rho(Si\oplus B_{i-1},B_i)] \\
&= \sum_{S_i,B_i,A_i} \rho(Si\oplus B_{i-1},B_i)P_{B_i|S_i,B_{i-1}}(b_i|s_i,b_{i-1})P_{S_i|B_{i-1}}(s_i|b_{i-1})P_{B_{i-1}}(b_{i-1}) \\
&= 1(1-\beta)(1-m)0.5+1(1-\alpha)m0.5+1(1-\alpha)m0.5+1(1-\beta)(1-m)0.5 \\
&= (1-\beta)(1-m)+(1-\alpha)m = D
\end{aligned}$$

Hence, the proposed semi-uncoded scheme achieves the average distortion. The channel input distribution achieves the optimal channel input distribution, since for $\kappa = m$, $A_i = X_i$ and $Y_i = B_i$, we have

$$P^*_{A_i|B_{i-1}} = P^*_{X_i|Y_{i-1}} = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} m & 1-m \\ 1-m & m \end{array}\right] \end{array} \tag{4.3.25}$$

and

$$P^*_{A_i|B_{i-1}} \mapsto P_{S_i=A_i\oplus B_{i-1}|B_{i-1}} = \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} m & m \\ 1-m & 1-m \end{array}\right] \end{array} \tag{4.3.26}$$

which are the optimal input distributionS subject to the transformation as defined by equation (4.3.24). For this input distribution the power constraint is also satisfied, as verified below.

$$\mathbb{E}_{P_{A_i|B_{i-1}}}\{c(A_i,B_{i-1})\} = 1.[0,5.P_{A_i|B_{i-1}}(0|0)+0,5.P_{A_i|B_{i-1}}(1|1)] = m \tag{4.3.27}$$

Remark 4.18. The form of the input distribution defined by (4.3.26), has independent property that the channel input S_i given the previous output symbol B_{i-1}, is independent of B_{i-1}, i.e., $P_{S_i|B_{i-1}} = P_{S_i}$. Thus, the S-Encoder by performing the modulo2 addition of the current source symbol and the previous output symbol, it generates an input to the channel which is independent of the previous output. This surprising result is similar to well known concept

Christos K. Kourtellaris

of the innovation encoder, which is widely applied in the Gaussian JSCC, where by sending the innovation, the probability distribution of the input symbol given the output symbol is independent from the previous output symbols [5].

4.3.2.3 No Feedback Realization

It is already shown in [3, 58], as well as in Chapter 3, that feedback does not increase the capacity of the BSSC (or POST) channel. The optimal input distribution for this channel is given by

$$P^*_{A_i|A_{i-1}}(a_i|a_{i-1}) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & \quad 1 \end{array} \\ \left[\begin{array}{cc} \dfrac{1-\kappa-\gamma}{1-2\gamma} & \dfrac{\kappa-\gamma}{1-2\gamma} \\ \dfrac{\kappa-\gamma}{1-2\gamma} & \dfrac{1-\kappa-\gamma}{1-2\gamma} \end{array}\right] \end{array}, \quad \gamma = \alpha_1\kappa + \beta_1(1-\kappa)$$

This input distribution satisfies the average cost constraint, even in the absence of feedback.

Setting $\kappa = m$, and $\alpha_1 = \alpha, \beta_1 = \beta$, yields $\gamma = 1-D$ and $\frac{1-\kappa-\gamma}{1-2\gamma} = 1-p$. Thus, the optimal input distribution of the BSSC without feedback that achieves the average cost constraint is given by

$$P^*_{A_i|A_{i-1}}(a_i|a_{i-1}) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & \quad 1 \end{array} \\ \left[\begin{array}{cc} 1-p & p \\ p & 1-p \end{array}\right] \end{array} \tag{4.3.28}$$

The above result is extremely convenient to design the encoder-decoder scheme, due to the fact that the optimal input distribution without feedback is equal to the distribution of the source, hence it eliminates the need of an encoder. Next, we check whether the average distortion, and the average cost constraint are satisfied in the absence of a decoder. If this holds, then uncoded transmission is indeed optimal.

Recall, that even in the absence of feedback the optimal input distribution without feedback, $P^*_{A_i|A_{i-1}}$ induces the capacity achieving input distribution of the feedback case $P^*_{A_i|B_{i-1}}$ (Proposition 3.50, Chapter 3). Thus, the average power constraint, $\mathbb{E}_{P_{A_i|B_{i-1}}}\{c(A_i,B_{i-1})\}$, is

Christos K. Kourtellaris

equal to

$$\mathbb{E}_{P_{A_i|B_{i-1}}}\{c(A_i,B_{i-1})\} = 1.[0,5.P_{A_i|B_{i-1}}(0|0)+0,5.P_{A_i|B_{i-1}}(1|1)] = m \qquad (4.3.29)$$

The average distortion between the source symbols and the reproduction symbols, Δ, is equal to

$$\begin{aligned}\Delta &= \mathbb{E}[\rho(X_i,Y_i)]\\ &= \mathbb{E}[\rho(A_i,B_i)]\\ &= \sum_{A_i,B_i,B_{i-1}} \rho(A_i,B_i)P_{B_i|A_i,B_{i-1}}(b_i|a_i,b_{i-1})P_{A_i|B_{i-1}}(a_i|b_{i-1})P_{B_{i-1}}(b_{i-1})\\ &= (1-\beta)(1-m)+(1-\alpha)m = D\end{aligned}$$

Therefore, the rate of the proposed uncoded scheme achieves its upper bound which is the capacity of the feedback channel while both average distortion and cost constraint are satisfied. Thus, uncoded transmission of a binary symmetric Markov source via a binary state symmetric channel subject to an average cost constraint, is indeed optimal.

Next, we evaluate the convergence of the distortion, via simulations. We construct a binary symmetric Markov source of length n and crossover probability p, encode it by applying modulo2 addition of the source symbol and the previous channel output, and send it via the channel $P_{B_i|S_i,B_{i-1}}$. The average distortion is then calculated by calculating $\sum_{i=1}^{n} X_i \oplus Y_i$. The results of a typical simulation are illustrated on Figure 4.3.3, and verify the expected convergence of the average distortion to D.

4.3.3 Communication over an Unmatched Realization

Due to data processing inequality, the capacity of the channel is always greater or equal to the rate distortion. For the uncoded transmission, equality holds if the optimal input distribution of the channel is equal to the distribution of the source. For the binary state symmetric channel this is achieved via a cost constraint which was explicitly addressed in the previous sections.

In this section, we drop the cost constraint ($\kappa = \kappa^*$) on the channel, and examine the uncoded transmission of a BSMS(p) over the binary state symmetric channel with parameters $\alpha_1 = \alpha$ and $\beta_1 = \beta$. We define the unmatched rate loss, as the excess amount of capacity that is lost,

Christos K. Kourtellaris

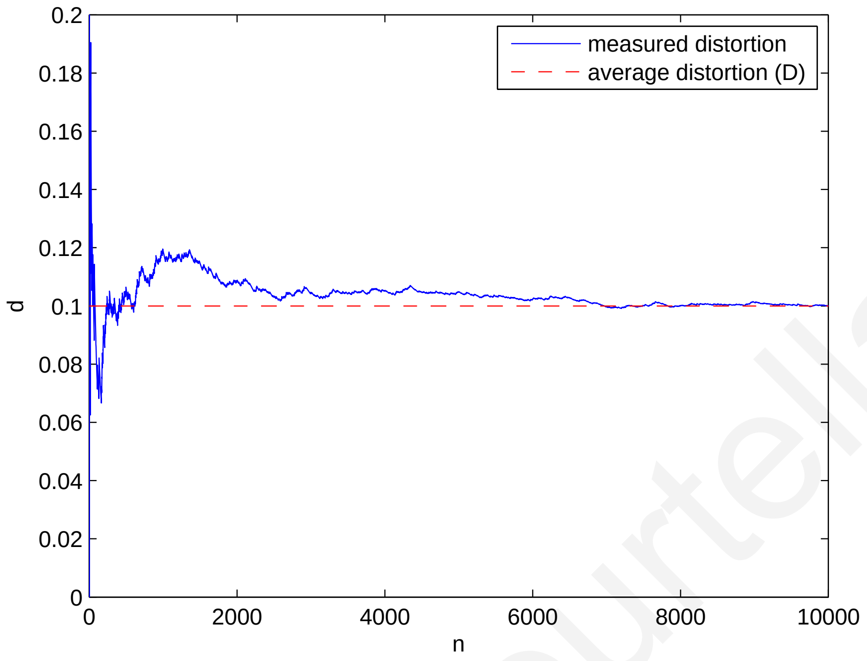

FIGURE 4.3.3: Simulation of a BSMS(p) via the feedback realization for $p = 0.3$ and $D = 0.1$

.

Christos K. Kourtellaris

and it is equal to the unconstrained capacity minus the nonanticipative RDF. The average distortion between the source symbols and the reproduction symbols, Δ, is equal to

$$\begin{aligned}\Delta &= \mathbb{E}[d(X_i,Y_i)]\\ &= \mathbb{E}[d(A_i,B_i)]\\ &= \sum_{A_i,B_i,B_{i-1}} \rho(A_i,B_i)P_{B_i|A_i,B_{i-1}}(b_i|a_i,b_{i-1})P_{A_i|B_{i-1}}(a_i|b_{i-1})P_{B_{i-1}}(b_{i-1})\\ &= (1-\beta)(1-m)+(1-\alpha)m = D\end{aligned}$$

The rate loss for a binary symmetric Markov source with crossover probability p, which is transmitted over a binary state symmetric channel with parameters α,β is defined by

$$RL^{un} \stackrel{\triangle}{=} \big(H(\lambda^*)-H(p)\big)-(\kappa^*-m)H(\alpha)-(m-\kappa^*)H(\beta)$$

The above transmission schemes satisfies the average distortion constraint, since the channel is equal to the optimal nonanticipative reproduction distribution of the Markov source. The cost of the rate loss is often balanced by the simplicity of the proposed scheme which is both

SbS (real-time transmission) and uncoded (absence of both encoder and decoder).

Remark 4.19. The unmatched rate loss decreases as p and/or $D \to 0.5$, and converges to 0 for $p = 0.5$, since this is the case of the uncoded transmission of an IID Bernoulli source transmitted over a binary symmetric channel with crossover probability $1 - D$.

4.3.4 Excess Distortion Probability

In previous section we discuss the average distortion, which is a performance measure suitable for infinite number of transmissions. Next, we discuss the excess distortion probability, a performance measure suitable for a finite number of transmissions.

The exact calculation of the excess distortion probability defined by $\mathbb{P}\Big\{d_{0,n}(x^n, y^n) > (n+1)d\Big\} \leq \varepsilon,\ \varepsilon \in (0,1),\ d \geq 0$, is not as straightforward as it is for the case of the IID Bernoulli source [44]. Thus, instead of evaluating exactly the probability of the distortion exceeding $(n+1)d$, we bound it by applying an extension of Hoeffding's inequality for Markov chains [32], which bounds the probability of a function of a Markov source. Therefore, we must show that the joint process defined by $(Y_i, X_i) : i = 1, 2, \ldots$ is Markov. Define

$$Z_i \triangleq (Y_i, X_i), \qquad S_n \triangleq \sum_{i=1}^{n} \rho(X_i \oplus Y_i)$$

Theorem 4.20. *For the optimal reproduction distribution characterized by* $P^*_{Y_i|Y^{i-1},X^i}(y_i|y^{i-1}, x^i)$ *of a Binary Symmetric Markov Source, the following MC holds*

$$Z_i \leftrightarrow Z_{i-1} \leftrightarrow Z^{i-2}, \quad i = 0, 1, \ldots$$

Proof.

$$\begin{aligned} P^*_{Y_i,X_i|Y^{i-1},X^{i-1}}(y_i, x_i|y^{i-1}, x^{i-1}) &= P^*_{Y_i|Y^{i-1},X^i}(y_i|y^{i-1}, x^i) P_{X_i|Y^{i-1},X^{i-1}}(x_i|y^{i-1}, x^{i-1}) \\ &= P^*_{Y_i|Y_{i-1},X_i}(y_i|y_{i-1}, x_i) P_{X_i|X_{i-1}}(x_i|x_{i-1}) \end{aligned}$$

Moreover,

$$\begin{aligned} P^*_{Y_i,X_i|Y_{i-1},X_{i-1}}(y_i, x_i|y_{i-1}, x_{i-1}) &= P^*_{Y_i|Y_{i-1},X_{i-1}}(y_i|y_{i-1}, x^i_{i-1}) P_{X_i|Y_{i-1},X_{i-1}}(x_i|y_{i-1}, x_{i-1}) \\ &= P^*_{Y_i|Y_{i-1},X_i}(y_i|y_{i-1}, x_i) P_{X_i|X_{i-1}}(x_i|x_{i-1}) \end{aligned}$$

Christos K. Kourtellaris

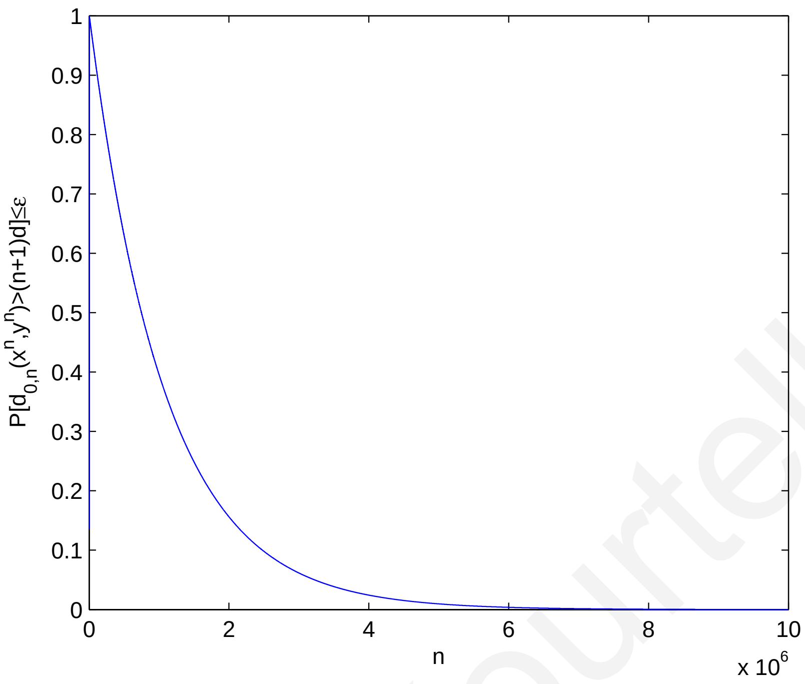

FIGURE 4.3.4: Hoeffding bound for excess distortion probability ($p=0.3, d=0.1, \varepsilon=0.1$)

Christos K. Kourtellaris

Thus,

$$P^*_{Y_i,X_i|Y^{i-1},X^{i-1}}(dy_i,dx_i|y^{i-1},x^{i-1}) = P^*_{Y_i,X_i|Y_{i-1},X_{i-1}}(dy_i,dx_i|y_{i-1},x_{i-1})$$

This shows that the joint process is Markov. □

The transition probabilities of the Markov process $\{Z_i : i=1,2,\ldots\}$ are given by

$$P_{Z_i|Z_{i-1}}(z_i|z_{i-1}) = \begin{bmatrix} \alpha(1-p) & (1-\beta)p & (1-\alpha)(1-p) & \beta p \\ \alpha p & (1-\beta)(1-p) & (1-\alpha)p & \beta(1-p) \\ \beta(1-p) & (1-\alpha)p & (1-\beta)(1-p) & \alpha p \\ \beta p & (1-\alpha)(1-p) & (1-\beta)p & \alpha(1-p) \end{bmatrix} \quad (4.3.30)$$

By applying the Hoeffding's inequality [32], the probability of the error is bounded by

$$P\Big[\frac{S_n-\mathbb{E}[S_n]}{n} \geq \varepsilon\Big] \leq exp\Big(-\frac{\lambda^2(n\varepsilon-2\|f\|m/\lambda)^2}{2n\|f\|^2m^2}\Big)$$

where $\|f\| \stackrel{\triangle}{=} \sup\{y_i : i = 1,2,\dots\} = 1$, $m = 1$ and $\lambda = \min\{p,1-p\}\min\{\alpha,\beta,1-\alpha,1-\beta\}$, and for $n > 2\|f\|m/(\lambda\varepsilon)$.

The probability of error for fixed values of p,d and ε is illustrated in Figure 4.3.4. For $p = 0.3$, $d = 0.1$, $\varepsilon = 0.1$, Figure 4.3.4 illustrates how the upper bound on the excess distortion probability changes as a function of the number of transmissions. It will be of interest to find tighter upper bounds to evaluate the excess distortion probability for finite "n".

4.4 Conclusions

In this chapter we put together the material of previous chapters to introduce the framework for nonanticipative transmission of general sources with memory via general channels with memory and feedback, derive noisy coding theorems based on nonanticipative and SbS code, and construct examples of JSCC based on SbS transmission with respect to average and excess distortion probability.

The theory is applied to analyze the SbS transmission of a binary symmetric Markov source with memory and Hamming distortion measure, transmitted over a binary state symmetric channel, subject to a state dependent cost constraint. For this example, we showed information matching among the nonanticipative RDF and the capacity, as well as optimal realizations with or without feedback. We additionally illustrated that unmatched transmission is also possible, and may be preferable in cases where the unmatched rate loss is negligible.

Christos K. Kourtellaris

Chapter 5

Nonanticipative RDF with Feedforward Information

5.1 Introduction

Lossy compression with side information at the encoder and/or decoder is investigated by Wyner-Ziv in the seminal paper[84]. Lossy compression with feedforward side information available at the decoder in terms of previous source symbols is investigated by Weissman and Merhav [82], and subsequently for Gaussian sources by Pradhan [61]. Recently, the OPTA by noncausal codes with feedforward side information, the so-called feedforward RDF, is characterized by Venkataramanan [79], in terms of the minimization of the directed information from the reproduction symbol to the source symbols subject to a fidelity constraint.

A lossy compression scheme with (causal) feedforward side information is illustrated in Figure 5.1.1. The signal is transmitted over two independent channels, where the first channel is noisy and has zero delay, while the second channel is noiseless but suffers from a delay. If the delay of the noiseless channel corresponds to a unit delay, then at each time instant i, the receiver knows the previously transmitted symbols. This model becomes more interesting if the source symbols are not packed into packets, but instead are sent form the transmitter to the receiver using nonanticipative processing. In such a formulation, the feedforward information is always causally known to the decoder, thus for the rest of this chapter we will refer to this scenario as feedforward information instead of (causal) feedforward information. The information measure introduced in [79], is not generally suitable to handle such nonanticipative coding.

Christos K. Kourtellaris

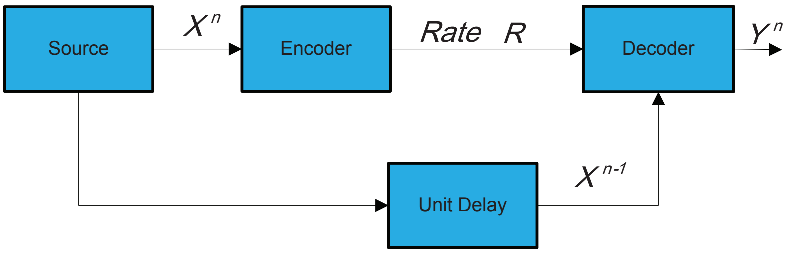

FIGURE 5.1.1: The feedforward rate distortion problem

This chapter consists of the following parts.

- Provides an information measure which is suitable for nonanticipative transmission with feedforward information at the decoder.
- Provides the relation between feedforward RDF, nonanticipative RDF with feedforward information, and relates them to Wyner-Ziv general formulation [84] for compression with side information, both at the encoder and the decoder. The later formulation utilizes mutual information between the source and its reproduction symbols, causally conditioned on the previous source symbols.
- Shows that for Markov sources and certain distortion criteria, all the above information measures are equivalent, and shows that for this case, the nonanticipative RDF with feedforward side information has an operational meaning. This is the Optimal Performance Theoretically Attainable (OPTA) by noncausal codes, with nonanticipative decoder side information.
- Computes the nonanticipative RDF with feedforward information at the decoder, the corresponding optimal reproduction distribution, and their respective bounds for the no feedforward case.
- Discusses the application on nonanticipative RDF with feedforward side information at the decoder in JSCC based on nonanticipative transmission.
- Compute examples.

5.2 Feedforward RDF

In this section we review certain definitions and results presented in [79], which we use in subsequent sections. Let $\mathbb{N} \triangleq \{0,1,2,\ldots,\}$. Throughout we assume that the source

Christos K. Kourtellaris

$\{X_i : i = 0, 1, \ldots\}$, and its reproduction $\{Y_i : i = 0, 1, \ldots\}$ take values in finite alphabets spaces $\mathscr{X}, \mathscr{Y}_i : i = 0, \ldots$ respectively, and they are jointly stationary ergodic. Given the source and the reproduction distribution denoted by $\{P_{X^n} : n \in \mathbb{N}\}$ and $\{P_{Y^n|X^n} : n \in \mathbb{N}\}$, respectively, the joint distribution is defined by $P_{X^n,Y^n} = P_{Y^n|X^n} \otimes P_{X^n}$, $\forall\, n \in \mathbb{N}$.

Consider a measurable, bounded, nonnegative distortion function denoted by $d_{0,n}(x^n, y^n) \overset{\triangle}{=} \sum_{i=0}^{n} \rho(T^i x^n, T^i y^n)$, where for each i, $T^i x^n$ is a causal mapping of x^n. The noncausal source code with feedforward information at the decoder is defined as follows.

Definition 5.1. An $(n, 2^{nR})$ feedforward source of code rate R and block length n, consists of the following encoder and decoder mappings:

$$\begin{aligned} &e : \mathscr{X}^n \mapsto \{1, 2, \ldots, 2^{nR}\} \\ &g_i : \{1, 2, \ldots, 2^{nR}\} \times \mathscr{X}_{0,i-1} \mapsto \mathscr{Y}_i, \;\; i = 0, 1, \ldots, n \end{aligned}$$

The decoder receives the index $\{1, 2, \ldots, 2^{nR}\}$ of the codeword and among with the available side information constructs the reproduction symbols at each time instant $i = 0, 1, \ldots, n$. The objective of this feedforward lossy compression scheme is to minimize the rate R, subject to the predefined distortion function $d_{0,n}(x^n, y^n)$. An achievable rate is defined as follows.

Definition 5.2. R is an achievable rate at expected distortion $D > 0$ if $\forall\, \varepsilon > 0$, and n sufficiently large, there exists an $(n, 2^{nR})$ feedforward code such that

$$\frac{1}{n+1} \mathbb{E}_{P_{X^n,Y^n}} \Big\{ d_{0,n}(X^n, Y^n) \Big\} \leq D + \varepsilon$$

The information measure which is used to derive direct and converse coding theorems is the directed information from the reproduction sequences to the source sequence, defined by [79]

$$I(Y^n \rightarrow X^n) = \sum_{i=0}^{n} I(Y^i; X_i | X^{i-1})$$

The following theorem is derived in [79].

Christos K. Kourtellaris

Theorem 5.3. *Define the average fidelity set by*

$$Q_{0,n}(D) \triangleq \{P_{Y^n|X^n} : \frac{1}{n+1}\mathbb{E}\{d_{0,n}(X^n,Y^n)\} \leq D\} \tag{5.2.1}$$

The finite-time information feedforward RDF is defined by

$$R^{ff}_{0,n}(D) \triangleq \inf_{Q_{0,n}(D)} I(Y^n \rightarrow X^n) \tag{5.2.2}$$

The OPTA by noncausal codes with feedforward information is

$$R^{ff}(D) \triangleq \lim_{n\rightarrow\infty} \frac{1}{n+1} R^{ff}_{0,n}(D) \tag{5.2.3}$$

provided the infimum exists and the limit is finite.

For Markov sources with finite memory m, it is shown in [79] that any family of distortion criteria, $d_{0,n}(x^n,y^n)$, satisfies

$$d_{0,n}(x^n,y^n) = -c\frac{1}{n}\log \frac{\otimes_{i=0}^{n} P^*_{Y_i|X^i_{i-m}} P_{X_i|X^{i-1}_{i-m}}}{\otimes_{i=0}^{n} \sum_{X_i} P^*_{Y_i|X^i_{i-m}} P_{X_i|X^{i-1}_{i-m}}} + d_0(x^n) \tag{5.2.4}$$

where $c > 0$ and $P^*_{Y_i|X^i_{i-m}}$ is the optimal reproduction distribution. The family of these distortion criteria, restricts the admissible set of joint probability distributions to

$$P^*_{X^n,Y^n} \triangleq \otimes_{i=0}^{n} P^*_{Y_i,X_i|X^{i-1}_{i-m}} = \otimes_{i=0}^{n} P^*_{Y_i|X^i_{i-m}} P_{X_i|X^{i-1}_{i-m}} \tag{5.2.5}$$

while the family of reproduction distributions satisfies

$$P^*_{Y^n|X^n} = \otimes_{i=0}^{n} P^*_{Y_i|Y^{i-1},X^n} = \otimes_{i=0}^{n} P_{Y_i|X^i_{i-m}} \tag{5.2.6}$$

The point to be made regarding (5.2.6), is that unlike the general RDF with feedforward information (which is not necessarily nonanticipative), this expression is a convolution of nonanticipative conditional distributions. However, computing $R^{ff}(D)$ appears to be difficult, and one has to introduce additional assumptions (see [79]).

Christos K. Kourtellaris

5.2.1 Information Measure Identities

In this section we show that the feedforward RDF can be derived from the general framework of Wyner-Ziv [84], and that it is equivalent to the maximization, over a fidelity set, of the mutual information causally conditioned on the previous source symbols. Moreover, we show that when side information is available both at the encoder and the decoder, then an equivalent information measure of (5.2.3) is the causally-conditioned mutual information. Hence, we identify another equivalent definition of the feedforward rate distortion function.

The causally conditioned mutual information is defined by

$$I(X^n;Y^n||X^{n-1}) \overset{\triangle}{=} \sum_{i=0}^{n} I(X^n;Y_i|Y^{i-1},X^{i-1}) = \sum_{i=0}^{n} \mathbb{E}\Big\{ \log \frac{P_{Y_i|X^n,Y^{i-1}}}{P_{Y_i|Y^{i-1},X^{i-1}}} \Big\}(5.2.11) \tag{5.2.7}$$

where $\mathbb{E}$ denotes the expectation over the joint probability distribution P_{X^n,Y^n}. Note that

$$P_{X^n,Y^n} = \otimes_{i=0}^{n}(P_{Y_i|Y^{i-1},X^n} \otimes P_{X_i|X^{i-1}}) \tag{5.2.8}$$
$$= \otimes_{i=0}^{n}(P_{Y_i|Y^{i-1},X^i} \otimes P_{X_i|X^{i-1},Y^{i-1}}) \tag{5.2.9}$$
$$= \otimes_{i=0}^{n}(P_{Y_i|Y^{i-1},X^{i-1}} \otimes P_{X_i|X^{i-1},Y^i}) \tag{5.2.10}$$

Define the following information measures.

$$I(Y^n \rightarrow X^n) \overset{\triangle}{=} \sum_{i=0}^{n} I(Y^i;X_i|X^{i-1}) \tag{5.2.11}$$
$$I(X^n \leftarrow Y^n) \overset{\triangle}{=} \sum_{i=0}^{n} I(Y^{i-1};X_i|X^{i-1}) \tag{5.2.12}$$
$$I(X^n \rightarrow Y^n||X^{n-1}) \overset{\triangle}{=} \sum_{i=0}^{n} I(X^i;Y_i|Y^{i-1},X^{i-1}) \tag{5.2.13}$$

Next, we presents theorems that relate the information measures (5.2.7), (5.2.11) and (5.2.11).

Theorem 5.4. *The mutual information causally conditioned defined by (5.2.7) is equivalent to the feedforward directed information defined by (5.2.11). Therefore,*

$$I(X^n;Y^n||X^{n-1}) = I(Y^n \rightarrow X^n)$$

Christos K. Kourtellaris

Proof. By combining (5.2.8) and (5.2.10), we obtain

$$\begin{aligned}
I(X^n;Y^n||X^{n-1}) &= \sum_{i=0}^{n} \mathbb{E}\Big\{ \log \frac{P_{Y_i|Y^{i-1},X^n}}{P_{Y_i|X^{i-1},Y^{i-1}}} \Big\} \\
&= \mathbb{E}\Big\{ \log \otimes_{i=0}^{n} \frac{P_{Y_i|Y^{i-1},X^n} \otimes P_{X_i|X^{i-1}}}{P_{Y_i|X^{i-1},Y^{i-1}} \otimes P_{X_i|X^{i-1}}} \Big\} \\
&= \mathbb{E}\Big\{ \log \frac{P_{X^n,Y^n}}{\otimes_{i=0}^{n} P_{Y_i|X^{i-1},Y^{i-1}} \otimes P_{X_i|X^{i-1}}} \Big\} \qquad (5.2.14) \\
&\overset{(\alpha)}{=} \mathbb{E}\Big\{ \log \otimes_{i=0}^{n} \frac{P_{Y_i|Y^{i-1},X^{i-1}} \otimes P_{X_i|X^{i-1},Y^i}}{P_{Y_i|Y^{i-1},X^{i-1}} \otimes P_{X_i|X^{i-1}}} \Big\} \\
&= \sum_{i=0}^{n} \mathbb{E}\Big\{ \log \frac{P_{Y_i|Y^{i-1},X^{i-1}} \otimes P_{X_i|X^{i-1},Y^i}}{P_{Y_i|Y^{i-1},X^{i-1}} \otimes P_{X_i|X^{i-1}}} \Big\} \\
&= \sum_{i=0}^{n} \mathbb{E}\Big\{ \log \frac{P_{X_i|X^{i-1},Y^i}}{P_{X_i|X^{i-1}}} \Big\} \\
&= I(Y^n \rightarrow X^n) \qquad (5.2.15)
\end{aligned}$$

where (α) follows from (5.2.10). Hence by (5.2.15), it follows that mutual information causally conditioned on the previous source symbols gives the information measure of the feedforward directed information. Thus, we establish the following equivalent representation of feedforward RDF. □

Corollary 5.5. *The feedforward RDF is alternatively defined as follows.*

$$R^{ff}(D) = \lim_{n \rightarrow \infty} \inf_{Q_{0,n}(D)} \frac{1}{n+1} I(X^n;Y^n||X^n) \qquad (5.2.16)$$

In general, the optimal reproduction distribution of (5.2.16) is not causal, Additionally, by combining (5.2.8) and (5.2.9) we have the following theorem.

Theorem 5.6. *The following information identity holds*

$$I(Y^n \rightarrow X^n) = I(X^n;Y^n||X^{n-1}) = I(X^n \rightarrow Y^n||X^{n-1}) + I(X^n \leftarrow Y^n)$$

Christos K. Kourtellaris

Proof.

$$
\begin{aligned}
I(X^n;Y^n||X^{n-1}) &= \sum_{i=0}^{n} \mathbb{E}\Big\{ \log \frac{P_{Y_i|Y^{i-1},X^n}}{P_{Y_i|Y^{i-1},X^{i-1}}} \Big\} \\
&\overset{(\alpha)}{=} \sum_{i=0}^{n} \mathbb{E}\Big\{ \log \frac{P_{Y_i|Y^{i-1},X^i} \otimes P_{X_i|X^{i-1},Y^{i-1}}}{P_{Y_i|Y^{i-1},X^{i-1}} \otimes P_{X_i|X^{i-1}}} \Big\} \\
&= \sum_{i=0}^{n} \mathbb{E}\Big\{ \log \frac{P_{Y_i|Y^{i-1},X^i}}{P_{Y_i|Y^{i-1},X^{i-1}}} \Big\} + \sum_{i=0}^{n} \mathbb{E}\Big\{ \log \frac{P_{X_i|X^{i-1},Y^{i-1}}}{P_{X_i|X^{i-1}}} \Big\} \\
&= I(X^n \rightarrow Y^n||X^{n-1}) + I(X^n \leftarrow Y^n) \qquad (5.2.17)
\end{aligned}
$$

where (α) is obtained from (5.2.14). Combining (5.2.15) and (5.2.17), results the provided information identity. □

The term $I(X^n \rightarrow Y^n||X^{n-1})$ appearing in (5.2.17), will be used in a subsequent section to characterize the nonanticipative RDF with feedforward side information.

Remark 5.7. Suppose the MC holds: $X_{i+1}^n \leftrightarrow X^i \leftrightarrow Y^i$, $i = 0,1,\ldots,n-1$. Then, from (5.2.17) we have $I(X^n;Y^n||X^{n-1}) = I(X^n \rightarrow Y^n||X^{n-1})$, in which the joint distribution is $P_{X^n,Y^n} = \otimes_{i=0}^n (P_{Y_i|Y^{i-1},X^i} \otimes P_{X_i|X^{i-1}})$. This leads to the definition of the nonanticipative RDF with feedforward information.

5.3 Nonanticipative RDF with FeedForward Side Information

In this section we modify the nonanticipative RDF of Chapter 3, to include feedforward side information. We also show that for m-order Markov sources and certain coupled letter distortion criteria, feedforward RDF and nonanticipative RDF with feedforward information are equivalent. This implies that the coding theorem derived in [79] is directly applicable to the nonanticipative RDF with feedforward information, and hence it is the OPTA by noncausal codes with causal side information. Moreover, we give the form of the optimal reconstruction distribution, and several of its properties which are important in solving examples. We utilize the optimal reproduction distribution to specific examples which are analysed in [79], to illustrate the simplicity by applying the general solution.

Christos K. Kourtellaris

Recall the nonanticipative RDF defined by

$$R^{na}_{0,n}(D) = \inf_{\substack{P_{Y^n|X^n} \in Q_{0,n}(D) \\ X^n_{i+1} \leftrightarrow X^i \leftrightarrow Y^i,\ i=0,1,\ldots,n-1}} I(X^n;Y^n)$$

Notice that the only difference between the classical RDF is that the following MC must hold.

$$X^n_{i+1} \leftrightarrow X^i \leftrightarrow Y^i \quad i=0,1,\ldots,n-1,\ n \geq 0 \tag{5.3.18}$$

Remark 5.8. The main point to be made regarding $R^{na}_{0,n}(D)$ is that the optimal reproduction is restricted by (5.3.18) to be nonanticipative. Hence, it is suitable for realization based on nonanticipative transmission, via an encoder-channel-decoder scheme, using real-time operations (causal). This property is necessary for joint source channel coding based on SbS or uncoded transmission schemes.

We recall the following Lemma derived in Section 2.3.1, Chapter 2.

Lemma 5.9. *The following statements are equivalent.*

1) $Y_i \leftrightarrow (X^i, Y^{i-1}) \leftrightarrow X^n_{i+1}$, $i = 0, 1\ldots, n-1$ *forms a MC*

2) $X_{i+1} \leftrightarrow X^i \leftrightarrow Y^i$, $i = 0, 1\ldots, n-1$ *forms a MC*

3) $X^n_{i+1} \leftrightarrow X^i \leftrightarrow Y^i$, $i = 0, 1\ldots, n-1$ *forms a MC*

By applying Lemma 5.9, we obtain the following result.

Theorem 5.10. *Suppose MC 3) of Lemma 5.9 holds.*
Then,

$$\begin{aligned}
I(Y^n \to X^n) &\overset{(\alpha)}{=} I(X^n;Y^n||X^{n-1}) \\
&= \sum_{i=0}^{n} \int \log\left(\frac{P_{Y_i|Y^{i-1},X^i}}{P_{Y_i|Y^{i-1},X^{i-1}}}\right) \\
&\quad \otimes_{j=0}^{i} P_{Y_j|Y^{j-1},X^j}(dy_j|y^{j-1}x^j) \otimes P_{X_j|X^{j-1}}(dx_j|x^{j-1}) \qquad (5.3.19) \\
&= I(X^n \to Y^n||X^{n-1}) \\
&\equiv \mathbb{I}_{X^n \to Y^n||X^{n-1}}(P_{X_i|X^{i-1}}, P_{Y_i|Y^{i-1},X^i} : i = 0, 1, \ldots, n) \qquad (5.3.20)
\end{aligned}$$

Christos K. Kourtellaris

Moreover,

$$\begin{aligned} P_{Y_i|Y^{i-1},X^{i-1}} &= \int_{\mathscr{X}_i} P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)\otimes P_{X_i|X^{i-1},Y^{i-1}}(dx_i|x^{i-1},y^{i-1}) \\ &\overset{(\beta)}{=} \int_{\mathscr{X}_i} P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)\otimes P_{X_i|X^{i-1}}(dx_i|x^{i-1}) \end{aligned}$$

Proof. Equality (α) holds due to (5.2.15). Directed information causally conditioned (without assuming any of the statements of Lemma 5.9) is defined by

$$\begin{aligned} I(X^n \rightarrow Y^n||X^{n-1}) = \sum_{i=0}^{n} \int \log\left(\frac{P_{Y_i|Y^{i-1},X^i}}{P_{Y_i|Y^{i-1},X^{i-1}}}\right) \otimes_{j=0}^{i} P_{Y_j|Y^{j-1},X^j}(dy_j|y^{j-1},x^j) \\ \otimes P_{X_j|X^{j-1},Y^{j-1}}(dx_j|x^{j-1},y^{j-1}) \end{aligned} \tag{5.3.21}$$

By MC 2), $P_{X_j|X^{j-1},Y^{j-1}}(dx_j|x^{j-1},y^{j-1}) = P_{X_j|X^{j-1}}(dx_j|x^{j-1})$, and by (5.3.21) we obtain (5.3.19), that is, $P_{X^n,Y^n} = \otimes_{i=0}^{n} P_{Y_i|X^i,Y^{i-1}} \otimes P_{X_i|X^{i-1}}$. Equality in (β) holds due to MC 2). This completes the derivation. □

The functional $\mathbb{I}_{X^n \rightarrow Y^n||X^{n-1}}(.,.)$ indicates the dependence on the conditional distributions $\{P_{X_i|X^{i-1}}, P_{Y_i|X^i,Y^{i-1}} : i = 0,1,\ldots,n\}$, when $\{P_{X_i|X^{i-1}} : i = 0,1,\ldots,n\}$ and $\{P_{Y_i|X^i,Y^{i-1}} : i = 0,1,\ldots,n\}$ is the reproduction conditional distribution.

We now proceed to define the nonanticipative RDF with feedforward information. Since $\{P_{Y_i|X^i,Y^{i-1}} : i=0,1,\ldots,n\}$ uniquely defines $\overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \overset{\triangle}{=} \otimes_{i=0}^{n} P_{Y_i|X^i,Y^{i-1}}$ and vice-versa, we have the following. Given a source distribution P_{X^n} and a causal $(n+1)$-fold convolution conditional distribution $\overrightarrow{P}_{Y^n|X^n}$, the joint distribution, P_{X^n,Y^n}, is well defined.

Definition 5.11. (Nonanticipative RDF with Feedforward Information)
The nonanticipative RDF with feedforward information is defined by

$$R^{na,ff}(D) \overset{\triangle}{=} \frac{1}{n+1} \lim_{n\rightarrow\infty} R_{0,n}^{na,ff}(D)$$

where

$$R_{0,n}^{na,ff}(D) \overset{\triangle}{=} \inf_{\substack{P_{Y^n|X^n} \in \mathcal{Q}_{0,n}(D) \\ X_{i+1}^n \leftrightarrow X^i \leftrightarrow Y^i,\ i=0,1,\ldots,n-1}} I(X^n;Y^n||X^{n-1}) \tag{5.3.22}$$

Christos K. Kourtellis

$$Q_{0,n}(D) \triangleq \Big\{ P_{Y^n|X^n} : \frac{1}{n+1} E(d_{0,n}(X^n,Y^n)) \leq D \Big\} \tag{5.3.23}$$

By Theorem 5.10, an equivalent definition is

$$\begin{aligned} R^{na,ff}_{0,n}(D) &= \inf_{\{P_{Y_i|Y^{i-1},X^i}:\ i=0,1,\ldots,n\}\in Q^{na}_{0,n}(D)} \mathbb{I}_{X^n\rightarrow Y^n||X^{n-1}}(P_{X_i|X^{i-1}}, P_{Y_i|Y^{i-1},X^i} : i=0,1,\ldots,n) \\ &\equiv \inf_{\overrightarrow{P}_{Y^n|X^n}\in Q^{na}_{0,n}(D)} \mathbb{I}_{X^n\rightarrow Y^n||X^{n-1}}(P_{X^n}, \overrightarrow{P}_{Y^n|X^n}) \end{aligned} \tag{5.3.24}$$

where

$$\begin{aligned} Q^{na}_{0,n}(D) &\triangleq \Big\{ P_{Y_i|Y^{i-1},X^i} : i=0,1,\ldots,n : \frac{1}{n+1} E(d_{0,n}(X^n,Y^n)) \leq D \Big\} \\ &\equiv \Big\{ \overrightarrow{P}_{Y^n|X^n} : \frac{1}{n+1} \int E(d_{0,n}(X^n,Y^n)) \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n) \otimes P_{X^n}(dx^n) \leq D \Big\} \end{aligned} \tag{5.3.25}$$

Next, we assume existence of the extremum solution of (5.3.24), and we give the optimal reproduction distribution $\overrightarrow{P}^*_{Y^n|X^n}$ which uniquely defines $\{P^*_{Y_i|X^i,Y^{i-1}} : i=0,1,\ldots,n\}$, and vice versa [73]. The existence of the optimal solution can be addressed following [73].

Theorem 5.12. *The optimal stationary reproduction conditional distribution for $R^{na,ff}(D)$ is given by*

$$\overrightarrow{P}^*_{Y^n|X^n} = \otimes_{i=0}^{n} \frac{e^{s\rho(T^i x^n, T^i y^n)} P^*_{Y_i|Y^{i-1},X^{i-1}}}{\int_{\mathscr{Y}_i} e^{s\rho(T^i x^n, T^i y^n)} P^*_{Y_i|Y^{i-1},X^{i-1}}} \tag{5.3.26}$$

$$P^*_{Y_i|Y^{i-1},X^{i-1}}(dy_i|y^{i-1},x^{i-1}) = \int_{\mathscr{X}_i} P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) P_{X_i|X^{i-1}}(dx_i|x^{i-1}) \tag{5.3.27}$$

and

$$R^{na,ff}_{0,n}(D) = sD(n+1) - \sum_{i=0}^{n} \int_{\mathscr{X}^i\times\mathscr{Y}^{i-1}} \log\Big(\int_{\mathscr{Y}_i} e^{s\rho(T^i x^n, T^i y^n)} P^*_{Y_i|Y^{i-1},X^{i-1}} \Big) P_{X_i|X^{i-1}} \otimes P^*_{X^{i-1},Y^{i-1}} \tag{5.3.28}$$

where s, denotes the Lagrange multiplier associated with the fidelity constraint (5.3.25), and $P^*_{X^{i-1},Y^{i-1}} = \overrightarrow{P}^*_{Y^{i-1}|X^{i-1}} P_{X^{i-1}}$.

Proof. The derivation is similar to Theorem 2.32, [73]. □

Christos K. Kourtellaris

By (5.3.26) we deduce that the stationary reproduction conditional distribution is given by

$$P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) = \frac{e^{s\rho(T^ix^n,T^iy^n)}P^*_{Y_i|Y^{i-1},X^{i-1}}(dy_i|y^{i-1},x^{i-1})}{\int_{\mathcal{Y}_i} e^{s\rho(T^ix^n,T^iy^n)}P^*_{Y_i|Y^{i-1},X^{i-1}}(dy_i|y^{i-1},x^{i-1})} \tag{5.3.29}$$

Hence, (5.3.29) is a nonlinear equation of the form

$$\xi^*(y^{i-1},x^i) = T(s,\rho(T^ix^n,T^iy^n),x^{i-1},\xi^*(y^{i-1},x^i)) \tag{5.3.30}$$

where $\xi^*(y^{i-1},x^i) \triangleq P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)$. Thus, if (5.3.30) has a unique solution, then the distortion function $\rho(T^ix^n,T^iy^n)$ and the source distribution $P_{X_i|X^{i-1}}(dx_i|x^{i-1})$ determine the dependence of $\xi^*(.,.)$ on (y^{i-1},x^i). Since the operator $T(s,\rho(T^ix^n,T^iy^n),\ x^{i-1},.)$ is nonlinear, the existence and uniqueness of solutions to (5.3.27) can be addressed by fixed point theorems.

The following Theorem gives some of the property of the nonanticipative RDF with feedforward information.

Theorem 5.13. *Suppose the optimal stationary reproduction conditional distribution (5.3.29) is unique. Then the following hold.*

1. $P^*_{Y_i|,X^i,Y^{i-1}}(dy_i|y^{i-1},x^i)$ *depends on the history* (y^{i-1},x^i) *through the source distribution* $P_{X_i|X^{i-1}}(dx_i|x^{i-1})$*, history* x^{i-1}*, and the distortion function* $\rho(T^ix^n,T^iy^n)$*.*

2. *If the source is m-order Markov process denoted by* $P_{X_i|X^{i-1}_{i-m}}(.|.) \triangleq P_{X_i|X_{i-m},X_{i-m+1},\ldots,X_{i-1}}(.|.)$*,* $m \in \{1,2,\ldots,M\}$*, and the distortion function is* $k \in \{1,2,\ldots,K\}$ *letter coupled, and* $l \in \{1,2,\ldots,L\}$ *letter coupled with respect to* T^ix^n *and* T^iy^n*, respectively, defined by*

$$\rho(T^ix^n,T^iy^n) \triangleq \bar{\rho}(x_{i-k+1},x_{i-k+2},\ldots,x_i,y_{i-l+1},y_{i-l+2},\ldots,y_i)$$

then

$$P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i) = P^*_{Y_i|Y^{i-1}_{i-l+1},X^i_{i-j}}(dy_i|y^{i-1}_{i-l+1},x^i_{i-j}) \tag{5.3.31}$$

where $j \triangleq \max\{k-1,m\}$ *and* $y^l_m = \{\emptyset\}$ *if* $l < m$*.*

Christos K. Kourtellaris

Proof. 1. This follows from (5.3.27). 2. Under the stated conditions, the operator $T(.,.,.,.)$ in (5.3.27) depends on $\{x_{i-k+1},x_{i-k+2},\ldots,x_i,y_{i-l+1},y_{i-l+2},\ldots,y_i\}$ through $\rho(.,.)$, and on $\{x_{i-m},x_{i-m+1},\ldots,x_{i-1}\}$ through the source $P_{X_i|X^{i-1}}(.|x^{i-1})$. Hence the claim. □

The next remark discusses several cases of m-order Markov sources and coupled distortion functions, and identifies fundamental differences between the optimal reproduction distribution of $R_{0,n}^{na,ff}(D)$ and $R_{0,n}^{na}(D)$

Remark 5.14. The following special cases follow directly from the previous Theorem.

1. If $\rho(T^ix^n,T^iy^n)=\bar{\rho}(x_i,y_i)$ and the source is Markov $P_{X_i|X^{i-1}}(dx_i|x^{i-1})=P_{X_i|X_{i-1}}(dx_i|x_{i-1})$, then

$$P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)=P^*_{Y_i|X_i,X_{i-1}}(dy_i|x_i,x_{i-1})$$

2. If $\rho(T^ix^n,T^iy^n)=\bar{\rho}(x_{i-1},x_i,y_i)$ and the source is m-order Markov $P_{X_i|X^{i-1}}(dx_i|x^{i-1})=P_{X_i|X^{i-1}_{i-m}}(dx_i|x^{i-1}_{i-m})$, then

$$P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)=P^*_{Y_i|X^i_{i-m}}(dy_i|x^i_{i-m})$$

3. If $\rho(T^ix^n,T^iy^n)=\bar{\rho}(x_{i-1},x_i,y_{i-1},y_i)$ and the source is m-order Markov $P_{X_i|X^{i-1}}(dx_i|x^{i-1})$ $=P_{X_i|X^{i-1}_{i-m}}(dx_i|x^{i-1}_{i-m})$, then

$$P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)=P^*_{Y_i|Y_{i-1},X^i_{i-m}}(dy_i|y_{i-1},x^i_{i-m})$$

Note that the optimal reproduction distribution corresponding to $R_{0,n}^{na,ff}(D)$ has the same form as the optimal reproduction distribution of $R_{0,n}^{na}(D)$, with $P_{Y_i|Y^{i-1}}$ replaced by $P_{Y_i|Y^{i-1},X^{i-1}}$. However, unlike the nonanticipative feedforward RDF, $R_{0,n}^{na,ff}(D)$, for the nonanticipative RDF, $R_{0,n}^{na}(D)$, even for $\rho(T^ix^n,T^iy^n)=\rho(x_i,y_i)$ we cannot determine á priori the dependence of the optimal reproduction conditional distribution on y^{i-1}, because its right hand side term depends on $P^*_{Y_i|Y^{i-1}}$. Specifically, from Theorem 2.32, Chapter 2, we have

$$P^*_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)\equiv P^*_{Y_i|Y^{i-1},X_i}(dy_i|y^{i-1},x_i)=\frac{e^{s\bar{\rho}(x_i,y_i)}P^*_{Y_i|Y^{i-1}}(dy_i|y^{i-1})}{\int_{\mathcal{Y}_i}e^{s\bar{\rho}(x_i,y_i)}P^*_{Y_i|Y^{i-1}}(dy_i|y^{i-1})}$$

Christos K. Kourtellaris

Upon conditioning, the right hand side conditional distribution gives

$$P^*_{Y_i|Y^{i-1}}(dy_i|y^{i-1}) = \int_{\mathscr{X}_i} P^*_{Y_i|Y^{i-1},X_i}(dy_i|y^{i-1},x_i) \otimes P_{X_i|Y^{i-1}}(dx_i|y^{i-1}) \tag{5.3.32}$$

Comparing (5.3.32) and (5.3.29) it is clear that feedforward side information reduces the computational complexity of the optimal reproduction conditional distribution of $R^{na,ff}(D)$.

In the next Theorem we provide lower and upper bounds for Markov sources using the nonanticipative RDF and the nonanticipative RDF with feedforward information.

Theorem 5.15. *Consider an m-oder Markov source and distortion criteria that satisfy (5.2.4). Then*

$$R^{ff}(D) = R^{na,ff}(D) \leq R(D) \leq R^{na}(D)$$

Proof. The first equality holds since the optimal reproduction distribution of the feedforward RDF is nonanticipative. The second inequality holds since side information both at the encoder and the decoder does not increase the classical RDF. Finally, the last inequality because the infimum is over a larger set since $Q^{na}_{o,n}(D) \subseteq Q_{o,n}(D)$. □

In general, noisy coding theorems for the nonanticipative RDF can be shown by nonanticipative JSCC, in the sense of Chapter 4. The optimal reproduction distribution of the nonanticipative RDF will define the form of the channel. For this channel and its respective average cost constraint P, one must design and encoder-decoder scheme that achieves the information matching, i.e. $R^{na,ff}(D) = C(P)$, and satisfies the average cost constraint and the average distortion.

5.4 Examples

In this section we apply Theorem 5.12 and Remark 5.14, for sources with feedforward information, to derive the expressions of the nonanticipative RDF with feedforward information, for specific Markov sources and distortion criteria.

Christos K Kourtellaris

5.4.1 Binary Symmetric Markov source

Consider a binary symmetric Markov source with transition probabilities $P_{X_i|X_{i-1}}(0|0) = P_{X_i|X_{i-1}}(1|1) = 1-p$ and $P_{X_i|X_{i-1}}(0|1) = P_{X_i|X_{i-1}}(1|0) = p,\ p \leq 0.5,\ i = 0,1,\ldots,n$, and a single letter Hamming distortion which is equal to 0, when the source symbol is identical to reproduction symbol, and 1 otherwise. A closed form expression for the classical RDF without side information, is only available for a small region of distortion, $D \leq D_c$, and is given by [37]

$$R(D) = H(p) - H(D), \quad \text{if} \quad 0 \leq D \leq \frac{1}{2}\left(1 - \sqrt{1 - \left(\frac{p}{q}\right)^2}\right) \equiv D_c \tag{5.4.33}$$

where $q = 1-p$. For any $D \in [D_c, D_{max}]$, (5.4.33) provides a lower bound on the classical RDF. Other lower and upper bounds are also known [5]. In the current example we assume that the decoder has feedforward information, which is the previous transmitted bit. The available side information will reduce the rate required to reconstruct the source subject to the fidelity criterion. By Theorem 5.15, the obtained result, $R^{na,ff}(D)$, is a lower bound on the classical RDF without side information, $R(D)$.

By using Theorem 5.12 and Remark 5.14.1, the optimal reproduction distribution of the nonanticipative RDF with feedforward is given by.

$$P^*_{Y_i|X_i,X_{i-1}} = \begin{array}{c|cccc|} & 0,0 & 0,1 & 1,0 & 1,1 \\ 0 & a & b & 1-b & 1-a \\ 1 & 1-a & 1-b & b & a \end{array} \tag{5.4.34}$$

$$P^*_{Y_i|X_{i-1}} = \begin{array}{c|cc|} & 0 & 1 \\ 0 & a & b \\ 1 & 1-a & 1-b \end{array} \tag{5.4.35}$$

where $a = \frac{(1-D)(1-p-D)}{(1-p)(1-2D)}$ and $b = \frac{(1-D)(p-D)}{p(1-2D)}$, while the Lagrange multiplier s, calculated via the average distortion constraint, is $s = \frac{D}{1-D}$. Substituting (5.4.34) and (5.4.35) to (5.3.28), we obtain

$$R^{na,ff}(D) = \begin{cases} H(p) - H(D) & \text{if} \quad 0 \leq D \leq D_{max} \\ 0 & \text{otherwise} \end{cases} \tag{5.4.36}$$

where $D_{max} = min\{p, 1-p\} = p$.

Christos K. Kourtellaris

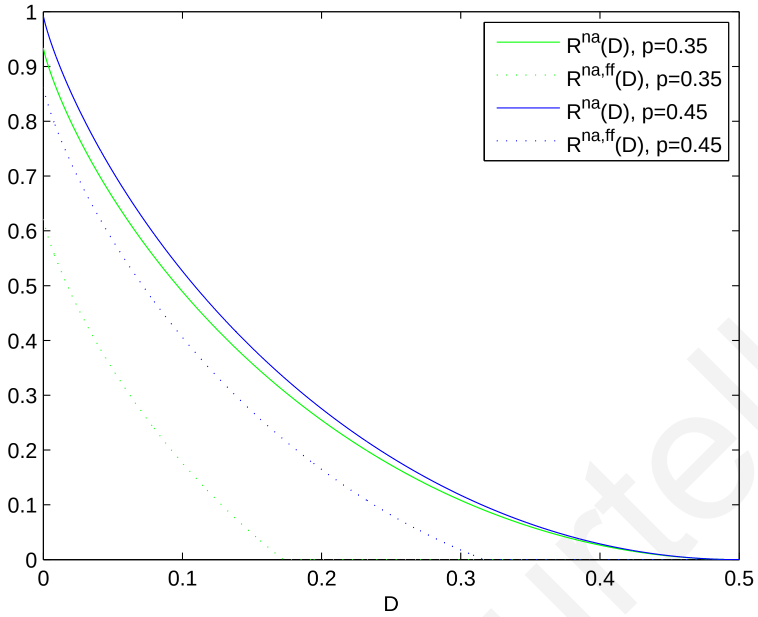

FIGURE 5.4.2: Upper and lower bounds for the classical RDF of a BSMSp.

It is surprising to observe that $R^{na,ff}(D)$, given by (5.4.36), is identical to the lower bound of the classical rate distortion problem without side information derived by Gray [37], which is tight for $D \leq \frac{1}{2}\left(1-\sqrt{1-\left(\frac{p}{q}\right)^2}\right)$, and identical to the RDF of a Bernoulli source. This interesting result is easily physically interpreted, since the knowledge of the previous transmitted symbol at the decoder converts the problem to its respective memoryless, Bernoulli $p, 1-p$ problem.

By using the solution of the nonanticipative RDF with feedforward information given by (5.4.36), and the result of the nonanticipative RDF without side information obtained in Section 2.5, Chapter 2, the classical RDF is bounded below by the nonanticipative RDF with feedforward and above by the nonanticipative RDF. Define $\{x\}^+ = \max(0,x)$. Then,

$$\{H(p)-H(D)\}^+ \leq R(D) \leq \{H(m)-H(D)\}^+ \tag{5.4.37}$$

The bounds defined by 5.4.37) for various values of p is illustrated in Figure. 5.4.2.

Christos K. Kourtellaris

5.4.2 Stock Market Problem

In this section we reproduce the solution of an example given in [79] using the expression of the optimal reproduction distribution. Assume the value of a stock X in the stock market over an n time period, is modelled by a Markov chain that takes k different values, as shown in Figure 5.4.3. Thus, by assuming that at a given time instant the value of the stock is j, the next day the value may be either increase to $j+1$ with probability p_j, either decrease to $j-1$ with probability q_j, or remain the same with probability $1-p_j-q_j$. Additionally, assume that the previous values of the stock are known, as side information. Our purpose is to send information when the value of the stock drops. Thus, $Y_i=1$ when the value drops from day $n-1$ to day n, and $Y_i=0$ otherwise. The distortion is modelled using a Hamming distortion measure as shown in Table 5.4.1.

		x_i, x_{i-1}		
		j+1 , j	j , j	j-1 , j
y_i	0	0	0	1
	1	1	1	0

TABLE 5.4.1: Distortion table: $\rho(y_i,x_i,x_{i-1})$

Summarizing, the objective of this problem is to estimate the minimum amount of information that needs to be sent in order to reconstruct the value of the source subject to a predefined average distortion D, taking into consideration the available feedforward information. By using Theorem 5.12 and Remark 5.14.2, the optimal reproduction distribution of the nonanticipative RDF with feedforward is given by.

$$R^{na,ff}(D) = \inf_{P_{Y_i|X_i,X_{i-1}} \in Q^{na}(D)} \sum_{Y_i,X_i,X_{i-1}} \log \frac{P_{Y_i|X_i,X_{i-1}}}{P_{Y_i|X_{i-1}}} P_{Y_i,X_i,X_{i-1}}$$

where $\{Q^{na}(D) = P_{Y_i|X_i,X_{i-1}} : \sum_{Y_i,X_i,X_{i-1}} \rho(Y_i,X_i,X_{i-1})\ P_{Y_i|X_i,X_{i-1}} P_{X_i|X_{i-1}} P_{X_{i-1}} \leq D\}$. The optimal reproduction distribution is given by

$$P_{Y_i|X_i,X_{i-1}} = \frac{e^{s\rho(y_i,x_i,x_{i-1})} P_{Y_i|X_{i-1}}}{\sum_{\mathscr{Y}_i} e^{s\rho(y_i,x_i,x_{i-1})} P_{Y_i|X_{i-1}}} \tag{5.4.38}$$

The results for the optimal reproduction distribution, are obtain iteratively by using (5.4.38). The distribution $P_{Y_i|X_{i-1}}$ is given by (5.4.39) where $\varepsilon = \frac{D}{1-\pi_j}$, and $\{\pi_j : j = 0,1,\ldots\}$ is the

Christos K. Kourtellaris

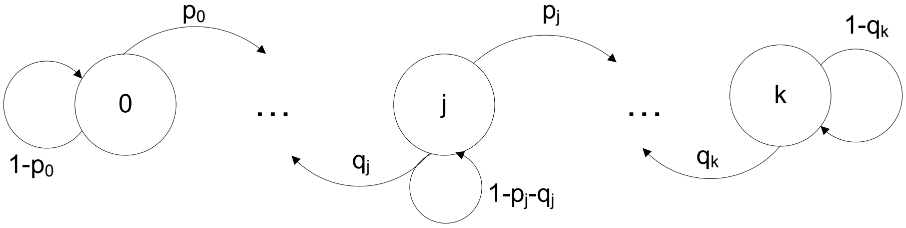

FIGURE 5.4.3: Stock value model

steady state probability of each state $j : j = 1,2,\dots,k$ of the source.

$$P_{Y_i|X_{i-1}}(y_i|x_{i-1}) = \begin{array}{c|cc} & 0 & j \\ \hline 0 & 1 & \frac{1-q_j-\varepsilon}{1-2\varepsilon} \\ 1 & 0 & \frac{q_j-\varepsilon}{1-2\varepsilon} \end{array} \tag{5.4.39}$$

From the solution of the causal rate distortion function we obtain

$$\begin{aligned} R^{na,ff}(D) &= sD - \sum_{x_i,x_i-1} \Big[\log \sum_{y_i} e^{sd(y_i,x_i,x_{i-1})} P_{Y_i|X_{i-1}}\Big] P_{X_i|X_{i-1}} P_{X_{i-1}} \\ &= \varepsilon(1-\pi_j)\log\frac{\varepsilon}{1-\varepsilon} - \sum_{j=1}^{k}\pi_j\Big[(1-q_j)\log\frac{1-q_j}{1-\varepsilon} + (q_j)\log\frac{q_j}{1-\varepsilon}\Big] \\ &= \sum_{j=1}^{k}\pi_j\Big[H(q_j) - H(\varepsilon)\Big] \end{aligned}$$

To calculate the maximum value of the distortion, D_{max}, we set the above equation equal to zero, thus we get $D_j = q_j(1-\pi_0)$ and $D_j = (1-q_j)(1-\pi_0)$. Consequently $D_{max} = \min[q_j(1-\pi_0),(1-q_j)(1-\pi_0)]$. This is an interesting result since the maximum value of the distortion depends on the previous value of the stock. For example, for some values of the stock, D_{max} might be overreached, thus sending any information for that value is useless. The solution of the nonanticipative RDF with feed-forward information is given by

$$R^{na,ff}(D) = \begin{cases} \sum_{j=1}^{k}\pi_j\Big[H(q_j) - H(\varepsilon)\Big] & \text{if } D \le D_{max} \\ 0 & \text{if } D > D_{max} \end{cases}$$

Christos K. Kourtellaris

5.4.3 Gaussian Markov Source

Consider a stationary ergodic Gaussian Markov source $\{X_i : i = 0,1,\ldots\}$ with mean 0 and variance σ^2 described by

$$X_i = \rho X_{i-1} + N_i \ \text{ where } \ N_i \sim \mathscr{N}(0,(1-\rho^2)\sigma^2), \ i = 1,2,\ldots \tag{5.4.40}$$

Suppose that we want to reconstruct a linear combination $aX_i + bX_{i-1}, \ i = 1,2,\ldots$, subject to the mean squared error distortion criterion $\frac{1}{n}\sum_{i=1}^{n}(Y_i - (aX_i + bX_{i-1}))^2$, where a and b are constants. Since the source is stationary we will assume time instant $n = 2$ for the rest of the problem. The optimal rate distortion function is given by

$$\begin{aligned}
R^{na,ff}(D) &= \lim_{n\to\infty} \frac{1}{n} I(X^n \to Y^n || X^{n-1}) \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} \sum_{\mathscr{X}^i,\mathscr{Y}^i} \log \frac{f_{Y_i|X^i,Y^{i-1}}}{f_{Y_i|X^{i-1},Y^{i-1}}} f_{X^i,Y^i} \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n} \sum_{\mathscr{X}^i,\mathscr{Y}^i} \log \frac{f_{Y_i|X_i,X_{i-1}}}{f_{Y_i|X_{i-1}}} f_{Y_i|X_i,X_{i-1}} P_{X_i,X_{i-1}} \\
&= \sum_{\mathscr{X}_2,\mathscr{X}_1,\mathscr{Y}_2} \log \frac{f_{Y_2|X_2,X_1}}{f_{Y_2|X_1}} f_{Y_2|X_2,X_1} f_{X_2,X_1} \\
&= \sum_{\mathscr{X}_2,\mathscr{X}_1,\mathscr{Y}_2} \log \frac{f_{Y_2,X_2,X_1} f_{X_1}}{f_{X_2,X_1} f_{Y_2,X_1}} f_{Y_2|X_2,X_1} f_{X_2,X_1} \\
&= \sum_{\mathscr{X}_2,\mathscr{X}_1,\mathscr{Y}_2} \log \frac{f_{X_2|Y_2,X_1}}{f_{X_2|X_1}} f_{Y_2|X_2,X_1} f_{X_2,X_1} \\
&= h(X_2|X_1) - h(X_2|Y_2,X_1)
\end{aligned} \tag{5.4.41}$$

Christos K. Kourtellaris

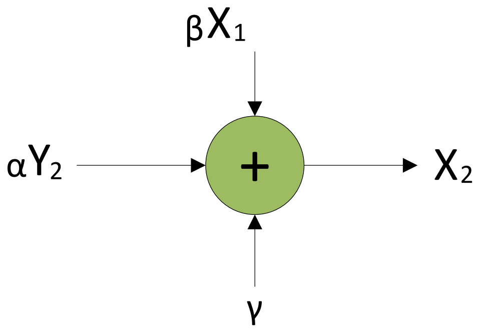

FIGURE 5.4.4: Test channel

Next, we will find a lower bound for the rate distortion function as defined in 5.4.41 and then prove that this is achievable.

$$
\begin{align*}
& h(X_2|X_1) - h(X_2|Y_2,X_1) \\
= \; & \frac{1}{2}\log 2\pi e(1-\rho^2)\sigma^2 - h(X_2|Y_2,X_1) \\
= \; & \frac{1}{2}\log 2\pi e(1-\rho^2)\sigma^2 - h((X_2 + \frac{b}{a}X_1 - \frac{1}{a}Y_2)|Y_2,X_1) \tag{5.4.42} \\
= \; & \frac{1}{2}\log 2\pi e(1-\rho^2)\sigma^2 - h(\frac{1}{a}(aX_2 + bX_1 - Y_2)|Y_2,X_1) \\
\overset{(1*)}{=} \; & \frac{1}{2}\log 2\pi e(1-\rho^2)\sigma^2 - h(aX_2 + bX_1 - Y_2|Y_2,X_1) - \log\frac{1}{a} \\
= \; & \frac{1}{2}\log 2\pi e(1-\rho^2)\sigma^2 - h(aX_2 + bX_1 - Y_2|Y_2,X_1) + \frac{1}{2}\log a^2 \\
\overset{(2*)}{\geq} \; & \frac{1}{2}\log 2\pi e(1-\rho^2)\sigma^2 - h(aX_2 + bX_1 - Y_2) + \frac{1}{2}\log a^2 \\
\overset{(3*)}{\geq} \; & \frac{1}{2}\log 2\pi e(1-\rho^2)\sigma^2 - h(\mathcal{N}(0, E((aX_2 + bX_i) - Y_2)^2)) + \frac{1}{2}\log a^2 \\
= \; & \frac{1}{2}\log 2\pi e(1-\rho^2)\sigma^2 - \frac{1}{2}\log(2\pi e)E((aX_2 + bX_i) - Y_2)^2 + \frac{1}{2}\log a^2 \\
\overset{(4*)}{\geq} \; & \frac{1}{2}\log 2\pi e(1-\rho^2)\sigma^2 - \frac{1}{2}\log(2\pi e)\frac{D}{a^2} \tag{5.4.43} \\
= \; & \frac{1}{2}\log\frac{2\pi e(1-\rho^2)\sigma^2 a^2}{D}
\end{align*}
$$

where (1*) follows from a property of relative entropy, (2*) holds since conditioning can not increase uncertainty, (3*) follows from the fact that Gaussian distributions for a given variance maximizes the entropy and (4*) follows from the average distortion constraint. After the calculation of the lower bound, what remains is to find a conditional density $f(Y_2|X_2,X_1)$ that achieves that lower bound. To do this, we first calculate the test channel, $f(X_2|Y_2,X_1)$ and then use Baye's rule to calculate the desired conditional density. Thus we assume that

Christos K. Kourtellaris

the test channel as described in Figure has the following form

$$X_2 = \alpha Y_2 + \beta X_1 + \gamma \;\Rightarrow\; X_2 - \alpha Y_2 - \beta X_1 = \gamma \tag{5.4.44}$$

where the parameters α, β and γ must equalize equations (5.4.42) and (5.4.43). This yields, $\gamma \sim \mathscr{N}(0, \frac{D}{a^2})$, $\alpha = \frac{1}{a}$ and $\beta = -\frac{b}{a}$. Since equation (5.4.44) is well defined both conditional density $f(Y_2|X_2, X_1)$ and the test channel are defined.

5.5 Conclusion

The equivalence between RDF with feedforward information [79] and Wyner's general formulation of lossy compression with side information is established, using causally conditioned mutual information [84] .

The nonanticipative RDF with feedforward information is formulated and its stationary solution is obtained. General properties of the optimal reproduction conditional distribution are identified, which are very important in solving such problems with coupled distortion functions and m-order Markov sources.

The nonanticipative RDF with feedforward information, is also shown to have an operational meaning for the case of m-order Markov sources and coupled distortion function, which follows from that of RDF with feedforward information [79]. The nonanticipative RDF with feedforward information is shown to be a lower bound on classical RDF (the OPTA by noncausal codes). The exact solution of the nonanticipative RDF with feedforward information is derived for several examples, using properties of the optimal reproduction conditional distribution.

The nonastationary solution of the nonanticipative RDF with feedforward information appears feasible and much simpler than the case of no feedforward information derived in [73].

Christos K. Kourtellaris

Chapter 6

Summary and Future Directions

Delayless or nonanticipative information transmission has several applications such as sensor networks, communication for control, and biomedical modelling and analysis. In this thesis, we introduced a theoretical framework of nonanticipative information theory for lossy compression of sources with memory, and we derived structural properties of encoders, for communication channels with memory and feedback. Subsequently, we integrated these results to provide the framework for symbol-by-symbol joint source channel coding. Finally, we discussed an extension of the nonanticipative rate distortion function, in situations where the decoder has access to previous transmitted symbols.

6.1 Summary and Concluding Remarks

Nonanticipative RDF for sources with memory

Summary:

We formulated the Nonanticipative RDF for general sources, by imposing a Markov chain constraint on the optimization problem, which does not allow current reproduction symbols to depend on future source symbols. We derived the stationary optimal reproduction distribution, corresponding to the nonanticipative RDF, a noisy coding theorem, and we elaborated on its relations to other compression schemes. Finally, we calculated the nonanticipative

Christos K. Kourtellaris

RDF of a binary symmetric Markov source.

Concluding Remarks:

- Nonanticipative lossy compression is operational over noisy coding theorems.
- The Rate Loss of nonanticipation defined by $RL \stackrel{\triangle}{=} R^{na}(D) - R(D)$, characterizes the excess amount of the information, with respect to non causal codes, due to nonanticipation.
- A closed form expression for the classical RDF for sources with memory, is not always available. Hence, we may apply the nonanticipative RDF to provide a realizable upper bound for the classical RDF.

Structural Properties of Extremum Problems of Capacity

Summary :

We derived structural encoder properties which maximize the directed information from the source to the channel output, and structural properties of capacity achieving distribution. We derived dynamic programming recursions to compute the encoder and the achieving distribution. Moreover, we generalized PMS from memoryless channels to channels with memory and feedback, to aid the design of encoders which achieve the information capacity. Finally, we calculated the capacity and optimal input distribution of the Binary State Symmetric Channel (BSSC) with or without feedback information at the encoder, and with and without imposing transmission cost constraint.

Concluding Remarks:

- For the class of channels with memory which are Markov with respect to the source, nothing can be gained by encoding blocks of source symbols, instead of encoding symbol-by-symbol.
- Posterior Matching Scheme, is feasible even for channels with memory and feedback.

Christos K. Kourtellaris

- Feedback does not increase the capacity of the BSSC, while the capacity achieving input distribution is Markov.

Nonanticipative Joint Source Channel Coding for Real-Time Transmission

Summary:

We introduced definitions of symbol-by-symbol code without anticipation, and minimum excess distortion, discussed the realization of the optimal non-anticipative reproduction distribution, and proved achievability of symbol-by-symbol code with memory and without anticipation via a noisy channel. We applied the framework to demonstrate optimality of the symbol-by-symbol JSCC for the Binary Symmetric Markov Source $BSMS(p)$, communicated over a Binary State Symmetric Channel, $BSSC(\alpha_1,\beta_1)$, while we discussed feedback, no feedback, and unmatched realizations. Finally, we provided a bound for the excess distortion for finite length case.

Concluding Remarks:

- Symbol-by-symbol JSCC shows achievability for the nonanticipative RDF.
- Uncoded transmission of a $BSMS(p)$ over a $BSSC(\alpha_1,\beta_1)$ is optimal in terms of symbol-by-symbol transmission.
- The feedback realization of a $BSMS(p)$ over a $BSSC(\alpha_1,\beta_1)$, yields a conditional input distribution which is independent of the previous output, showing optimality of an innovation encoder.
- Unmatched realization is achievable. The unmatched rate loss is a useful tool to compare the performance of uncoded transmission compared to classical coding approaches.

Nonanticipative RDF with Feedforward Information

Summary:

Christos K. Kourtellaris

We provided an information measure suitable for nonanticipative transmission with feedforward information at the decoder, compared it with feedforward RDF, and discussed its operation meaning for Markov sources. Finally, we computed the nonanticipative RDF with feedforward information, for a Binary Markov source, and provided the corresponding optimal reproduction distribution.

Concluding Remarks:

- The nonanticipative RDF with feedforward information, provides an upper bound on the feedfoward RDF.
- For Markov sources with certain disortion criteria, nonanticipative RDF with feedforward information and feedfoward RDF are equivalent.
- The nonanticipative RDF with feedforward of Markov sources, provides a lower bound on classical RDF without feedforward information.

6.2 Future Directions and Open Problems

6.2.1 Future Directions

Control over Communication Constraints

The general framework of Chapter 3, where the source may depend on previous channel outputs, is suitable to develop the subject of communication for control [15, 16]. One can apply the framework to address analysis and synthesis questions, related to stochastic optimal control over finite rate noisy channels, under general conditions on the channel and the controller. The main problem is to optimize a control pay-off subject to rate constraints on the feedback link, between the controlled system outputs or/and the input to the controller. The feedback link is often subject to limited rate, and the interest is to design encoders, decoders and controllers for general channels with memory and feedback, for reliable communication. It would be interesting and design optimal encoders, decoders and controllers, which minimize a control pay-off, subject to a rate constraint, and understand the trade-off

Christos K. Kourtellaris

between control and communication performance.

The Gilbert-Elliot Channel

The Gilber-Elliot channel is a time varying binary symmetric channel with crossover probabilities determined by a binary-state Markov process, and its capacity is obtained via a limiting expression [54]. Its capacity is often obtained via algorithms [64]. In Chapter 3, we discussed the binary state symmetric channel, where we showed that subject to the proposed transformation, it decomposes into two binary symmetric channels. Moreover, we applied an average cost constraint, by fixing the state of the channel, and calculated its capacity. If, instead of fixing the steady state distribution of the state, we fix the transition probabilities of the states, then the channel transforms to the Gilbert-Elliot channel. It would be interesting to investigate whether it is possible to provide a closed form expression for some special cases of the Gilbert-Elliot channel, based on the proposed approach.

Symbol-by-symbol Joint Source Channel Coding in the Presence of Feedforward Information at the Decoder

In Chapter 5, we discussed the nonanticipative RDF with feedforward side information of a binary Markov source with single letter distortion criterion. The optimal reproduction distribution depends on the current and previous channel input. It would be interesting to provide a symbol-by-symbol JSCC for a channel that has the form of the optimal reproduction distribution. This preassumes a closed form expression for the capacity of this channel. To calculate the capacity of this channel we may consider different schemes such as: adopting the approach of Section 3.6, Chapter 3, considering it as a special case of the Multiple Access Channel (MAC), in which the second input is a delay version of the first input, or even approaching it as a relay network in which the relay imposes a unit delay to its respective output. A similar approach to third alternative (relay network), is adopted in [30], and regards a multiple Gaussian network where each relay node i introduces i-delay.

Christos K. Kourtellaris

Nonanticipative Transmission for Networks

As illustrated explicitly in the binary symmetric Markov source example in the presence of feedforward information, the reproduction distribution represents a specific network channel. It is on our interest to calculate the nonanticipative RDF of other sources with memory, not necessarily symmetric or Markov, and provide symbol-by-symbol JSCC schemes that perform optimally in terms of symbol-by-symbol transmission. Initially, we will restrict our interest in cases where the respective capacities are known, in order to provide matching conditions and the possible unmatched rate loss. We aim also to find cases where uncoded transmission schemes perform optimally, in terms of symbol-by-symbol transmission.

6.2.2 Open Problems

Additional general open problems are the following.

- The solution of the nonstationary nonanticipative RDF is of interest in limited length reliable communications.
- Characterizing the realizability conditions, for JSSC is of interest in designing computational algorithms for symbol-by-symbol transmission.
- Generalizing nonanticipative RDF with general side information available to the decoder, which is complementary to the Wyner-Ziv formulation for non causal codes.
- Applying the framework to network information theory is an open challenge
- Trade-off between rate and minimum number of transmissions, for finite length symbol-by-symbol transmission.

Christos K. Kourtellaris

Appendix A

Proofs of Chapter 2

A.1 Proof of Lemma. 2.21

First, recall that given the RV's X, Y, Z, we say X and Y are conditionally independent given Z if $P_{X,Y|Z}(dx,dy|z) = P_{X|Z}(dx|z)P_{Y|Z}(dy|z) - a.s.$ This statement is equivalent to $P_{X|Y,Z}(dx|y,z) = P_{X|Z}(dx|z) - a.s.$ and $P_{Y|X,Z}(dy|x,z) = P_{Y|Z}(dy|z) - a.s.$ or $X \leftrightarrow Y \leftrightarrow Z$ forms a MC in both directions. Now we proceed with the derivation, by often assuming existence of densities to avoid lengthy measure theoretic arguments.

MC1 $\Longrightarrow$ **MC2**: Since **MC1** states that $P_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^n P_{Y_i|Y^{i-1},X^n}(dy_i|y^{i-1},x^n) = \otimes_{i=0}^n P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)$, which is valid if and only if $P_{Y_i|Y^{i-1},X^n}(dy_i|y^{i-1},x^n) = P_{Y_i|Y^{i-1},X^i}(dy_i|y^{i-1},x^i)$, $i = 0,1,\ldots,n$, or equivalently, $X_{i+1}^n \leftrightarrow (X^i,Y^{i-1}) \leftrightarrow Y_i$ forms a MC for each $i = 0,1,\ldots,n-1$, then **MC2** is obtained.

MC1 $\Longleftarrow$ **MC2**: By Bayes' rule $P_{Y^n|X^n}(dy^n|x^n) = \otimes_{i=0}^n P_{Y_i|Y^{i-1},X^n}(dy_i|y^{i-1},x^n) \overset{(a)}{=} \overrightarrow{P}_{Y^n|X^n}(dy^n|x^n)$, where equality in (a) holds due to **MC2**, hence **MC1**.

MC1 $\Longrightarrow$ **MC3**: We need to show that $P_{X_{i+1}|X^i,Y^i}(dx_{i+1}|x^i,y^i) = P_{X_{i+1}|X^i}(dx_{i+1}|x^i)$, $i =$

Christos K Kourtellaris

$0,\ldots,n-1$. Note that

$$\begin{aligned} P_{X_{i+1}|X^i,Y^i}(dx_{i+1}|x^i,y^i) &= \int_{\mathcal{X}_{i+2,n}} P_{X_{i+1},X^n_{X+2}|X^i,Y^i}(dx_{i+1},dx^n_{i+2}|x^i,y^i) \\ &= \int_{\mathcal{X}_{i+2,n}} P_{X^n_{i+1}|X^i,Y^i}(dx^n_{i+1}|x^i,y^i) \\ &\overset{(b)}{=} \int_{\mathcal{X}_{i+2,n}} P_{X^n_{i+1}|X^i,Y^{i-1}}(dx^n_{i+1}|x^i,y^{i-1}) \\ &= P_{X_{i+1}|X^i,Y^{i-1}}(dx_{i+1}|x^i,y^{i-1}),\ i=0,\ldots,n-1 \end{aligned} \tag{A.1.1}$$

where (b) follows because **MC1** and **MC2** are equivalent. We show the rest of the derivation by assuming existence of density functions which are denoted by lower case letters $\bar{p}(\cdot|\cdot)$. From (A.1.1), we have

$$\begin{aligned} \bar{p}(x_{i+1}|x^i,y^i) = \bar{p}(x_{i+1}|x^i,y^{i-1}) = \frac{\bar{p}(x^{i+1},y^{i-1})}{\bar{p}(x^i,y^{i-1})} &= \frac{\bar{p}(y^{i-1}|x^{i+1})\bar{p}(x^{i+1})}{\bar{p}(y^{i-1}|x^i)\bar{p}(x^i)} \\ = \frac{\times_{j=0}^{i-1}\bar{p}(y_j|y^{j-1},x^{i+1})\bar{p}(x^{i+1})}{\times_{j=0}^{i-1}\bar{p}(y_j|y^{j-1},x^i)\bar{p}(x^i)} &\overset{(c)}{=} \frac{\times_{j=0}^{i-1}\bar{p}(y_j|y^{j-1},x^j)}{\otimes_{j=0}^{i-1}\bar{p}(y_j|y^{j-1},x^j)}.\frac{\bar{p}(x^{i+1})}{\bar{p}(dx^i)} \\ = \bar{p}(x_{i+1}|x^i),\ i=0,\ldots,n-1 & \end{aligned}$$

where (c) follows because **MC1** and **MC2** are equivalent. This shows **MC1** $\Longrightarrow$ **MC3**.

MC2 $\Longleftarrow$ **MC3**: We need to show that $\bar{p}(y_i|y^{i-1},x^i,x^n_{i+1}) = \bar{p}(y_i|y^{i-1},x^i)$, $i=0,\ldots,n-1$. But

$$\begin{aligned} \bar{p}(y_i|y^{i-1},x^i,x^n_{i+1}) &= \frac{\bar{p}(y^i,x^i,x^n_{i+1})}{\bar{p}(y^{i-1},x^i,x^n_{i+1})} \\ &= \frac{\bar{p}(x_n|x^{n-1},y^i)\bar{p}(x_{n-1}|x^{n-2},y^i)\ldots\bar{p}(x_{i+1}|x^i,y^i)\bar{p}(x^i,y^i)}{\bar{p}(x_n|x^{n-1},y^{i-1})\bar{p}(x_{n-1}|x^{n-2},y^{i-1})\ldots\bar{p}(x_{i+1}|x^i,y^{i-1})\bar{p}(x^i,y^{i-1})} \\ &\overset{(d)}{=} \frac{\bar{p}(x^i,y^i)}{\bar{p}(x^i,y^{i-1})} = \frac{\bar{p}(y_i|y^{i-1},x^i)\bar{p}(x^i,y^{i-1})}{\bar{p}(x^i,y^{i-1})} = \bar{p}(y_i|y^{i-1},x^i),\ i=0,\ldots,n-1 \end{aligned}$$

where (d) follows from **MC3**. This shows **MC3** $\Longrightarrow$ **MC2**. Thus, we have the equivalence **MC1** $\Longleftrightarrow$ **MC2** $\Longleftrightarrow$ **MC3**.

MC4 $\Longrightarrow$ **MC3**: Since for $i=0,\ldots,n-1$, by **MC4** we have

$$P_{X^n_{i+1}|X^i,Y^i}(dx^n_{i+1}|x^i,y^i) = P_{X^n_{i+1}|X^i}(dx^n_{i+1}|x^i)$$

Christos K. Kourtellaris

then by integrating over $\mathscr{X}_{i+2,n}$ both sides of the previous identity we obtain **MC3**.

MC4 $\Longleftarrow$ **MC3**: Since **MC3** $\Longleftrightarrow$ **MC2**, we show that if $X_{i+1}^n \leftrightarrow (X^i, Y^{i-1}) \leftrightarrow Y_i$ forms a MC for $i = 0, 1, \ldots, n-1$, then $X_{i+1}^n \leftrightarrow X^i \leftrightarrow Y^i$ forms a MC for $i = 0, 1, \ldots, n-1$. We show this by induction. First, we show that $(X_{i+1}, X_{i+2}) \leftrightarrow X^i \leftrightarrow Y^i$ forms a MC, or equivalently, $\bar{p}(x_{i+1}, x_{i+2}|x^i, y^i) = \bar{p}(x_{i+1}, x_{i+2}|x^i)$. Since

$$\begin{aligned}
\bar{p}(x_{i+1},x_{i+2}|x^i,y^i) &= \frac{\bar{p}(x^i,x_{i+1},x_{i+2},y^i)}{\bar{p}(x^i,y^i)} = \frac{\bar{p}(y_i|y^{i-1},x^{i+2})\bar{p}(y^{i-1},x^{i+2})}{\bar{p}(x^i,y^i)}\\
&= \frac{\underbrace{\bar{p}(y_i|y^{i-1},x^i)}_{(e)}\bar{p}(x_{i+2}|x^{i+1},y^{i-1})\bar{p}(x^{i+1},y^{i-1})}{\bar{p}(x^i,y^i)}\\
&= \frac{\bar{p}(y_i|y^{i-1},x^i)\underbrace{\bar{p}(x_{i+2}|x^{i+1})}_{(f)}\bar{p}(x_{i+1}|x^i,y^{i-1})\bar{p}(x^i,y^{i-1})}{\bar{p}(y_i|y^{i-1},x^i)\bar{p}(x^i,y^{i-1})}\\
&= \bar{p}(x_{i+2}|x^{i+1})\underbrace{\bar{p}(x_{i+1}|x^i)}_{(g)}\\
&= \bar{p}(x_{i+2},x_{i+1}|x^i).
\end{aligned}$$

where (e) is implied from **MC2**, while (f), (g) follows from **MC3** $\Longleftrightarrow$ **MC2**. Hence, **MC4** holds for $n = i+2$.

Suppose $X_{i+1}^k \leftrightarrow X^i \leftrightarrow Y^i$ forms a MC, for some $i+2 \leq k < n-1$. We show that it holds for $k \longrightarrow k+1$.

$$\begin{aligned}
\bar{p}(x_{i+1}^{k+1}|x^i,y^i) &= \frac{\bar{p}(x_{i+1}^{k+1},x^i,y^i)}{\bar{p}(x^i,y^i)} = \frac{\bar{p}(x_{k+1}|x_{i+1}^k,x^i,y^i)\bar{p}(x_{i+1}^k,x^i,y^i)}{\bar{p}(x^i,y^i)}\\
&= \frac{\underbrace{\bar{p}(x_{k+1}|x^k)}_{(h)}\bar{p}(x_{i+1}^k|x^i,y^i)\bar{p}(x^i,y^i)}{\bar{p}(x^i,y^i)}\\
&= \bar{p}(x_{k+1}|x^k)\underbrace{\bar{p}(x_{i+1}^k|x^i)}_{(i)}\\
&= \bar{p}(x_{i+1}^{k+1}|x^i)
\end{aligned}$$

where (h), (i) follow from **MC3** $\Longleftrightarrow$ **MC2**. This completes the derivation. □

Christos K. Kourtellaris

Appendix B

Proofs of Chapter 3

B.1 Proof of Theorem. 3.9

We first address part (b).

(b) We give a derivation based on stochastic optimal control techniques. Let $\{e^*_i(x^i, a^{i-1}, b^{i-1}) : i = 0, 1, \ldots, n\}$ denote the optimal encoder strategy. Consider the pay-off

$$I(X^n \rightarrow B^n) = \sum_{i=0}^{n} \mathbb{E}^e \Big\{ \int_{\mathscr{B}_i} \log \frac{P^e_{B_i|B^{i-1},X^i}(db_i|B^{i-1}, X^i)}{P^e_{B_i|B^{i-1}}(db_i|B^{i-1})} P^e_{B_i|B^{i-1},X^i}(db_i|B^{i-1}, X^i) \Big\} \tag{B.1.1}$$

$$= \mathbb{E}^e \Big\{ \sum_{i=0}^{n} \log \frac{P^e_{B_i|B^{i-1},X^i}(dB_i|B^{i-1}, X^i)}{P^e_{B_i|B^{i-1}}(dB_i|B^{i-1})} \Big\} \tag{B.1.2}$$

where superscripts emphasize the dependence on the policies. Since $a_i = e_i(x^j, a^{i-1}, b^{j-1})$ then $P^e_{B_i|B^{i-1},X^i}(db_i; b^{i-1}, x^i) = P_i(db_i; b^{i-1}, x^i, a^i)$, $i = 0, 1, \ldots, n$. Moreover by Assumption 3.8 then $P_i(db_i; b^{i-1}, x^i, a^i) = q_i(db_i; b^{i-1}, x_i, a_i)$-a.a.$(b^{i-1}, x^i, a^i)$, $i = 0, 1, \ldots, n$. Hence under Assumption 3.8 the pay-off becomes

$$I(X^n \rightarrow B^n) = \mathbb{E}^e \Big\{ \sum_{i=0}^{n} \log \frac{q_i(dB_i; B^{i-1}, X_i, A_i)}{\nu^e_i(dB_i; B^{i-1})} \Big\} = \sum_{i=0}^{n} I(X_i; B_i|B^{i-1}) \tag{B.1.3}$$

where $\nu^e(db_i; b^{i-1}) \in \mathscr{K}(\mathscr{B}_i; \mathscr{B}_{0,i-1})$ is the conditional distribution obtained via integration

$$\nu^e(db_i; b^{i-1}) = \int_{\mathscr{X}_i} q_i(db_i; b^{i-1}, x_i, a_i) P^e_i(dx_i; b^{i-1})$$

Christos K. Kourtellaris

By the property of conditional expectation, we have

$$\sum_{i=0}^{n} I(X_i;B_i|B^{i-1}) = \mathbb{E}^e\Big\{\sum_{i=0}^{n}\mathbb{E}^e\Big(\log\frac{q_i(dB_i;B^{i-1},X_i,A_i)}{\nu_i^e(dB_i;B^{i-1})}|X^i,B^{i-1},A^i\Big)\Big\}$$

By Assumption.3.8, the inner expectation is with respect to the conditional measure $P^e_{B_i|B^{i-1},X^i}(db_i|b^{i-1},x^i) = q_i(db_i;b^{i-1},x_i,a_i)$, $i = 0,1,\ldots,n$. Thus

$$\begin{aligned} I(X^n \rightarrow B^n) &= \sum_{i=0}^{n} I(X_i;B_i|B^{i-1}) \\ &= \mathbb{E}^e\Big\{\sum_{i=0}^{n}\mathbb{E}^e\Big(\log\frac{q_i(dB_i;B^{i-1},X_i,A_i)}{\nu_i^e(dB_i;B^{i-1})}\Big|X_i,B^{i-1},A_i\Big)\Big\} \\ &\equiv \mathbb{E}^e\Big\{\sum_{i=0}^{n}\ell(X_i,B^{i-1},A_i)\Big\} \end{aligned} \tag{B.1.4}$$

where

$$\ell(X_i,B^{i-1},A_i) \triangleq \int_{\mathscr{B}_i}\log\Big(\frac{q_i(db_i;B^{i-1},X_i,A_i)}{\nu_i^e(db_i;B^{i-1})}\Big)q_i(db_i;B^{i-1},X_i,A_i)$$

Define $Z_i \triangleq (X_i,B^{i-1}), i = 0,1,\ldots,n$. Clearly the maximization of $I(X^n \rightarrow B^n)$ over $\{a_i = e_i(x^i,a^{i-1},b^{i-1}) : i = 0,1,..,n\}$ is performed by choosing the encoder output sequence $\{a_i = e_i(x^i,a^{i-1},b^{i-1}) : i = 0,1,..,n\}$ to control the joint process $\{Z_i \triangleq (X_i,B^{i-1}) : i = 0,1,..,n\}$. Thus

$$I(X^n \rightarrow B^n) = \mathbb{E}^e\Big\{\sum_{i=0}^{n}\ell(Z_i,A_i)\Big\}$$

.

From stochastic optimal control theory, it is known that if $\{Z_i \triangleq (X_i,B^{i-1}) : i = 0,1,\ldots,n\}$ is a Markov process controlled by the encoder process $\{A_i : i = 0,1,..n\}$, then the optimal encoder strategies $\{e_j^*(a^{j-1},x^j,b^{j-1}) : i = 0,1,..n\}$ will reduce to the simplified form $\{\bar{g}_i(z_i) = g_i(x_i,b^{i-1}) : i = 0,1,..,n\}$. Under Assumption.3.7 is sufficient to show that $\{Z_i : i = 0,1,...,n\}$ is a Markov process controlled by $\{A_i : i = 0,1,...,n\}$, that is

$$P_i(dz_i;a^{i-1},z^{i-1}) = P_i(dz_i;a_{i-1},z_{i-1}) - a.a.\ (a^{i-1},z^{i-1}), \forall\, i \in \mathbb{N}_+^n$$

Next, we show that $\{Z_i : i = 0,1,\ldots,n\}$ is a Markov process controlled by the channel input process $\{A_i : i = 0,1,\ldots,n\}$. The rigorous way to prove this is by using the a.s. definition of conditional independence. However,without loss of generality we assume existence of

Christos K. Kourtellaris

probability density functions

$$P_{i+1}(dz_{i+1};z^i,a^i) = p_{i+1}(z_{i+1}|z^i,a^i)dz_{i+1}$$
$$P_{i+1}(dz_{i+1};z_i,a_i) = p_{i+1}(z_{i+1}|z_i,a_i)dz_{i+1}$$

and we show that $P_{i+1}(dz_{i+1};z^i,a^i) = p_{i+1}(dz_{i+1};z_i,a_i)dz_{i+1}$, for almost all $(a^i,z^i) : i = 0,1,\ldots,n-1$. First we note that

$$\begin{aligned} p_{i+1}(z_{i+1}|z^i,a^i) &= p_{i+1}(x_{i+1},b^i|x^i,b^{i-1},a^i) \\ &= p_{i+1}(x_{i+1}|x^i,b^i,a^i)p_{i+1}(b_i|x^i,b^{i-1},a^i) \end{aligned}$$

By Assumptions 3.7 and 3.8, then

$$\begin{aligned} p_{i+1}(z_{i+1}|z^i,a^i) &= p_{i+1}(x_{i+1}|x_i,b^i,a_i)p_{i+1}(b_i|x_i,b^{i-1},a_i) \\ &= p_{i+1}(x_{i+1},b_i|x_i,b^{i-1},a_i) \\ &= p_{i+1}(z_{i+1}|z_i,a_i) \end{aligned} \tag{B.1.5}$$

Now, from (B.1.4) we have

$$\begin{aligned} I(X^n \to Y^n) &= \mathbb{E}^A\Big\{\sum_{i=0}^n \ell(X_i,B^{i-1},A_i)\Big\} \\ &\triangleq \mathbb{E}^A\Big\{\sum_{i=0}^n \bar{\ell}(Z_i,A_i)\Big\} \end{aligned} \tag{B.1.6}$$

and by (B.1.5), the process $\{Z_i : i = 0,1,\ldots,n\}$ is a Markov process controlled by $\{A_i : i = 0,1,\ldots,n\}$. Since the payoff in (B.1.6) is additive, the minimization of $I(X^n \to B^n)$ over $\{A_i : i = 0,1,\ldots,n\}$ is done by choosing $\{A_i : i = 0,1,\ldots,n\}$ to control the Markov process $\{Z_i : i = 0,1,\ldots,n\}$. Hence, from optimal control theory the encoder structure should be of the form $a_i = g_i(x_i,b^{i-1})$, for some measurable function $g(.,.)$.

(a) The derivation for randomized strategies is similar, hence it is omitted. It follows from the fact that for general complete separable spaces and strategies based on classical nested information over time, that randomized strategies and deterministic strategies are equivalent in terms of optimal performance.

Christos K. Kourtellaris

B.2 Proof of Theorem. 3.39.1

The derivation utilizes from the above discussion and Fano's inequality as follows. Suppose rate R is achievable and hence there exists an (n, M_n, ε_n) code, $\mathscr{C}_n \overset{\triangle}{=} \{u_1, u_2, \ldots, u_{M_n}\}$, $u_j \in \mathscr{A}_{0,j-1}$ satisfying

$$\lim_{n\to\infty} \varepsilon_n = 0 \;\text{ and }\; \liminf_{n\to\infty} \frac{1}{n}\log M_n \geq R$$

Let $A^{n-1} \in \mathscr{A}_{0,n-1}$ be a random variable which is uniformly distributed over the code $\mathscr{C}_n$. Denote by $B^{n-1} \in \mathscr{B}_{0,n-1}$ the channel output corresponding to the channel input $A^{n-1} \in \mathscr{A}_{0,n-1}, A_i = \varphi_i(X, B^{i-1}), i = 0, 1, \ldots, n-1, x \in \mathscr{M}_n$. Define another random variable $\hat{A}^{n-1} \in \mathscr{C}_n$ such that $\hat{A}^{n-1} = u_x$ if $x = d_n(B^{n-1})$. The probability of error is expressed as

$$\varepsilon_n = \text{Prob}\Big\{\hat{A}^{n-1} \neq A^{n-1}\Big\}$$

Then

$$\begin{aligned}
\log M_n &= H(\mathscr{M}_n)\\
&= H(X|B^n) + I(X; B^{n-1})\\
&\overset{(a)}{\leq} H(\varepsilon_n) + nR\varepsilon_n + I(X; B^{n-1})\\
&= H(\varepsilon_n) + nR\varepsilon_n + H(B^{n-1}) - \sum_{i=0}^{n-1} H(B_i|B^{i-1}, X)\\
&\overset{(b)}{=} H(\varepsilon_n) + nR\varepsilon_n + H(B^{n-1}) - \sum_{i=0}^{n-1} H(B_i|B^{i-1}, X, A^i)\\
&\overset{(c)}{=} H(\varepsilon_n) + nR\varepsilon_n + \sum_{i=0}^{n-1} \Big\{H(B^{n-1}) - H(B_i|B^{i-1}, A^i)\Big\}\\
&\overset{(d)}{=} H(\varepsilon_n) + nR\varepsilon_n + \sum_{i=0}^{n-1} \Big\{H(B^{n-1}) - H(B_i|B^{i-1}, A_i)\Big\}\\
&= H(\varepsilon_n) + nR\varepsilon_n + \sum_{i=0}^{n-1} I(A_i; B_i|B^{i-1})
\end{aligned} \tag{B.2.7}$$

where (a) follows from Fano's inequality, (b) knowing the codebook implies knowing the encoder law, and knowing the encoder law among with the message and previous channel outputs specifies the current value of the encoder, since $A_i = \varphi_i(X, B^{i-1}), i = 0, 1, \ldots, n-1$, (c) follows from the Markov chain $X \leftrightarrow (A^i, B^{i-1}) \leftrightarrow B_i, , i = 0, 1, \ldots, n-1$ and (d) follows from Markov chain $A^{i-1} \leftrightarrow (A_i, B^{i-1}) \leftrightarrow B_i, , i = 0, 1, \ldots, n-1$. Further, maximizing the right side

Christos K. Kourtellaris

of (B.2.7) over all deterministic feedback encoders is less than or equal to maximizing the same quantity over all randomized strategies $\{P_i(da_i;a^{i-1},b^{i-1}) \in \mathscr{K}(\mathscr{A}_i;\mathscr{A}_{0,i-1} \times \mathscr{B}_{0,i-1}) : i=0,1,\ldots n-1\}$, and by Theorem 3.31, this maximization (if it exists) it is achieved over randomized strategies $\{P^*_i(da_i;b^{i-1}) \in \mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}) : i=0,1,\ldots n-1\}$. Taking the supremum over all such randomized strategies and using $\lim_{n\to\infty} \varepsilon_n \to 0$ implies $\lim_{n\to\infty} H(\varepsilon_n) = 0$, then

$$R \leq \liminf_{n\to\infty} \frac{1}{n} \log M_n \leq \liminf_{n\to\infty} \sup_{\{P_i(da_i;b^{i-1}) \in \mathscr{K}(\mathscr{A}_i;\mathscr{B}_{0,i-1}):i=0,1,\ldots n-1\}} \frac{1}{n} \sum_{i=0}^{n-1} I(A_i;B_i|B^{i-1}) \tag{B.2.8}$$

This completes the derivation.

Christos K. Kourtellaris

Bibliography

[1] N. Ahmed and C. Charalambous. Stochastic minimum principle for partially observed systems subject to continuous and jump diffusion processes and driven by relaxed controls. *SIAM Journal on Control and Optimization*, 51(4):3235–3257, 2013. URL http://epubs.siam.org/doi/abs/10.1137/120885656.

[2] E. Akyol, K. Viswanatha, K. Rose, and T. A. Ramstad. On zero delay source-channel coding. *submitted to Information Theory, IEEE Transactions on*, 2013. URL http://arxiv.org/abs/1302.3660.

[3] H. Asnani, H. H. Permuter, and T. Weissman. Capacity of a post channel with and without feedback. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, 2538–2542, 2013.

[4] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[5] T. Berger. Explicit bounds to r(d) for a binary symmetric markov source. *Information Theory, IEEE Transactions on*, 23(1):52 – 59, January 1977.

[6] T. Berger. Living Information Theory. *IEEE Information Theory Society Newsletter*, 53(1), March 2003. ISSN 1059-2362.

[7] P. E. Caines. *Linear Stochastic Systems*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York, 1988.

[8] C. Charalambous, P.A. Stavrou, and N.U. Ahmed. Nonanticipative rate distortion function and relations to filtering theory. *Automatic Control, IEEE Transactions on*, PP(99): 1–1, 2013.

[9] C. D. Charalambous and A. Farhadi. LQG optimality and separation principle for general discrete time partially observed stochastic systems over finite capacity communication channels. *Automatica*, 44(12):3181–3188, 2008.

Christos K. Kourtellaris

[10] C. D. Charalambous and P. A. Stavrou. Directed information on abstract spaces: properties and extremum problems. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 518–522, July 1-6 2012.

[11] C. D. Charalambous, C. K. Kourtellaris, and P. Stavrou. Stochastic control over finite capacity channels: Causality, feedback and uncertainty. In *Proceedings of the 48th IEEE Conference on Decision and Control, held jointly with the 28th Chinese Control Conference (CDC/CCC)*, 5889–5894, 2009.

[12] C. D. Charalambous, P. A. Stavrou, and N. U. Ahmed. Nonanticipative rate distortion function and relations to filtering theory. *Automatic Control, IEEE Transactions on*, 42 (4):937 –952, April 2014.

[13] C.D. Charalambous and R.J. Elliott. Certain nonlinear partially observable stochastic optimal control problems with explicit control laws equivalent to leqg/lqg problems. *Automatic Control, IEEE Transactions on*, 42(4):482 –497, April 1997.

[14] C.D. Charalambous and F. Rezaei. Stochastic uncertain systems subject to relative entropy constraints: Induced norms and monotonicity properties of minimax games. *Automatic Control, IEEE Transactions on*, 52(4):647 –663, April 2007.

[15] C.D. Charalambous, C.K. Kourtellaris, and C. Hadjicostis. Capacity of channels with memory and feedback: Encoder properties and dynamic programming. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, 1450–1457, September 2010.

[16] C.D. Charalambous, C.K. Kourtellaris, and P.A. Stavrou. Rate distortion function with causal decoding. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, 6445–6450, December 2010.

[17] C.D. Charalambous, C.K. Kourtellaris, and C. Hadjicostis. Optimal encoder and control strategies in stochastic control subject to rate constraints for channels with memory and feedback. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, 4522 –4527, December. 2011.

[18] J. Chen and T. Berger. The capacity of finite-state markov channels with feedback. *Information Theory, IEEE Transactions on*, 55(6):780–798, 2005.

[19] T. M. Cover and S. Pombra. Gaussian feedback capacity. *Information Theory, IEEE Transactions on*, 35(1):37–43, 1989.

Christos K. Kourtellaris

[20] I. Csiszár. On an extremum problem of information theory. *Studia Scientiarum Mathematicarum Hungarica*, 9:57–71, 1974.

[21] I. Csiszar. On the error exponent of source-channel transmission with a distortion threshold. *Information Theory, IEEE Transactions on*, 28(6):823–828, 1982.

[22] M. S. Derpich and J. Østergaard. Improved upper bounds to the causal quadratic rate-distortion function for gaussian stationary sources. *Information Theory, IEEE Transactions on*, 58(5):3131–3152, May 2012.

[23] R. L. Dobrushin. Information transmission in channel with feedback. *Theory of Probability and its Applications*, 3(4):367–383, 1958.

[24] P. Ebert. The capacity of the gaussian channel with feedback. *Bell Systems Technical Journal*, 47:1705–1712, 1970.

[25] T. Ericson. A result on delay-less information transmission. In *IEEE International Symposium on Information Theory (ISIT)*, 1979.

[26] A. Farhadi and C. D. Charalambous. Stability and reliable data reconstruction of uncertain dynamic systems over finite capacity channels. *Automatica*, 46(5):889–896, 2010.

[27] N. Gaarder and D. Slepian. On optimal finite-state digital transmission systems. *Information Theory, IEEE Transactions on*, 28(2):167–186, 1982.

[28] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., New York, NY, USA, 1968. ISBN 0471290483.

[29] M. Gastpar. *To code or not to code*. PhD thesis, Ecole Polytechnique Fédérale (EPFL), Lausanne, 2002.

[30] M. Gastpar and M. Vetterli. On the capacity of large gaussian relay networks. *Information Theory, IEEE Transactions on*, 51(3):765–779, March 2005. ISSN 0018-9448.

[31] M. Gastpar, B. Rimoldi, and M. Vetterli. To code, or not to code: Lossy source-channel communication revisited. *Information Theory, IEEE Transactions on*, 49(5): 1147–1158, May 2003.

[32] P. W. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic markov chains. *Statistics & Probability Letters*, 56(2):143 – 146, 2002.

Christos K. Kourtellaris

[33] S.K. Gorantla and T.P. Coleman. On reversible markov chains and maximization of directed information. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, 216 –220, June 2010.

[34] A. K. Gorbunov and M. S. Pinsker. Nonanticipatory and prognostic epsilon entropies and message generation rates. *Problems of Information Transmission*, 9(3):184–191, July-September 1973. Translation from Problemy Peredachi Informatsii, vol. 9, no. 3, 12-21, July-September 1973.

[35] A. K. Gorbunov and M. S. Pinsker. Prognostic epsilon entropy of a Gaussian message and a Gaussian source. *Problems of Information Transmission*, 10(2):93–109, Apr.-June 1974. Translation from Problemy Peredachi Informatsii, vol. 10, no. 2, 5-25, April-June 1974.

[36] A. K. Gorbunov and M. S. Pinsker. Asymptotic behavior of nonanticipative epsilon-entropy for Gaussian processes. *Problems of Information Transmission*, 27(4):361–365, 1991. Translation from Problemy Peredachi Informatsii, vol. 27, no. 4, 100-104, October-December 1991.

[37] R. Gray. Information rates of stationary ergodic finite-alphabet sources. *Information Theory, IEEE Transactions on*, 17(5):516 – 523, September 1971.

[38] T. S. Han. *Information-Spectrum Methods in Information Theory*. Springer-Verlag, Berlin, Heidelberg, New York, second edition, 2003.

[39] T. S. Han and S. Verdu. Approximation theory of output statistics. *IEEE Trans. Inform. Theory*, 39:752–772, 1993.

[40] M. Horstein. Sequential transmission using noiseless feedback. *Information Theory, IEEE Transactions on*, 9(3):136 – 143, July 1963.

[41] S. Ihara. *Information theory - for continuous systems*. World Scientific, 1993. ISBN 978-981-02-0985-8.

[42] S. Jalali and T. Weissman. New bounds on the rate-distortion function of a binary markov source. In *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*, 571 –575, June 2007.

[43] Y. H. Kim. A coding theorem for a class of stationary channels with feedback. *Information Theory, IEEE Transactions on*, 54(4):1488–1499, April 2008.

Christos K. Kourtellaris

[44] V. Kostina and S. Verdu. Fixed-length lossy compression in the finite blocklength regime. *Information Theory, IEEE Transactions on*, 58(6):3309–3338, 2012.

[45] V. Kostina and S. Verdu. Lossy joint source-channel coding in the finite blocklength regime. *Information Theory, IEEE Transactions on*, 59(5):2545–2575, 2013.

[46] G. Kramer. Causality, feedback and directed information. *Ph.D Thesis, Swiss Federal Institute of Technology*, (Diss. ETH No.12656), 1998.

[47] F. H. Lin, K. Hara, V. Solo, M. Vangel, J. W. Belliveau, S. T. Stufflebeam, and H am al ainen M. S. Dynamic Granger-Geweke causality modeling with application to interictal spike propagation. *Human Brain Mapping*, 30(6):1877–1886, June 2009.

[48] T. Linder and G. Lagosi. A zero-delay sequential scheme for lossy coding of individual sequences. *Information Theory, IEEE Transactions on*, 47(6):2533–2538, 2001.

[49] T. Linder and R. Zamir. Causal coding of stationary sources and individual sequences with high resolution. *Information Theory, IEEE Transactions on*, 52(2):662–680, February 2006.

[50] Nan M. and P. Ishwar. On delayed sequential coding of correlated sources. *Information Theory, IEEE Transactions on*, 57(6):3763–3782, 2011.

[51] H. Marko. The bidirectional communication theory–a generalization of information theory. *Communications, IEEE Transactions on*, 21(12):1345 – 1351, December 1973.

[52] J. Massey. Causality, feedback and directed information. *IEEE International Symposium on Information Theory and its Applicationss*, 72:303–305, November 2001.

[53] J. L. Massey. Causality, feedback and directed information. In *International Symposium on Information Theory and its Applications (ISITA '90)*, 303–305, November 27-30 1990.

[54] M. Mushkin and I. Bar-David. Capacity and coding for the gilbert-elliott channels. *Information Theory, IEEE Transactions on*, 35(6):1277–1290, November 1989. ISSN 0018-9448.

[55] G. N. Nair and R. J. Evans. Stabilizability of Stochastic Linear Systems with Finite Feedback Data Rates. *SIAM Journal on Control and Optimization*, 43(2):413–436, 2004.

Christos K. Kourtellaris

[56] D. Neuhoff and R. Gilbert. Causal source codes. *Information Theory, IEEE Transactions on*, 28(5):701–713, September 1982. ISSN 0018-9448.

[57] J. Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, New York, NY, USA, second edition, 2009. ISBN 0-521-77362-8.

[58] H. H. Permuter, H. Asnani, and T. Weissman. Capacity of a post channel with and without feedback. *CoRR*, abs/1309.5440, 2013.

[59] H.H. Permuter, T. Weissman, and A.J. Goldsmith. Finite state channels with time-invariant deterministic feedback. *Information Theory, IEEE Transactions on*, 55(2): 644–662, February 2009.

[60] M. S. Pinsker and A. K. Gorbunov. Epsilon-entropy with delay from small mean-square reproduction error. *Problems of Information Transmission*, 23(2):91–95, April-June 1987. Translation from Problemy Peredachi Informatsii, vol. 23, no. 2, 3-8, April-June 1987.

[61] S. S. Pradhan. On the role of feedforward in gaussian sources: Point-to-point source coding and multiple description source coding. *Information Theory, IEEE Transactions on*, 53(1):331–349, January 2007.

[62] C. van Putten and J. H. van Schuppen. Invariance properties of the conditional independence relation. *The Annals of Probability*, 13(3):934–945, 1985. URL http://www.jstor.org/stable/2243720.

[63] F. Rezaei, N. U. Ahmed, and C. D. Charalambous. Rate distortion theory for general sources with potential application to image processing. *International Journal of Applied Mathematical Sciences*, 3(2):141–165, 2006.

[64] M. Rezaeian. Computation of capacity for gilbert-elliott channels, using a statistical method. In *Communications Theory Workshop, 2005. Proceedings. 6th Australian*, 56–61, February 2005.

[65] C. E. Shannon. A mathematical theory on communication. *Bell System Technical Journal*, (27):379–423, October 1948.

[66] C. E. Shannon. The zero error capacity of a noisy channel. *IRE Transactions on Information Theory*, 2(3):112–124, 1956.

Christos K. Kourtellaris

[67] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE Nat. Conv. Rec., Pt. 4*, 142–163. 1959.

[68] O. Shayevitz and M. Feder. Communication with feedback via posterior matching. In *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*, 391 –395, June 2007.

[69] O. Shayevitz and M. Feder. The posterior matching feedback scheme: Capacity achieving and error analysis. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, 900 –904, July 2008.

[70] O. Shayevitz and M. Feder. Achieving the empirical capacity using feedback: Memoryless additive models. *Information Theory, IEEE Transactions on*, 55(3):1269 –1295, march 2009.

[71] V. Solo. On causality and mutual information. In *47th IEEE Conference on Decision and Control (CDC '08)*, 4939–4944, December 2008.

[72] P. A. Stavrou and C. D. Charalambous. Variational equalities of directed information and applications. In *IEEE International Symposium on Information Theory (ISIT)*, 2577–2581, July 7-12 2013.

[73] P. A. Stavrou, C. K. Kourtelalris, and C. D. Charalambous. Nonanticipative rate distortion function and its application. *submitted on Information Information Theory, IEEE Transactions on*.

[74] P.A. Stavrou, C. Kourtellaris, and D.C. Charalambous. accepted in *Information Theory Proceedings (ISIT), 2014 IEEE International Symposium on*.

[75] S. Tatikonda. *Control Over Communication Constraints*. Ph.d. thesis, M.I.T, Cambridge, MA, 2000.

[76] S. Tatikonda and S. Mitter. Control over noisy channels. *IEEE Transactions on Automatic Control*, 49(7):1196–1201, July 2004.

[77] S. Tatikonda and S. Mitter. The capacity of channels with feedback. *Information Theory, IEEE Transactions on*, 55(1):323 –349, January 2009.

[78] D. Teneketzis. On the structure of optimal real-time encoders and decoders in noisy communication. *Information Theory, IEEE Transactions on*, 52(9):4017–4035, 2006.

Christos K. Kourtellaris

[79] R. Venkataramanan. *Information-Theoretic Results on Communication Problems with Feed-forward and Feedback*. PhD thesis, University of Michigan-Ann Arbor, December 2008.

[80] H. Viswanathan and T. Berger. Sequential coding of correlated sources. *Information Theory, IEEE Transactions on*, 46(1):236–246, 2000.

[81] J. Walrand and P. Varaiya. Optimal causal coding - decoding problems. *Information Theory, IEEE Transactions on*, 29(6):814–820, 1983.

[82] T. Weissman and N. Merhav. On competitive prediction and its relation to rate-distortion theory. *Information Theory, IEEE Transactions on*, 49(12):3185–3194, December 2003.

[83] T. Weissman and N. Merhav. On causal source codes with side information. *Information Theory, IEEE Transactions on*, 51(11):4003–4013, 2005.

[84] A. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *Information Theory, IEEE Transactions on*, 22(1):1–10, January 1976.

[85] S. Yang, A. Kavcic, and S. Tatikonda. Feedback capacity of finite-state machine channels. *Information Theory, IEEE Transactions on*, 51(3):799–810, 2005.

Christos K. Kourtellaris

PUBLICATIONS

Photios Stavrou, Christos K. Kourtellaris, Charalambos D. Charalambous, *" Nonanticipative rate distortion function and its application "*, submitted to IEEE Transactions on Information Theory.

.

Photios Stavrou, Christos K. Kourtellaris, Charalambos D. Charalambous, *"Applications of Information Nonanticipative Rate Distortion Function"*, Proceedings of IEEE International Symposium on Information Theory (under revision), May 29-04, 2014, Honolulu, HI, USA.

Photios Stavrou, Charalambos D. Charalambous, Christos K. Kourtellaris, *"Optimal Nonstationary Reproduction Distribution for Nonanticipative RDF on Abstract Alphabets"*, CoRR abs/1301.6522.

C. D. Charalambous, C. Kourtellaris, and C. N. Hadjicostis , *"Optimal Encoders Maximizing Directed Information of Channels with Memory and Feedback: Stochastic Control and Dynamic Programming"*, Proceedings of 2011 IFAC World Congress, Milan, Italy.

C. D. Charalambous, C. K. Kourtellaris and P. A. Stavrou, *"Rate Distortion Function with Causal Decoding"*, in proceedings of 49th IEEE Conference on Decision and Control (CDC), 2010.

C. D. Charalambous, C. K. Kourtellaris and P. Stavrou, *"On the Optimal Reconstruction Kernel of Causal Rate Distortion Function,"*, in proceedings of 19th International Symposium on Mathematical Theory of Networks and Systems (MTNS). 2010 (Invited Paper)

F. Rezaei, C. D. Charalambous, P. A. Stavrou and C. K. Kourtellaris, *"Minimax Rate Distortion for a Class of Sources," in proceedings of in 4th IEEE Inter- national Symposium on Communications*, Control and Signal Processing (ISCCSP). 2010. (Invited Paper)

C. D. Charalambous, C. K. Kourtellaris and P. Stavrou, *"Rate Distortion with Causal Feedback on Abstract Spaces"*, in proceedings of in 19th International Symposium on Network and Systems (MTNS), Budapest, Hungary, Juluy 5-July 9 2010

C. D. Charalambous, C. Kourtellaris, and C. N. Hadjicostis, *"Capacity of Channels with Memory and Feedback: Encoder Properties and Dynamic Programming"*, Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing, pp. 1450-1457, Allerton House, IL, 2010. (invited)

Christos K. Kourtellaris

C.D. Charalambous and C. Kourtellaris, *"Stochastic Control Over Finite Ca- pacity Channels: Causality, Feedback and Uncertainty,"*, in proceedings of 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference(CDC-CCC), Shanghai, China, December 16- December 18 2009.

C.D. Charalambous and C. Kourtellaris, *"Information Theory for Control Sys- tems: Causality and Feedback"*, in proceedings of 2009 European Control Conference (ECC), Budapest,Hungary.

C. D. Charalambous, C. K. Kourtellaris and P. Stavrou, *"Stochastic Control Over Finite Capacity Channels: Causality and Feedback"*, in proceedings of European Control Conference (ECC), Budapest, Hungary.

C.D. Charalambous, S.M. Djouadi and C. Kourtellaris , *"Statistical Analysis of Multipath Fading Channels using Generalizations of Shot-Noise"*, EURASIP on Wireless Communications and Networking, pages-20.

Christos K. Kourtellaris